THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques Appliquées
Unités de recherche : Laboratoire d'Informatique de Grenoble et GIPSA-lab

Modèles de Matrices et Tenseurs Aléatoires pour le Traitement de Données Massives

Random Matrix and Tensor Models for Large Data Processing

Présentée par :

Hugo LEBEAU

Direction de thèse :

Romain COUILLET PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Directeur de thèse
Florent CHATELAIN MAITRE DE CONFERENCES, GRENOBLE INP - UGA	Co-encadrant de thèse

Rapporteurs :

Philippe LOUBATON
PROFESSEUR DES UNIVERSITES, UNIVERSITE GUSTAVE EIFFEL
Rémi BARDENET
DIRECTEUR DE RECHERCHE, CNRS DELEGATION HAUTS-DE-FRANCE

Thèse soutenue publiquement le **11 mars 2025**, devant le jury composé de :

Olivier MICHEL, PROFESSEUR DES UNIVERSITES, GRENOBLE INP - UGA	Président
Romain COUILLET, PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Directeur de thèse
Philippe LOUBATON, PROFESSEUR DES UNIVERSITES, UNIVERSITE GUSTAVE EIFFEL	Rapporteur
Rémi BARDENET, DIRECTEUR DE RECHERCHE, CNRS DELEGATION HAUTS-DE-FRANCE	Rapporteur
Walid HACHEM, DIRECTEUR DE RECHERCHE, UNIVERSITE GUSTAVE EIFFEL	Examinateur
Mylène MAÏDA, PROFESSEURE DES UNIVERSITES, UNIVERSITE DE LILLE	Examinatrice
Pierre COMON, DIRECTEUR DE RECHERCHE EMERITE, CNRS DELEGATION ALPES	Examinateur

Invités :

Florent CHATELAIN
MAITRE DE CONFERENCES, GRENOBLE INP - UGA

Abstract / Résumé

The exponential growth of computing power has enabled the widespread deployment of machine learning, which has in turn given rise to new challenges in data processing. The sheer volume of data now being generated means that the standard statistical assumption of a number of samples far greater than their dimension is no longer tenable. In the paradigm of the *Big Data* era, datasets are typically of very large dimension and may also comprise several modes, indicating a variety of sources, modalities, domains, and so on. Furthermore, the advancement of technologies required to develop models capable of processing vast quantities of data results in significant environmental and human costs. In light of these concerns, it is imperative to promote a more clever and prudent use of our resources.

Random matrix theory provides powerful tools to precisely study the statistical and computational limitations associated with the processing of large and multidimensional data. Through this lens, we examine several learning approaches to identify the relevant parameters influencing the success of a task and thereby facilitate an informed use.

First of all, we establish a "central limit theorem" on the behavior of the entries of *spike* eigenvectors of a Gram kernel matrix. This is an essential result to predict the performances of spectral clustering, which has so far been missing from the literature.

Then, we study an extension of spectral clustering to data streams. This approach makes it possible to cluster a potentially very large dataset with controlled and limited memory usage. In addition to revealing the exotic spectral behavior of the associated matrix model, our results characterize the reconstruction performances of an observed noisy signal in a data stream. In addition, we show that with an astute management of the available memory, it is possible to achieve performances comparable to those obtained without resource constraints. This means a significant reduction in memory costs compared with standard spectral clustering, with negligible loss of performance.

Finally, we turn our attention to the computational limits to tensor estimation, with a particular focus on low-rank approximation. Through the study of matrices obtained by unfolding a random tensor, we describe precisely the reconstruction performances of a noisy tensor signal by means of a truncated MLSVD (which generalizes the concept of truncated SVD to tensors). In contrast to the matrix case, this estimate is only *quasi-optimal*, and we then investigate the computation of the best low-multilinear-rank tensor approximation with the HOOI algorithm. Using a similar approach, we examine the multi-view clustering problem from the perspective of a rank-one tensor approximation. Our results highlight and precisely quantify the pivotal role of view informativeness in the quality of the estimation. Furthermore, this study sheds light on a central phenomenon in tensor approximation: the statistical-to-computational gap, that is, the fundamental inability to achieve algorithmically the performances theoretically attainable by statistical estimation.

Face au déploiement massif de l'apprentissage machine permis par la puissance exponentiellement grandissante des ressources de calcul, nous devons faire face à de nouveaux défis en matière de traitement des données. En effet, la dimension de ces dernières atteint des tailles désormais si grandes que l'hypothèse standard en statistiques affirmant un nombre d'individus bien plus grand que leur dimension n'est plus acceptable. Dans le paradigme de l'ère du *Big Data*, les données collectées sont généralement de très grande dimension et peuvent également comporter plusieurs modes, indiquant une variété de sources, modalités, domaines, etc. De plus, le développement des technologies nécessaires à la mise en œuvre de modèles capables de traiter des quantités pharamineuses de données implique un coût environnemental et humain catastrophique face auquel il est nécessaire de promouvoir un usage plus raisonné et intelligent de nos ressources.

La théorie des matrices aléatoires fournit des outils puissants pour étudier précisément les limites statistiques et computationnelles associées au traitement de données volumineuses et multidimensionnelles. À travers ce prisme, nous explorons plusieurs approches d'apprentissage afin d'en caractériser les paramètres pertinents pour la réussite d'une tâche et ainsi en permettre un usage informé.

Nous établissons en tout premier lieu un "théorème central limite" sur le comportement des entrées des vecteurs propres *spikes* d'une matrice à noyau de Gram. Il s'agit d'un résultat essentiel pour prédire les performances du clustering spectral qui manquait, jusqu'à présent, dans la littérature.

Ensuite, nous étudions une extension du clustering spectral aux flux de données. Cette approche permet de partitionner un jeu de données potentiellement très grand avec un usage mémoire contrôlé et limité. En plus de dévoiler le comportement spectral exotique du modèle matriciel associé, nos résultats précisent les performances de reconstruction d'un signal observé à travers un flux de données. De plus, nous montrons qu'avec une gestion astucieuse de la mémoire disponible, il est possible d'atteindre des performances comparables à celles obtenues sans contraintes de ressources. Cela permet donc une réduction importante du coût en mémoire par rapport à un clustering spectral standard, pour une perte de performance négligeable.

Enfin, nous nous intéressons aux limites computationnelles de l'estimation tensorielle et, en particulier, de l'approximation de petit rang. À travers l'étude des matrices obtenues en dépliant un tenseur aléatoire, nous décrivons précisément les performances de reconstruction d'un signal tensoriel bruité au moyen d'une MLSVD tronquée (qui généralise aux tenseurs le concept de SVD tronquée). Contrairement au cas matriciel, cette estimation n'est que *quasi-optimale* et nous étudions donc ensuite le calcul de la meilleure approximation tensorielle de petit rang multilinéaire au moyen de l'algorithme HOOI. Par une approche similaire, nous examinons le problème du clustering multi-vues sous l'angle d'une approximation tensorielle de rang un. Nos résultats mettent en évidence et quantifient précisément l'importance de l'informativité des vues dans la qualité de l'estimation. De plus, cette étude permet de mettre en lumière un phénomène central de l'approximation tensorielle : le fossé statistico-computationnel, c'est-à-dire l'incapacité fondamentale d'atteindre algorithmiquement les performances théoriquement atteignables par estimation statistique.

Remerciements

Les prémices de cette thèse se situent lors de mon master, à l'époque où, Covid obligeant, nous ne nous voyions plus qu'à travers des écrans interposés. J'étais à la recherche d'un stage de fin d'études avec l'idée d'enchaîner sur une thèse. La théorie des matrices aléatoires, qui était le sujet d'un des cours du master MVA, m'apparaissait alors comme un outil intriguant et élégant pour apporter une compréhension rigoureuse des méthodes d'apprentissage en grande dimension. C'est ainsi que je suis naturellement entré en contact avec Romain et me suis retrouvé catapulté, quelques mois plus tard, dans la capitale des Alpes où j'ai pu m'épanouir à la fois scientifiquement et personnellement.

En tout premier lieu, je souhaite bien évidemment remercier mes deux encadrants : Romain et Florent. Merci pour votre accompagnement et tout ce que j'ai pu apprendre grâce à vous depuis mon stage de fin d'études jusqu'à l'aboutissement de cette thèse. J'ai pu découvrir le monde de la recherche, la persévérance qu'elle demande mais aussi le plaisir qu'elle procure. Merci Florent, pour tes précieux conseils et les nombreuses fois où tu as su trouver l'argument juste pour débloquer des problèmes sur lesquels je m'embourbais ! Merci Romain, pour m'avoir appris à toujours questionner la pertinence de nos recherches et à « relever la tête du guidon » pour les comprendre dans un contexte global.

Je remercie également tous les membres de mon jury : Rémi, Philippe, Mylène, Walid, Pierre et Olivier. Merci pour toute l'attention que vous avez porté à mon travail et pour nos échanges stimulants et fructueux. Je suis notamment très reconnaissant envers mes rapporteurs, Philippe et Rémi, pour le temps que vous avez dédié à l'étude de mon manuscrit et les retours précieux que vous m'avez donné. Merci également à Walid pour les nombreux échanges que nous avons eu tout au long de ma thèse.

Je dois en outre beaucoup à toutes les personnes qui m'ont entouré et que j'ai rencontrées lors de cette thèse. Pour les bons moments partagés autour d'un repas, d'un palet breton ou même d'une carte de l'Europe, merci Victor, Rémi, Pierre et Julien. Je n'oublie pas les moments passés avec Achille, Charles et tous les doctorants, post-doctorants, stagiaires et permanents des équipes POLARIS et GAIA que j'ai eu la chance de rencontrer. J'ai notamment une pensée pour Annie et Arnaud, pour votre gentillesse et votre disponibilité sans faille.

Cette aventure aurait été bien différente sans les nombreuses sorties trail ou vélo partagées avec Pierrick, ainsi que les séances course à pied de l'ALE Triathlon qui rythmaient mes mercredis soir.

Enfin, je remercie mes parents, c'est grâce à votre soutien inconditionnel depuis mon enfance que j'ai eu la chance formidable de poursuivre les études qui me passionnaient et de pouvoir me lancer sans crainte dans cette thèse. Et bien sûr, je remercie Agathe, ta rencontre a changé ma vie, merci pour le bonheur que tu m'apportes au quotidien.

Contents

Symbols and Notations

Objects

a	A scalar.
$\boldsymbol{a}$, a_i	A vector and its i-entry.
$\boldsymbol{A}$, $A_{i,j}$	A matrix and its (i,j)-entry.
$\boldsymbol{\mathcal{A}}$, $\mathcal{A}_{i_1,\ldots,i_d}$	A tensor and its $(i_1,\ldots,i_d)$-entry.

Norms and Inner Products

$\|\cdot\|_p$	Vector p-norm ($p \in [1,+\infty]$) or the induced matrix norm ($p=2$ by default).
$\|\cdot\| = \|\cdot\|_2$	Standard Euclidean norm for vectors, or the corresponding operator norm (spectral norm) for matrices and tensors.
$\langle\cdot,\cdot\rangle$	Euclidean inner product (between vectors).
$\|\cdot\|_{\mathrm{F}}$	Frobenius norm of matrices and tensors.
$\langle\cdot,\cdot\rangle_{\mathrm{F}}$	Frobenius inner product (between matrices or tensors).
$\|\cdot\|_{\max}$	Matrix max norm.

Sets

$[n] = \{1,\ldots,n\}$	Positive integers smaller or equal to n.
$\mathfrak{S}_n$	Symmetric group on $[n]$.
$\lvert\mathscr{S}\rvert$	Cardinality of the finite set $\mathscr{S}$.
$\mathbb{Z} = \{\ldots,-2,-1,0,1,2,\ldots\}$	Integers.
$\mathbb{R}$	Real numbers.
$\mathbb{C}$	Complex numbers.
$\mathbb{C}^+ = \{z \in \mathbb{C} \mid \Im z > 0\}$	Complex numbers with positive imaginary part.
$\mathbb{R}^n$	n-dimensional real vectors.
$\mathbb{S}^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$	Unit sphere in $\mathbb{R}^n$.
$\mathbb{R}^{p\times n}$	$p \times n$ real matrices.
$V_k(\mathbb{R}^n) = \{\boldsymbol{A} \in \mathbb{R}^{n\times k} \mid \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_k\}$	Orthonormal k-frames in $\mathbb{R}^n$ (Stiefel manifold).
$\mathscr{O}_n(\mathbb{R}) = V_n(\mathbb{R}^n)$	$n \times n$ orthogonal matrices (orthogonal group).
$\mathbb{R}^{n_1\times\ldots\times n_d}$	$n_1 \times \ldots \times n_d$ real order-d tensors.
$[a,b]$, $[a,b[$, $]a,b]$, $]a,b[$	Closed, right-open, left-open and open interval from a to b.

Real and Complex Numbers

$[x]^+ = \max(0, x)$	Positive part of $x \in \mathbb{R}$.
i	Imaginary unit.
$\Re z$, $\Im z$	Real and imaginary parts of $z \in \mathbb{C}$.
$\bar{z} = \Re z - \mathrm{i}\Im z$	Complex conjugate of z.
$\lvert z\rvert = \sqrt{(\Re z)^2 + (\Im z)^2}$	Modulus of z (or absolute value if $z \in \mathbb{R}$).
$\mathrm{Dist}(z, \mathscr{S}) = \min_{x\in\mathscr{S}} \lvert x - z\rvert$	Distance of $z \in \mathbb{C}$ to the *closed* subset $\mathscr{S} \subset \mathbb{C}$.

Probability Measures

$\mathbb{P}(A)$	Probability of the event $A \in \mathscr{A}$ with respect to the underlying probability space $(\Omega, \mathscr{A}, \mathbb{P})$.
δ_x	Dirac measure at x.
$\mathrm{Supp}\,\mu$	Support of the probability measure μ.

Random Variables

$X \sim \mathscr{L}$	The random variable X follows the law $\mathscr{L}$.
$X_i \overset{\text{i.i.d.}}{\sim} \mathscr{L}$	The X_i's are independent and identically distributed according to $\mathscr{L}$.
$\mathbb{E}[X]$	Expectation of X.
$\mathrm{Var}(X)$	Variance of X.
$\mathrm{Cov}(X, Y)$	Covariance of X and Y.
$X_n \xrightarrow[n\to+\infty]{\mathscr{D}} \mathscr{L}$	The sequence of random variables $(X_n)_{n\geqslant 0}$ converges in distribution to $\mathscr{L}$.
$X_n \xrightarrow[n\to+\infty]{\text{a.s.}} L$	The sequence of random variables $(X_n)_{n\geqslant 0}$ converges almost surely to L.
$\mathscr{B}(p)$	Bernoulli distribution with parameter $p \in [0, 1]$.
$\chi^2(d)$	χ^2 distribution with d degrees of freedom.
$\mathscr{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,\mathrm{d}t$	Gaussian cumulative distribution function evaluated at $x \in \mathbb{R}$.

Vectors, Matrices and Tensors

$\delta_{i,j} = \mathbf{1}_{\{i=j\}}$	Kronecker delta.
$\mathbf{0}_n$	n-dimensional vector with all its entries equal to 0.
$\mathbf{1}_n$	n-dimensional vector with all its entries equal to 1.
$\boldsymbol{e}_i^{(n)}$	i-th vector of the canonical basis of $\mathbb{R}^n$.
$\mathbf{0}_{n\times n}$	$n \times n$ matrix with all its entries equal to 0.
$\mathbf{1}_{n\times n}$	$n \times n$ matrix with all its entries equal to 1.
$\boldsymbol{I}_n$	$n \times n$ identity matrix.
$\boldsymbol{A}^\top$	Transpose of $\boldsymbol{A}$.
$\boldsymbol{A}^*$	Conjugate transpose of $\boldsymbol{A}$.
$\boldsymbol{A}_{i,\cdot}$	The i-th row of $\boldsymbol{A}$ as a *row* vector.

$\boldsymbol{A}_{\cdot,j}$	The j-th column of $\boldsymbol{A}$ as a *column* vector.
$\operatorname{Span}\boldsymbol{A} = \{\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^n\} \subset \mathbb{R}^p$	Span of the $p \times n$ real matrix $\boldsymbol{A}$.
$\operatorname{Rank}\boldsymbol{A}$	Rank of $\boldsymbol{A}$, i.e., the dimension of $\operatorname{Span}\boldsymbol{A}$.
$s_1(\boldsymbol{A}) \geqslant s_2(\boldsymbol{A}) \geqslant \dots$	Singular values of the matrix $\boldsymbol{A}$ in *non-increasing* order.
$\operatorname{Tr}\boldsymbol{B} = \sum_{i=1}^n B_{i,i}$	Trace of the $n \times n$ matrix $\boldsymbol{B}$.
$\operatorname{Sp}\boldsymbol{B}$	Spectrum of $\boldsymbol{B}$, i.e., the set of all its eigenvalues.
$\lambda_1(\boldsymbol{S}) \geqslant \lambda_2(\boldsymbol{S}) \geqslant \dots$	Eigenvalues of the *symmetric* matrix $\boldsymbol{S}$ in *non-increasing* order.
$\boldsymbol{u}_k(\boldsymbol{S})$	Unit-norm eigenvector of the *symmetric* matrix $\boldsymbol{S}$ associated to $\lambda_k(\boldsymbol{S})$.
$\operatorname{Diag}(v_1, \dots, v_n)$	$n \times n$ diagonal matrix with diagonal values $v_1, \dots, v_n$.
$\operatorname{Vec}\boldsymbol{A}$	Vector of $\mathbb{R}^{np}$ obtained by stacking the columns of $\boldsymbol{A} \in \mathbb{R}^{p \times n}$.
$\mathcal{A} \odot \mathcal{B}$	Hadamard (pointwise) product of $\mathcal{A}$ and $\mathcal{B}$.
$\boldsymbol{A} \boxtimes \boldsymbol{B}$	Kronecker product of the two matrices $\boldsymbol{A}, \boldsymbol{B}$.
$\boldsymbol{a} \otimes \boldsymbol{b}$	Outer product of the two vectors $\boldsymbol{a}, \boldsymbol{b}$.
$\operatorname{Rank}\mathcal{T}$	Rank or multilinear rank (depending on the context) of the tensor $\mathcal{T}$.
$\mathcal{T}(\boldsymbol{A}^{(1)}, \dots, \boldsymbol{A}^{(d)})$	Contraction of $\mathcal{T}$ on $\boldsymbol{A}^{(1)}, \dots, \boldsymbol{A}^{(d)}$.
$[\![\mathcal{G}; \boldsymbol{U}^{(1)}, \dots, \boldsymbol{U}^{(d)}]\!]$	Tucker decomposition with core tensor $\mathcal{G}$ and factor matrices $\boldsymbol{U}^{(1)}, \dots, \boldsymbol{U}^{(d)}$.

Asymptotic Notation

$u_n = \mathcal{O}(v_n)$	There exist $C > 0$ and an integer n_0 such that $\lvert u_n \rvert \leqslant C \lvert v_n \rvert$ as soon as $n \geqslant n_0$.
$u_n(z) = \mathcal{O}_z(v_n)$	There exist two polynomials P, Q with nonnegative coefficients and an integer n_0 such that $\lvert u_n(z) \rvert \leqslant \frac{P(\lvert z \rvert)}{\lvert \Im z \rvert Q(\lvert \Im z \rvert)} \lvert v_n \rvert$ for all $z \in \mathbb{C} \setminus \mathbb{R}$ as soon as $n \geqslant n_0$.
$u_n = \mathrm{o}(v_n)$	For all $\varepsilon > 0$, there exists an integer n_0 such that $\lvert u_n \rvert \leqslant \varepsilon \lvert v_n \rvert$ as soon as $n \geqslant n_0$.
$u_n \ll v_n$	Alternative notation for $u_n = \mathrm{o}(v_n)$.
$v_n \gg u_n$	Alternative notation for $u_n = \mathrm{o}(v_n)$.
$u_n = \Theta(v_n)$	There exist $c, C > 0$ and an integer n_0 such that $c \lvert v_n \rvert \leqslant \lvert u_n \rvert \leqslant C \lvert v_n \rvert$ as soon as $n \geqslant n_0$.

Given a norm $N(\cdot)$ (usually the spectral norm $\lVert \cdot \rVert$), we also write

- $\boldsymbol{A}_n = \mathcal{O}^{N(\cdot)}(v_n)$ if $N(\boldsymbol{A}_n) = \mathcal{O}(v_n)$,
- $\boldsymbol{A}_n(z) = \mathcal{O}_z^{N(\cdot)}(v_n)$ if $N(\boldsymbol{A}_n(z)) = \mathcal{O}_z(v_n)$,
- $\boldsymbol{A}_n = \mathrm{o}^{N(\cdot)}(v_n)$ if $N(\boldsymbol{A}_n) = \mathrm{o}(v_n)$.

Chapter 1

Introduction

1.1 Learning from Large Multidimensional Data: A New Paradigm

STATISTICS are defined in the Cambridge English Dictionary as "the science of using information discovered from studying numbers". The word itself comes from the German *Statistik* introduced in 1749 by Gottfried Achenwall in its *Abriß der neuen Staatswissenschaft der vornehmen Europäischen Reiche und Republiken*. He described in this work the conditions of the leading European countries with numbers: their agriculture, manufactures and commerce. Initially, "statistics" only referred to numerical information about states, such as demographics in particular, and later encompassed any collected numerical information as well as its analysis and interpretation. It was not until the advent of the computer in the mid-20$^{\text{th}}$ century that larger amounts of data could be efficiently processed and information extracted from them.

Standard statistical models usually assume a number n of individuals for each of which p numerical values are given. As an example, n individuals of a population could be represented, for medical purposes, by p-dimensional feature vectors $\boldsymbol{x}_i$ whose entries are their age, weight, height, blood pressures, etc. Thus, it is often the case that the number of individuals is much larger (several orders of magnitude) than the number of features by which they are represented. A natural assumption when studying these statistical models is therefore that $n \to +\infty$ while p remains fixed. In fact, most standard results in statistics (van der Vaart, 1998) are given in this setting, which has indeed long been judicious.

Let us look at the following elementary example as an illustration of such results. Consider n independent and identically distributed random vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^p$ with zero mean, $\mathbb{E}[\boldsymbol{x}_i] = \boldsymbol{0}_p$, and a positive-definite covariance matrix $\boldsymbol{C} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top]$. The sample covariance matrix is defined as

$$\hat{\boldsymbol{C}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top .$$

It is a convergent estimator of the covariance $\boldsymbol{C}$. Indeed, by the strong law of large numbers, we have the almost sure convergence

$$\max_{1 \leqslant j,k \leqslant p} \left| \hat{C}_{j,k} - C_{j,k} \right| \xrightarrow[n \to +\infty]{\text{a.s.}} 0 \tag{1.1}$$

and the convergence in spectral norm $\|\hat{\boldsymbol{C}} - \boldsymbol{C}\| \to 0$ almost surely as $n \to +\infty$. This shows that $\hat{\boldsymbol{C}}$ is an accurate estimator of the covariance in the regime where $n \gg p$.

However, since the beginning of the digital age at the start of the 21$^{\text{st}}$ century, the amount of available data and their size has rapidly risen. Notably, in biology, the emergence of DNA microarrays in

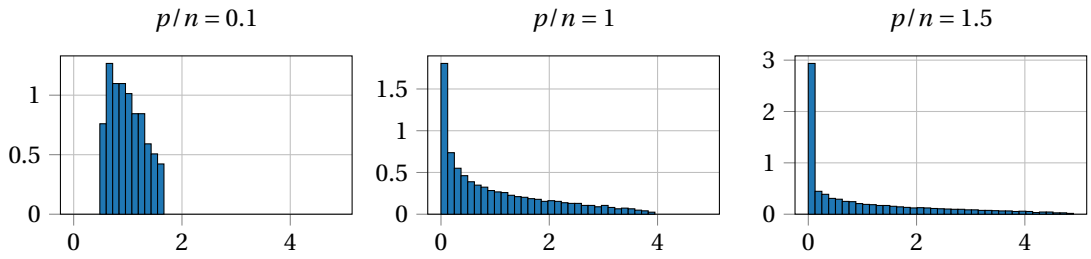


Figure 1.1: Empirical distribution of eigenvalues of the sample covariance matrix $\hat{\boldsymbol{C}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ with $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$, $n = 1000$ and, from left to right, $p/n = 0.1, 1, 1.5$.

the late 1990s and the publication of the human genome in 2001 launched a debate on how to manage and analyze this "data deluge" (Gershon, 2002). The "big data era" in which we live today provides tremendous amounts of information in many different forms. Combined with the power of modern computers, these resources allow solid statistical inference and a rapid flowering of machine learning methods such as deep neural networks, which thrive since the "AlexNet surprise" (Krizhevsky et al., 2012). This new paradigm, where data is no longer scarce but superabundant, invites us to process data with *large* and *multiple* dimensions.

1.1.1 The Large-Dimensional Regime

The old standard assumption $n \gg p$ is no longer verified and new statistical models — requiring new tools — must be considered. In the covariance estimation example presented above, if *both* n and p are large, that is, if $p, n \to +\infty$ with the ratio p/n neither vanishing nor diverging, it is no longer true that $\|\hat{\boldsymbol{C}} - \boldsymbol{C}\| \to 0$, even if the convergence (1.1) still holds! Indeed, this is easily seen in the case where $p > n$ since this implies that $\hat{\boldsymbol{C}}$ must have at least $p - n$ zero eigenvalues, therefore $\|\hat{\boldsymbol{C}} - \boldsymbol{C}\| \not\to 0$. Because $\hat{\boldsymbol{C}}$ is a $p \times p$ matrix, the fact that $p \to +\infty$ with n hinders the transition from the convergence of the entries (1.1) to the "global" convergence of the matrix (in spectral norm).

This phenomenon is illustrated in Figure 1.1. Empirical distributions of eigenvalues of $\hat{\boldsymbol{C}}$ are represented for $n = 1000$ and various p/n ratios when $\boldsymbol{C} = \boldsymbol{I}_p$. The larger p is (relatively to n), the more spread out are the eigenvalues. If $p < n$, there are no zero eigenvalues, but they still occupy a rather large interval. Clearly, it is hard to infer from these empirical distributions that the underlying covariance has all its eigenvalues equal to 1.

Thus we see that the ratio p/n is a central parameter in statistical models dealing with large-dimensional data. As a rule of thumb, standard statistical results where $n \to +\infty$ but p is fixed may no longer hold as soon as $p/n \gtrsim 0.01$ (Couillet and Liao, 2022, Remark 1.2). Consider, e.g., VGG features (Simonyan and Zisserman, 2015) of n images. Each image is represented by a vector of dimension $p = 4\,096$. Therefore, n must be at least $409\,600$ to avoid potential large-dimensional fallacies. Then, if one is given a kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ and wishes to compute a Gram kernel matrix $\boldsymbol{K}$ with entries $K_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (e.g., for clustering purposes), there are, due to symmetry, $\frac{1}{2}n(n+1)$ entries to store. Considering each floating-point entry stored with half precision on 16 bits, the storage of the kernel matrix would require 168 gigabytes! This is much more than the available amount of random-access memory on most current computers. The assumption $n \gg p$ is hence conspicuously not reasonable in a realistic modeling of this setting. In Chapter 4, we propose a efficient way to deal with such very large kernel matrices with a limited amount of memory resources.

1.1.2 The Multidimensional Regime

Besides being of increasingly larger dimension, data could also come with several *modes* representing various sources, modalities, domains, and so on. Consequently, the representation of an individual $\boldsymbol{x}_i$ as a p-dimensional vector may not properly represent the available information. As an illustration, consider the problem of localizing the origin of an epileptic seizure through an analysis of electroencephalogram (EEG), as it was done by Acar et al. (2007). The collected data correspond to wavelet coefficients at each time step, each scale (frequency) and each electrode. Therefore the simplest way to represent the dataset while still preserving the structure of the problem is with a 3-dimensional array whose (t,s,e)-entry is the coefficient at time t, scale s and electrode e.

d-dimensional arrays are commonly represented with *tensors*, which are objects describing multilinear relationships between vectors, matrices and even tensors themselves (in fact, vectors and matrices can be seen as particular cases of tensors). In particular, just as a matrix (order-2 tensor) $\boldsymbol{M}$ describes a bilinear form $(\boldsymbol{x},\boldsymbol{y}) \mapsto \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{y}$, an order-$d$ tensor $\mathcal{T}$ describes a d-linear form $(\boldsymbol{x}_1,\dots,\boldsymbol{x}_d) \mapsto \mathcal{T}(\boldsymbol{x}_1,\dots,\boldsymbol{x}_d)$. They appear in multiple areas such as brain imaging (Zhou et al., 2013), neurophysiological measurements (Rabinowitz et al., 2015; Seely et al., 2016), community detection (Anandkumar et al., 2013), hyperspectral imaging (Li and Li, 2010; Zhang et al., 2013; Kanatsoulis et al., 2018), spatio-temporal gene expression (Liu et al., 2022), recommender systems (Karatzoglou et al., 2010; Rendle and Schmidt-Thieme, 2010; Frolov and Oseledets, 2017) and topic modeling (Anandkumar et al., 2014). Hence, modern data analysis is not only *large-dimensional* but also *multidimensional*.

A major concern of machine learning and signal processing is to discover and exploit underlying low-dimensional structures in data. It is therefore a natural sparsity assumption to model the sought information as being represented by a few algebraic terms in a certain tensor decomposition (Kadmon and Ganguli, 2018; Anandkumar et al., 2014). In their fMRI study, Hunyadi et al. (2017) perform a blind source separation via a joint tensor decomposition on a *channel* × *time* × *patient* array, whereas Williams et al. (2018) use a low-rank tensor approximation on a *neuron* × *time* × *trial* array as a dimensionality reduction technique to study neural dynamics.

The reconstruction of such low-dimensional information is however more challenging with tensors than matrices as some notions and properties pertaining to the latter do not easily generalize to the former. A most striking fact is the non-closure of the set of rank-R order-d tensors[1] as soon as $R \geqslant 2$ and $d \geqslant 3$. In the order-2 case, it is well-known that if the sequence of rank-R matrices $(\boldsymbol{A}_n)_{n\geqslant 0}$ converges to $\boldsymbol{A}_\infty$, then the rank of $\boldsymbol{A}_\infty$ is at most R. In other words, the set $\{\boldsymbol{A} \in \mathbb{R}^{p\times n} \mid \operatorname{Rank}\boldsymbol{A} \leqslant R\}$ is closed. For higher-order tensors ($d \geqslant 3$), this property is lost as soon as $R \geqslant 2$. As an example, de Silva and Lim (2008) give the following sequence of order-3 tensors,

$$\mathcal{A}_n = n\left(\boldsymbol{x}_1 + \frac{1}{n}\boldsymbol{y}_1\right) \otimes \left(\boldsymbol{x}_2 + \frac{1}{n}\boldsymbol{y}_2\right) \otimes \left(\boldsymbol{x}_3 + \frac{1}{n}\boldsymbol{y}_3\right) - n\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3$$

where $\boldsymbol{x}_1, \boldsymbol{y}_1 \in \mathbb{R}^{n_1}$, $\boldsymbol{x}_2, \boldsymbol{y}_2 \in \mathbb{R}^{n_2}$ and $\boldsymbol{x}_3, \boldsymbol{y}_3 \in \mathbb{R}^{n_3}$. Although each $\mathcal{A}_n$ has rank 2 (except for some particular choices of $\boldsymbol{x}_1, \boldsymbol{y}_1, \boldsymbol{x}_2, \boldsymbol{y}_2, \boldsymbol{x}_3, \boldsymbol{y}_3$ not considered here), the limit $\mathcal{A}_\infty = \boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{y}_3 + \boldsymbol{x}_1 \otimes \boldsymbol{y}_2 \otimes \boldsymbol{x}_3 + \boldsymbol{y}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3$ has rank 3! Comon et al. (2009) give another insightful example of a sequence of rank-3 tensors with a rank-5 limit. In fact, a sequence of fixed-rank tensors can converge to a tensor of arbitrarily higher tensor rank (de Silva and Lim, 2008).

Within the framework of the reconstruction of a low-dimensional information from large data, the low-rank approximation is a commonly-used technique. It is a well-posed problem when working with matrices, but the previous example expressly shows that it is ill-posed for tensors of general order. Indeed, $\mathcal{A}_\infty$ is an example of tensor having no best rank-2 approximation. More broadly, the generalization of a matrix-related problem to a tensor-related one raises many difficulties. The *prima*

[1] The rank of a tensor is defined as the minimum number of rank-one terms $\bigotimes_{\ell=1}^{d} \boldsymbol{a}^{(\ell)}$ in which it can be exactly decomposed.

facie disheartening observation made by Hillar and Lim (2013) that most tensor problems are NP-hard presents in fact a variety of exciting, new, and sometimes unexpected challenges.

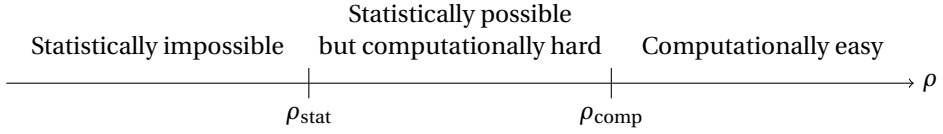
1.1.3 Inference Problems and Phase Transitions

The reconstruction of a low-dimensional information from large data mentioned previously falls into the realm of a particular class of *statistical inference* problems, which we define as follows: given an observation X modeled as a noisy measurement of a sought information, i.e., $X = S + N$ where S is a signal and N an additive noise, the inference problem consists in reconstructing S from the observation of X, that is, finding a transformation f such that $\hat{S} = f(X)$ maximizes a certain measure of affinity with S. Although we will be mostly interested in the case where X, S and N are matrices or tensors, they could be any object as long as our operations make sense. The model usually comes with a sparsity assumption on S, i.e., it is expected to break down in a few simple terms. Harnessing this decomposition is key to finding a good transformation f.

The reconstruction performance of $\hat{S} = f(X)$ depends on the *signal-to-noise ratio* ρ which is generally defined as the ratio between the powers of the signal and the noise, $\rho = \frac{\mathscr{P}_{\text{signal}}}{\mathscr{P}_{\text{noise}}}$, where the *power* may have different definitions depending on the type of objects and the problem considered (for matrices, it is commonly defined as the squared norm). In any case, the signal-to-noise ratio measures the distinguishability of the signal from the noise. It is therefore natural to ask the following questions.

- What is the minimum signal-to-noise ratio value at which the information in X is theoretically sufficient for the recovery of S?
- What is the minimum signal-to-noise ratio value at which S can be efficiently reconstructed from X?
- At a given signal-to-noise ratio, what is the best reconstruction performance achievable with $\hat{S} = f(X)$?

At first glance, it may seem that the first two questions should have the same answer. That is, if it is *statistically* possible to estimate S from X as soon as $\rho > \rho_{\text{stat}}$, then it should also be possible to design an algorithm which is able to do so as soon as $\rho > \rho_{\text{stat}}$. However, new inference problems of the big data era face a *computational* barrier: there are various problems for which it is fundamentally impossible for an algorithm to recover efficiently (i.e., in polynomial time) the information of interest if $\rho < \rho_{\text{comp}}$. This exhibits two fundamental thresholds: a statistical threshold ρ_{stat} and a computational threshold $\rho_{\text{comp}} \geqslant \rho_{\text{stat}}$. In some problems, they are equal, meaning that there exists an algorithm able to (partially) reconstruct S from X in polynomial time as soon as this is statistically possible. Yet, there exists a myriad of problems displaying a "hard phase" where the inference is statistically possible but cannot be performed in polynomial time, this is referred to as a *computational-to-statistical gap* (Bandeira et al., 2018; Zdeborová and Krzakala, 2016; Gamarnik et al., 2022).



The two thresholds ρ_{stat} and ρ_{comp} indicate the positions of abrupt changes in the complexity of the problem. The latter are usually called *phase transitions* — a term borrowed from statistical physics, which have many deep connections with statistical inference.

The third question is concerned with the quality of the reconstruction $\hat{S}$ achieved with a certain algorithm f. Its performance is expected to be an increasing function of the signal-to-noise ratio but

its precise characterization would be an asset in understanding the inner workings of algorithms and comparing them to potential information-theoretic optimal curves. Although values of practical interest are only for $\rho \geqslant \rho_{\text{comp}}$ since the estimation of S is not feasible below the computational threshold, this function should also have definite values in the hard phase $\rho_{\text{stat}} < \rho < \rho_{\text{comp}}$, representing the theoretically-achievable reconstruction performance. In "practice", the latter can be attained by initializing the algorithm near the sought solution.

Let us use an example to illustrate what has just been said. Consider the observation of an order-3 $n \times n \times n$ symmetric tensor $\mathcal{T}$ which is known to be the noisy observation of a rank-one signal,

$$\mathcal{T} = \beta \boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x} + \frac{1}{\sqrt{n}} \mathcal{N}, \tag{1.2}$$

where $\beta > 0$, $\boldsymbol{x} \in \mathbb{S}^{n-1}$ and $\mathcal{N}$ is a random symmetric tensor from the Gaussian orthogonal ensemble, i.e., it has a density g given by

$$g(\mathcal{N}) = \frac{1}{Z_n} e^{-\frac{1}{2}\|\mathcal{N}\|_{\text{F}}^2}$$

with Z_n such that $\int g(\mathcal{N}) \mathrm{d}\mathcal{N} = 1$ (integration is over the set of symmetric tensors). More specifically, for all $i, j, k \in \{1, \dots, n\}$ such that $i < j < k$,

$$\mathcal{N}_{i,i,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad \mathcal{N}_{i,i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{3}\right), \quad \mathcal{N}_{i,j,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{6}\right),$$

and $\mathcal{N}_{\sigma(u,v,w)} = \mathcal{N}_{u,v,w}$ for all permutation $\sigma \in \mathfrak{S}_3$ and $u, v, w \in \{1, \dots, n\}$.

The inference problem consists in reconstructing the *planted* signal $(\beta, \boldsymbol{x})$ from the observation $\mathcal{T}$. Therefore, we consider the maximization of the likelihood associated with Model (1.2), which reduces to the best rank-one approximation problem

$$(\hat{\beta}, \hat{\boldsymbol{x}}) \in \underset{\gamma > 0,\ \boldsymbol{y} \in \mathbb{S}^{n-1}}{\operatorname{argmin}} \left\| \mathcal{T} - \gamma \boldsymbol{y}^{\otimes 3} \right\|_{\text{F}}^2. \tag{1.3}$$

Here, the *power* of the signal and the noise should be understood as the squared operator norms[2] $\|\beta \boldsymbol{x}^{\otimes 3}\|^2 = \beta^2$ and $\|\frac{1}{\sqrt{n}}\mathcal{N}\|^2$. Since the latter is expected to concentrate as $n \to +\infty$ (Tomioka and Suzuki, 2014), the signal-to-noise ratio is proportional to β^2. Hence, we ask the following questions:

- statistical threshold — what is the minimal value of β at which Problem (1.3) has a solution positively correlated with $(\beta, \boldsymbol{x})$?
- computational threshold — what is the minimal value of β at which this solution can be computed in polynomial time?
- how good is the reconstruction $\hat{\beta}\hat{\boldsymbol{x}}^{\otimes 3}$? i.e., how close is $\hat{\beta}$ to β and how correlated is $\hat{\boldsymbol{x}}$ to $\boldsymbol{x}$?

These questions find answers in several works, among which we notably refer to Jagannath et al. (2020); Ben Arous et al. (2020); Goulart et al. (2022), respectively for each question. Summarizing their results very shortly, Problem (1.3) presents a statistical-to-computational gap, i.e., there exists a hard phase $[\beta_{\text{stat}}, \beta_{\text{comp}}]$ where it is statistically possible to reconstruct $(\beta, \boldsymbol{x})$ from $\mathcal{T}$ but no known algorithm is able to do so in polynomial time. Contrary to β_{stat}, the position of the computational threshold diverges with n. As of now, the best algorithms compute $(\hat{\beta}, \hat{\boldsymbol{x}})$ above a threshold behaving like $Cn^{1/4}$ as $n \to +\infty$, and it is conjectured that no algorithm can do better.

Relying on powerful random matrix tools, Goulart et al. (2022) precisely characterize the value of $\hat{\beta}$ and the alignment $\langle \boldsymbol{x}, \hat{\boldsymbol{x}} \rangle^2$ in the limit where $n \to +\infty$. Their results are depicted in Figure 1.2

[2] The spectral norm of an order-3 tensor $\mathcal{X}$ is $\|\mathcal{X}\| = \sup_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w} \neq \boldsymbol{0}} \frac{|\mathcal{X}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})|}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|\|\boldsymbol{w}\|}$. This naturally extends to any order $d \geqslant 2$.

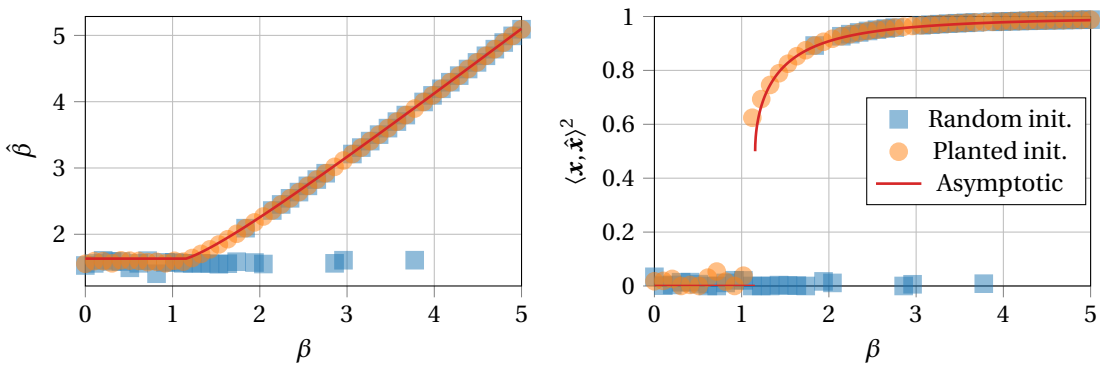


Figure 1.2: Reconstruction performance of the maximum likelihood estimation (1.3) as a function of β (controlling the signal-to-noise ratio). Red curves are the values predicted by Goulart et al. (2022) in the large-dimensional regime $n \to +\infty$. They are compared with simulation results at $n = 100$ of a tensor power iteration initialized at random uniformly on the sphere (blue squares) or at the planted signal (orange circles).

(red curves) and compared to simulation results when $n = 100$ obtained with a tensor power iteration (De Lathauwer et al., 2000a), estimating a solution to Problem (1.3), initialized at random with a uniform distribution on the sphere $\mathbb{S}^{n-1}$ (blue squares) or at the planted vector $\boldsymbol{x}$ (orange circles).

If β is too small, it is theoretically impossible to recover any signal: $\hat{\beta} = \mathbb{E}[\|\frac{1}{\sqrt{n}}\mathcal{N}\|] = \frac{4}{\sqrt{6}}$ (the plateau on the left part of Figure 1.2) and $\langle \boldsymbol{x}, \hat{\boldsymbol{x}} \rangle^2 = 0$, the observed tensor $\mathcal{T}$ is statistically indistinguishable from the noise $\frac{1}{\sqrt{n}}\mathcal{N}$. This suddenly changes as soon as β crosses the value $\frac{2}{\sqrt{3}}$: $\hat{\beta}$ becomes greater than $\mathbb{E}[\|\frac{1}{\sqrt{n}}\mathcal{N}\|]$ and $\hat{\boldsymbol{x}}$ is positively correlated with $\boldsymbol{x}$. In particular, the value of $\langle \boldsymbol{x}, \hat{\boldsymbol{x}} \rangle^2$ presents a discontinuity at $\beta = \frac{2}{\sqrt{3}}$, it abruptly jumps from 0 to $\frac{1}{2}$. Yet, just above this critical value of β, the power iteration is not able to reconstruct the planted signal and converges to an *uninformative* optimum, unless it is initialized very near the sought solution (which, obviously, is not possible in practice). This is the hard phase: although an informative solution exists, its computation is too hard and we are bound to fall into an uninformative one. As β increases, the power iteration eventually converges to the informative optimum — it is the easy phase.

Remark 1.1 (Phase transitions). Strictly speaking, phase transitions do not exist at finite system size (here, n). Indeed, they are characterized by abrupt changes (in statistical physics, they correspond to non-analyticities in the free energy density) which can only appear when the limit $n \to +\infty$ is taken. Thus, we do not observe a sharp transition from the hard phase to the easy phase in the finite-n simulations of Figure 1.2.

Remark 1.2 (Statistical threshold). In fact, Jagannath et al. (2020) reveal that the behavior of the maximum likelihood estimator $(\hat{\beta}, \hat{\boldsymbol{x}})$ near the statistical threshold is trickier than what is presented above. There are two close but distinct values $\beta_s = \frac{2}{\sqrt{3}} \approx 1.1547$ and $\beta_c \approx 1.2066$ such that, for values of β lying between them, the informative optimum of Problem (1.3) is not the global one. In other words, if $\beta < \beta_s$, there is no informative optimum, if $\beta > \beta_c$, the informative optimum is the global one, but if $\beta_s < \beta < \beta_c$, there exists an informative *local* optimum, although the global optimum (the maximum likelihood estimator) is not informative. Strictly speaking, the statistical threshold corresponds to β_c (be careful that the subscripts s and c do not refer to "statistical" and "computational") and β_s is a "sub-statistical" threshold. The results of Goulart et al. (2022) specify the reconstruction performance of this informative optimum as soon as $\beta \geqslant \beta_s$ but they do not give any insight into the value of β_c.

1.2 Deciphering Large-Multidimensional Statistics with Random Matrix Theory

A central tool to understand the behavior of large and multidimensional arrays of data, such as those mentioned above, is the theory of large random matrices. The study of these objects and, in particular, their eigenvalues and eigenvectors emerged in the early days of statistical sciences with Wishart (1928) who was interested in empirical covariances. Then, it resurfaced thirty years later when the physicist Wigner (1955, 1958) used them as statistical models for heavy nuclei atoms — for which he was awarded the Nobel prize in physics in 1963. Since then, the field has known a wide expansion in various scientific domains such as statistics (El Karoui, 2010), finance (Potters and Bouchaud, 2020, Chapter 20), quantum mechanics (Guhr et al., 1998), free probability (Voiculescu, 1991). It is even linked with the Riemann hypothesis as the zeros of the zeta function on the critical line $\frac{1}{2} + \mathrm{i}t$ are believed to share the same distribution as that of the eigenvalues of a GOE matrix inside the bulk (Montgomery, 1973; Keating, 2002).

In a nutshell, random matrix theory is concerned with the behavior of eigenvalues and eigenvectors of matrices with random entries at *macroscopic* and *microscopic* scales. While the former involves questions regarding the limiting eigenvalue distribution, linear statistics of eigenvalues or eigenvectors, outlying eigenvalues, etc., the latter deals with the limiting distribution of a specific eigenvalue, the point process governing the position of its neighbors and its link with the corresponding eigenvectors, etc. The present work is only concerned with the study of the proposed models at the macroscopic scale, that is, the *global regime* (as opposed to the *local regime*).

1.2.1 Limiting Spectral Distribution of Random Matrices

The spectrum of the empirical covariance matrix (Figure 1.1) was characterized by Marčenko and Pastur (1967). In order to study the asymptotic behavior of the *empirical spectral distribution* $\mu_p = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(\hat{\boldsymbol{C}})}$ they exploited a powerful tool, namely its *Stieltjes transform*,

$$m_p : z \in \mathbb{C} \setminus \mathbb{R} \mapsto \int_{\mathbb{R}} \frac{\mathrm{d}\mu_p(\lambda)}{\lambda - z}.$$

Given the particular expression of μ_p, the expression of its Stieltjes transform simplifies into $m_p(z) = \frac{1}{p}\sum_{i=1}^{p} (\lambda_i(\hat{\boldsymbol{C}}) - z)^{-1}$. As $p, n \to +\infty$ with $p/n \overset{\text{def}}{=} c \in]0, +\infty[$, Marčenko and Pastur (1967) showed the almost sure pointwise convergence of m_p to a Stieltjes transform m_{MP} such that

$$z c m_{\mathrm{MP}}^2(z) - (1 - c - z) m_{\mathrm{MP}}(z) + 1 = 0 \qquad \text{for all } z \in \mathbb{C} \setminus \mathbb{R}, \tag{1.4}$$

which implies the almost sure weak convergence of μ_p to μ_{MP} given by the *inverse Stieltjes transform* of m_{MP}. This limiting measure is nowadays known as the Marčenko-Pastur distribution with parameter c and is expressed as

$$\mu_{\mathrm{MP}} = \left[1 - \frac{1}{c}\right]^+ \delta_0 + \nu \qquad \text{with} \qquad \mathrm{d}\nu(\lambda) = \frac{\sqrt{[\lambda - E_-]^+ [E_+ - \lambda]^+}}{2\pi c \lambda} \mathrm{d}\lambda \tag{1.5}$$

where $E_\pm = (1 \pm \sqrt{c})^2$.

The density of the Marčenko-Pastur distribution is plotted in Figure 1.3 with the same empirical spectral distributions (histograms) as in Figure 1.1. Notice that the particular case $p/n = 1$ presents a *hard edge* as $\lambda \downarrow 0$ which is due to a singularity in the limiting density,

$$\frac{\sqrt{[\lambda]^+ [4 - \lambda]^+}}{2\pi\lambda} = \frac{1}{\pi\sqrt{\lambda}} + \underset{\lambda \downarrow 0}{\mathrm{o}}(1).$$

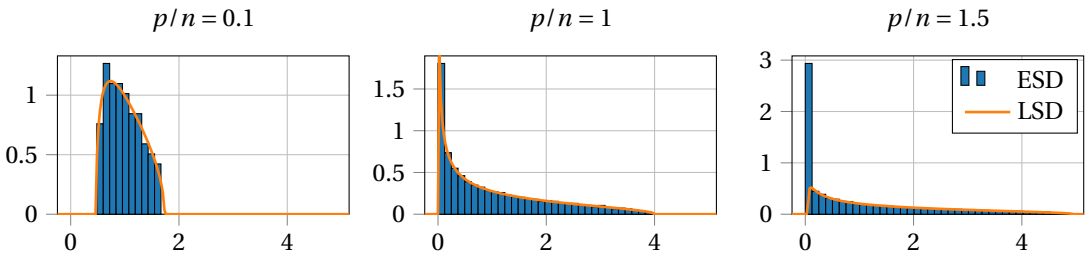


Figure 1.3: Empirical spectral distribution (ESD) and limiting spectral distribution (LSD) of the sample covariance matrix $\hat{\boldsymbol{C}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ with $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$, $n = 1000$ and, from left to right, $p/n = 0.1, 1, 1.5$.

This hard edge scenario marks a transition from the absence ($c < 1$) to the presence ($c > 1$) of an atom around 0 in the Marčenko-Pastur distribution.

In fact, the "Stieltjes transform approach" is a very versatile method to determine the limiting spectral distribution of a random matrix model and its was extensively used by Silverstein and Bai (1995); Bai and Silverstein (1998, 2010). It usually results in a fixed-point equation of the form $m(z) = f(m(z))$ which allows to numerically compute m and to reconstruct the corresponding probability distribution μ. In the case of the sample covariance matrix considered by Marčenko and Pastur (1967), f is simple (see Equation (1.4)) and analytic expressions of m_{MP} as well as μ_{MP} can be found.

1.2.2 Outlying Eigenvalues and Eigenvectors in Spiked Models

In a statistical data analysis context, we are not interested in "mere random matrices" (although they are interesting mathematical objects) but rather in related *spiked models*. This terminology refers to low-rank perturbations of (usually standard) random matrix models.

For example, Baik and Silverstein (2006), following a model initially proposed by Johnstone (2001), were interested in the eigenvalue distribution of the sample covariance matrix $\hat{\boldsymbol{C}}$ when the underlying true covariance is $\boldsymbol{C} = \boldsymbol{I}_p + \boldsymbol{P}$ where $\boldsymbol{P}$ has a small rank compared to p, n (usually, it is kept fixed while $p, n \to +\infty$). Thus, $\hat{\boldsymbol{C}} = \frac{1}{n}(\boldsymbol{I}_p + \boldsymbol{P})^{\frac{1}{2}} \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{I}_p + \boldsymbol{P})^{\frac{1}{2}\top}$ where $\boldsymbol{X}$ is a $p \times n$ matrix with i.i.d. entries having zero mean, unit variance and finite fourth moment[3]. It is a low-rank perturbation of the original random matrix model $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ whose empirical spectral distribution is known to converge weakly almost surely to the Marčenko-Pastur distribution. The limiting spectral distribution of $\hat{\boldsymbol{C}}$ is the same but, depending on the "strength" (the eigenvalues) of the perturbation $\boldsymbol{P}$, a finite number of eigenvalues of $\hat{\boldsymbol{C}}$ may be isolated on the right-side of the *bulk* described by ν in Equation (1.5). Paul (2007) further precised that the corresponding eigenvectors are correlated with those of $\boldsymbol{P}$ in a manner which also depends on the strength of the perturbation.

Later, Benaych-Georges and Nadakuditi (2012) studied a different spiked model, which is of particular interest to us, namely $\boldsymbol{X} = \boldsymbol{P} + \boldsymbol{N}$ where $\boldsymbol{P}$ is a low-rank matrix modeling a *signal* and $\boldsymbol{N}$ is a random matrix modeling a *noise*. With a free probability approach, they characterize the asymptotic behavior of the largest and smallest singular values as well as that of their corresponding singular vectors. This is equivalent to considering the eigenvalues and eigenvectors of the sample covariance matrix $\hat{\boldsymbol{C}} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ or the Gram kernel matrix $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$. Given that the noise has, e.g., i.i.d. Gaussian $\mathcal{N}(0,1)$ entries, both $\hat{\boldsymbol{C}}$ and $\boldsymbol{K}$ have their empirical spectral distributions converging weakly almost surely to a Marčenko-Pastur distribution (with parameters c and c^{-1} respectively). Moreover, depend-

[3] This last assumption is needed for technical reasons, see Theorem 2.31.

ing on the strength (the singular values) of the perturbation $\boldsymbol{P}$, there may be a few isolated eigenvalues with corresponding eigenvectors correlated with the singular vectors of $\boldsymbol{P}$.

In both cases just discussed, the spiked model has the same limiting spectral distribution as the original unperturbed model. The only influence that has the addition of a low-rank perturbation $\boldsymbol{P}$ is that it may cause the emergence of a few isolated eigenvalues whose positions (and presence) depend on the strength of the perturbation. Moreover, some information regarding the singular subspaces of $\boldsymbol{P}$ are also contained in the corresponding eigenvectors. In a statistical inference setting, spiked models are particularly relevant objects of study to understand the main limiting factors in the reconstruction of the signal $\boldsymbol{P}$ from the observation $\boldsymbol{X}$ and it can thus provide answers to the questions raised in Section 1.1.3. Here, it is the singular values of $\boldsymbol{P}$ which control the signal-to-noise ratio.

Remark 1.3 (No prior information on $\boldsymbol{P}$). Some studies on spiked models, such as, e.g., Lelarge and Miolane (2017), consider that the perturbation $\boldsymbol{P}$ is also random and follows a certain prior distribution. We shall not be concerned with such considerations: in our settings, no prior information is given on $\boldsymbol{P}$ besides its structure. Concretely, if $\boldsymbol{P}$ is a rank-one symmetric matrix of the form $\beta \boldsymbol{x}\boldsymbol{x}^\top$ with $\|\boldsymbol{x}\| = 1$, then we assume no further knowledge on the direction $\boldsymbol{x}$ (which is equivalent to saying that it follows a uniform prior on $\mathbb{S}^{n-1}$) and the scale β becomes a parameter of the problem.

As an illustration of the interest of spiked models for statistical learning, consider the task of separating n individuals represented by p-dimensional feature vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ into two clusters. This is a very general problem as the $\boldsymbol{x}_i$'s could represent, e.g., images of cats and dogs, healthy and sick patients or sentences about sports and politics. Assuming, for simplicity, that each cluster has mean $\pm\boldsymbol{\mu} \in \mathbb{R}^p$ (which states that the origin $\boldsymbol{0}_p$ is in the middle of both clusters), the feature vectors can be expressed as

$$\boldsymbol{x}_i = j_i \boldsymbol{\mu} + \boldsymbol{n}_i$$

where $j_i = \pm 1$ depending on the cluster which $\boldsymbol{x}_i$ belongs to and $\boldsymbol{n}_i \in \mathbb{R}^p$ represents the dispersion of $\boldsymbol{x}_i$ around $j_i\boldsymbol{\mu}$. The $\boldsymbol{n}_i$'s are modeled as i.i.d. $\mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$ random vectors (this choice is discussed in detail in Section 2.2.3). Hence, the *data matrix* $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{p\times n}$ decomposes into

$$\boldsymbol{X} = \boldsymbol{\mu}\boldsymbol{j}^\top + \boldsymbol{N} \tag{1.6}$$

where $\boldsymbol{j} = \begin{bmatrix} j_1 & \ldots & j_n \end{bmatrix}^\top$ and $\boldsymbol{N}$ is a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries. This is precisely a signal-plus-noise model. Since we want to perform clustering of the $\boldsymbol{x}_i$'s, our goal is to reconstruct $\boldsymbol{j}$ from $\boldsymbol{X}$. Thus, we consider the Gram kernel matrix $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$.

A realization of the spectrum of $\boldsymbol{K}$ is depicted in the left panel of Figure 1.4 when the first half of the $\boldsymbol{x}_i$'s is assigned to the "-1 cluster" while the other half is assigned to the "$+1$ cluster" and $\|\boldsymbol{\mu}\|^2 = 4$ (note that $\|\boldsymbol{\mu}\|^2$ controls the signal-to-noise ratio $\|\boldsymbol{\mu}\boldsymbol{j}^\top\|^2/\mathbb{E}[\|\boldsymbol{N}\|^2]$). Because the addition of a low-rank perturbation does not change the limiting spectral distribution, it is not surprising to see $\frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(\boldsymbol{K})}$ approaching the Marčenko-Pastur distribution (here, with parameter $\frac{1}{2}$). The only display of the rank-one perturbation $\boldsymbol{\mu}\boldsymbol{j}^\top$ is in the fact that the dominant eigenvalue $\lambda_1(\boldsymbol{K})$ has left the bulk. Yet, the latter has a negligible weight in the empirical spectral distribution as $p, n \to +\infty$. Thus, we must introduce a new tool, borrowed to the theory of linear operator in Hilbert space (Akhiezer and Glazman, 1993): the *resolvent matrix*. Specifically, the resolvent of $\boldsymbol{K}$ is defined for all $z \in \mathbb{C} \setminus \operatorname{Sp}\boldsymbol{K}$ as

$$\boldsymbol{Q}_{\boldsymbol{K}}(z) = (\boldsymbol{K} - z\boldsymbol{I}_n)^{-1}.$$

The isolated eigenvalue is characterized by the fact that it is the only point in $\mathbb{C} \setminus [E_-, E_+]$ for which $\boldsymbol{Q}_{\boldsymbol{K}}(z)$ is undefined as $p, n \to +\infty$ with $p/n \overset{\text{def}}{=} c \in]0, +\infty[$. Hence, a careful asymptotic analysis of the singular points of the resolvent shows that

$$\lambda_1(\boldsymbol{K}) \xrightarrow[p,n\to+\infty]{\text{a.s.}} \xi \overset{\text{def}}{=} \begin{cases} \left(1+\sqrt{c^{-1}}\right)^2 & \text{if } \|\boldsymbol{\mu}\|^2 \leqslant \sqrt{c} \\ \frac{(\|\boldsymbol{\mu}\|^2+1)(\|\boldsymbol{\mu}\|^2+c)}{c\|\boldsymbol{\mu}\|^2} & \text{if } \|\boldsymbol{\mu}\|^2 > \sqrt{c} \end{cases}. \tag{1.7}$$

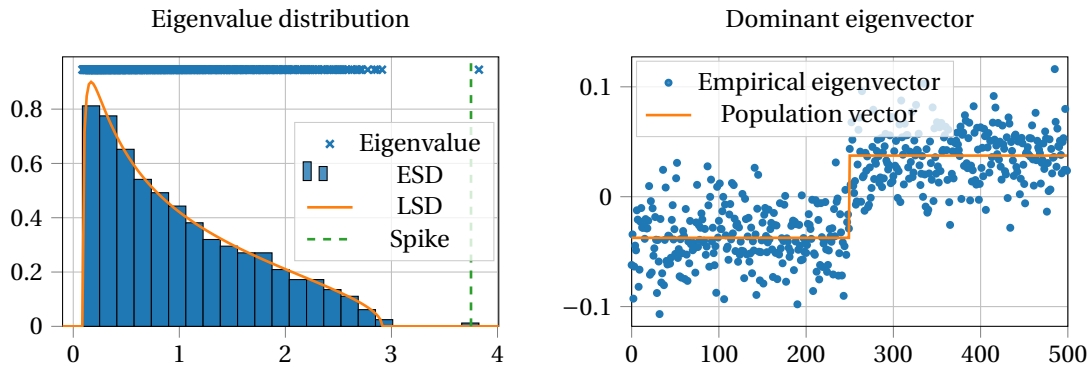


Figure 1.4: **Left:** Empirical spectral distribution (ESD) and limiting spectral distribution (LSD) of $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$ with $\boldsymbol{X}$ following Equation (1.6). The green dashed line indicates the almost sure asymptotic position of an isolated eigenvalue ξ defined in Equation (1.7). **Right:** Coordinates of the dominant eigenvector of $\boldsymbol{K}$ compared with the population vector $\sqrt{\zeta/n}\,\boldsymbol{j}$ with ζ defined in Equation (1.8). **Experimental setting:** $p = 1000$, $n = 500$, $\|\boldsymbol{\mu}\|^2 = 4$, equal class sizes.

The asymptotic position ξ depends only on $\|\boldsymbol{\mu}\|^2$ (the strength of the perturbation) and the parameter c. If $\|\boldsymbol{\mu}\|^2 \leqslant \sqrt{c}$ (the signal-to-noise ratio is too small), then ξ is simply the right edge of the bulk E_+. On the other hand, as soon as $\|\boldsymbol{\mu}\|^2 > \sqrt{c}$ (the signal-to-noise ratio is large enough), the dominant eigenvalue escapes the bulk, $\xi > E_+$. Its position is represented by the green dashed line.

The coordinates of the eigenvector $\boldsymbol{u}_1(\boldsymbol{K})$ associated to $\lambda_1(\boldsymbol{K})$ are plotted in the right panel of Figure 1.4. We can see that they are correlated with the sought vector $\boldsymbol{j}$ (recall that $j_i = -1$ if $i \leqslant n/2$ and $j_i = +1$ otherwise). Indeed, this eigenvector carries some information about the perturbation $\boldsymbol{\mu}\boldsymbol{j}^\top$. More precisely, it is aligned with the single eigenvector of $\frac{1}{n}[\boldsymbol{\mu}\boldsymbol{j}^\top]^\top[\boldsymbol{\mu}\boldsymbol{j}^\top] = \frac{\|\boldsymbol{\mu}\|^2}{n}\boldsymbol{j}\boldsymbol{j}^\top$, namely $\frac{1}{\sqrt{n}}\boldsymbol{j}$. The resolvent formalism is also useful here, in that it allows to compute these alignments: denoting $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1(\boldsymbol{K}), \ldots, \lambda_n(\boldsymbol{K})) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{U} = \big[\boldsymbol{u}_1(\boldsymbol{K}) \quad \ldots \quad \boldsymbol{u}_n(\boldsymbol{K})\big] \in \mathcal{O}_n(\mathbb{R})$, we have,

$$\boldsymbol{Q}_{\boldsymbol{K}}(z) = \big(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top - z\boldsymbol{I}_p\big)^{-1} = \boldsymbol{U}\big(\boldsymbol{\Lambda} - z\boldsymbol{I}_p\big)^{-1}\boldsymbol{U}^\top = \sum_{i=1}^{n} \frac{\boldsymbol{u}_i(\boldsymbol{K})\boldsymbol{u}_i(\boldsymbol{K})^\top}{\lambda_i(\boldsymbol{K}) - z},$$

and, choosing a positively-oriented simple closed complex contour γ circling around $\lambda_1(\boldsymbol{K})$ and leaving all the other eigenvalues outside, Cauchy's integral formula (Proposition 2.15) yields,

$$\langle \boldsymbol{a}, \boldsymbol{u}_1(\boldsymbol{K})\rangle^2 = -\frac{1}{2\mathrm{i}\pi}\oint_\gamma \boldsymbol{a}^\top \boldsymbol{Q}_{\boldsymbol{K}}(z)\boldsymbol{a}\,\mathrm{d}z$$

for all deterministic vector $\boldsymbol{a} \in \mathbb{R}^n$. Again, a careful asymptotic analysis of the previous integral when $\boldsymbol{a} = \frac{1}{\sqrt{n}}\boldsymbol{j}$ shows that

$$\frac{1}{n}\big\langle \boldsymbol{j}, \boldsymbol{u}_1(\boldsymbol{K})\big\rangle^2 \xrightarrow[p,n\to+\infty]{\text{a.s.}} \zeta \overset{\text{def}}{=} \begin{cases} 0 & \text{if } \|\boldsymbol{\mu}\|^2 \leqslant \sqrt{c} \\ 1 - \frac{\|\boldsymbol{\mu}\|^2 + c}{\|\boldsymbol{\mu}\|^2(\|\boldsymbol{\mu}\|^2+1)} & \text{if } \|\boldsymbol{\mu}\|^2 > \sqrt{c} \end{cases}. \tag{1.8}$$

As for ξ, the quantity ζ only depends on $\|\boldsymbol{\mu}\|^2$ and c. Moreover, the alignment of $\boldsymbol{u}_1(\boldsymbol{K})$ with $\boldsymbol{j}$ is non-zero as soon as $\lambda_1(\boldsymbol{K})$ leaves the bulk ($\|\boldsymbol{\mu}\|^2 > \sqrt{c}$). Notice that ζ is an increasing function of $\|\boldsymbol{\mu}\|^2$ on $[\sqrt{c}, +\infty[$ starting at 0 (no alignment) and approaching 1 (perfect alignment) as $\|\boldsymbol{\mu}\|^2 \to +\infty$.

Thus, the resolvent matrix allows the study of spiked models by describing the asymptotic position of isolated eigenvalues and the alignment of the corresponding eigenvectors with the sought information (the eigenspace of the perturbation). We should also mention the following elementary property of the resolvent,

$$\frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_{\boldsymbol{K}}(z)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i(\boldsymbol{K})-z}=\int_{\mathbb{R}}\frac{1}{\lambda-z}\,\mathrm{d}\left[\frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i(\boldsymbol{K})}\right](\lambda).$$

In words, *the normalized trace of the resolvent is the Stieltjes transform of the empirical spectral distribution*. Hence, the resolvent is the central tool we shall be concerned with in order to study both the limiting spectral distribution (via the Stieltjes transform) as well as the spike eigenvalues and eigenvectors of our models.

In light of the questions raised in Section 1.1.3, the previous analysis of the spiked model $\boldsymbol{X}=\boldsymbol{\mu}\boldsymbol{j}^{\top}+\boldsymbol{N}$ gave insightful results regarding the original clustering problem. Firstly, it exhibited a threshold value $\|\boldsymbol{\mu}\|^2=\sqrt{c}$ marking a transition between two distinct phases of the inference problem. If $\|\boldsymbol{\mu}\|^2\leqslant\sqrt{c}$, the statistical information contained in $\boldsymbol{X}$ is not sufficient to reconstruct $\boldsymbol{j}$ and spectral clustering fails to properly separate the $\boldsymbol{x}_i$'s ($\zeta=0$). Conversely, if $\|\boldsymbol{\mu}\|^2>\sqrt{c}$, the strength of the signal is sufficient to distinguish it from the noise and the partial reconstruction of $\boldsymbol{j}$ from the observation $\boldsymbol{X}$ is possible ($\zeta>0$). Moreover, the "amount of accessible information" in $\boldsymbol{X}$ is explicitly given by ζ as a function of $\|\boldsymbol{\mu}\|^2$. Note that spectral clustering is information-theoretically optimal in this setting (Onatski et al., 2013; Löffler et al., 2021), meaning that it achieves the best clustering performance as soon as it is theoretically possible. In other words, there is no *hard phase* and we jump directly from the *impossible phase* to the *easy phase* at $\|\boldsymbol{\mu}\|^2=\sqrt{c}$. The study of such spiked random matrix models related to clustering problems is the main topic of Part I of this thesis.

1.2.3 Random Tensors and Random Matrices

All the tools introduced so far are designed for the study of large random *matrix* models. Yet, modern challenges in statistical learning concern not only large but also *multidimensional* data. Thus, we are also interested in large random *tensors*. The statistical analysis of the latter is rather new — it essentially started with the work of Richard and Montanari (2014) — and there is still no proper "random tensor theory". Nonetheless, we can rely on the powerful concepts and tools of the well-established random matrix theory to provide a better theoretical understanding of statistical inference on tensor data.

The careful reader may have noticed that the random tensor model in Equation (1.2) corresponds to what we could (and will) call a *spiked tensor model*. Once again, we shall consider signal-plus-noise models: the observed tensor is expressed as $\mathcal{T}=\mathcal{P}+\mathcal{N}\in\mathbb{R}^{n_1\times\ldots\times n_d}$, modeling a low-rank signal $\mathcal{P}$ perturbed by an additive noise $\mathcal{N}$. Reconstructing $\mathcal{P}$ from the observation $\mathcal{T}$ consists in finding a solution to

$$\min_{\mathcal{X}\in\mathscr{M}}\|\mathcal{T}-\mathcal{X}\|_{\mathrm{F}}^2 \tag{1.9}$$

where $\mathscr{M}$ is a set of low-rank tensors which shall be properly defined in Section 2.4.2. How can we tackle this optimization problem involving a random tensor with our set of tools from the theory of large random matrices?

A first idea is to study the *unfolded tensor*, that is, a matrix constructed by "flattening" the original tensor $\mathcal{T}$. This operation is usually performed along a given mode. For example, the unfolding of $\mathcal{T}$ along the first mode, denoted $\boldsymbol{T}^{(1)}$, is an $n_1\times\prod_{\ell=2}^{d}n_\ell$ matrix whose i-th line is made of all the entries of $\mathcal{T}$, in a certain predefined order, obtained by fixing the first index to i. Therefore, we are left to study a random matrix instead of a random tensor. However, this operation comes at a high cost, for it essentially wipes out the structural information in the data, which was precisely our main motivation

for considering multidimensional arrays! Moreover, as the tensor is "large", we assume that all dimensions grow at a similar rate $n_1, \dots, n_d = \Theta(N) \to +\infty$. Hence, the second dimension of $\boldsymbol{T}^{(1)}$ grows at a faster polynomial rate $\Theta(N^{d-1})$ than the first one $\Theta(N)$, which is different from the standard random matrix regime where both dimensions have similar sizes.

In fact, the structural information can be preserved if we consider all the unfoldings $\boldsymbol{T}^{(1)}, \dots, \boldsymbol{T}^{(d)}$ as a whole. Hence, instead of a random tensor we study a set of *d long* random matrices (Ben Arous et al., 2023). This approach will allow us to analyze the reconstruction performance of standard techniques designed to provide low-rank tensor approximations, namely a truncated version of the multilinear singular value decomposition (De Lathauwer et al., 2000b) — which is an extension of the common singular value decomposition for matrices — and the higher order orthogonal iteration (HOOI, De Lathauwer et al., 2000a). Among the various algorithms designed to estimate a solution to Problem (1.9), HOOI is among the ones having the lowest critical signal-to-noise ratio value required for a non-trivial estimation — i.e., with non-zero alignment — of the perturbation $\mathcal{P}$ (Richard and Montanari, 2014; Hopkins et al., 2015). However, it does not reach the information-theoretic lower bound established by Perry et al. (2020) and it is therefore conjectured that this problem exhibits a *hard phase* and that HOOI (among other algorithms) actually reaches the computational threshold (Ben Arous et al., 2019). The "tensor unfolding approach", which we develop in Part II of this thesis, is thus useful to study the reconstruction performance near the transition between the hard and easy phases.

In order to complete the picture, the following natural question concerns the analysis of Problem (1.9) near the statistical threshold, that is, near the transition between the impossible and hard phases. The recent works of Goulart et al. (2022) and Seddik et al. (2022) showed that the study of a spiked tensor model can be reduced to that of an equivalent spiked random matrix model. In particular, Seddik et al. (2022) considered the rank-one asymmetric spiked model consisting in a perturbation $\mathcal{P} = \sigma \boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)} \otimes \boldsymbol{x}^{(3)} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, with $\sigma > 0$ controlling the signal-to-noise ratio and $\boldsymbol{x}^{(\ell)} \in \mathbb{S}^{n_\ell - 1}$, $\ell \in \{1,2,3\}$, observed as $\mathcal{T} = \sigma \boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)} \otimes \boldsymbol{x}^{(3)} + \frac{1}{\sqrt{N}} \mathcal{N}$ where $\mathcal{N}_{i,j,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $N = n_1 + n_2 + n_3$. In this particular setting, Problem (1.9) becomes

$$\min_{\varsigma > 0,\ \boldsymbol{y}^{(\ell)} \in \mathbb{S}^{n_\ell - 1}} \left\| \mathcal{T} - \varsigma \boldsymbol{y}^{(1)} \otimes \boldsymbol{y}^{(2)} \otimes \boldsymbol{y}^{(3)} \right\|_{\mathrm{F}}^2.$$

Given a solution $(\hat{\sigma}, \hat{\boldsymbol{x}}^{(1)}, \hat{\boldsymbol{x}}^{(2)}, \hat{\boldsymbol{x}}^{(3)})$ to this problem, Seddik et al. (2022) are able to recover the asymptotic behavior of $\hat{\sigma}$ and the alignments $\langle \boldsymbol{x}^{(\ell)}, \hat{\boldsymbol{x}}^{(\ell)} \rangle^2$, through, notably, the analysis of an associated symmetric random matrix $\boldsymbol{\Phi}$ defined by blocks as

$$\boldsymbol{\Phi} = \begin{bmatrix} \mathbf{0}_{n_1 \times n_1} & \mathcal{T}(\cdot, \cdot, \hat{\boldsymbol{x}}^{(3)}) & \mathcal{T}(\cdot, \hat{\boldsymbol{x}}^{(2)}, \cdot) \\ \mathcal{T}(\cdot, \cdot, \hat{\boldsymbol{x}}^{(3)})^\top & \mathbf{0}_{n_2 \times n_2} & \mathcal{T}(\hat{\boldsymbol{x}}^{(1)}, \cdot, \cdot) \\ \mathcal{T}(\cdot, \hat{\boldsymbol{x}}^{(2)}, \cdot)^\top & \mathcal{T}(\hat{\boldsymbol{x}}^{(1)}, \cdot, \cdot)^\top & \mathbf{0}_{n_3 \times n_3} \end{bmatrix}.$$

Recall that, as $\mathcal{T}$ describes a 3-linear form on $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^{n_3}$, the contraction of $\mathcal{T}$ on a single vector describes a bilinear form, which we identify as a matrix. This model has a complicated structure because the vectors $\hat{\boldsymbol{x}}^{(1)}, \hat{\boldsymbol{x}}^{(2)}, \hat{\boldsymbol{x}}^{(3)}$, on which $\mathcal{T}$ is contracted, actually depend on the noise $\mathcal{N}$. Yet, relying on the Stieltjes transform and resolvent approach, its analysis is still amenable to precise asymptotic results near the statistical threshold. This interesting approach leading to unusual new random matrix models is a promising direction to complete the results of Part II.

Remark 1.4. In fact, as it was mentioned in Remark 1.2, the results found with this approach are valid for all values of σ such that there exist an informative local optimum, even if this is not the global one. Hence, it predicts positive alignments even (slightly) below the actual statistical threshold! The precise characterization of the latter requires different tools which we shall not discuss here, see Jagannath et al. (2020); Zdeborová and Krzakala (2016).

1.3 Practical, Societal and Environmental Challenges of the Big Data Era

Before delving further into mathematical considerations, let us take a step back in order to reflect on the physical reality of what is at stake here. This (relatively short) digression is meant to serve as a warning against potential excessive usage of the technology, on which machine learning inherently relies, in the context of drastic environmental changes caused by human activities since the industrial revolution. We do not intend to argue in favor or against the use of machine learning, nor to propose solutions, but simply to provide some elements of the background behind the mathematical questions in this thesis. For a general presentation of the current trends of global warming, we naturally refer to the *2023 Synthesis Report on Climate Change* of the Intergovernmental Panel on Climate Change (IPCC).

In 2014, at the request of the United Nations Secretary-General, the *Independent Expert Advisory Group on a Data Revolution for Sustainable Development* prepared a report gathering recommendations on how the data revolution can be mobilized for sustainable development. In particular, the introducing words of its executive summary state:

> Data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time; designing, monitoring and evaluating effective policies becomes almost impossible.

The document shows how the increasing volume of data should be used by policy makers to achieve their "sustainable development goals" announced in the 2030 agenda of the United Nations. Recalling the origins of statistics, the report claims that demographic data allow to "know more about the state of the world". However, it never mentions the high environmental footprint of the data revolution, which undeniably threatens sustainable development (Meadows et al., 1972).

1.3.1 The Physical Reality of Big Data

The past decades have seen a rapid development of Artificial Intelligence (AI), and machine learning in particular. This is allowed by the dazzling enhancement of computational resources and the massive growth of available data, notably via numerous sensors, mobiles phones, digitization of processes and services. The fast processing and intensive storage of data has been ground-breaking in many societal domains — the processing of transport and travel preferences allows to predict and avoid road congestion and accidents, the analysis of consumers' preferences allows more targeted and effective marketing strategies, information about students' progress and learning habits allow the understanding of learning patterns and the design of more effective educational programs, etc.

Yet, the unstoppable pursuit of enhancement of AI models creates a voracious demand in computing power, which has significant consequences on global climate change (Wu et al., 2022). Although data appears to the user as a virtual object stored in a "cloud", it has a very real physical existence, notably via invisible *data centers*, "the central nervous system of the 21st century" (Whitehead et al., 2014). This equipment stores, manages and processes digital data, as well as it provides applications and services for data processing. Thus, data centers consume an increasing amount of energy to run their operations and cool down their servers, resulting in a heavy environmental footprint (Whitehead et al., 2014; Williams, 2011). The growing use of internet, online services and connected objects raises the demand for data and service availability which in turn requires that data be stored in servers in multiple sites and centers. Hence, there are also broader ethical questions about the desirability of such a data revolution requiring attention, see Lucivero (2020).

Moreover, the increasing demand for new hardware puts pressure on the construction of new devices, which also represents a threat for the environment, and human rights as well. Indeed, the min-

ing of ore needed for high-tech devices, benefiting for the most part to the richest people on Earth, is essentially performed in poor countries, where working conditions are less restrictive. The quest for these precious resources causes widespread deforestation and the chemicals used for mineral separation pollute streams and rivers, releasing substances harmful to aquatic life and human health. Children are not exempted from the labor, which is performed under dangerous conditions. Moreover, it often contributes to fomenting tensions in the concerned countries, as in the Democratic Republic of Congo, where the land grab of Chinese companies for coltan mining nourishes a war between rebel groups and the Congolese army, perpetuating a cycle of poverty and violence (Constantine and Wolff, 2023). Manufacturing of new hardware suffers from similar issues as it is also conducted in poor countries (mostly south-east Asia), with a low cost of manpower, in tragic working conditions.

But a larger production of devices also means a larger amount of waste to handle. Here again, poor countries are at the mercy of the richer ones. In 2022, 62 billion kg of electronic waste were produced worldwide, among which only 13.8 billion were formally collected and recycled in an environmentally sound manner (Baldé et al., 2024). The rest is either handled outside formal systems in high- and upper-middle-income countries (16 billion), in low- and lower-middle-income countries (18 billion) or disposed as residual waste (14 billion). The biggest producers of electronic waste are in Europe, Oceania and Americas but they are not those who suffer the most from it. Indeed, the disposal of computing hardware produces harmful emissions and is environmentally detrimental, particularly in low- and middle-income countries, which have lower environmental controls and where recovery of valuable materials such as copper and gold in hardware through practices like incineration results in aggravated environmental pollution (Williams, 2011; Baldé et al., 2024).

The insatiable hunger for new technology and capitalistic growth fuels an environmental and human crisis worldwide. While investment in AI development keeps soaring, research in data science cannot pretend to be independent from this disastrous situation.

1.3.2 Sustainable Big Data Initiatives: Myth or Reality?

In view of this alarming reality, several works propose new tools and methods for a *frugal AI* relying, e.g., on tools monitoring energy consumption (Jay et al., 2023), edge computing promoting locally distributed computations (Angelelli et al., 2023) or compressive learning (Gribonval et al., 2021). In essence, frugal AI seeks to minimize resource usage but under the constraint that it achieves a certain performance. This allows to execute competitive learning tasks under limited memory or computing resources. In other words, performance is no longer the only metric to evaluate a given model, it has to be combined with an evaluation of its need for resources. For example, even if GPT-4 (OpenAI, 2024) is among the best current large language models according to the sole measure of performance, its exorbitant size (1.76 trillion parameters) and the enormous cost of its training[4] certainly place it *among the worst* models according to the resource-usage metric.

However, simply reducing the energy consumption of AI models does not necessarily mean that less energy is spent for the same result after all. In fact, it is more commonplace that this improvement is used to *do more* with the same amount of energy as before — this is the well-known *rebound effect* (Thiesen et al., 2008). Too often, such user-related effects are neglected in assessments of the environmental impacts of information and communication technologies (Pohl et al., 2019). By naively seeking to do more with less, the basic conception of frugal AI may therefore only promote a business-

[4] Although this is not yet known, studies on the previous GPT-3 model give an idea of what this could represent. The training of GPT-3 required 1 287 MWh of energy, that is 552 tons of net CO2eq emissions (Patterson et al., 2021). Knowing that it had "only" 175 billion parameters, a simple application of the rule of three gives an estimated 13 000 MWh of energy (6 000 tons of net CO2eq emissions) for the training of GPT-4, which is equivalent to the energy consumption of almost 3 000 "typical" French households *during a year*! In fact, this quick calculation assumes that the energy consumption depends linearly on the number of parameters, this dependence is more realistically expected to be exponential. Thus the actual training cost should be even bigger.

as-usual scenario, where the central place of technology nourishes a human and environmental crisis. Despite its admirable motivating intentions, frugal AI fails to trigger a profound paradigm shift.

From a broader perspective, AI is not just a very powerful tool which is unfortunately also very energy-demanding — it deeply changes our way to live and relate to others. Relying on high-tech devices, the technology required to run AI models leaves users with a very low power of action on their own tool and it creates a high level of dependence. Moreover, technological devices are so sophisticated that they are often replaced by new ones instead of being repaired, which creates a greater need for resources. In this context, Couillet et al. (2022) promote a *resilient AI* that takes into account a broader set of metrics — in addition to performance and resource usage, it also considers robustness and repairability, power of action of the user, level of dependence to the tool and environment preservation and accessibility. In essence, resilient AI seeks a certain degree of simplicity in the final tool. As a striking example, instead of a black-box neural network working on a computer, a simple decision tree written on a sheet of paper may achieve a (slightly) smaller performance but it is a more versatile tool which gives a much bigger power of action to the user.

1.3.3 The Ambiguous Position of This Work

This thesis is concerned with the mathematical aspects of random matrix and tensor models motivated by machine learning and statistical signal processing applications. After what has been said, the strange situation in which the reader may find himself with regard to the latter gives a glimpse of the psychological context in which the author was during the preparation of this thesis. Nevertheless, it is a crying passion for mathematics which carried it to this result, in the hope that it will do more good than harm at a time where information could become the new deadly weapon of the 21st century (Nguyen Hoang and El Mhamdi, 2019; Nguyen Hoang and Fourquet, 2024).

Now I am become Death, the destroyer of worlds.
Bhagavad Gita, Chapter 11, Verse 32

1.4 Outline and Contributions

Chapter 2 presents the main technical tools used in the following chapters. In particular, the central tools and results of random matrix theory are stated and followed by an introduction to the "Gaussian method" which is extensively used throughout this thesis. Notably, we present a comprehensive demonstration of the spectral analysis of a standard signal-plus-noise model (convergence to the Marčenko-Pastur distribution, confinement of the spectrum, position of spike eigenvalues and alignments of spike eigenvectors) *in the real case*, which lacks in the current literature. This chapter ends with basic notions on tensors and their decompositions.

Part I: Resource-Efficient Spectral Clustering

Firstly, we consider random matrix models related to spectral clustering. Chapter 3 establishes a central limit theorem on the distribution of the entries of spike eigenvectors. This result is crucial to accurately predict the clustering performance of spectral clustering methods and we rely on it in the following chapters. Its proof relies solely on the rotational invariance of the noise and can easily be extended to any standard spiked model other than the signal-plus-noise model.

In Chapter 4, a new random matrix model — namely, a "banded" Gram kernel matrix — is introduced to study the performances of spectral methods on data stream clustering. Our random matrix analysis allows to characterize the spectral behavior of the model and therefore the detectability of a signal in a data stream. Besides the study of a non-standard random matrix model exhibiting an exotic

spectral behavior, these results give precious insights into the online clustering problem. Then, relying on this analysis, we describe the reconstruction performance of this spectral method and show how, with an astute memory management, it outperforms batch clustering under limited memory constraints. These results are supported by numerical experiments.

Lebeau, Chatelain, and Couillet (2024a) *Asymptotic Gaussian Fluctuations of Eigenvectors in Spectral Clustering* (IEEE Signal Processing Letters)

Lebeau, Couillet, and Chatelain (2022a) *Une analyse par matrices aléatoires du clustering en ligne : comprendre l'impact des limitations en mémoire* (GRETSI)

Lebeau, Couillet, and Chatelain (2022b) *A Random Matrix Analysis of Data Stream Clustering: Coping With Limited Memory Resources* (ICML)

Part II: Algorithms for Tensor Approximation

In the second part of this thesis, we study random tensor models through the random matrices corresponding to their unfoldings. In particular, Chapter 5 proposes an analysis of a simple low-rank tensor approximation which generalizes the truncated singular value decomposition on matrices to higher-order tensors. The main results show that, after a proper rescaling, the unfoldings behave like the well-known spiked Wigner model (see Section 2.2.1). Therefore, we can precisely characterize the reconstruction performance of the truncated *multilinear* singular value decomposition on a signal-plus-noise tensor model. Despite not being optimal on tensors of order $d \geqslant 3$, this *quasi-optimal* solution is easy to compute and is an efficient initialization for a numerical scheme estimating the optimal solution: the higher order orthogonal iteration (De Lathauwer et al., 2000a), whose performance we characterize as well. In particular, we prove that this algorithm is asymptotically optimal, in the sense that its number of iterations converges to 1 as the size of the tensor grows large.

Relying on a similar approach, we study in Chapter 6 the performance of multi-view clustering via a nested matrix-tensor model. This newly-introduced model generalizes the standard rank-one spiked tensor model in that it adds a third term mixing a signal *vector* with a noise *matrix*. Our random matrix analysis of the unfoldings describes under which setting a reconstruction of the signal is possible, depending on the difficulty of the problem and the informativeness of the views. These results are compared with the performance of the maximum likelihood estimator derived by Seddik et al. (2023a). Thereby, we bring to light and precisely quantify the computational-to-statistical gap in this problem.

Lebeau, Couillet, and Chatelain (2023) *HOSVD Tronquée : Analyse d'une Approximation Tensorielle Rapide* (GRETSI)

Lebeau, Chatelain, and Couillet (2025) *A Random Matrix Approach to Low-Multilinear-Rank Tensor Approximation* (JMLR)

Lebeau, Seddik, and Goulart (2024b) *Performance Gaps in Multi-view Clustering under the Nested Matrix-Tensor Model* (ICLR)

Chapter 2

Technical Tools

FUNDAMENTAL random matrix tools and the basic concepts needed for the study of random tensors are presented in this chapter. We start by introducing the basic objects and identities on which we rely throughout this thesis in Section 2.1. Central random matrix results such as the Wigner semi-circle law and the Marčenko-Pastur distribution are presented in Section 2.2. Then, Section 2.3 illustrates the "Gaussian method" on a standard spiked signal-plus-noise model. This case study gives the elementary intuitions behind the analysis of more involved spiked models. Finally, Section 2.4 is a short introduction to tensors as multi-way arrays and their main decompositions, namely the canonical polyadic decomposition (CPD) and the multilinear singular value decomposition (MLSVD).

For a general introduction to random matrices, we refer to any of the following books: Anderson et al. (2009); Bai and Silverstein (2010); Pastur and Shcherbina (2011); Tao (2012); Potters and Bouchaud (2020); Couillet and Liao (2022).

2.1 Elements of Random Matrix Theory

Throughout this section, we consider a random $n \times n$ *symmetric* matrix $\boldsymbol{S}$. The set of all its eigenvalues is its *spectrum* $\operatorname{Sp} \boldsymbol{S} \subset \mathbb{R}$. We denote $\lambda_1(\boldsymbol{S}) \geqslant \ldots \geqslant \lambda_n(\boldsymbol{S})$ its eigenvalues in non-increasing order and $\boldsymbol{u}_1(\boldsymbol{S}), \ldots, \boldsymbol{u}_n(\boldsymbol{S}) \in \mathbb{R}^n$ the columns of $\boldsymbol{U} \in \mathcal{O}_n(\mathbb{R})$ such that $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{\Lambda} = \operatorname{Diag}(\lambda_1(\boldsymbol{S}), \ldots, \lambda_n(\boldsymbol{S}))$.

We start by defining the central object to study the spectral behavior of $\boldsymbol{S}$ in the *global* regime.

Definition 2.1. The resolvent of $\boldsymbol{S}$ is $\boldsymbol{Q}_{\boldsymbol{S}} : z \in \mathbb{C} \setminus \operatorname{Sp} \boldsymbol{S} \mapsto (\boldsymbol{S} - z\boldsymbol{I}_n)^{-1}$.

Notice that $\boldsymbol{Q}_{\boldsymbol{S}}(z)$ is a symmetric (but not Hermitian unless $z \in \mathbb{R} \setminus \operatorname{Sp} \boldsymbol{S}$) matrix. It also has the useful property that its spectral norm is bounded.

Proposition 2.2. *For all* $z \in \mathbb{C} \setminus \operatorname{Sp} \boldsymbol{S}$, $\|\boldsymbol{Q}_{\boldsymbol{S}}(z)\| = 1/\operatorname{Dist}(z, \operatorname{Sp} \boldsymbol{S}) \leqslant 1/|\Im z|$.

Proof. From the definition of the resolvent, $\|\boldsymbol{Q}_{\boldsymbol{S}}(z)\|^2 = \max_{\lambda \in \operatorname{Sp} \boldsymbol{S}} |\lambda - z|^{-2} \leqslant |\operatorname{Dist}(z, \operatorname{Sp} \boldsymbol{S})|^{-2} \leqslant |\Im z|^{-2}$. □

Remark 2.3 (Sequences of matrices). Each time we consider the limiting behavior of a random matrix, we implicitly deal with a *sequence* of random matrices indexed by their size. The same goes for the related objects (resolvent, Stieltjes transform, eigenvalues, etc.). This has the double advantage to ease the notation (by removing the index n) and to be closer to what this large-dimensional regime actually models (in practice, matrices are large but remain finite).

2.1.1 Global Behavior of the Eigenvalues

Often, we want to characterize the limiting global behavior of the eigenvalues of $\boldsymbol{S}$ as $n \to +\infty$. Thus, we consider its *empirical spectral distribution* (ESD).

Definition 2.4. The empirical spectral distribution of $\boldsymbol{S}$ is $\frac{1}{n}\sum_{\lambda\in\operatorname{Sp}\boldsymbol{S}}\delta_\lambda$.

Probability distributions are studied through their Stieltjes transforms.

Definition 2.5. The Stieltjes transform of a real probability measure μ is $m : z \in \mathbb{C}\setminus\operatorname{Supp}\mu \mapsto \int_{\mathbb{R}}\frac{\mathrm{d}\mu(t)}{t-z}$ where $\operatorname{Supp}\mu \subset \mathbb{R}$ is the support of μ.

An important property is that the Stieltjes transform of the empirical spectral distribution is the normalized trace of the resolvent.

Proposition 2.6. *The Stieltjes transform of* $\frac{1}{n}\sum_{\lambda\in\operatorname{Sp}\boldsymbol{S}}\delta_\lambda$ *is* $\frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_{\boldsymbol{S}}$.

Proof. Let $\mu = \frac{1}{n}\sum_{\lambda\in\operatorname{Sp}\boldsymbol{S}}\delta_\lambda$. Its support is the spectrum of $\boldsymbol{S}$. Hence, for all $z \in \mathbb{C}\setminus\operatorname{Sp}\boldsymbol{S}$,

$$\int_{\mathbb{R}}\frac{\mathrm{d}\mu(t)}{t-z} = \frac{1}{n}\sum_{\lambda\in\operatorname{Sp}\boldsymbol{S}}\frac{1}{\lambda-z} = \frac{1}{n}\operatorname{Tr}\operatorname{Diag}\left(\frac{1}{\lambda_i(\boldsymbol{S})-z}\right)_{i\in[n]} = \frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_{\boldsymbol{S}}(z).$$

□

We give the following interesting properties of the Stieltjes transform without proof.

Proposition 2.7 (Properties of the Stieltjes transform)**.** *Let μ be a real probability measure and m be its Stieltjes transform.*

1. *m is an analytic function on $\mathbb{C}\setminus\operatorname{Supp}\mu$.*
2. *For all integer $k \geqslant 0$ and $z \in \mathbb{C}\setminus\operatorname{Supp}\mu$, $|m^{(k)}(z)| \leqslant \frac{k!}{\operatorname{Dist}(z,\operatorname{Supp}\mu)^{k+1}} \leqslant \frac{k!}{|\Im z|^{k+1}}$.*
3. $\operatorname{Sign}(\Im[m(z)]) = \operatorname{Sign}(\Im z)$.
4. $\operatorname{Supp}\mu \subset [0,+\infty[\iff \operatorname{Sign}(\Im[zm(z)]) = \operatorname{Sign}(\Im z)$.
5. $\lim_{y\to\pm\infty} -\mathrm{i}ym(\mathrm{i}y) = 1$.
6. *$x \mapsto m(x)$ is an increasing function on all connected components of $\mathbb{R}\setminus\operatorname{Supp}\mu$.*

The following theorem gives criteria to identify a complex-valued function as the Stieltjes transform of a probability measure.

Theorem 2.8 (Herglotz)**.** *If the function m is analytic on $\mathbb{C}^+ \overset{def}{=} \{z \in \mathbb{C} \mid \Im z > 0\}$ and satisfies*

1. *$\Im m(z) \geqslant 0$ for all $z \in \mathbb{C}^+$,*
2. *$|m(z)| \leqslant 1/\Im z$ for all $z \in \mathbb{C}^+$,*
3. *$\lim_{y\to+\infty} -\mathrm{i}ym(\mathrm{i}y) = 1$,*
4. *$m(\bar{z}) = \overline{m(z)}$ for all $z \in \mathbb{C}^+$,*

then m is the Stieltjes transform of a unique probability measure on $\mathbb{R}$.

Proof. See Weidmann (1980, Theorem B.3). □

As the Stieltjes transform of the empirical spectral measure of $\boldsymbol{S}$ is a random quantity, it is often easier to study its expectation. We have the following important corollary to Theorem 2.8.

Corollary 2.9. $\mathbb{E}[\frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_{\boldsymbol{S}}]$ *is the Stieltjes transform of a unique probability measure on* $\mathbb{R}$.

Proof. $m_n \overset{\text{def}}{=} \frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_{\boldsymbol{S}}$ is the Stieltjes transform of the empirical spectral distribution of $\boldsymbol{S}$ (Proposition 2.6). Let us show that $\mathbb{E}[m_n]$ is analytic. Let $z_0 \in \mathbb{C}^+$ and $\varepsilon \in]0, \Im z_0[$. Our choice of ε is such that $\mathscr{B}(z_0,\varepsilon) \overset{\text{def}}{=} \{z \in \mathbb{C} \mid |z - z_0| < \varepsilon\} \subset \mathbb{C}^+$. We have $\sup_{z\in\mathscr{B}(z_0,\varepsilon)} |m_n(z)| \leqslant (\Im z_0 - \varepsilon)^{-1}$ and, because $m_n' = \frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_{\boldsymbol{S}}^2$, $\sup_{z\in\mathscr{B}(z_0,\varepsilon)} |m_n'(z)| \leqslant (\Im z_0 - \varepsilon)^{-2}$. With these two upper bounds, we can apply the standard theorems of continuity and differentiation of integrals depending on a parameter (see, e.g., Theorem 2.12 and Theorem 2.13 of Le Gall, 2022). Hence $\mathbb{E}[m_n]$ is differentiable for all $z \in \mathscr{B}(z_0,\varepsilon)$ and thus analytic on $\mathbb{C}^+$. Then, from the properties of m_n, it is easy to check that $\Im\mathbb{E}[m_n(z)] \geqslant 0$ and $|\mathbb{E}[m_n(z)]| \leqslant 1/\Im z$ for all $z \in \mathbb{C}^+$. Finally, $\lim_{y\to+\infty} -\mathrm{i}y\mathbb{E}[m_n(\mathrm{i}y)] = 1$ by the dominated convergence theorem (Le Gall, 2022, Theorem 2.11) and $\mathbb{E}[m_n(\bar{z})] = \overline{\mathbb{E}[m_n(z)]}$ for all $z \in \mathbb{C}^+$. □

The weak convergence[1] of the empirical spectral distribution of $\boldsymbol{S}$ can be characterized via the pointwise convergence of its Stieltjes transform.

Proposition 2.10 (Convergence of Stieltjes transforms)**.** *Let* $(\mu_n)_{n\geqslant 0}$ *be a sequence of real probability measures and, for all* $n \geqslant 0$, m_n *be the Stieltjes transform of* μ_n.

(a) If there exists a subset $\mathscr{S} \subset \mathbb{C}^+ \overset{def}{=} \{z \in \mathbb{C} \mid \Im z > 0\}$ *containing an accumulation point*[2] *such that, for all* $z \in \mathscr{S}$, $m_n(z) \to \bar{m}(z)$ *as* $n \to +\infty$, *then there exists a measure* $\bar{\mu}$ *such that* $\bar{\mu}(\mathbb{R}) \leqslant 1$ *and*

$$\bar{m}(z) = \int_{\mathbb{R}} \frac{\mathrm{d}\bar{\mu}(t)}{t - z} \qquad \text{for all } z \in \mathscr{S}.$$

(b) Moreover, if $\lim_{y\to+\infty} -\mathrm{i}y\bar{m}(\mathrm{i}y) = 1$, *then* $\bar{\mu}(\mathbb{R}) = 1$ *and* $\mu_n \to \bar{\mu}$ *weakly as* $n \to +\infty$.

Proof. See Geronimo and Hill (2003). □

When μ_n is the empirical spectral distribution (ESD) of $\boldsymbol{S}$, $\bar{\mu}$ is called its limiting spectral distribution (LSD). It can be recovered thanks to the inverse Stieltjes transform.

Proposition 2.11 (Inverse Stieltjes transform)**.** *Let* μ *be a real probability measure and* m *be its Stieltjes transform.*

1. *If* $a, b \in \mathbb{R}$, $a < b$, *are continuity points of* μ, *i.e.,* $\mu(\{a\}) = \mu(\{b\}) = 0$, *then,*

$$\mu([a,b]) = \lim_{y\downarrow 0} \frac{1}{\pi} \int_a^b \Im[m(x + \mathrm{i}y)]\, \mathrm{d}x.$$

2. *If* μ *has a density at* $x \in \mathbb{R}$ *then,*

$$\frac{\mathrm{d}\mu}{\mathrm{d}x}(x) = \frac{1}{\pi} \lim_{y\downarrow 0} \Im[m(x + \mathrm{i}y)].$$

3. *If* μ *has an atom at* $x \in \mathbb{R}$ *then,*

$$\mu(\{x\}) = \lim_{y\downarrow 0} -\mathrm{i}y m(x + \mathrm{i}y).$$

[1] A sequence of probability measures $(\mu_n)_{n\geqslant 0}$ is said to converge weakly to $\bar{\mu}$ if $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\bar{\mu}$ as $n \to +\infty$ for all bounded continuous functions f.

[2] $z_0 \in \mathbb{C}$ is an accumulation point of $\mathscr{S}$ if, for all $\varepsilon > 0$, there exists $z \in \mathscr{S} \setminus \{z_0\}$ such that $|z_0 - z| < \varepsilon$.

Proof. See Couillet and Liao (2022, Theorem 2.1). □

A common way to study the asymptotic behavior of the spectrum of a random symmetric matrix $\boldsymbol{S}$ is to show the existence of a limiting spectral distribution by proving the pointwise convergence of the Stieltjes transform of the empirical spectral distribution (Proposition 2.10). Typically, the asymptotic analysis of the resolvent $\boldsymbol{Q}_{\boldsymbol{S}}$ and Proposition 2.6 give a fixed point equation characterizing the Stieltjes transform of the limiting spectral distribution. The sought probability measure can then be reconstructed thanks to Proposition 2.11.

However, the existence of a limiting spectral distribution $\bar{\mu}$ does not necessarily mean that *every* eigenvalue of $\boldsymbol{S}$ must lie within $\operatorname{Supp}\bar{\mu}$ for n large enough. Indeed, a $\mathrm{o}(n)$ number of eigenvalues can stay outside the limiting measure without affecting it since it has a $\mathrm{o}(1)$ mass. In order to prove that no eigenvalue stays outside the support of the limiting spectral distribution, a powerful tool is often the following formula.

Proposition 2.12 (Helffer-Sjöstrand formula)**.** *Let μ be a probability measure on $\mathbb{R}$ and m be its Stieltjes transform. Let $f:\mathbb{R}\to\mathbb{R}$ be a* compactly supported *function which has a continuous $(k+1)$-th derivative ($k\geqslant 1$). We define $\Phi_k[f]$, the quasi-analytic extension of f on $\mathbb{C}^+ \overset{def}{=} \{z\in\mathbb{C}\mid \Im z>0\}$ as*

$$\Phi_k[f](z)=\sum_{l=0}^{k}\frac{(\mathrm{i}\Im z)^l}{l!}f^{(l)}(\Re z)\chi(\Im z)$$

where $\chi:\mathbb{R}\to[0,1]$ is an infinitely differentiable even function such that[3] *$\chi(y)=0$ if $|y|\geqslant 1$ and $\chi(y)=1$ if $|y|\leqslant\delta$ for some $\delta\in]0,1[$. Then,*

$$\int_{\mathbb{R}} f\,\mathrm{d}\mu=\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_k[f]}{\partial\bar{z}}(z)m(z)\,\mathrm{d}z$$

where $\dfrac{\partial}{\partial\bar{z}}=\dfrac{1}{2}\left(\dfrac{\partial}{\partial\Re z}+\mathrm{i}\dfrac{\partial}{\partial\Im z}\right)$ is the Wirtinger derivative.

Proof. The support of $\Phi_k[f]$ is compact therefore an integration by parts gives

$$\begin{aligned}\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_k[f]}{\partial\bar{z}}(z)m(z)\,\mathrm{d}z&=\lim_{\eta\downarrow 0}\frac{2}{\pi}\Re\left[\frac{1}{2}\int_{\eta}^{+\infty}\left(\int_{-\infty}^{+\infty}\frac{\partial\Phi_k[f]}{\partial x}(x+\mathrm{i}y)m(x+\mathrm{i}y)\,\mathrm{d}x\right)\mathrm{d}y\right.\\&\qquad\left.+\frac{\mathrm{i}}{2}\int_{-\infty}^{+\infty}\left(\int_{\eta}^{+\infty}\frac{\partial\Phi_k[f]}{\partial y}(x+\mathrm{i}y)m(x+\mathrm{i}y)\,\mathrm{d}y\right)\mathrm{d}x\right]\\&=\frac{2}{\pi}\Re\left[\frac{-\mathrm{i}}{2}\lim_{\eta\downarrow 0}\int_{\mathbb{R}}\Phi_k[f](x+\mathrm{i}\eta)m(x+\mathrm{i}\eta)\,\mathrm{d}x-\int_{\mathbb{C}^+}\Phi_k[f](z)\frac{\partial m}{\partial\bar{z}}(z)\,\mathrm{d}z\right].\end{aligned}$$

Since m is an analytic function, the Cauchy-Riemann equations give $\partial m/\partial\bar{z}=0$ and the second integral in the right-hand side is zero. Then, for $0<\eta\leqslant\delta$, we have $\Phi_k[f](x+\mathrm{i}\eta)=\sum_{l=0}^{k}\frac{\mathrm{i}^l\eta^l}{l!}f^{(l)}(x)$, thus,

$$\begin{aligned}\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_k[f]}{\partial\bar{z}}(z)m(z)\,\mathrm{d}z&=\lim_{\eta\downarrow 0}\frac{1}{\pi}\Im\left[\int_{\mathbb{R}}f(x)m(x+\mathrm{i}\eta)\,\mathrm{d}x\right]+\lim_{\eta\downarrow 0}\frac{1}{\pi}\Re\left[\eta\int_{\mathbb{R}}f'(x)m(x+\mathrm{i}\eta)\,\mathrm{d}x\right]\\&\qquad+\lim_{\eta\downarrow 0}\frac{1}{\pi}\Im\left[\sum_{l=2}^{k}\frac{\mathrm{i}^l\eta^l}{l!}\int_{\mathbb{R}}f^{(l)}(x)m(x+\mathrm{i}\eta)\,\mathrm{d}x\right].\end{aligned}$$

[3] An example of such function is $\chi:y\mapsto g\left(\frac{1+y}{1-\delta}\right)g\left(\frac{1-y}{1-\delta}\right)$ with $g(u)=\frac{f(u)}{f(u)+f(1-u)}$ and $f(u)=\begin{cases}\exp(-1/u) & \text{if } u>0\\ 0 & \text{if } u\leqslant 0\end{cases}$.

From Proposition 2.7, $|m(x+\mathrm{i}\eta)| \leqslant \frac{1}{\eta}$ therefore $\left|\frac{\mathrm{i}^l\eta^l}{l!}\int_{\mathbb{R}} f^{(l)}(x)m(x+\mathrm{i}\eta)\,\mathrm{d}x\right| \leqslant \frac{\eta^{l-1}}{l!}\int_{\mathbb{R}}|f^{(l)}(x)|\,\mathrm{d}x$ and the last limit is 0. For the second limit, fix $a<b$ such that the interval $[a,b]$ contains the support of f and observe that

$$\Re\left[\eta\int_{\mathbb{R}} f'(x)m(x+\mathrm{i}\eta)\,\mathrm{d}x\right] = \eta\int_a^b f'(x)\Re\left[\int_{\mathbb{R}}\frac{\mathrm{d}\mu(t)}{t-(x+\mathrm{i}\eta)}\right]\mathrm{d}x = \int_{\mathbb{R}}\left(\int_a^b f'(x)\frac{\eta(t-x)}{(t-x)^2+\eta^2}\,\mathrm{d}x\right)\mathrm{d}\mu(t)$$

where the interchange of integrals is justified by the fact that $\int_a^b|f'(x)\frac{\eta(t-x)}{(t-x)^2+\eta^2}|\,\mathrm{d}x$ is finite (and therefore μ-integrable) since f' is continuous on $[a,b]$. We can now apply the dominated convergence theorem (Le Gall, 2022, Theorem 2.11) to $(t,x)\in\mathbb{R}\times[a,b]\mapsto f'(x)\frac{\eta(t-x)}{(t-x)^2+\eta^2}$, which is integrable with respect to the product of μ and the Lebesgue measure on $[a,b]$. Since $\frac{\eta(t-x)}{(t-x)^2+\eta^2}\to 0$ as $\eta\downarrow 0$ for all $(t,x)\in\mathbb{R}\times[a,b]$ (even if $t=x$), we find that $\lim_{\eta\downarrow 0}\frac{1}{\pi}\Re[\eta\int_{\mathbb{R}} f'(x)m(x+\mathrm{i}\eta)\,\mathrm{d}x]=0$. Eventually, we are left with

$$\begin{aligned}
\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_k[f]}{\partial\bar{z}}(z)m(z)\,\mathrm{d}z &= \lim_{\eta\downarrow 0}\frac{1}{\pi}\Im\left[\int_{\mathbb{R}} f(x)m(x+\mathrm{i}\eta)\,\mathrm{d}x\right]\\
&= \lim_{\eta\downarrow 0}\frac{1}{\pi}\int_{\mathbb{R}}\left(\int_{\mathbb{R}} f(x)\frac{\eta}{(t-x)^2+\eta^2}\,\mathrm{d}x\right)\mathrm{d}\mu(t)\\
&= \lim_{\eta\downarrow 0}\frac{1}{\pi}\int_{\mathbb{R}}\left(\int_{\mathbb{R}} f(\eta u+t)\frac{\mathrm{d}u}{u^2+1}\right)\mathrm{d}\mu(t)\\
\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_k[f]}{\partial\bar{z}}(z)m(z)\,\mathrm{d}z &= \int_{\mathbb{R}} f(t)\,\mathrm{d}\mu(t).
\end{aligned}$$

The interchange of integrals is again justified by the fact that $\int_{\mathbb{R}}|f(x)\frac{\eta}{(t-x)^2+\eta^2}|\,\mathrm{d}x$ is finite (and therefore μ-integrable) since f is continuous and compactly supported. The last equality follows from the dominated convergence theorem (Le Gall, 2022, Theorem 2.11) and $\int_{\mathbb{R}}\frac{\mathrm{d}u}{u^2+1}=\pi$. □

Remark 2.13. Since χ is constant on $]0,\delta]$, $0<\Im z\leqslant\delta \implies \frac{\partial\Phi_k[f]}{\partial\bar{z}}(z)=\frac{1}{2}\frac{(\mathrm{i}\Im z)^k}{k!}f^{(k+1)}(\Re z)$. The greater k, the faster the convergence to zero of $\frac{\partial\Phi_k[f]}{\partial\bar{z}}$ as $\Im z\downarrow 0$. This is particularly useful to evaluate quantities such as $\int f\mathrm{d}\mu_1-\int f\mathrm{d}\mu_2$ from the knowledge that the Stieltjes transforms m_1,m_2 of μ_1,μ_2 (respectively) satisfy $|m_1(z)-m_2(z)|=\mathcal{O}(|\Im z|^{-k})$ as $\Im z\downarrow 0$:

$$\int f\,\mathrm{d}\mu_1-\int f\,\mathrm{d}\mu_2 = \frac{2}{\pi}\Re\int_{\mathbb{C}^+}\underbrace{\frac{\partial\Phi_k[f]}{\partial\bar{z}}(z)(m_1(z)-m_2(z))}_{=\mathcal{O}(1)\text{ as }\Im z\downarrow 0}\,\mathrm{d}z.$$

The smoothness of f compensates for the divergence of $m_1(z)-m_2(z)$ near the real axis.

The next lemma is an extension of the previous Helffer-Sjöstrand formula and is a handy result to evaluate some integrals which appear after using the latter.

Lemma 2.14. *Let $\mathcal{K}$ be a compact subset of $\mathbb{R}$ and h be an analytic function on $\mathbb{C}\setminus\mathcal{K}$ such that*

1. $\lim_{|z|\to+\infty}h(z)=0$,
2. $h(\bar{z})=\overline{h(z)}$ *for all* $z\in\mathbb{C}\setminus\mathcal{K}$,
3. *there exist an integer* $n_0\geqslant 1$ *and a constant* $C_{\mathcal{K}}>0$ *such that* $|h(z)|\leqslant C_{\mathcal{K}}\max(\mathrm{Dist}(z,\mathcal{K})^{-n_0},1)$ *for all* $z\in\mathbb{C}\setminus\mathcal{K}$.

Then, for all compactly supported infinitely differentiable function $\varphi : \mathbb{R} \to \mathbb{R}$ and integer $k \geqslant n_0$,

$$\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_k[\varphi]}{\partial\bar{z}}(z)h(z)\,\mathrm{d}z = \lim_{y\downarrow 0}\frac{1}{\pi}\int_{\mathscr{K}}\varphi(x)\Im[h(x+\mathrm{i}y)]\,\mathrm{d}x$$

Moreover, if $\lim_{y\to+\infty} -\mathrm{i}yh(\mathrm{i}y) = 0$ *then it holds that* $\lim_{y\downarrow 0}\frac{1}{\pi}\int_{\mathscr{K}}\Im[h(x+\mathrm{i}y)]\,\mathrm{d}x = 0$.

Proof. According to Theorem 5.4 of Schultz (2005) (see also Capitaine et al., 2009, Theorem 4.3; Loubaton, 2016, Lemma 9.1), h is the Stieltjes transform of a compactly-supported Schwartz distribution (Rudin, 1991, Chapter 6) whose support is in $\mathscr{K}$ and therefore satisfies $\lim_{y\downarrow 0}\frac{1}{\pi}\int_{\mathbb{R}}\Im[h(x+\mathrm{i}y)]\,\mathrm{d}x = \lim_{y\downarrow 0}\frac{1}{\pi}\int_{\mathscr{K}}\Im[h(x+\mathrm{i}y)]\,\mathrm{d}x = 0$ if $\lim_{y\to+\infty} -\mathrm{i}yh(\mathrm{i}y) = 0$ (Schultz, 2005, Lemma 5.6). Furthermore, the Helffer-Sjöstrand formula (Proposition 2.12) can be extended to such Stieltjes transforms as shown in Loubaton (2016, Lemma 9.3). □

2.1.2 Isolated Eigenvalues and Eigenvectors

In the perturbed random matrix models we will consider, a finite number of eigenvalues may converge outside the support of the limiting spectral distribution. Once the latter is accurately characterized, our attention moves to these *isolated* eigenvalues in the spectrum of $\boldsymbol{S}$ and their corresponding eigenvectors. Given that $\boldsymbol{Q}_{\boldsymbol{S}} : z \mapsto (\boldsymbol{S} - z\boldsymbol{I}_n)^{-1}$ is defined for all $z \in \mathbb{C}$ which is *not* an eigenvalue of $\boldsymbol{S}$, the set of asymptotic positions of isolated eigenvalues is

$$\left[\limsup_{n\to+\infty}\operatorname{Sp}\boldsymbol{S}\right] \setminus \operatorname{Supp}\bar{\mu}$$

where $\bar{\mu}$ is the limiting spectral distribution of $\boldsymbol{S}$.

Let us recall two useful results to study the corresponding eigenvectors of $\boldsymbol{S}$.

Proposition 2.15 (Cauchy's integral formula)**.** *Let U be a simply-connected open subset of $\mathbb{C}$, $f : U \to \mathbb{C}$ be a* holomorphic *function and $\gamma \subset U$ be a positively-oriented simple closed contour. Denote $\mathring{\gamma}$ the open subset of U delimited by γ. Then,*

$$-\frac{1}{2\mathrm{i}\pi}\oint_{\gamma}\frac{f(z)}{z_0 - z}\mathrm{d}z = f(z_0)\mathbf{1}_{\{z_0\in\mathring{\gamma}\}}.$$

Proposition 2.16 (Residue theorem)**.** *Let U be a simply-connected open subset of $\mathbb{C}$, $f : U \to \mathbb{C}$ be a* meromorphic *function and $\gamma \subset U$ be a positively-oriented simple closed contour which does not contain any singular point of f. Let $\mathscr{S}$ be the set of singular points of f inside the open set delimited by γ. Then,*

$$\frac{1}{2\mathrm{i}\pi}\oint_{\gamma} f(z)\,\mathrm{d}z = \sum_{s\in\mathscr{S}}\operatorname{Res}(f,s)$$

where $\operatorname{Res}(f,s)$ *is the residue of f at the order-n pole s,*

$$\operatorname{Res}(f,s) = \frac{1}{(n-1)!}\lim_{z\to s}\frac{\partial^{n-1}}{\partial z^{n-1}}\left[(z-s)^n f(z)\right].$$

To see why these results are useful, observe that

$$\boldsymbol{Q}_{\boldsymbol{S}}(z) = \left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top - z\boldsymbol{I}_n\right)^{-1} = \boldsymbol{U}(\boldsymbol{\Lambda} - z\boldsymbol{I}_n)^{-1}\boldsymbol{U}^\top = \sum_{i=1}^{n}\frac{\boldsymbol{u}_i(\boldsymbol{S})\boldsymbol{u}_i(\boldsymbol{S})^\top}{\lambda_i(\boldsymbol{S}) - z}.$$

Thus, according to Proposition 2.15, for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$,

$$\boldsymbol{a}^\top \boldsymbol{\Pi} \boldsymbol{b} = -\frac{1}{2\mathrm{i}\pi} \oint_{\gamma_i} \boldsymbol{a}^\top \boldsymbol{Q}_{\boldsymbol{S}}(z) \boldsymbol{b} \, \mathrm{d}z \qquad \text{with} \qquad \boldsymbol{\Pi} = \sum_{\substack{1 \leqslant j \leqslant n \\ \lambda_j(\boldsymbol{S}) = \lambda_i(\boldsymbol{S})}} \boldsymbol{u}_j(\boldsymbol{S}) \boldsymbol{u}_j(\boldsymbol{S})^\top$$

where $\boldsymbol{\Pi}$ is the projection matrix on the eigenspace associated to $\lambda_i(\boldsymbol{S})$ (if the eigenvalue has multiplicity one, this is simply $\boldsymbol{u}_i(\boldsymbol{S})\boldsymbol{u}_i(\boldsymbol{S})^\top$) and γ_i is a well-chosen positively-oriented simple closed (complex) contour circling around $\lambda_i(\boldsymbol{S})$ and leaving all the other eigenvalues outside. Furthermore, the integral on the right-hand side can be evaluated using residue calculus (Proposition 2.16),

$$\frac{1}{2\mathrm{i}\pi} \oint_{\gamma_i} \boldsymbol{a}^\top \boldsymbol{Q}_{\boldsymbol{S}}(z) \boldsymbol{b} \, \mathrm{d}z = \lim_{z \to \lambda_i(\boldsymbol{S})} (z - \lambda_i(\boldsymbol{S})) \boldsymbol{a}^\top \boldsymbol{Q}_{\boldsymbol{S}}(z) \boldsymbol{b}.$$

At finite n, this is the limit at a random point of a random quantity. However, if $\lambda_i(\boldsymbol{S})$ is an isolated eigenvalue with a deterministic limit as $n \to +\infty$, the asymptotic analysis of the resolvent $\boldsymbol{Q}_{\boldsymbol{S}}$ allows to characterize the limiting behavior of $\boldsymbol{a}^\top \boldsymbol{\Pi} \boldsymbol{b}$ for all deterministic $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$.

2.1.3 Asymptotic Analysis of the Resolvent

It appears that the asymptotic analysis of $\boldsymbol{Q}_{\boldsymbol{S}}$ is crucial to characterize the limiting spectral behavior of $\boldsymbol{S}$. To that end, we introduce the central notion of *deterministic equivalent*.

Definition 2.17 (Deterministic equivalent)**.** Let $\boldsymbol{X}$ be a *random* $n \times n$ matrix and $\bar{\boldsymbol{X}}$ be a *deterministic* $n \times n$ matrix. We write $\boldsymbol{X} \leftrightarrow \bar{\boldsymbol{X}}$ if, for all (sequences of) deterministic $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ of bounded norm (respectively, Euclidean and spectral norm), we have,

$$\boldsymbol{a}^\top \big(\boldsymbol{X} - \bar{\boldsymbol{X}}\big) \boldsymbol{b} \xrightarrow[n \to +\infty]{\text{a.s.}} 0 \qquad \text{and} \qquad \frac{1}{n} \operatorname{Tr} \boldsymbol{A} \big(\boldsymbol{X} - \bar{\boldsymbol{X}}\big) \xrightarrow[n \to +\infty]{\text{a.s.}} 0.$$

The matrix $\bar{\boldsymbol{X}}$ is called a *deterministic equivalent* of $\boldsymbol{X}$.

A deterministic equivalent $\bar{\boldsymbol{Q}}_{\boldsymbol{S}}(z)$ of the (random) resolvent $\boldsymbol{Q}_{\boldsymbol{S}}(z)$ has the same *scalar observations* as its random counterpart. Therefore, it is entirely sufficient to study the limiting spectral distribution ($\frac{1}{n}\operatorname{Tr}(\boldsymbol{Q}_{\boldsymbol{S}} - \bar{\boldsymbol{Q}}_{\boldsymbol{S}}) \to 0$) as well as the alignments of the isolated eigenvectors ($\oint_\gamma \boldsymbol{a}^\top (\boldsymbol{Q}_{\boldsymbol{S}} - \bar{\boldsymbol{Q}}_{\boldsymbol{S}}) \boldsymbol{b} \, \mathrm{d}z \to 0$) of $\boldsymbol{S}$. Hence, exhibiting a deterministic equivalent of the resolvent is key to characterize the limiting spectral behavior of $\boldsymbol{S}$. The "Gaussian method" exposed in Section 2.3 presents a powerful approach to derive such deterministic equivalents.

In particular, the following "Gaussian integration by parts" lemma, due to Stein (1981), is extensively used to show the convergence in mean of the resolvent.

Lemma 2.18 (Stein, 1981)**.** *Let $Z \sim \mathcal{N}(0,1)$ and $f : \mathbb{R} \to \mathbb{C}$ be a polynomially bounded differentiable function such that $\mathbb{E}[|f'(Z)|]$ exists and is finite. Then,*

$$\mathbb{E}\big[Z f(Z)\big] = \mathbb{E}\big[f'(Z)\big].$$

Proof. Since $\lim_{x \to \pm\infty} |f(x) e^{-\frac{x^2}{2}}| = 0$, an integration by parts yields

$$\mathbb{E}\big[Z f(Z)\big] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x f(x) e^{-\frac{x^2}{2}} \, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f'(x) e^{-\frac{x^2}{2}} \, \mathrm{d}x = \mathbb{E}\big[f'(Z)\big].$$

□

The almost sure convergence of bilinear forms and traces of the resolvent (Definition 2.17) is justified by the following powerful Poincaré-Nash inequality (Chen, 1982; Ledoux, 2001) together with Lemma 2.20, which is a standard way to prove almost sure convergence relying on Markov's inequality and the first Borel-Cantelli lemma.

Lemma 2.19 (Poincaré-Nash inequality)**.** *Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$ and $f : \mathbb{R}^p \to \mathbb{C}$ be a differentiable function with polynomially bounded partial derivatives $\partial_1 f, \ldots, \partial_p f$. Then,*

$$\operatorname{Var}(f(\boldsymbol{z})) \leqslant \mathbb{E}\left[\left\|\nabla f(\boldsymbol{z})\right\|^2\right] = \sum_{i=1}^{p} \mathbb{E}\left[\left|\partial_i f(\boldsymbol{z})\right|^2\right]$$

where $\nabla = \begin{bmatrix}\partial_1 & \ldots & \partial_p\end{bmatrix}^\top$.

Proof. See Pastur and Shcherbina (2011, Proposition 2.1.6). □

Lemma 2.20. *Let $(X_n)_{n\geqslant 0}$ be a sequence of random variables on $\mathbb{C}$. If $\mathbb{E}[X_n] \to L \in \mathbb{C}$ as $n \to +\infty$ and there exists an integer $\kappa \geqslant 1$ such that $\sum_{n\geqslant 0} \mathbb{E}[|X_n - \mathbb{E}[X_n]|^\kappa] < +\infty$, then $X_n \to L$ almost surely as $n \to +\infty$.*

Proof. Let $\varepsilon > 0$. By Markov's inequality (Billingsley, 2012, Equation 5.31),

$$\sum_{n\geqslant 0} \mathbb{P}(|X_n - \mathbb{E}[X_n]| \geqslant \varepsilon) \leqslant \frac{1}{\varepsilon^\kappa} \sum_{n\geqslant 0} \mathbb{E}\left[|X_n - \mathbb{E}[X_n]|^\kappa\right] < +\infty.$$

Thus, by the first Borel-Cantelli lemma[4] (Billingsley, 2012, Theorem 4.3),

$$\mathbb{P}\left(\limsup_{n\to+\infty} |X_n - \mathbb{E}[X_n]| \geqslant \varepsilon\right) = 0.$$

This implies the almost sure convergence of $(X_n)_{n\geqslant 0}$ to L:

$$\mathbb{P}\left(\limsup_{n\to+\infty} |X_n - L| \geqslant \varepsilon\right) \leqslant \mathbb{P}\left(\limsup_{n\to+\infty} [|X_n - \mathbb{E}[X_n]| + |\mathbb{E}[X_n] - L|] \geqslant \varepsilon\right) = 0.$$

□

The next elementary result is useful to compare inverse matrices and, in particular, resolvents.

Proposition 2.21 (Resolvent identity)**.** *Let $\boldsymbol{A}, \boldsymbol{B}$ be two invertible matrices. $\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{B}^{-1}$.*

Proof. $\boldsymbol{A}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{B}^{-1}$. □

We state another elementary result allowing to turn an $n \times n$ determinant into a $k \times k$ one. This is especially interesting when n is large but k is small.

Proposition 2.22 (Sylvester's identity)**.** *Let $\boldsymbol{A} \in \mathbb{C}^{n\times k}$ and $\boldsymbol{B} \in \mathbb{C}^{k\times n}$.* $\det(\boldsymbol{I}_n + \boldsymbol{A}\boldsymbol{B}) = \det(\boldsymbol{I}_k + \boldsymbol{B}\boldsymbol{A})$.

Proof. $\det(\boldsymbol{I}_n + \boldsymbol{A}\boldsymbol{B}) = \det\begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{A} \\ -\boldsymbol{B} & \boldsymbol{I}_k\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{0}_{n\times k} \\ \boldsymbol{B} & \boldsymbol{I}_k\end{bmatrix} = \det\begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{0}_{n\times k} \\ \boldsymbol{B} & \boldsymbol{I}_k\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{A} \\ -\boldsymbol{B} & \boldsymbol{I}_k\end{bmatrix} = \det(\boldsymbol{I}_k + \boldsymbol{B}\boldsymbol{A})$. □

Then, the Woodbury matrix identity (Guttman, 1946, Equation (13)) is particularly convenient to express inverse of *perturbed* matrices as, notably, in signal-plus-noise models.

[4] Given a sequence of events $(A_n)_{n\geqslant 0}$ in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, if $\sum_{n\geqslant 0} \mathbb{P}(A_n)$ converges then $\mathbb{P}(\limsup_{n\geqslant 0} A_n) = 0$.

Proposition 2.23 (Woodbury identity)**.** *Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, $\boldsymbol{C} \in \mathbb{C}^{k\times k}$ and $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{C}^{n\times k}$ be such that $\boldsymbol{A}$, $\boldsymbol{C}$ and $\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top$ are invertible. Then,*

$$\left(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top\right)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\left(\boldsymbol{C}^{-1} + \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-1}.$$

Proof. Simply check that $\left(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top\right)\left[\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\left(\boldsymbol{C}^{-1} + \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}^\top\boldsymbol{A}^{-1}\right] = \boldsymbol{I}_n$. □

In essence, the Woodbury identity states that the inverse of a perturbation of $\boldsymbol{A}$ is the inverse of $\boldsymbol{A}$ plus a perturbation whose expression is explicitly given.

We conclude this section with an important theorem which informally states that knowing an analytic function locally is enough to know it globally.

Theorem 2.24 (Vitali)**.** *Let $(f_n)_{n\geqslant 0}$ be a locally bounded*[5] *sequence of analytic functions on a connected open subset $\mathscr{D} \subset \mathbb{C}$. Assume that there exists a countable set of points $\mathscr{E} \subset \mathscr{D}$ having an accumulation point in $\mathscr{D}$ such that the limit of the complex sequence $(f_n(z))_{n\geqslant 0}$ exists for all $z \in \mathscr{E}$. Then, $(f_n)_{n\geqslant 0}$ converges uniformly on compact subsets of $\mathscr{D}$ to an analytic function.*

Proof. See Titchmarsh (1939, 5.21) or Schiff (1993, 2.4). □

Theorem 2.24 can be used to extend the domain of a deterministic equivalent. If the resolvent $\boldsymbol{Q}_{\boldsymbol{S}}$ is defined on a domain $\mathscr{D}$ (usually $\mathbb{C}\setminus\mathscr{K}$ where $\mathscr{K}$ is a compact subset of $\mathbb{R}$) and its deterministic equivalent $\bar{\boldsymbol{Q}}_{\boldsymbol{S}}$ is only defined on a subset $\mathscr{S} \subset \mathscr{D}$ with an accumulation point, then, provided that the functionals $\boldsymbol{a}^\top(\boldsymbol{Q}_{\boldsymbol{S}} - \bar{\boldsymbol{Q}}_{\boldsymbol{S}})\boldsymbol{b}$ and $\frac{1}{n}\operatorname{Tr}\boldsymbol{A}(\boldsymbol{Q}_{\boldsymbol{S}} - \bar{\boldsymbol{Q}}_{\boldsymbol{S}})$ are analytic and bounded on $\mathscr{D}$, their almost sure convergence on $\mathscr{S}$ can be extended to any compact subset of $\mathscr{D}$ by Vitali's convergence theorem and the domain of the deterministic equivalent $\bar{\boldsymbol{Q}}_{\boldsymbol{S}}$ can be similarly analytically-extended.

2.2 Central Random Matrix Results

In this section, we present some well-known results regarding the limiting spectral behavior of standard random matrix models, which we shall occasionally refer to in the following chapters.

2.2.1 Wigner Semicircle Law and Spiked Wigner Model

We call a *Wigner matrix* an $n \times n$ symmetric matrix $\boldsymbol{W}$ such that, for all $i, j \in [n]$ with $i < j$,

$$W_{i,i} \overset{\text{i.i.d.}}{\sim} \frac{1}{\sqrt{n}}\mathscr{L}_{\text{d}} \qquad \text{and} \qquad W_{i,j} = W_{j,i} \overset{\text{i.i.d.}}{\sim} \frac{1}{\sqrt{n}}\mathscr{L}_{\text{od}}$$

where $\mathbb{E}[\mathscr{L}_{\text{d}}] = 0$, $\operatorname{Var}(\mathscr{L}_{\text{d}}) < +\infty$ and $\mathbb{E}[\mathscr{L}_{\text{od}}] = 0$, $\operatorname{Var}(\mathscr{L}_{\text{od}}) = 1$. The particular case where $\mathscr{L}_{\text{d}} = \mathscr{N}(0,2)$ and $\mathscr{L}_{\text{od}} = \mathscr{N}(0,1)$ is of particular importance and is known as the *Gaussian orthogonal ensemble* (GOE). Such matrices have a density g over the set of symmetric matrices given by

$$g(\boldsymbol{W}) = \frac{1}{Z_n}e^{-\frac{1}{4}\operatorname{Tr}\boldsymbol{W}^2}$$

where Z_n is a normalizing constant such that $\int g(\boldsymbol{W})\mathrm{d}\boldsymbol{W} = 1$.

The following theorem, originally due to Wigner (1955), can be considered as the starting point of random matrix theory.

[5] That is, for all $z_0 \in \mathscr{D}$, there exists $M > 0$ and a neighborhood $\mathscr{V} \subset \mathscr{D}$ of z_0 such that $|f(z)| \leqslant M$ for all $z \in \mathscr{V}$.

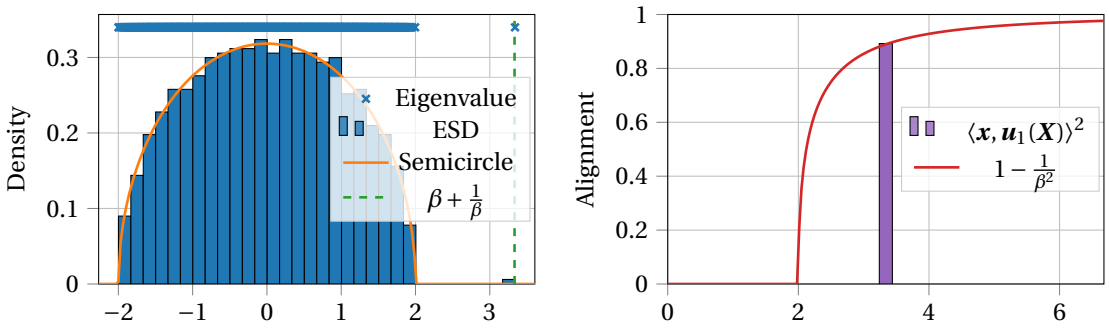


Figure 2.1: **Left:** Empirical spectral distribution (ESD) of $\boldsymbol{X} = \beta \boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{W}$, the semicircle distribution and the almost sure limit of $\lambda_1(\boldsymbol{X})$. **Right:** Asymptotic alignment between $\boldsymbol{x}$ and $\boldsymbol{u}_1(\boldsymbol{X})$ as a function of the position of the dominant eigenvalue of $\boldsymbol{X}$. **Experimental setting:** $n = 1000$ and $\beta = 3$.

Theorem 2.25 (Wigner, 1955)**.** *Let μ_{SC} be the semicircle distribution given by*

$$\mathrm{d}\mu_{\mathrm{SC}}(x) = \frac{1}{2\pi}\sqrt{\left[4 - x^2\right]^+}\mathrm{d}x.$$

The empirical spectral distribution of a Wigner matrix converges weakly in probability to μ_{SC}, that is, for all bounded continuous function f and $\varepsilon > 0$,

$$\lim_{n \to +\infty} \mathbb{P}\left(\left| \frac{1}{n} \sum_{\lambda \in \mathrm{Sp}\, \boldsymbol{W}} f(\lambda) - \int f \, \mathrm{d}\mu_{\mathrm{SC}} \right| > \varepsilon \right) = 0.$$

Remark 2.26. The proof of Wigner (1955) assumes that all moments of $\mathscr{L}_{\mathrm{d}}$ and $\mathscr{L}_{\mathrm{od}}$ are finite but this is not needed for the result (Anderson et al., 2009, Theorem 2.1.21).

Of especial interest to us is the corresponding *spiked Wigner model*

$$\boldsymbol{X} = \beta \boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{W} \in \mathbb{R}^{n \times n}$$

where, without loss of generality, $\beta > 0$ (otherwise multiply by -1), $\boldsymbol{x} \in \mathbb{S}^{n-1}$ and $\boldsymbol{W}$ is a Gaussian Wigner matrix with $\mathscr{L}_{\mathrm{d}} = \mathscr{L}_{\mathrm{od}} = \mathcal{N}(0,1)$. $\boldsymbol{X}$ is a rank-one perturbation of the random matrix $\boldsymbol{W}$. As it was emphasized in the introduction and the previous section, our favorite approach to study the spectral properties of this kind of signal-plus-noise model is through its resolvent.

Theorem 2.27. *The resolvent $\boldsymbol{Q}_{\boldsymbol{X}} : z \mapsto (\boldsymbol{X} - z\boldsymbol{I}_n)^{-1}$ has a deterministic equivalent $\bar{\boldsymbol{Q}}_{\boldsymbol{X}}$ as $n \to +\infty$ which satisfies*

$$m_{\mathrm{SC}}(z)\bar{\boldsymbol{Q}}_{\boldsymbol{X}}(z) + z\bar{\boldsymbol{Q}}_{\boldsymbol{X}}(z) + \boldsymbol{I}_n = \beta \boldsymbol{x}\boldsymbol{x}^\top \bar{\boldsymbol{Q}}_{\boldsymbol{X}}(z) \qquad \text{for all } z \in \mathbb{C} \setminus [-2,2]$$

where m_{SC} is the Stieltjes transform of μ_{SC}.

Proof. This is proven with the same method as in Section 2.3. □

Since the limiting spectral distribution of $\boldsymbol{X}$ is μ_{SC} (the LSD is not altered by the addition of a low-rank perturbation[6]), we find, using Proposition 2.6 and the fact that $\frac{1}{n}\mathrm{Tr}(\boldsymbol{Q}_{\boldsymbol{X}}(z) - \bar{\boldsymbol{Q}}_{\boldsymbol{X}}(z)) \to 0$ almost surely as $n \to +\infty$ (Definition 2.17), that its Stieltjes transform satisfies

$$m_{\mathrm{SC}}^2(z) + z m_{\mathrm{SC}}(z) + 1 = 0 \qquad \text{for all } z \in \mathbb{C} \setminus [-2,2]. \tag{2.1}$$

[6] This can be verified by comparing $\frac{1}{n}\mathrm{Tr}\,\boldsymbol{Q}_{\boldsymbol{X}}$ and $\frac{1}{n}\mathrm{Tr}\,\boldsymbol{Q}_{\boldsymbol{W}}$ with the resolvent identity (Proposition 2.21).

Moreover, we have the following explicit expression of the deterministic equivalent,

$$\bar{\boldsymbol{Q}}_{\boldsymbol{X}} = \left(\beta \boldsymbol{x}\boldsymbol{x}^\top + \frac{1}{m_{\mathrm{SC}}(z)} \boldsymbol{I}_n\right)^{-1}$$

where we have used the fact that $m(z) + z = -\frac{1}{m_{\mathrm{SC}}(z)}$ according to Equation (2.1).

Corollary 2.28. *If $\beta > 1$,*

$$\lambda_1(\boldsymbol{X}) \xrightarrow[n\to+\infty]{a.s.} \beta + \frac{1}{\beta} \qquad \textit{and} \qquad \langle \boldsymbol{x}, \boldsymbol{u}_1(\boldsymbol{X}) \rangle^2 \xrightarrow[n\to+\infty]{a.s.} 1 - \frac{1}{\beta^2}.$$

Proof. This can be proven with the method presented in Section 2.3 and we just give a sketch here. With a confinement of the spectrum argument, we show that the only $\xi \in \mathbb{R} \setminus [-2,2]$ such that $\det(\boldsymbol{X} - \xi \boldsymbol{I}_n) = 0$ satisfies $\beta = -\frac{1}{m_{\mathrm{SC}}(\xi)}$ as $n \to +\infty$. The asymptotic position of $\lambda_1(\boldsymbol{X})$ is found by injecting this expression in Equation (2.1). The asymptotic alignment $-\frac{1}{2\mathrm{i}\pi} \oint_\gamma \boldsymbol{x}^\top \bar{\boldsymbol{Q}}_{\boldsymbol{X}}(z) \boldsymbol{x} \, \mathrm{d}z$ is computed by residue calculus with a positively-oriented simple closed complex contour γ circling around $\lambda_1(\boldsymbol{X})$ and leaving all other eigenvalues outside. □

If $\beta \leqslant 1$, the random matrix model $\boldsymbol{X}$ is statistically equivalent to $\boldsymbol{W}$ (Péché, 2006; Féral and Péché, 2007), that is, the perturbation $\beta \boldsymbol{x}\boldsymbol{x}^\top$ has no impact on the limiting spectral behavior. Hence, Corollary 2.28 exhibits a *phase transition phenomenon*: as soon as $\beta > 1$, it explicitly gives the asymptotic position of the isolated eigenvalue and the alignment of the corresponding eigenvector with $\boldsymbol{x}$.

2.2.2 Sample Covariance Matrix Model: Marčenko-Pastur and Friends

Recall the sample covariance matrix model considered in the introduction: $\hat{\boldsymbol{C}} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ is a $p \times p$ matrix where $\boldsymbol{X}$ is a $p \times n$ matrix with i.i.d., zero-mean and unit-variance entries. Here, we formally state the results regarding its limiting spectral behavior. $\hat{\boldsymbol{C}}$ is a symmetric positive semidefinite matrix thus its eigenvalues are real and non-negative. Its limiting spectral distribution is characterized by the following theorem.

Theorem 2.29 (Marčenko and Pastur, 1967)**.** *As $p, n \to +\infty$ with $p/n \overset{def}{=} c \in]0, +\infty[$, the empirical spectral distribution of $\hat{\boldsymbol{C}}$ converges weakly in probability to the Marčenko-Pastur distribution*

$$\mu_{\mathrm{MP}} = \left[1 - \frac{1}{c}\right]^+ \delta_0 + \nu \qquad \textit{with} \qquad \mathrm{d}\nu(x) = \frac{\sqrt{[x - E_-]^+ [E_+ - x]^+}}{2\pi c x} \mathrm{d}x$$

and $E_\pm = (1 \pm \sqrt{c})^2$.

The Marčenko-Pastur distribution is composed of an absolutely continuous part ν and a Dirac mass at the origin which only exists if $c > 1$. As explained in the introduction, it describes the law of the eigenvalues of a sample covariance matrix $\frac{1}{p}\boldsymbol{X}\boldsymbol{X}^\top$ or a Gram kernel matrix $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$ in the particular case where $\boldsymbol{X}$ has i.i.d., zero-mean and unit variance entries. The following theorem generalizes this result to a broader class of matrices.

Theorem 2.30 (Silverstein and Bai, 1995; Couillet and Liao, 2022, Theorem 2.6)**.** *Let $\boldsymbol{X}$ be a $p \times n$ matrix with i.i.d., zero-mean and unit variance entries. Let $\boldsymbol{C}$ be an $n \times n$ symmetric positive semidefinite matrix with bounded spectral norm and whose empirical spectral distribution converges weakly almost surely to a probability distribution η as $p \to +\infty$. Then, as $p, n \to +\infty$ with $p/n \overset{def}{=} c \in]0, +\infty[$, the empirical spectral distribution of $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}$ converges weakly almost surely to a probability distribution $\bar{\mu}$ whose Stieltjes transform $\bar{m}$ satisfies*

$$\bar{m}(z) = \left[-z + c \int_{\mathbb{R}} \frac{t}{1 + t\bar{m}(z)} \mathrm{d}\eta(t)\right]^{-1} \qquad \textit{for all } z \in \mathbb{C} \setminus \mathrm{Supp}\, \bar{\mu}.$$

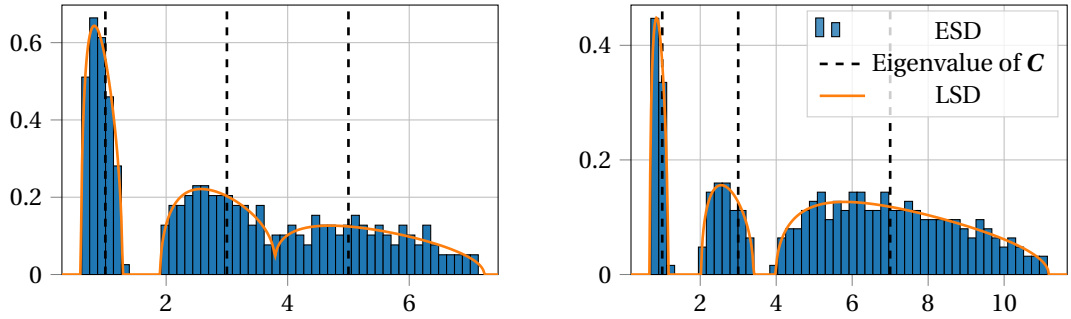


Figure 2.2: Empirical spectral distribution (ESD) and limiting spectral distribution (LSD) of $\frac{1}{n}\boldsymbol{C}^{1/2}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{C}^{1/2}$ computed by inverting the Stieltjes transform found with Theorem 2.30. **Experimental setting:** $p = 300$, $n = 3000$, $\boldsymbol{X}$ has i.i.d. $\mathcal{N}(0,1)$ entries and $\boldsymbol{C}$ is a diagonal matrix with ESD $\frac{1}{3}\delta_1 + \frac{1}{3}\delta_3 + \frac{1}{3}\delta_5$ (**left**) or $\frac{1}{6}\delta_1 + \frac{1}{6}\delta_3 + \frac{2}{3}\delta_7$ (**right**).

This result defines the Stieltjes transform $\bar{m}$ of the limiting spectral distribution of $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}$ via a *fixed-point equation* depending on η, the limiting spectral distribution of $\boldsymbol{C}$. From Theorem 2.30, the corresponding Stieltjes transforms for the Gram kernel matrix $\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}$ and the sample covariance matrix $\frac{1}{n}\boldsymbol{C}^{1/2}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{C}^{1/2}$ can be deduced. Indeed, for the Gram kernel matrix, notice that

$$\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X} - z\boldsymbol{I}_n\right)^{-1} = \frac{n}{p}\left(\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X} - \frac{n}{p}z\boldsymbol{I}_n\right)^{-1}$$

and, since the Stieltjes transform is the normalized trace of the resolvent (Proposition 2.6), we find that the limiting spectral distribution of $\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}$ has its Stieltjes transform given by $z \mapsto \frac{1}{c}\bar{m}(\frac{z}{c})$. Then, for the sample covariance matrix, observe that, denoting μ the empirical spectral distribution of $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}$, that of $\frac{1}{n}\boldsymbol{C}^{1/2}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{C}^{1/2}$ is given by $\frac{n}{p}\mu - \frac{n-p}{p}\delta_0$. Thus, the Stieltjes transform of the limiting spectral distribution of the sample covariance matrix is $z \mapsto \frac{1}{c}\bar{m}(z) + \frac{1-c}{c}\frac{1}{z}$. In particular, if $\boldsymbol{C} = \boldsymbol{I}_p$, the empirical spectral distribution of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ converges to the Marčenko-Pastur distribution μ_{MP} (Theorem 2.29). Taking $\eta = \delta_1$ in Theorem 2.30, we find that its Stieltjes transform m_{MP} satisfies

$$zcm_{\mathrm{MP}}^2(z) - (1-c-z)m_{\mathrm{MP}}(z) + 1 = 0 \qquad \text{for all } z \in \mathbb{C} \setminus \operatorname{Supp}\mu_{\mathrm{MP}}. \tag{2.2}$$

Figure 2.2 depicts the empirical and limiting spectral distributions of a sample covariance matrix $\frac{1}{n}\boldsymbol{C}^{1/2}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{C}^{1/2}$ when $\boldsymbol{C}$ has 3 distinct eigenvalues. The numerical computation of the LSD is performed by computing $\bar{m}$ with a fixed-point iteration on the equation given in Theorem 2.30 near the real axis so that the corresponding density can be computed with the inversion formula (Proposition 2.11). We see that each eigenvalue of $\boldsymbol{C}$ creates its own "bulk" in the spectrum of $\frac{1}{n}\boldsymbol{C}^{1/2}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{C}^{1/2}$ and the size of the latter is proportional to the multiplicity of the corresponding eigenvalue. These bulks may merge if they are too close, as in the left panel of Figure 2.2.

The results of Theorem 2.29 and Theorem 2.30 describe the asymptotic global behavior of the eigenvalues by specifying their limiting distribution. Yet, they say nothing about the possibility that a few eigenvalues lie outside it. Indeed, since only the weak convergence of the empirical spectral distribution is guaranteed, there may be a $o(n)$ number of eigenvalues staying outside the limiting support. The following result indicates when such a phenomenon can occur.

Theorem 2.31 (Bai et al., 1988; Bai and Silverstein, 1998)**.** *Under the setting of Theorem 2.30, further assume that*

$$\max_{1 \leqslant i \leqslant p} \operatorname{Dist}(\lambda_i(\boldsymbol{C}), \operatorname{Supp}\eta) \xrightarrow[p \to +\infty]{a.s.} 0.$$

1. *If* $\mathbb{E}[|X_{i,j}|^4] < +\infty$, *then for all connected component* $\mathscr{I} \subset \mathbb{R} \setminus \operatorname{Supp} \bar{\mu}$, $\left|\mathscr{I} \cap \operatorname{Sp} \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X}\right| \xrightarrow[p,n\to+\infty]{a.s.} 0$.

2. *If* $\mathbb{E}[|X_{i,j}|^4] = +\infty$, *then* $\lambda_1\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X}\right) \xrightarrow[p,n\to+\infty]{a.s.} +\infty$.

This theorem states that the presence of eigenvalues outside the limiting support in the large p, n regime is completely determined by the finiteness of the fourth moment of the entries of $\boldsymbol{X}$. If $\mathbb{E}[|X_{i,j}|^4]$ is finite, there is asymptotically no eigenvalue outside the support of $\bar{\mu}$. Conversely, if $X_{i,j}$ has no fourth moment, there is at least one eigenvalue outside $\operatorname{Supp} \bar{\mu}$ and the dominant eigenvalue of $\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X}$ diverges as $p, n \to +\infty$.

Remark 2.32 (Exact separation). This result can be refined in the case of finite fourth moment by stating that the number of eigenvalues of $\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X}$ lying in each bulk is exactly proportional to the number of eigenvalues of $\boldsymbol{C}$ (repeated with multiplicity) that "generated" this bulk, see Bai and Silverstein (1999).

2.2.3 Signal-plus-Noise Models

A signal-plus-noise matrix model is a particular case of spiked model in the form $\boldsymbol{X} = \boldsymbol{P} + \boldsymbol{N}$ where $\boldsymbol{N}$ is a random matrix (the noise) and $\boldsymbol{P}$ is a finite-rank perturbation (the signal). Since the rank of $\boldsymbol{P}$ remains finite while the size of $\boldsymbol{X}$ grows, we say that $\boldsymbol{P}$ is a low-rank perturbation (comparatively to the rank of $\boldsymbol{N}$ which diverges almost surely as the matrix grows).

In this thesis, we mostly consider settings where the sought information is contained in the dominant right singular subspace of $\boldsymbol{P}$. Thus, it is convenient to represent $\boldsymbol{P}$ as $\boldsymbol{L}\boldsymbol{V}^\top$ where $\boldsymbol{L}$ is a $p \times K$ matrix (K is the rank of $\boldsymbol{P}$) and $\boldsymbol{V}$ is an $n \times K$ semi-orthogonal matrix[7] whose columns are the right singular vectors of $\boldsymbol{P}$.

Then, we are concerned with the study of the random matrix $\boldsymbol{X}$ consisting of a zero-mean random matrix $\boldsymbol{N}$ perturbed by a low-rank matrix $\boldsymbol{P} = \boldsymbol{L}\boldsymbol{V}^\top$ or, equivalently, a low-rank signal corrupted by additive noise. In particular, depending on the strength of the perturbation — the singular values of $\boldsymbol{P}$ —, we are interested in the alignment between the right singular subspace of $\boldsymbol{P}$ and the K-dimensional dominant right singular subspace of $\boldsymbol{X}$, that is, the dominant eigenspace of the Gram kernel matrix

$$\boldsymbol{K} \overset{\text{def}}{=} \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X} = \frac{1}{p}\left[\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{P}^\top \boldsymbol{N} + \boldsymbol{N}^\top \boldsymbol{P} + \boldsymbol{N}^\top \boldsymbol{N}\right].$$

This matrix has the great advantage that it is *symmetric* and the tools presented in Section 2.1 directly apply to it.

Regarding the noise, the random matrix $\boldsymbol{N}$ is often assumed having i.i.d. Gaussian entries (see the discussion below). In this case, the "Gaussian method", which we detail in the next section, allows the precise characterization of the limiting spectral behavior of $\boldsymbol{K}$. In particular, we have the following (now expected) result.

Theorem 2.33 (Limiting spectral distribution)**.** *The empirical spectral distribution of* $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$ *converges weakly almost surely to* μ_{MP} *defined in Theorem 2.29 with parameter* $c^{-1} = n/p$ *(instead of* c*). Moreover, if* $\boldsymbol{P} = \boldsymbol{0}_{p\times n}$ *(no signal), no eigenvalue stays outside* $\operatorname{Supp} \mu_{\text{MP}}$ *almost surely as* $p, n \to +\infty$.

Proof. See Section 2.3. □

Then, denoting $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K$ the columns of $\boldsymbol{V}$ sorted in decreasing order of the corresponding singular value of $\boldsymbol{P}$, we can describe the limiting behavior of spike eigenvalues and eigenvectors. We

[7] That is, $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_K$. The set of such matrices is the Stiefel manifold $V_K(\mathbb{R}^n)$.

make the additional assumption that there exists a *diagonal* matrix $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \dots, \lambda_K)$ with non-degenerate eigenvalues[8] such that $\left\| \frac{1}{n} \boldsymbol{L}^\top \boldsymbol{L} - \boldsymbol{\Lambda} \right\| \to 0$.

Theorem 2.34 (Spike behavior)**.** *For all $k \in [K]$,*

$$\lambda_k(\boldsymbol{K}) \xrightarrow[p,n\to+\infty]{a.s.} \frac{(\ell_k + c)(\ell_k + 1)}{\ell_k c} \qquad \textit{and} \qquad \langle \boldsymbol{v}_k, \boldsymbol{u}_k(\boldsymbol{K}) \rangle^2 \xrightarrow[p,n\to+\infty]{a.s.} 1 - \frac{\ell_k + c}{\ell_k(\ell_k + 1)}$$

where $\ell_k = \max(\sqrt{c}, \lambda_k(\boldsymbol{\Lambda}))$.

Proof. See Section 2.3. □

Remark 2.35 (Non-trivial regime). The assumption $\frac{1}{n} \boldsymbol{L}^\top \boldsymbol{L} \to \boldsymbol{\Lambda}$ implicitly states that the spectral norm of $\frac{1}{\sqrt{p}} \boldsymbol{P}$ is bounded. This is an important point because it precisely places us in the *non-trivial regime* where the problem is neither too easy (the signal can be perfectly recovered) nor too hard (it is impossible to reconstruct the signal), which is just the interesting regime in practice. Indeed, Theorem 2.34 exhibits a phase transition between impossible and possible reconstruction of $\boldsymbol{P}$. The "too hard" and "too easy" regimes correspond to $\lambda_k(\boldsymbol{\Lambda}) \to 0$ and $\lambda_k(\boldsymbol{\Lambda}) \to +\infty$ respectively.

These results are illustrated in Chapter 3 (see Figure 3.1 and Figure 3.2).

On the i.i.d. Gaussian Noise Assumption

One could easily criticize our choice of *Gaussian* noise by saying that this is not a realistic assumption because Gaussian noise is rarely observed in practice and therefore the results found with the Gaussian method are too specific to be really useful. Still, there are several reasons which make us confident with our assumption and make us believe that the results derived with it are, in fact, very general.

Firstly — and this is probably the most important point —, following the lines of Jaynes (2003), the Gaussian probability assignment is *not* an assumption on the real physical phenomenon behind our observation but it is a description of our *state of knowledge* about the noise. Given that the only prior information we have about the noise is its mean and variance (and nothing else), the Gaussian distribution is the only maximizer of Shannon's entropy — which measures our "level of uncertainty" — under the constraint that the mean and variance are known. Moreover, the fact that the entries of $\boldsymbol{N}$ are independent is not an assumption that no correlation exists in the data but it is simply our way to recognize that we have no prior information about such correlation. There may well be a correlation but, without any prior knowledge about it, an i.i.d. noise is the most reasonable choice of model. It is the only way to honestly acknowledge our ignorance.

Thus, by saying that the entries of $\boldsymbol{N}$ are i.i.d. $\mathcal{N}(0,1)$ random variables, we are only making an assumption on the mean and the variance of the noise. While the zero-mean assumption on the noise is fairly acceptable (otherwise there is a signal that our model does not take into account), there is, however, no reason that the variance be 1 and we should instead assume $\mathcal{N}(0, \sigma^2)$ entries, that is, $\boldsymbol{X} = \boldsymbol{P} + \sigma \boldsymbol{N}$. But then σ becomes a parameter of our model (recall Remark 1.3) and we could as well consider $\frac{1}{\sigma} \boldsymbol{X} = \frac{1}{\sigma} \boldsymbol{P} + \boldsymbol{N}$. Therefore we are back to a model with $\mathcal{N}(0,1)$ noise where the "strength" of the perturbation $s_1^2(\boldsymbol{P})/\sigma^2$ rightfully appears as the signal-to-noise ratio.

Although this is not the focus of this thesis, there may be situations where further prior knowledge leads us to consider different distributions for the noise. In such cases, the results of the Gaussian method can be generalized to i.i.d. entries up to a control on the moments of the distribution thanks to an "interpolation trick" (Lytova and Pastur, 2009, Corollary 3.1), as it is done, e.g., in Pastur and Shcherbina (2011, Theorems 18.4.2 and 19.2.1) and Merlevède et al. (2015); Banna et al. (2015). In the

[8] That is, $k \neq k' \implies \lambda_k \neq \lambda_{k'}$. This assumption is not needed but often verified in practice and simplifies the result.

non-i.i.d. setting, the works of Benaych-Georges and Nadakuditi (2011, 2012) show very general results with the milder assumption that the noise is rotationally invariant. Similar spectral behaviors are observed under the very broad assumption that the noise is concentrated (Talagrand, 1995; Ledoux, 2001). More general results can thus be proved relying on measure concentration tools, such as those developed by El Karoui (2009); Louart and Couillet (2021). We should also mention that, in a machine learning context, concentrated random vectors encompass a large set of very realistic data (Seddik et al., 2020).

2.3 A Primer on the "Gaussian Method"

In order to illustrate concretely how the tools presented in Section 2.1 can be used to analyze a signal-plus-noise matrix model, we prove in this section the results of Theorem 2.33 and Theorem 2.34 on the model $\boldsymbol{X} = \boldsymbol{P} + \boldsymbol{N} \in \mathbb{R}^{p\times n}$ with $\boldsymbol{N}$ having i.i.d. $\mathcal{N}(0,1)$ entries and $\boldsymbol{P} = \boldsymbol{L}\boldsymbol{V}^\top$ as described in Section 2.2.3.

The proof is performed in three steps.

1. Firstly, we show the concentration around its expectation of the resolvent $\boldsymbol{Q}(z) = (\boldsymbol{K} - z\boldsymbol{I}_n)^{-1}$, where $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$ is the Gram kernel matrix, and we find an equation giving access to the asymptotic behavior of $\mathbb{E}[\boldsymbol{Q}(z)]$. To achieve this, we prove that $\mathbb{E}[\boldsymbol{Q}(z)]$ is a deterministic equivalent of $\boldsymbol{Q}(z)$ with the Poincaré-Nash inequality (Lemma 2.19) and Lemma 2.20. Then, we rely on Stein's lemma (Lemma 2.18) to exhibit a simple equation for $\mathbb{E}[\boldsymbol{Q}(z)]$.

2. The previous equation allows us to characterize the limiting spectral distribution of $\boldsymbol{K}$ through the deterministic quantity $\mathbb{E}[\frac{1}{n}\operatorname{Tr}\boldsymbol{Q}(z)]$ as $p,n \to +\infty$ (Proposition 2.6). At this point we obtain the result of Theorem 2.33. We also show that no eigenvalue of $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}$ (the model without signal) stays outside the support of the limiting spectral distribution almost surely as $p,n \to +\infty$ thanks to an expansion of $\mathbb{E}[\frac{1}{n}\operatorname{Tr}\boldsymbol{Q}(z)]$ and the Helffer-Sjöstrand formula (Proposition 2.12).

3. Finally, we exhibit the asymptotic positions of the spike eigenvalues in the spectrum of $\boldsymbol{K}$ by studying the singular points of $z \mapsto (\boldsymbol{K} - z\boldsymbol{I}_n)^{-1}$ outside the limiting support and we formulate a simple deterministic equivalent $\bar{\boldsymbol{Q}}(z)$ of the resolvent. With Cauchy's integral formula (Proposition 2.15) and residue calculus (Proposition 2.16), the alignments between the spike eigenvectors of $\boldsymbol{K}$ and the columns of $\boldsymbol{V}$ can be characterized as well. This results in Theorem 2.34.

2.3.1 Asymptotic Behavior of the Resolvent

Let $\boldsymbol{Q} : z \in \mathbb{C}\setminus\operatorname{Sp}\boldsymbol{K} \mapsto (\boldsymbol{K} - z\boldsymbol{I}_n)^{-1} \in \mathbb{C}^{n\times n}$ be the resolvent of $\boldsymbol{K} \overset{\text{def}}{=} \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$. In order to ease the notation we drop the subscript $\boldsymbol{K}$ and we will also often drop the dependence in z so we simply write $\boldsymbol{Q}$ instead of $\boldsymbol{Q}_{\boldsymbol{K}}(z)$. In Sections 2.3.1 and 2.3.2, we consider $z \in \mathbb{C}\setminus\mathbb{R}$, i.e., $\Im z \neq 0$. Moreover, we use the notation $u_n(z) = \mathcal{O}_z(v_n)$ if there exist two polynomials P, Q with nonnegative coefficients and an integer n_0 such that $|u_n(z)| \leqslant \frac{P(|z|)}{|\Im z|Q(|\Im z|)}|v_n|$ for all $z \in \mathbb{C}\setminus\mathbb{R}$ as soon as $n \geqslant n_0$.. This allows the control of possible divergences of resolvents and Stieltjes transforms near the real axis.

Preliminaries

As derivatives of $\boldsymbol{Q}$ with respect to the entries of $\boldsymbol{N}$ will appear regularly in our computations, we derive their expression here once and for all. Since $\boldsymbol{Q}^{-1}\boldsymbol{Q} = \boldsymbol{I}_n$, we have $\partial\boldsymbol{Q} = -\boldsymbol{Q}(\partial\boldsymbol{Q}^{-1})\boldsymbol{Q}$ and, since

$\boldsymbol{Q}^{-1} = \boldsymbol{K} + z\boldsymbol{I}_n$, this gives

$$\begin{aligned}
\frac{\partial Q_{a,b}}{\partial N_{c,d}} &= -\sum_{e=1}^{n}\sum_{f=1}^{p} Q_{a,e}\frac{\partial K_{e,f}}{\partial N_{c,d}}Q_{f,b} \\
&= -\frac{1}{p}\sum_{e=1}^{n}\sum_{f=1}^{p}\sum_{g=1}^{p} Q_{a,e}\left(\frac{\partial X_{g,e}}{\partial N_{c,d}}X_{g,f} + X_{g,e}\frac{\partial X_{g,f}}{\partial N_{c,d}}\right)Q_{f,b} \\
&= -\frac{1}{p}\sum_{e=1}^{n}\sum_{f=1}^{p}\sum_{g=1}^{p} Q_{a,e}\big(\delta_{g,c}\delta_{e,d}X_{g,f} + X_{g,e}\delta_{g,c}\delta_{f,d}\big)Q_{f,b} \\
&= -\frac{1}{p}\sum_{f=1}^{p} Q_{a,d}X_{c,f}Q_{f,b} - \frac{1}{p}\sum_{e=1}^{n} Q_{a,e}X_{c,e}Q_{d,b} \\
\frac{\partial Q_{a,b}}{\partial N_{c,d}} &= -\frac{1}{p}\big[Q_{a,d}[\boldsymbol{XQ}]_{c,b} + Q_{d,b}[\boldsymbol{XQ}]_{c,a}\big]. \qquad (2.3)
\end{aligned}$$

Concentration around $\mathbb{E}[\boldsymbol{Q}]$

It is more convenient to show first that $\mathbb{E}[\boldsymbol{Q}]$ is a deterministic equivalent of $\boldsymbol{Q}$, so that we can rely on this afterwards.

Firstly, for two deterministic vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ with bounded norm we show the almost sure convergence to 0 of the quadratic form $\boldsymbol{a}^\top(\boldsymbol{Q} - \mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}$. In order to apply Lemma 2.20, we need to find an integer $\kappa \geqslant 2$ such that $\mathbb{E}[|\boldsymbol{a}^\top(\boldsymbol{Q} - \mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}|^\kappa] = \mathcal{O}_z(n^{-2})$ as $p, n \to +\infty$. To that end, we use the Poincaré-Nash inequality (Lemma 2.19). We start with $\kappa = 2$.

$$\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\boldsymbol{Q} - \mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big|^2\Big] = \mathrm{Var}\big(\boldsymbol{a}^\top\boldsymbol{Q}\boldsymbol{b}\big) \leqslant \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\left|\frac{\partial \boldsymbol{a}^\top\boldsymbol{Q}\boldsymbol{b}}{\partial N_{i,j}}\right|^2\right].$$

The derivatives can be computed using Equation (2.3),

$$\begin{aligned}
\frac{\partial \boldsymbol{a}^\top\boldsymbol{Q}\boldsymbol{b}}{\partial N_{i,j}} &= \sum_{k=1}^{n}\sum_{l=1}^{n} a_k\frac{\partial Q_{k,l}}{\partial N_{i,j}}b_l \\
&= -\frac{1}{p}\sum_{k=1}^{n}\sum_{l=1}^{n} a_k\big[Q_{k,j}[\boldsymbol{XQ}]_{i,l} + Q_{j,l}[\boldsymbol{XQ}]_{i,k}\big]b_l \\
&= -\frac{1}{p}\sum_{k=1}^{n}\sum_{l=1}^{n}\big[[\boldsymbol{XQ}]_{i,k}a_kb_lQ_{l,j} + [\boldsymbol{XQ}]_{i,l}b_la_kQ_{k,j}\big] \qquad \text{since } Q_{j,l} = Q_{l,j} \\
&= -\frac{1}{p}\big[\boldsymbol{XQ}\big(\boldsymbol{a}\boldsymbol{b}^\top + \boldsymbol{b}\boldsymbol{a}^\top\big)\boldsymbol{Q}\big]_{i,j}.
\end{aligned}$$

Thus, using the fact that $|a+b|^2 \leqslant 2(|a|^2 + |b|^2)$, we have,

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\left|\frac{\partial \boldsymbol{a}^\top\boldsymbol{Q}\boldsymbol{b}}{\partial N_{i,j}}\right|^2\right] &\leqslant \frac{2}{p^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\left|\big[\boldsymbol{XQ}\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{Q}\big]_{i,j}\right|^2 + \left|\big[\boldsymbol{XQ}\boldsymbol{b}\boldsymbol{a}^\top\boldsymbol{Q}\big]_{i,j}\right|^2\right] \\
&= \frac{2}{p^2}\mathbb{E}\Big[\big\|\boldsymbol{XQ}\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{Q}\big\|_{\mathrm{F}}^2 + \big\|\boldsymbol{XQ}\boldsymbol{b}\boldsymbol{a}^\top\boldsymbol{Q}\big\|_{\mathrm{F}}^2\Big] \leqslant \frac{4}{p^2}\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\big[\|\boldsymbol{Q}\|^4\|\boldsymbol{X}\|^2\big]
\end{aligned}$$

where the last inequality is obtained using the property $\|\boldsymbol{AB}\|_{\mathrm{F}} \leqslant \|\boldsymbol{A}\|\|\boldsymbol{B}\|_{\mathrm{F}}$, the sub-multiplicative property $\|\boldsymbol{AB}\| \leqslant \|\boldsymbol{A}\|\|\boldsymbol{B}\|$ and the fact that $\|\boldsymbol{a}\boldsymbol{b}^\top\|_{\mathrm{F}} = \|\boldsymbol{a}\|\|\boldsymbol{b}\|$. Since $\|\boldsymbol{a}\|\|\boldsymbol{b}\| = \mathcal{O}(1)$ from our assump-

tions, $\|\boldsymbol{Q}\| \leqslant |\Im z|^{-1}$ from Proposition 2.2 and $\frac{1}{\sqrt{p}}\|\boldsymbol{X}\| = \mathcal{O}(1)$ in the non-trivial regime[9] , we find that $\mathbb{E}[|\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}|^2] = \mathcal{O}_z(n^{-1})$. Unfortunately, this is not enough to apply Lemma 2.20 so we need to evaluate the moment of order $\kappa = 4$.

$$\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big|^4\Big] = \mathrm{Var}\Big(\big(\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big)^2\Big) + \Big|\mathbb{E}\Big[\big(\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big)^2\Big]\Big|^2.$$

The rightmost term is upper bounded by $\mathbb{E}[|\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}|^2]^2 = \mathcal{O}_z(n^{-2})$ from Jensen's inequality (Le Gall, 2022, Theorem 4.3). We use again the Poincaré-Nash inequality (Lemma 2.19) on the variance of $(\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b})^2$.

$$\mathrm{Var}\Big(\big(\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big)^2\Big) \leqslant \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[\left|2\big(\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big)\frac{\partial \boldsymbol{a}^\top\boldsymbol{Q}\boldsymbol{b}}{\partial N_{i,j}}\right|^2\right].$$

Similarly, we find that

$$\begin{aligned}
\mathrm{Var}\Big(\big(\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big)^2\Big) &\leqslant \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[\left|-\frac{2}{p}\big(\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big)\big[\boldsymbol{X}\boldsymbol{Q}\big(\boldsymbol{a}\boldsymbol{b}^\top+\boldsymbol{b}\boldsymbol{a}^\top\big)\boldsymbol{Q}\big]_{i,j}\right|^2\right] \\
&\leqslant \frac{8}{p^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[\big|\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big|^2\left(\big|\big[\boldsymbol{X}\boldsymbol{Q}\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{Q}\big]_{i,j}\big|^2 + \big|\big[\boldsymbol{X}\boldsymbol{Q}\boldsymbol{b}\boldsymbol{a}^\top\boldsymbol{Q}\big]_{i,j}\big|^2\right)\right] \\
&\leqslant \frac{16}{p^2}\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big|^2\|\boldsymbol{Q}\|^4\|\boldsymbol{X}\|^2\Big]
\end{aligned}$$

where the last inequality is obtained similarly to the case $\kappa = 2$. Then, with the Cauchy-Schwarz inequality,

$$\mathrm{Var}\Big(\big(\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big)^2\Big)^2 \leqslant \underbrace{\left(\frac{16}{p^2}\right)^2\|\boldsymbol{a}\|^4\|\boldsymbol{b}\|^4\mathbb{E}\big[\|\boldsymbol{Q}\|^8\|\boldsymbol{X}\|^4\big]}_{=\mathcal{O}_z(n^{-2})}\Big(\mathrm{Var}\Big(\big(\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b}\big)^2\Big) + \mathcal{O}_z(n^{-2})\Big)$$

which shows that $\mathrm{Var}((\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b})^2) = \mathcal{O}_z(n^{-2})$. Therefore $\boldsymbol{a}^\top(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\boldsymbol{b} \to 0$ almost surely for all $z \in \mathbb{C}\setminus\mathbb{R}$ as $p, n \to +\infty$ by Lemma 2.20.

Secondly, given a deterministic matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ of bounded norm, we show the almost sure convergence to 0 of the normalized trace $\frac{1}{n}\mathrm{Tr}\,\boldsymbol{A}\boldsymbol{Q}$. Once again, this is performed with the Poincaré-Nash inequality (Lemma 2.19).

$$\mathbb{E}\left[\left|\frac{1}{n}\mathrm{Tr}\,\boldsymbol{A}(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}])\right|^2\right] = \mathrm{Var}\left(\frac{1}{n}\mathrm{Tr}\,\boldsymbol{A}\boldsymbol{Q}\right) \leqslant \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[\left|\frac{\partial \mathrm{Tr}\,\boldsymbol{A}\boldsymbol{Q}}{\partial N_{i,j}}\right|^2\right].$$

With Equation (2.3), we have,

$$\begin{aligned}
\frac{\partial \mathrm{Tr}\,\boldsymbol{A}\boldsymbol{Q}}{\partial N_{i,j}} &= \sum_{k=1}^n\sum_{l=1}^n A_{k,l}\frac{\partial Q_{l,k}}{\partial N_{i,j}} \\
&= -\frac{1}{p}\sum_{k=1}^n\sum_{l=1}^n A_{k,l}\big[Q_{l,j}[\boldsymbol{X}\boldsymbol{Q}]_{i,k} + Q_{j,k}[\boldsymbol{X}\boldsymbol{Q}]_{i,l}\big]
\end{aligned}$$

[9] We have $\frac{\|\boldsymbol{X}\|}{\sqrt{p}} \leqslant \frac{\|\boldsymbol{P}\|}{\sqrt{p}} + \frac{\|\boldsymbol{N}\|}{\sqrt{p}}$ with $\frac{\|\boldsymbol{P}\|}{\sqrt{p}} = \mathcal{O}(1)$ by assumption (see Remark 2.35) and $\frac{\|\boldsymbol{N}\|}{\sqrt{p}} = \mathcal{O}(1)$ almost surely as proven by Geman (1980) or as a particular case of a general concentration result proven independently in Chapter 5 (Lemma 5.10). Note that we can avoid the necessity of this result by writing $\frac{1}{p^2}\|\boldsymbol{X}\boldsymbol{Q}\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{Q}\|_{\mathrm{F}}^2 = \frac{1}{p}\mathrm{Tr}(\boldsymbol{Q}\boldsymbol{b}\boldsymbol{a}^\top\boldsymbol{Q}(z\boldsymbol{Q}+\boldsymbol{I}_n)\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{Q})$ since $(\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X} - z\boldsymbol{I}_n)\boldsymbol{Q} = \boldsymbol{I}_n$. Thus we have an upper bound depending only on $\|\boldsymbol{a}\|$, $\|\boldsymbol{b}\|$ and $\|\boldsymbol{Q}\|$.

$$= -\frac{1}{p}\big[\boldsymbol{XQAQ} + \boldsymbol{XQA}^\top\boldsymbol{Q}\big]_{i,j}.$$

Hence,

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[\left|\frac{\partial \operatorname{Tr}\boldsymbol{AQ}}{\partial N_{i,j}}\right|^2\right] &\leqslant \frac{2}{n^2p^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left[\left|[\boldsymbol{XQAQ}]_{i,j}\right|^2 + \left|\big[\boldsymbol{XQA}^\top\boldsymbol{Q}\big]_{i,j}\right|^2\right] \\ &\leqslant \frac{4}{n^2p^2}\|\boldsymbol{A}\|_{\mathrm{F}}^2\mathbb{E}\big[\|\boldsymbol{Q}\|^4\|\boldsymbol{X}\|^2\big] \leqslant \frac{4}{np^2}\|\boldsymbol{A}\|^2\mathbb{E}\big[\|\boldsymbol{Q}\|^4\|\boldsymbol{X}\|^2\big] = \mathcal{O}_z(n^{-2})\end{aligned}$$

and, with Lemma 2.20, we find that $\frac{1}{n}\operatorname{Tr}\boldsymbol{A}(\boldsymbol{Q}-\mathbb{E}[\boldsymbol{Q}]) \to 0$ almost surely for all $z \in \mathbb{C}\setminus\mathbb{R}$ as $p, n \to +\infty$.

Remark 2.36. In passing, we have shown that $\operatorname{Var}(\boldsymbol{a}^\top\boldsymbol{Qb}) = \mathcal{O}_z(n^{-1})$ and $\operatorname{Var}(\frac{1}{n}\operatorname{Tr}\boldsymbol{AQ}) = \mathcal{O}_z(n^{-2})$.

Derivations with Stein's lemma

From the observation that $\boldsymbol{Q}^{-1}\boldsymbol{Q} = \boldsymbol{I}_n$, i.e., $\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{XQ} - z\boldsymbol{Q} = \boldsymbol{I}_n$ with $\boldsymbol{X} = \boldsymbol{P} + \boldsymbol{N}$, our starting point is the following equation,

$$\frac{1}{p}\boldsymbol{P}^\top\boldsymbol{XQ} + \frac{1}{p}\boldsymbol{N}^\top\boldsymbol{XQ} - z\boldsymbol{Q} = \boldsymbol{I}_n. \tag{2.4}$$

Using Stein's lemma (Lemma 2.18) and Equation (2.3) for the derivatives of $\boldsymbol{Q}$, we seek expressions for the expectations of $\frac{1}{p}\boldsymbol{P}^\top\boldsymbol{XQ}$ and $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{XQ}$.

Let us start with $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{XQ}$.

$$\begin{aligned}\frac{1}{p}\mathbb{E}\big[\boldsymbol{N}^\top\boldsymbol{XQ}\big]_{i,j} &= \frac{1}{p}\sum_{k=1}^p\sum_{l=1}^n \mathbb{E}\big[N_{k,i}X_{k,l}Q_{l,j}\big] \\ &= \frac{1}{p}\sum_{k=1}^p\sum_{l=1}^n \mathbb{E}\left[\frac{\partial X_{k,l}}{\partial N_{k,i}}Q_{l,j} + X_{k,l}\frac{\partial Q_{l,j}}{\partial N_{k,i}}\right] \\ &= \frac{1}{p}\sum_{k=1}^p\sum_{l=1}^n \mathbb{E}\big[\delta_{l,i}Q_{l,j}\big] - \frac{1}{p^2}\sum_{k=1}^p\sum_{l=1}^n \mathbb{E}\big[X_{k,l}\big[Q_{l,i}[\boldsymbol{XQ}]_{k,j} + Q_{i,j}[\boldsymbol{XQ}]_{k,l}\big]\big] \\ &= \mathbb{E}[\boldsymbol{Q}]_{i,j} - \frac{1}{p^2}\mathbb{E}\big[\boldsymbol{QX}^\top\boldsymbol{XQ} + \boldsymbol{Q}\operatorname{Tr}\big(\boldsymbol{X}^\top\boldsymbol{XQ}\big)\big]_{i,j}.\end{aligned}$$

Here, recall that, from $\boldsymbol{Q}^{-1}\boldsymbol{Q} = \boldsymbol{I}_n$, we have $\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{XQ} = z\boldsymbol{Q} + \boldsymbol{I}_n$. Hence,

$$\begin{aligned}\frac{1}{p}\mathbb{E}\big[\boldsymbol{N}^\top\boldsymbol{XQ}\big] &= \mathbb{E}\left[\boldsymbol{Q} - \frac{1}{p}\boldsymbol{Q}(z\boldsymbol{Q}+\boldsymbol{I}_n) - \frac{1}{p}\boldsymbol{Q}\operatorname{Tr}(z\boldsymbol{Q}+\boldsymbol{I}_n)\right] \\ &= \mathbb{E}\left[\left(1-\frac{1}{p}-\frac{n}{p}\right)\boldsymbol{Q} - z\left(\frac{1}{p}\boldsymbol{Q}^2 + \frac{1}{p}\boldsymbol{Q}\operatorname{Tr}\boldsymbol{Q}\right)\right].\end{aligned}$$

Thus, the expectation of Equation (2.4) becomes

$$\frac{1}{p}\mathbb{E}\big[\boldsymbol{P}^\top\boldsymbol{XQ}\big] - \mathbb{E}\big[zc^{-1}m_n(z)\boldsymbol{Q} - \big(1-c^{-1}-z\big)\boldsymbol{Q}\big] + \frac{1}{p}\mathbb{E}\big[\boldsymbol{Q} + z\boldsymbol{Q}^2\big] = \boldsymbol{I}_n \tag{2.5}$$

where $m_n(z) \overset{\text{def}}{=} \frac{1}{n}\operatorname{Tr}\boldsymbol{Q}(z)$ is the Stieltjes transform of the spectral distribution of $\boldsymbol{K}$.

Then, we consider $\frac{1}{p}\boldsymbol{P}^\top\boldsymbol{XQ}$ using, again, Stein's lemma and Equation (2.3).

$$\frac{1}{p}\mathbb{E}\big[\boldsymbol{P}^\top\boldsymbol{XQ}\big]_{i,j} = \frac{1}{p}\mathbb{E}\big[\boldsymbol{P}^\top\boldsymbol{PQ}\big]_{i,j} + \frac{1}{p}\sum_{k=1}^p\sum_{l=1}^n \mathbb{E}\big[P_{k,i}N_{k,l}Q_{l,j}\big]$$

$$
\begin{aligned}
&= \frac{1}{p}\mathbb{E}\big[\boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{Q}\big]_{i,j} + \frac{1}{p}\sum_{k=1}^{p}\sum_{l=1}^{n}\mathbb{E}\left[P_{k,i}\frac{\partial Q_{l,j}}{\partial N_{k,l}}\right] \\
&= \frac{1}{p}\mathbb{E}\big[\boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{Q}\big]_{i,j} - \frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{n}\mathbb{E}\big[P_{k,i}\big[Q_{l,l}[\boldsymbol{X}\boldsymbol{Q}]_{k,j} + Q_{l,j}[\boldsymbol{X}\boldsymbol{Q}]_{k,l}\big]\big] \\
&= \mathbb{E}\left[\frac{1}{p}\boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{Q} - \frac{1}{p^2}\boldsymbol{P}^\top \boldsymbol{X}\boldsymbol{Q}\operatorname{Tr}\boldsymbol{Q} - \frac{1}{p^2}\boldsymbol{P}^\top \boldsymbol{X}\boldsymbol{Q}^2\right]_{i,j}.
\end{aligned}
$$

Hence $\frac{1}{p}\mathbb{E}[(1+c^{-1}m_n(z))\boldsymbol{P}^\top\boldsymbol{X}\boldsymbol{Q}] = \frac{1}{p}\mathbb{E}[\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{Q}] - \frac{1}{p^2}\mathbb{E}[\boldsymbol{P}^\top\boldsymbol{X}\boldsymbol{Q}^2]$ and, recalling that $|m_n(z)-\mathbb{E}[m_n(z)]| \to 0$ almost surely for all $z \in \mathbb{C}\setminus\mathbb{R}$, we have $\|\frac{1}{p}(1+c^{-1}\mathbb{E}[m_n(z)])\mathbb{E}[\boldsymbol{P}^\top\boldsymbol{X}\boldsymbol{Q}] - \frac{1}{p}\mathbb{E}[\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{Q}]\| \to 0$ as $p,n \to +\infty$. Moreover, since $\mathbb{E}[m_n]$ is the Stieltjes transform of a probability distribution on $[0,+\infty[$ (Corollary 2.9 and Proposition 2.7), we can see that $\frac{1}{|z(1+c^{-1}\mathbb{E}[m_n(z)])|} \leqslant \frac{1}{|\Im[z+c^{-1}z\mathbb{E}[m_n(z)]]|} \leqslant \frac{1}{|\Im z|}$ thus $\frac{1}{|1+c^{-1}\mathbb{E}[m_n(z)]|} \leqslant \frac{|z|}{|\Im z|}$ and

$$
\left\| \frac{1}{p}\mathbb{E}\big[\boldsymbol{P}^\top\boldsymbol{X}\boldsymbol{Q}\big] - \frac{1}{1+c^{-1}\mathbb{E}[m_n(z)]}\frac{1}{p}\boldsymbol{P}^\top\boldsymbol{P}\mathbb{E}[\boldsymbol{Q}] \right\| \xrightarrow[p,n\to+\infty]{} 0. \tag{2.6}
$$

2.3.2 Limiting Spectral Distribution

Expansion of $\mathbb{E}[m_n(z)]$

In order to study the limiting spectral distribution (LSD) of $\boldsymbol{K}$, we consider the simpler setting where $\boldsymbol{P} = \boldsymbol{0}_{p\times n}$. Indeed, since $\boldsymbol{P}$ is a deterministic low-rank perturbation, its addition does not change the LSD: denoting $\boldsymbol{Q}_0$ the resolvent of $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}$, the resolvent identity (Proposition 2.21) yields

$$
\frac{1}{n}\operatorname{Tr}\boldsymbol{Q}_0 - \frac{1}{n}\operatorname{Tr}\boldsymbol{Q} = \frac{1}{np}\operatorname{Tr}\big[\boldsymbol{Q}_0\big(\boldsymbol{P}^\top\boldsymbol{P} + \boldsymbol{P}^\top\boldsymbol{N} + \boldsymbol{N}\boldsymbol{P}^\top\big)\boldsymbol{Q}\big] = \mathcal{O}_z(n^{-1}) \qquad \text{almost surely.}
$$

Then, applying $\frac{1}{n}\operatorname{Tr}$ to Equation (2.5), we find that $\mathbb{E}[m_n(z)]$ satisfies

$$
zc^{-1}\mathbb{E}[m_n(z)]^2 - \big(1-c^{-1}-z\big)\mathbb{E}[m_n(z)] + 1 = -\frac{1}{p}\mathbb{E}[m_n(z)] - \frac{z}{p}\mathbb{E}\big[m_n'(z)\big] + \mathcal{O}_z(n^{-2})
$$

where we have used the fact that $\frac{1}{n}\operatorname{Tr}\boldsymbol{Q}^2 = m_n'(z)$ and $\mathbb{E}[m_n^2(z)] = \mathbb{E}[m_n(z)]^2 + \mathcal{O}_z(n^{-2})$ (recall Remark 2.36). We know that the Stieltjes transform m_{MP} of the Marčenko-Pastur distribution is characterized by

$$
zc^{-1}m_{\mathrm{MP}}^2(z) - \big(1-c^{-1}-z\big)m_{\mathrm{MP}}(z) + 1 = 0. \tag{2.7}
$$

Subtracting the last two equations and factorizing $\mathbb{E}[m_n(z)] - m_{\mathrm{MP}}(z)$ yields

$$
(\mathbb{E}[m_n(z)] - m_{\mathrm{MP}}(z))\big(zc^{-1}(\mathbb{E}[m_n(z)] + m_{\mathrm{MP}}(z)) - \big(1-c^{-1}-z\big)\big) = -\frac{1}{p}\mathbb{E}[m_n(z)] - \frac{z}{p}\mathbb{E}\big[m_n'(z)\big] + \mathcal{O}_z(n^{-2}).
$$

Hence, $\mathbb{E}[m_n(z)] - m_{\mathrm{MP}}(z) = \frac{1}{p}g_n(z)\mathbb{E}[m_n(z) + zm_n'(z)] + \mathcal{O}_z(n^{-2})$ where

$$
g_n : z \mapsto \frac{-1}{z\Big(1 + c^{-1}\mathbb{E}[m_n(z)] + c^{-1}m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\Big)}
$$

satisfies $|g_n(z)| \leqslant |\Im z|^{-1}$ because $\Im z$ and $\Im[z(c^{-1}\mathbb{E}[m_n(z)] + c^{-1}m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z})]$ have the same sign since $\mathbb{E}[m_n]$ and m_{MP} are Stieltjes transforms of probability distributions on $[0,+\infty[$ (see Corollary 2.9 and Proposition 2.7).

Lemma 2.37. $\mathbb{E}[m_n(z)] = m_{\mathrm{MP}}(z) + \frac{1}{p}h(z) + \mathcal{O}_z(n^{-2})$ *with* $h(z) = \frac{-z m_{\mathrm{MP}}(z)\left(c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\right)}{z^2\left(1+2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\right)^2}$.

Proof. Our goal is to show that $\frac{1}{p} g_n(z)\mathbb{E}[m_n(z) + z m_n'(z)] = \frac{1}{p}h(z) + \mathcal{O}_z(n^{-2})$. Since $|g_n(z)|, |m_n(z)| \leqslant |\Im z|^{-1}$ and $|m_n'(z)| \leqslant |\Im z|^{-2}$, we have $\frac{1}{p} g_n(z)\mathbb{E}[m_n(z) + z m_n'(z)] = \mathcal{O}_z(n^{-1})$ and $\mathbb{E}[m_n(z)] = m_{\mathrm{MP}}(z) + \mathcal{O}_z(n^{-1})$. Moreover, we have the following results, which we prove below.

$$g_n(z) = \frac{-1}{z\left(1 + 2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\right)} + \mathcal{O}_z(n^{-1}), \tag{2.8}$$

$$\mathbb{E}\big[m_n'(z)\big] = \frac{-\big(1 + c^{-1} m_{\mathrm{MP}}(z)\big) m_{\mathrm{MP}}(z)}{z\left(1 + 2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\right)} + \mathcal{O}_z(n^{-1}). \tag{2.9}$$

Then, we find the stated result by observing that $|z(1 + 2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z})|^{-1} \leqslant |\Im z|^{-1}$ because $\Im z$ and $\Im[z(2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z})]$ have the same sign since m_{MP} is the Stieltjes transform of a probability distribution on $[0, +\infty[$ (Proposition 2.7). Indeed,

$$\begin{aligned} g_n(z)\mathbb{E}\big[m_n(z) + z m_n'(z)\big] &= -\frac{m_{\mathrm{MP}}(z)}{z\left(1 + 2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\right)} + \frac{z\big(1 + c^{-1} m_{\mathrm{MP}}(z)\big) m_{\mathrm{MP}}(z)}{z^2\left(1 + 2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\right)^2} + \mathcal{O}_z(n^{-1}) \\ &= h(z) + \mathcal{O}_z(n^{-1}). \end{aligned}$$

Proof of Equation (2.8)**.** Since $\mathbb{E}[m_n(z)] = m_{\mathrm{MP}}(z) + \mathcal{O}_z(n^{-1})$, we have

$$g_n(z) = \frac{-1}{z\left(1 + 2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\right)} + \frac{\mathcal{O}_z(n^{-1})}{z\left(1 + 2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\right)^2}$$

and the modulus of $\frac{-1}{z\left(1+2c^{-1} m_{\mathrm{MP}}(z) - \frac{1-c^{-1}}{z}\right)}$ is bounded by $|\Im z|^{-1}$ therefore the last term is $\mathcal{O}_z(n^{-1})$.

Proof of Equation (2.9)**.** Since $m_n' = \frac{1}{n}\operatorname{Tr} \boldsymbol{Q}^2$, using the Poincaré-Nash inequality (Lemma 2.19) we find that

$$\operatorname{Var}(m_n'(z)) \leqslant \frac{16}{np^2}\mathbb{E}\big[\|\boldsymbol{X}\|^2 \|\boldsymbol{Q}\|^6\big] = \mathcal{O}_z(n^{-2}).$$

Taking $\frac{1}{n}\operatorname{Tr}$ of Equation (2.5) with $\boldsymbol{P} = \boldsymbol{0}_{p\times n}$ yields

$$\mathbb{E}\left[z c^{-1} m_n^2(z) - \big(1 - c^{-1} - z\big) m_n(z) + 1 + \frac{1}{p} m_n(z) + \frac{z}{p} m_n'(z)\right] = 0.$$

In order to differentiate this expression, we must swap $\frac{\partial}{\partial z}$ and $\mathbb{E}$, which is possible because $z \mapsto |z c^{-1} m_n^2(z) - (1 - c^{-1} - z) m_n(z) + 1 + \frac{1}{p} m_n(z) + \frac{z}{p} m_n'(z)|$ can be upper bounded on every compact subset of $\mathbb{C} \setminus \mathbb{R}$. Hence, we find that, for all $z \in \mathbb{C} \setminus \mathbb{R}$,

$$\mathbb{E}\left[c^{-1} m_n^2(z) + 2z c^{-1} m_n(z) m_n'(z) + m_n(z) - \big(1 - c^{-1} - z\big) m_n'(z) + \frac{2}{p} m_n'(z) + \frac{z}{p} m_n''(z)\right] = 0.$$

Then, we use the relations $\mathbb{E}[\frac{2}{p} m_n'(z) + \frac{z}{p} m_n''(z)] = \mathcal{O}_z(n^{-1})$, $\mathbb{E}[m_n^2(z)] = \mathbb{E}[m_n(z)]^2 + \mathcal{O}_z(n^{-2})$ as well as $\mathbb{E}[m_n(z) m_n'(z)] = \mathbb{E}[m_n(z)]\mathbb{E}[m_n'(z)] + \mathcal{O}_z(n^{-2})$ — from the Cauchy-Schwarz inequality with the fact that $\operatorname{Var}(m_n(z))$ and $\operatorname{Var}(m_n'(z))$ are $\mathcal{O}_z(n^{-2})$ — and our previous equation becomes

$$\big(2z c^{-1}\mathbb{E}[m_n(z)] - \big(1 - c^{-1} - z\big)\big)\mathbb{E}\big[m_n'(z)\big] = -\big(1 + c^{-1}\mathbb{E}[m_n(z)]\big)\mathbb{E}[m_n(z)] + \mathcal{O}_z(n^{-1}).$$

Finally, noticing that $|z(1+2c^{-1}\mathbb{E}[m_n(z)]-\frac{1-c^{-1}}{z})| \leqslant |\Im z|^{-1}$ because $\Im z$ and $\Im[z(2c^{-1}\mathbb{E}[m_n(z)]-\frac{1-c^{-1}}{z})]$ have the same sign since $\mathbb{E}[m_n]$ is the Stieltjes transform of a probability distribution on $[0,+\infty[$ (Corollary 2.9 and Proposition 2.7), we have $\mathbb{E}[m_n'(z)] = \frac{-(1+c^{-1}\mathbb{E}[m_n(z)])\mathbb{E}[m_n(z)]}{z\left(1+2c^{-1}\mathbb{E}[m_n(z)]-\frac{1-c^{-1}}{z}\right)} + \mathscr{O}_z(n^{-1})$ and we just need to recall that $\mathbb{E}[m_n(z)] = m_{\mathrm{MP}}(z) + \mathscr{O}_z(n^{-1})$ to conclude. □

Remark 2.38. We can check that the function $z \mapsto \frac{-1}{z\left(1+2c^{-1}m_{\mathrm{MP}}(z)-\frac{1-c^{-1}}{z}\right)}$ is the Stieltjes transform of a probability measure with the same support as μ_{MP} (Proposition 2.7 and Theorem 2.8). Thus, h can be analytically extended on $\mathbb{C}\setminus\operatorname{Supp}\mu_{\mathrm{MP}}$ and $h(x)\in\mathbb{R}$ for all $x\in\mathbb{R}\setminus\operatorname{Supp}\mu_{\mathrm{MP}}$.

Confinement of the Spectrum

We have just found the following expansion: $\mathbb{E}[m_n(z)] = m_{\mathrm{MP}}(z) + \frac{1}{p}h(z) + \mathscr{O}_z(n^{-2})$. Note that this shows that the limiting spectral distribution of $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}$ (and therefore that of $\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$) is the Marčenko-Pastur distribution with parameter c^{-1}. We will now use this expansion to prove that, almost surely, no eigenvalue of $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}$ stays outside the support of the Marčenko-Pastur distribution as $p,n\to+\infty$. That is, for all $\varepsilon>0$, there exists an integer n_0 such that $\max_{\lambda\in\mathrm{Sp}\,\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}}\mathrm{Dist}(\lambda,\operatorname{Supp}\mu_{\mathrm{MP}})\leqslant\varepsilon$ as soon as $n\geqslant n_0$.

Let $\varepsilon>0$ and $\mathscr{S}_\varepsilon=\{x\in\mathbb{R}\mid\mathrm{Dist}(x,\operatorname{Supp}\mu_{\mathrm{MP}})<\varepsilon\}$. Let $\varphi:\mathbb{R}\to[0,1]$ be an infinitely differentiable function which equals 1 on $\operatorname{Supp}\mu_{\mathrm{MP}}$ and 0 on $\mathbb{R}\setminus\mathscr{S}_\varepsilon$. We also define $\psi=1-\varphi$. Then, we must show that $\operatorname{Tr}\psi(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}) = \sum_{\lambda\in\mathrm{Sp}\,\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}}\psi(\lambda)\to 0$ almost surely as $p,n\to+\infty$. This implies the confinement of the spectrum (otherwise, we would have $\sum_{\lambda\in\mathrm{Sp}\,\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}}\psi(\lambda)\geqslant 1$).

From the Helffer-Sjöstrand formula (Proposition 2.12), we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\operatorname{Tr}\varphi\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\right] &= \frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar z}(z)\mathbb{E}[m_n(z)]\,\mathrm{d}z\\ &= \frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar z}(z)m_{\mathrm{MP}}(z)\,\mathrm{d}z + \frac{1}{p}\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar z}(z)h(z)\,\mathrm{d}z + \frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar z}(z)\times\mathscr{O}_z(n^{-2})\,\mathrm{d}z.\end{aligned}$$

The first integral is $\int_{\mathbb{R}}\varphi\,\mathrm{d}\mu_{\mathrm{MP}}=1$ (from Proposition 2.12 and the fact that φ is constant equal to 1 on $\operatorname{Supp}\mu_{\mathrm{MP}}$) and we can chose q large enough so that $(\mathrm{i}\Im z)^q$ in the expression of $\frac{\partial\Phi_q[\varphi]}{\partial\bar z}$ compensates for the divergence of the $\mathscr{O}_z(n^{-2})$ in the last integral (see Remark 2.13), which is therefore a $\mathscr{O}(n^{-2})$. It remains to evaluate the second integral, which we can do with Lemma 2.14. Indeed, h is analytic on $\mathbb{C}\setminus\operatorname{Supp}\mu_{\mathrm{MP}}$, $h(z)\to 0$ as $|z|\to+\infty$ and $h(\bar z)=\overline{h(z)}$. Furthermore, since $z\mapsto c^{-1}m_{\mathrm{MP}}(z)-\frac{1-c^{-1}}{z}$ and $z\mapsto\frac{-1}{z\left(1+2c^{-1}m_{\mathrm{MP}}(z)-\frac{1-c^{-1}}{z}\right)}$ are two Stieltjes transforms of probability distributions on $\operatorname{Supp}\mu_{\mathrm{MP}}$, we have

$$|h(z)|\leqslant\frac{|z|}{\mathrm{Dist}(z,\operatorname{Supp}\mu_{\mathrm{MP}})^4}\leqslant\frac{E_+ + \mathrm{Dist}(z,\operatorname{Supp}\mu_{\mathrm{MP}})}{\mathrm{Dist}(z,\operatorname{Supp}\mu_{\mathrm{MP}})^4}\leqslant(E_+ + 1)\max\left(\mathrm{Dist}(z,\operatorname{Supp}\mu_{\mathrm{MP}})^{-4},1\right)$$

where E_+ is the right-edge of the Marčenko-Pastur distribution. As a consequence, for $q\geqslant 4$,

$$\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar z}(z)h(z)\,\mathrm{d}z = \lim_{y\downarrow 0}\frac{1}{\pi}\int_{\operatorname{Supp}\mu_{\mathrm{MP}}}\varphi(x)\Im[h(x+\mathrm{i}y)]\,\mathrm{d}x$$

and, as per Lemma 2.14, this is equal to 0 since $\varphi(x)=1$ for all $x\in\operatorname{Supp}\mu_{\mathrm{MP}}$ and $\lim_{y\to+\infty}-\mathrm{i}yh(\mathrm{i}y)=0$. Thus, we have $\mathbb{E}[\frac{1}{n}\operatorname{Tr}\varphi(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N})] = 1+\mathscr{O}(n^{-2})$ and therefore $\mathbb{E}[\operatorname{Tr}\psi(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N})] = \mathscr{O}(n^{-1})$.

In order to conclude on the almost sure convergence of $\operatorname{Tr}\psi(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N})$ to 0, we just need to show that $\operatorname{Var}(\operatorname{Tr}\psi(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N})) = \mathcal{O}(n^{-2})$ and apply Lemma 2.20. As usual, this is performed with the Poincaré-Nash inequality (Lemma 2.19).

$$\begin{aligned}
\operatorname{Var}\left(\operatorname{Tr}\psi\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\right) = \operatorname{Var}\left(\operatorname{Tr}\varphi\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\right) &\leqslant \sum_{i=1}^{p}\sum_{j=1}^{n}\mathbb{E}\left[\left|\frac{\partial \operatorname{Tr}\varphi\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)}{\partial N_{i,j}}\right|^2\right] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{n}\mathbb{E}\left[\left|\operatorname{Tr}\left(\varphi'\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\frac{\partial}{\partial N_{i,j}}\left[\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right]\right)\right|^2\right] \\
&= \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{n}\mathbb{E}\left[\left|\operatorname{Tr}\left(\varphi'\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\left[\boldsymbol{e}_j^{(n)}\boldsymbol{e}_i^{(p)\top}\boldsymbol{N} + \boldsymbol{N}^\top\boldsymbol{e}_i^{(p)}\boldsymbol{e}_j^{(n)\top}\right]\right)\right|^2\right] \\
&= \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{n}\mathbb{E}\left[\left|\left[2\boldsymbol{N}\varphi'\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\right]_{i,j}\right|^2\right] \\
&= \frac{4}{p^2}\mathbb{E}\left[\operatorname{Tr}\left(\varphi'\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\boldsymbol{N}^\top\boldsymbol{N}\varphi'\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\right)\right] \\
&= \frac{4}{p}\mathbb{E}\left[\operatorname{Tr}u\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\right]
\end{aligned}$$

where $u : x \mapsto x[\varphi'(x)]^2$ is an infinitely differentiable function with compact support which equals 0 on $\operatorname{Supp}\mu_{\mathrm{MP}}$. Hence, applying similarly the Helffer-Sjöstrand formula (Proposition 2.12), we have

$$\frac{4}{p}\mathbb{E}\left[\operatorname{Tr}u\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}\right)\right] = \frac{8c^{-1}}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[u]}{\partial\bar{z}}(z)\mathbb{E}[m_n(z)]\,\mathrm{d}z = \mathcal{O}(n^{-2})$$

for q chosen sufficiently large.

Intermediate Conclusion

The limiting spectral distribution of $\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$ is the Marčenko-Pastur distribution with parameter c^{-1} and, almost surely, no eigenvalue of $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}$ stays outside the support of this limiting spectral distribution as $p, n \to +\infty$.

2.3.3 Spike Behavior

We are now interested in the "spike" eigenvalues and eigenvectors. That is, the eigenvalues of $\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$ (and their corresponding eigenvectors) staying outside $\operatorname{Supp}\mu_{\mathrm{MP}}$ and which are due to the presence of a signal $\boldsymbol{P} = \boldsymbol{L}\boldsymbol{V}^\top$.

Isotropic Marčenko-Pastur

As preliminaries, we must show a useful result for the upcoming analysis — the almost sure convergence $\boldsymbol{a}^\top\boldsymbol{Q}_0\boldsymbol{b} - m_{\mathrm{MP}}(z)\langle\boldsymbol{a},\boldsymbol{b}\rangle \to 0$ as $p, n \to +\infty$ for all (sequences of) deterministic vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ with bounded norm and $z \in \mathbb{C}\setminus\operatorname{Supp}\mu_{\mathrm{MP}}$. Recall that $\boldsymbol{Q}_0$ is the resolvent of $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}$ therefore, from the confinement of the spectrum of the latter, if $x \in \mathbb{R}\setminus\operatorname{Supp}\mu_{\mathrm{MP}}$, $\boldsymbol{Q}_0(x)$ is properly defined for p, n sufficiently large. Given the concentration of the resolvent proven in Section 2.3.1, we only need to show that $\boldsymbol{a}^\top\mathbb{E}[\boldsymbol{Q}_0]\boldsymbol{b} - m_{\mathrm{MP}}(z)\langle\boldsymbol{a},\boldsymbol{b}\rangle \to 0$.

From Equation (2.5), we have

$$-zc^{-1}\mathbb{E}\big[m_n(z)\boldsymbol{a}^\top\boldsymbol{Q}_0\boldsymbol{b}\big]+\big(1-c^{-1}-z\big)\boldsymbol{a}^\top\mathbb{E}[\boldsymbol{Q}_0]\boldsymbol{b}+\mathscr{O}_z(n^{-1})=\langle\boldsymbol{a},\boldsymbol{b}\rangle.$$

Then, since $\mathrm{Cov}(m_n(z),\boldsymbol{a}^\top\boldsymbol{Q}_0\boldsymbol{b})=\mathscr{O}_z(n^{-3/2})$ and $\mathbb{E}[m_n(z)]=m_{\mathrm{MP}}(z)+\mathscr{O}_z(n^{-1})$, this simplifies into

$$-\big(zc^{-1}m_{\mathrm{MP}}(z)-\big(1-c^{-1}-z\big)\big)\boldsymbol{a}^\top\mathbb{E}[\boldsymbol{Q}_0]\boldsymbol{b}=\langle\boldsymbol{a},\boldsymbol{b}\rangle+\mathscr{O}_z(n^{-1}).$$

Finally, notice that $-(zc^{-1}m_{\mathrm{MP}}(z)-(1-c^{-1}-z))=\frac{1}{m_{\mathrm{MP}}(z)}$, hence $\boldsymbol{a}^\top\mathbb{E}[\boldsymbol{Q}_0]\boldsymbol{b}=m_{\mathrm{MP}}(z)\langle\boldsymbol{a},\boldsymbol{b}\rangle+\mathscr{O}_z(n^{-1})$.

Isolated Eigenvalues

An eigenvalue of $\boldsymbol{K}=\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$ is isolated if it stays outside $\mathrm{Supp}\,\mu_{\mathrm{MP}}$ as $p,n\to+\infty$. Thus, let us seek $\xi\in\mathbb{R}\setminus\mathrm{Supp}\,\mu_{\mathrm{MP}}$ such that $\det(\boldsymbol{K}-\xi\boldsymbol{I}_n)=0$. After developing $\boldsymbol{K}$ in terms of $\boldsymbol{P}$ and $\boldsymbol{N}$, we find that ξ is characterized by

$$\det\left(\boldsymbol{I}_n+\frac{1}{p}\big[\boldsymbol{P}^\top\boldsymbol{P}+\boldsymbol{P}^\top\boldsymbol{N}+\boldsymbol{N}^\top\boldsymbol{P}\big]\boldsymbol{Q}_0(\xi)\right)\det\left(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}-\xi\boldsymbol{I}_n\right)=0$$

and, from the confinement of the spectrum of $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}$, we can assume p,n sufficiently large so that $\det(\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}-\xi\boldsymbol{I}_n)\neq0$. Using the decomposition $\boldsymbol{P}=\boldsymbol{L}\boldsymbol{V}^\top$, the sum of matrix products $\boldsymbol{P}^\top\boldsymbol{P}+\boldsymbol{P}^\top\boldsymbol{N}+\boldsymbol{N}^\top\boldsymbol{P}$ can be written as $\begin{bmatrix}\sqrt{n}\boldsymbol{V} & \sqrt{n}\boldsymbol{V} & \frac{1}{\sqrt{n}}\boldsymbol{N}^\top\boldsymbol{L}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{n}}\boldsymbol{L}^\top\boldsymbol{L}\boldsymbol{V}^\top\\ \frac{1}{\sqrt{n}}\boldsymbol{L}^\top\boldsymbol{N}\\ \sqrt{n}\boldsymbol{V}^\top\end{bmatrix}$ and we can use Sylvester's identity (Proposition 2.22) to turn the $n\times n$ determinant into a $3K\times3K$ determinant:

$$\det(\boldsymbol{I}_{3K}+\boldsymbol{M})=0\qquad\text{with}\qquad\boldsymbol{M}=\frac{1}{p}\begin{bmatrix}\boldsymbol{L}^\top\boldsymbol{L}\boldsymbol{V}^\top\boldsymbol{Q}_0(\xi)\boldsymbol{V} & \boldsymbol{L}^\top\boldsymbol{L}\boldsymbol{V}^\top\boldsymbol{Q}_0(\xi)\boldsymbol{V} & \frac{1}{n}\boldsymbol{L}^\top\boldsymbol{L}\boldsymbol{V}^\top\boldsymbol{Q}_0(\xi)\boldsymbol{N}^\top\boldsymbol{L}\\ \boldsymbol{L}^\top\boldsymbol{N}\boldsymbol{Q}_0(\xi)\boldsymbol{V} & \boldsymbol{L}^\top\boldsymbol{N}\boldsymbol{Q}_0(\xi)\boldsymbol{V} & \frac{1}{n}\boldsymbol{L}^\top\boldsymbol{N}\boldsymbol{Q}_0(\xi)\boldsymbol{N}^\top\boldsymbol{L}\\ n\boldsymbol{V}^\top\boldsymbol{Q}_0(\xi)\boldsymbol{V} & n\boldsymbol{V}^\top\boldsymbol{Q}_0(\xi)\boldsymbol{V} & \boldsymbol{V}^\top\boldsymbol{Q}_0(\xi)\boldsymbol{N}^\top\boldsymbol{L}\end{bmatrix}.$$

We have $\boldsymbol{V}^\top\boldsymbol{Q}_0(\xi)\boldsymbol{V}\to m_{\mathrm{MP}}(\xi)\boldsymbol{I}_K$ almost surely as $p,n\to+\infty$ (from the convergence of bilinear forms proven in the previous paragraph and the fact that $\boldsymbol{V}^\top\boldsymbol{V}=\boldsymbol{I}_K$ with K finite) therefore the blocks $(3,1)$ and $(3,2)$ of $\boldsymbol{M}$ converge to $c^{-1}m_{\mathrm{MP}}(\xi)\boldsymbol{I}_K$ and the blocks $(1,1)$ and $(1,2)$ converge to $c^{-1}m_{\mathrm{MP}}(\xi)\boldsymbol{\Lambda}$ (recall that $\frac{1}{n}\boldsymbol{L}^\top\boldsymbol{L}\to\boldsymbol{\Lambda}$). The block $(2,3)$ can be handled with the co-resolvent $\widetilde{\boldsymbol{Q}}_0(\xi)=(\frac{1}{p}\boldsymbol{N}\boldsymbol{N}^\top-\xi\boldsymbol{I}_p)$:

$$\frac{1}{p}\frac{1}{n}\boldsymbol{L}^\top\boldsymbol{N}\boldsymbol{Q}_0(\xi)\boldsymbol{N}^\top\boldsymbol{L}=\frac{1}{n}\boldsymbol{L}^\top\left(\frac{1}{p}\boldsymbol{N}\boldsymbol{N}^\top\widetilde{\boldsymbol{Q}}_0(\xi)\right)\boldsymbol{L}=\frac{1}{n}\boldsymbol{L}^\top\big(\xi\widetilde{\boldsymbol{Q}}_0(\xi)+\boldsymbol{I}_p\big)\boldsymbol{L}\xrightarrow[p,n\to+\infty]{\text{a.s.}}(\xi\widetilde{m}_{\mathrm{MP}}(\xi)+1)\boldsymbol{\Lambda}$$

where $\widetilde{m}_{\mathrm{MP}}(\xi)=c^{-1}m_{\mathrm{MP}}(\xi)-\frac{1-c^{-1}}{\xi}$ hence the limit is equal to $c^{-1}(\xi m_{\mathrm{MP}}(\xi)+1)\boldsymbol{\Lambda}$. The blocks $(1,3)$, $(2,1)$, $(2,2)$ and $(3,3)$ of $\boldsymbol{M}$ vanish almost surely as $p,n\to+\infty$ as per the following lemma.

Lemma 2.39. *For all bounded (sequences of) vectors $\boldsymbol{a}\in\mathbb{R}^p$, $\boldsymbol{b}\in\mathbb{R}^n$, $\frac{1}{\sqrt{p}}|\boldsymbol{a}^\top\boldsymbol{N}\boldsymbol{Q}_0(\xi)\boldsymbol{b}|\to0$ almost surely as $p,n\to+\infty$.*

Proof. Let $\boldsymbol{N}=\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{H}^\top$ be the singular value decomposition of $\boldsymbol{N}$ with $\boldsymbol{\Sigma}=\mathrm{Diag}(s_1(\boldsymbol{N}),\ldots,s_{\min(p,n)}(\boldsymbol{N}))$. Then $\boldsymbol{N}\boldsymbol{Q}_0(\xi)=\boldsymbol{G}\boldsymbol{\Sigma}(\frac{1}{p}\boldsymbol{\Sigma}^2-\xi\boldsymbol{I}_{\min(p,n)})^{-1}\boldsymbol{H}^\top$. Hence, $\frac{1}{\sqrt{p}}|\boldsymbol{a}^\top\boldsymbol{N}\boldsymbol{Q}_0(\xi)\boldsymbol{b}|\leqslant\frac{1}{\sqrt{p}}\|\boldsymbol{N}\boldsymbol{Q}_0(\xi)\|\,|\boldsymbol{a}^\top\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{b}|$ where $\frac{1}{\sqrt{p}}\|\boldsymbol{N}\boldsymbol{Q}_0(\xi)\|=\mathscr{O}(1)$ almost surely so our goal is to show that $|\boldsymbol{a}^\top\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{b}|\to0$ almost surely. According to Chikuse (2003, Theorem 2.2.1), $\boldsymbol{G}$ (resp. $\boldsymbol{H}$) follows a uniform distribution on the Stiefel manifold $V_{\min(p,n)}(\mathbb{R}^p)$ (resp. $V_{\min(p,n)}(\mathbb{R}^n)$), that is, the set of orthonormal $\min(p,n)$-frames in $\mathbb{R}^p$ (resp. $\mathbb{R}^n$). Therefore, $\boldsymbol{a}^\top\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{b}=\boldsymbol{a}^\top\boldsymbol{O}_a^\top\boldsymbol{O}_a\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{O}_b^\top\boldsymbol{O}_b\boldsymbol{b}=\|\boldsymbol{a}\|\|\boldsymbol{b}\|\boldsymbol{e}_1^{(p)\top}\boldsymbol{O}_a\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{O}_b^\top\boldsymbol{e}_1^{(n)}$ for two well-chosen orthogonal matrices $\boldsymbol{O}_a\in\mathscr{O}_p(\mathbb{R})$, $\boldsymbol{O}_b\in\mathscr{O}_n(\mathbb{R})$, and this quantity is identically distributed to

$\|\boldsymbol{a}\|\|\boldsymbol{b}\|\boldsymbol{e}_1^{(p)\top}\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{e}_1^{(n)}$ by the properties of the uniform distribution on the Stiefel manifold (Chikuse, 2003, Theorem 2.2.1). Moreover, if $n \leqslant p$, then $V_{\min(p,n)}(\mathbb{R}^n) = \mathcal{O}_n(\mathbb{R})$ thus $\boldsymbol{G}\boldsymbol{H}^\top$ is uniformly distributed on $V_n(\mathbb{R}^p)$ so $\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{e}_1^{(n)}$ is uniformly distributed on $V_1(\mathbb{R}^p) = \mathbb{S}^{p-1}$ and $(\boldsymbol{e}_1^{(p)\top}\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{e}_1^{(n)})^2$ follows a beta distribution[10] with parameters $\frac{1}{2}, \frac{p-1}{2}$. Consequently, we have $\mathbb{E}[(\boldsymbol{e}_1^{(p)\top}\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{e}_1^{(n)})^2] = \frac{1}{p}$ and $\mathrm{Var}((\boldsymbol{e}_1^{(p)\top}\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{e}_1^{(n)})^2) = \frac{2(p-1)}{p^2(p+2)}$, which shows, by Lemma 2.20, that $(\boldsymbol{e}_1^{(p)\top}\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{e}_1^{(n)})^2 \to 0$ almost surely and therefore $|\boldsymbol{a}^\top\boldsymbol{G}\boldsymbol{H}^\top\boldsymbol{b}| \to 0$ almost surely. If $p \leqslant n$, then with the same reasoning we find that $(\boldsymbol{e}_1^{(n)\top}\boldsymbol{H}\boldsymbol{G}^\top\boldsymbol{e}_1^{(p)})^2$ follows a beta distribution with parameters $\frac{1}{2}, \frac{n-1}{2}$ hence the same result. □

Eventually, from the continuity of $\boldsymbol{A} \mapsto \det \boldsymbol{A}$ in the entries of $\boldsymbol{A}$, we have asymptotically

$$\det\begin{bmatrix} \boldsymbol{I}_K + c^{-1}m_{\mathrm{MP}}(\xi)\boldsymbol{\Lambda} & c^{-1}m_{\mathrm{MP}}(\xi)\boldsymbol{\Lambda} & \boldsymbol{0}_{K\times K} \\ \boldsymbol{0}_{K\times K} & \boldsymbol{I}_K & c^{-1}(\xi m_{\mathrm{MP}}(\xi)+1)\boldsymbol{\Lambda} \\ c^{-1}m_{\mathrm{MP}}(\xi)\boldsymbol{I}_K & c^{-1}m_{\mathrm{MP}}(\xi)\boldsymbol{I}_K & \boldsymbol{I}_K \end{bmatrix} = 0$$

and, using twice the fact that $\det\left[\begin{smallmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{smallmatrix}\right] = \det(\boldsymbol{A}\boldsymbol{D} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{D})$ when $\boldsymbol{D}$ is invertible, we find

$$\det\big(\boldsymbol{I}_K + c^{-1}m_{\mathrm{MP}}(\xi)\boldsymbol{\Lambda} - c^{-2}m_{\mathrm{MP}}(\xi)(\xi m_{\mathrm{MP}}(\xi)+1)\boldsymbol{\Lambda}\big) = 0.$$

Then, notice that Equation (2.7) can be written as $(zm_z+1)(1+c^{-1}m_{\mathrm{MP}}(z)) - m_{\mathrm{MP}}(z) = 0$ therefore $zm_{\mathrm{MP}}(z)+1 = \frac{m_{\mathrm{MP}}(z)}{1+c^{-1}m_{\mathrm{MP}}(z)}$ and the previous equation becomes

$$\det\left(\boldsymbol{I}_K + \frac{c^{-1}m_{\mathrm{MP}}(\xi)}{1+c^{-1}m_{\mathrm{MP}}(\xi)}\boldsymbol{\Lambda}\right) = 0. \tag{2.10}$$

Hence we find that $1 + \frac{c^{-1}m_{\mathrm{MP}}(\xi)}{1+c^{-1}m_{\mathrm{MP}}(\xi)}\lambda_k(\boldsymbol{\Lambda}) = 0 \iff m_{\mathrm{MP}}(\xi) = \frac{-c}{\lambda_k(\boldsymbol{\Lambda})+1}$ for some $k \in [K]$. At this point, we must ask for which values of $\lambda_k(\boldsymbol{\Lambda})$ a ξ verifying this relation may exist. From the properties of the Stieltjes transform (Proposition 2.7), m_{MP} is an increasing function on all connected components of $\mathbb{R}\setminus\mathrm{Supp}\,\mu_{\mathrm{MP}}$. Furthermore, $m_{\mathrm{MP}}(x) \to 0$ as $x \to \pm\infty$. Thus, if $c^{-1} \leqslant 1$, then $\mathrm{Supp}\,\mu_{\mathrm{MP}} = [E_-, E_+]$ and m_{MP} must be positive on $]-\infty, E_-[$ and negative on $]E_+, +\infty[$. If $c^{-1} > 1$, then $\mathrm{Supp}\,\mu_{\mathrm{MP}} = \{0\}\cup[E_-, E_+]$ and m_{MP} can also be negative on $]0, E_-[$ but $\lim_{x\uparrow E_-} m_{\mathrm{MP}}(x) = -\sqrt{c}/(1-\sqrt{c})$ hence, for all $x \in]0, E_-[$, $m_{\mathrm{MP}}(x) < -c \leqslant -c/(\lambda_k(\boldsymbol{\Lambda})+1)$. As a result, an isolated eigenvalue can only lie on the right side of the spectrum, i.e., $\xi > E_+$. But, since $\lim_{x\downarrow E_+} m_{\mathrm{MP}}(x) = -c/(\sqrt{c}+1)$, m_{MP} maps $]E_+, +\infty[$ onto $]\frac{-c}{\sqrt{c}+1}, 0[$ and $m_{\mathrm{MP}}(\xi) = \frac{-c}{\lambda_k(\boldsymbol{\Lambda})+1}$ has a solution if, and only if, $\lambda_k(\boldsymbol{\Lambda}) > \sqrt{c}$. Assuming this is verified, we can inject the expression of $m_{\mathrm{MP}}(\xi)$ into Equation (2.7),

$$\xi c^{-1}\left(\frac{c}{\lambda_k(\boldsymbol{\Lambda})+1}\right)^2 + \left(1-c^{-1}-\xi\right)\frac{c}{\lambda_k(\boldsymbol{\Lambda})+1} + 1 = 0 \iff \xi = \frac{(\lambda_k(\boldsymbol{\Lambda})+c)(\lambda_k(\boldsymbol{\Lambda})+1)}{\lambda_k(\boldsymbol{\Lambda})c}.$$

Hence, for all $k \in [K]$,

$$\lambda_k(\boldsymbol{K}) \xrightarrow[p,n\to+\infty]{\text{a.s.}} \xi_k \overset{\text{def}}{=} \frac{(\ell_k+c)(\ell_k+1)}{\ell_k c} \qquad \text{with} \qquad \ell_k = \max(\sqrt{c}, \lambda_k(\boldsymbol{\Lambda})).$$

This shows the first convergence[11] of Theorem 2.34.

[10] This is because if $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$ then $\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}$ is uniformly distributed on $\mathbb{S}^{p-1}$ and $\frac{z_1^2}{\|\boldsymbol{z}\|^2} \sim \frac{X}{X+Y}$ with $X \sim \chi^2(1)$ and $Y \sim \chi^2(p-1)$ thus $\frac{X}{X+Y} \sim \mathrm{Beta}(\frac{1}{2}, \frac{p-1}{2})$.

[11] To verify that the solutions to $\det(\boldsymbol{K} - \xi\boldsymbol{I}_n) = 0$ indeed converge to the solutions to Equation (2.10), one can follow the same line of arguments as Chapon et al. (2014). The idea is to show that the number of solutions to each equation coincide in any interval outside the bulk.

Deterministic Equivalent of the Resolvent

From Equation (2.5), Equation (2.6) and the fact that $\mathbb{E}[m_n(z)] = m_{\mathrm{MP}}(z) + \mathcal{O}_z(n^{-1})$, we have,

$$\left\| \frac{1}{1+c^{-1}m_{\mathrm{MP}}(z)} \frac{1}{p} \boldsymbol{P}^\top \boldsymbol{P} \mathbb{E}[\boldsymbol{Q}] - \big(zc^{-1}m_{\mathrm{MP}}(z) - \big(1-c^{-1}-z\big)\big)\mathbb{E}[\boldsymbol{Q}] - \boldsymbol{I}_n \right\| \xrightarrow[p,n\to+\infty]{} 0$$

and $(zc^{-1}m_{\mathrm{MP}}(z) - (1-c^{-1}-z)) = -\frac{1}{m_{\mathrm{MP}}(z)}$. Therefore, from the decomposition $\boldsymbol{P} = \boldsymbol{L}\boldsymbol{V}^\top$ with $\frac{1}{n}\boldsymbol{L}^\top\boldsymbol{L} \to \boldsymbol{\Lambda}$ we can define the following *deterministic equivalent* (Definition 2.17) of $\boldsymbol{Q}(z)$:

$$\bar{\boldsymbol{Q}}(z) = m_{\mathrm{MP}}(z)\left(\frac{c^{-1}m_{\mathrm{MP}}(z)}{1+c^{-1}m_{\mathrm{MP}}(z)} \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top + \boldsymbol{I}_n \right)^{-1}.$$

Remark 2.40. Note that the convergences in Definition 2.17 hold for $\bar{\boldsymbol{Q}}(z)$ *uniformly* in $z \in \mathbb{C}\backslash \mathrm{Supp}\,\mu_{\mathrm{MP}}$.

Remark 2.41. From the definition of $\bar{\boldsymbol{Q}}$ and Equation (2.10) governing the asymptotic position of spike eigenvalues, we see that their is an exact correspondence between the later and the singular points of $z \mapsto \bar{\boldsymbol{Q}}(z)$. In particular Equation (2.10) can be written as $m_{\mathrm{MP}}(\xi)\det(\boldsymbol{V}^\top\bar{\boldsymbol{Q}}^{-1}(\xi)\boldsymbol{V}) = 0$.

Isolated Eigenvectors

Since $\boldsymbol{Q}(z) = \sum_{i=1}^n \frac{\boldsymbol{u}_i(\boldsymbol{K})\boldsymbol{u}_i(\boldsymbol{K})^\top}{\lambda_i(\boldsymbol{K})-z}$, with Cauchy's integral formula (Proposition 2.15), we have, for $k \in [K]$,

$$\langle \boldsymbol{v}_k, \boldsymbol{u}_k(\boldsymbol{K})\rangle^2 = -\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_k} \boldsymbol{v}_k^\top \boldsymbol{Q}(z)\boldsymbol{v}_k \,\mathrm{d}z \xrightarrow[p,n\to+\infty]{\text{a.s.}} -\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_k} \boldsymbol{v}_k^\top \bar{\boldsymbol{Q}}(z)\boldsymbol{v}_k \,\mathrm{d}z$$

where γ_k is a positively-oriented simple closed complex contour circling around $\lambda_k(\boldsymbol{K})$ and leaving all other eigenvalues outside (which is possible by the confinement of the spectrum of $\frac{1}{p}\boldsymbol{N}^\top\boldsymbol{N}$). Assuming $\lambda_k(\boldsymbol{\Lambda}) > \sqrt{c}$, the limiting integral can be computed with residue calculus (Proposition 2.16),

$$\begin{aligned} -\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_k} \boldsymbol{v}_k^\top \bar{\boldsymbol{Q}}(z)\boldsymbol{v}_k \,\mathrm{d}z &= -\lim_{z\to\xi_k} (z-\xi_k)\,\boldsymbol{v}_k^\top \bar{\boldsymbol{Q}}(z)\boldsymbol{v}_k \\ &= -\lim_{z\to\xi_k} (z-\xi_k) m_{\mathrm{MP}}(z)\left[\frac{c^{-1}m_{\mathrm{MP}}(z)}{1+c^{-1}m_{\mathrm{MP}}(z)}\ell_k + 1 \right]^{-1}. \end{aligned}$$

This last limit can be computed using L'Hôpital's rule,

$$\begin{aligned} -\lim_{z\to\xi_k} (z-\xi_k) m_{\mathrm{MP}}(z)\left[\frac{c^{-1}m_{\mathrm{MP}}(z)}{1+c^{-1}m_{\mathrm{MP}}(z)}\ell_k + 1 \right]^{-1} &= -\left(\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{c^{-1}}{1+c^{-1}m_{\mathrm{MP}}(z)}\ell_k + \frac{1}{m_{\mathrm{MP}}(z)} \right]_{z=\xi_k} \right)^{-1} \\ &= \left(\frac{c^{-2}m'_{\mathrm{MP}}(\xi_k)}{\big(1+c^{-1}m_{\mathrm{MP}}(\xi_k)\big)^2}\ell_k + \frac{m'_{\mathrm{MP}}(\xi_k)}{m^2_{\mathrm{MP}}(\xi_k)} \right)^{-1}. \end{aligned}$$

We already know that $m_{\mathrm{MP}}(\xi_k) = -\frac{c}{\ell_k+1}$. In order to find an expression for $m'_{\mathrm{MP}}(\xi_k)$, we differentiate Equation (2.7),

$$\begin{aligned} c^{-1}m^2_{\mathrm{MP}}(z) + 2zc^{-1}m_{\mathrm{MP}}(z)m'_{\mathrm{MP}}(z) + m_{\mathrm{MP}}(z) - \big(1-c^{-1}-z\big)m'_{\mathrm{MP}}(z) = 0 \\ \iff m'_{\mathrm{MP}}(z) = -\frac{c^{-1}m^2_{\mathrm{MP}}(z) + m_{\mathrm{MP}}(z)}{2zc^{-1}m_{\mathrm{MP}}(z) - \big(1-c^{-1}-z\big)}. \end{aligned}$$

With $z = \xi_k$, this simplifies into $m'_{\mathrm{MP}}(\xi_k) = \frac{1}{\ell_k^2-c}\left(\frac{\ell_k c}{\ell_k+1}\right)^2$ and, skipping straightforward calculations, we find that

$$\langle \boldsymbol{v}_k, \boldsymbol{u}_k(\boldsymbol{K})\rangle^2 \xrightarrow[p,n\to+\infty]{\text{a.s.}} 1 - \frac{\ell_k + c}{\ell_k(\ell_k+1)}.$$

This concludes the proof of Theorem 2.34.

2.4 Basic Concepts on Tensors and their Decompositions

We end this chapter by presenting basic notions on tensors and their decompositions. For a broader introduction to tensors, we refer to Comon (2014); Landsberg (2011); Hackbusch (2012); Sun et al. (2021); Bi et al. (2021).

2.4.1 The ABCs of Tensors

An order-d tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ is a d-way array composed of elements $\mathcal{T}_{i_1,\ldots,i_d} \in \mathbb{R}$ with $i_\ell \in [n_\ell]$ for all $\ell \in [d]$. Just as vectors (order-1 tensors) and matrices (order-2 tensors) define linear and bilinear forms, $\mathcal{T}$ defines a d-linear form

$$\mathcal{T} : \left\{ \begin{array}{ccc} \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d} & \to & \mathbb{R} \\ \left(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(d)}\right) & \mapsto & \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \mathcal{T}_{i_1,\ldots,i_d} \prod_{\ell=1}^{d} u_{i_\ell}^{(\ell)} \end{array} \right. .$$

This operation is called the *contraction* of $\mathcal{T}$ on $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(d)}$. It can be extended to any d-uplet of matrices: the contraction of $\mathcal{T}$ on $\boldsymbol{A}^{(\ell)} \in \mathbb{R}^{n_\ell \times p_\ell}$, $\ell \in [d]$, gives a new tensor $\mathcal{T}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)}) \in \mathbb{R}^{p_1 \times \ldots \times p_d}$ whose $(j_1, \ldots, j_d)$-entry is

$$[\mathcal{T}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)})]_{j_1,\ldots,j_d} = \mathcal{T}(\boldsymbol{A}^{(1)}_{\cdot,j_1}, \ldots, \boldsymbol{A}^{(d)}_{\cdot,j_d}) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \mathcal{T}_{i_1,\ldots,i_d} \prod_{\ell=1}^{d} A^{(\ell)}_{i_\ell,j_\ell}.$$

The tensor contraction generalizes the matrix operation $\boldsymbol{A}^{(1)\top} \boldsymbol{M} \boldsymbol{A}^{(2)} = \boldsymbol{M}(\boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)})$.

We can define the Frobenius inner product between two tensors $\mathcal{T}, \mathcal{T}' \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ as

$$\left\langle \mathcal{T}, \mathcal{T}' \right\rangle_{\mathrm{F}} \overset{\text{def}}{=} \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \mathcal{T}_{i_1,\ldots,i_d} \mathcal{T}'_{i_1,\ldots,i_d}$$

and the corresponding norm is the Frobenius norm $\|\mathcal{T}\|_{\mathrm{F}} \overset{\text{def}}{=} \sqrt{\langle \mathcal{T}, \mathcal{T} \rangle_{\mathrm{F}}}$. A second tensor norm is the spectral norm, that is, the operator norm induced by the Euclidean norm on vectors

$$\|\mathcal{T}\| \overset{\text{def}}{=} \max_{\boldsymbol{u}^{(\ell)} \in \mathbb{S}^{n_\ell - 1},\ \ell \in [d]} \left| \mathcal{T}(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(d)}) \right|.$$

An *ℓ-fiber* of $\mathcal{T}$ is the vector of $\mathbb{R}^{n_\ell}$ obtained by fixing all indices of $\mathcal{T}$ but the ℓ-th. This is the generalization of rows and columns of matrices, which are respectively 1- and 2-fibers. Similarly, an *ℓ-slice* of $\mathcal{T}$ is the order-$(d-1)$ tensor obtained by fixing only the ℓ-th index of $\mathcal{T}$.

Unfolding (or matricization) is the process by which a matrix is built from a tensor. It is usually performed along a given mode: $\boldsymbol{T}^{(\ell)}$ denotes the unfolding of $\mathcal{T}$ along mode $\ell \in [d]$ — it is an $n_\ell \times \prod_{\ell' \neq \ell} n_{\ell'}$ matrix whose columns are ℓ-fibers of $\mathcal{T}$ in a certain predefined order[12].

The *outer product* $\otimes$ of d vectors is an order-d tensor. This generalizes the operation by which two vectors yield a rank-one matrix $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \boldsymbol{x}\boldsymbol{y}^\top \overset{\text{def}}{=} \boldsymbol{x} \otimes \boldsymbol{y}$. Given $\boldsymbol{x}^{(1)} \in \mathbb{R}^{n_1}, \ldots, \boldsymbol{x}^{(d)} \in \mathbb{R}^{n_d}$, their outer product is the $n_1 \times \ldots \times n_d$ tensor denoted $\boldsymbol{x}^{(1)} \otimes \ldots \otimes \boldsymbol{x}^{(d)}$ or $\bigotimes_{\ell=1}^{d} \boldsymbol{x}^{(\ell)}$ whose $(i_1, \ldots, i_d)$-entry is $\prod_{\ell=1}^{d} x_{i_\ell}^{(\ell)}$. Such a tensor is also said to have rank one.

Given two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of respective sizes $n_1 \times n_2$ and $p_1 \times p_2$, their *Kronecker product*, denoted $\boldsymbol{A} \boxtimes \boldsymbol{B}$, is the $n_1 p_1 \times n_2 p_2$ matrix such that $[\boldsymbol{A} \boxtimes \boldsymbol{B}]_{p_1(i_1-1)+j_1, p_2(i_2-1)+j_2} = A_{i_1,i_2} B_{j_1,j_2}$ for all $(i_1, i_2) \in [n_1] \times [n_2]$ and $(j_1, j_2) \in [p_1] \times [p_2]$. This product also applies to n-dimensional vectors, seen

[12] This order does not matter as long as other operations, such as the Kronecker product, are defined in a consistent manner.

as $n \times 1$ matrices. The unfoldings of a tensor defined with outer products can easily be expressed with Kronecker products. For instance,

$$[\boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z}]^{(1)} = \boldsymbol{x}(\boldsymbol{y} \boxtimes \boldsymbol{z})^\top, \qquad [\boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z}]^{(2)} = \boldsymbol{y}(\boldsymbol{x} \boxtimes \boldsymbol{z})^\top, \qquad [\boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z}]^{(3)} = \boldsymbol{z}(\boldsymbol{x} \boxtimes \boldsymbol{y})^\top. \tag{2.11}$$

Among the various properties of the Kronecker product (Abadir and Magnus, 2005, Chapter 10), we highlight that it is bilinear, associative, *non-commutative*, $(\boldsymbol{A} \boxtimes \boldsymbol{B})^\top = \boldsymbol{A}^\top \boxtimes \boldsymbol{B}^\top$ and $(\boldsymbol{A} \boxtimes \boldsymbol{B})(\boldsymbol{C} \boxtimes \boldsymbol{D}) = (\boldsymbol{AC}) \boxtimes (\boldsymbol{BD})$ when the matrix products $\boldsymbol{AC}$ and $\boldsymbol{BD}$ are defined (this is known as the mixed-product property).

2.4.2 Tensor Decompositions and Tensor Ranks

Tensor decompositions are necessary to reveal information hiding in large multi-way arrays. We review briefly two fundamental decompositions, namely the canonical polyadic decomposition and the multilinear singular value decomposition. Both generalize, in a sense, the well-known matrix singular value decomposition. We also discuss how they relate to the low-rank tensor approximation problem under a signal-plus-noise model. More details about tensor decompositions and low-rank approximations can be found in Sidiropoulos et al. (2017); Kolda and Bader (2009); Cichocki et al. (2015); Rabanser et al. (2017); Vervliet et al. (2014, 2019).

The singular value decomposition (SVD) of a rank-R matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ is $\sum_{q=1}^{R} \sigma_q \boldsymbol{u}_q \boldsymbol{v}_q^\top$ where $\sigma_q = s_q(\boldsymbol{M}) > 0$ is the q-th singular value and $\boldsymbol{u}_q \in \mathbb{S}^{n_1-1}$, $\boldsymbol{v}_q \in \mathbb{S}^{n_2-1}$ are the corresponding left and right singular vectors. Can we perform a similar decomposition with a tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$?

Canonical Polyadic Decomposition (CPD)

A natural way to extend the previous matrix decomposition to tensors is to express $\mathcal{T}$ as

$$\mathcal{T} = \sum_{q=1}^{R} \sigma_q \bigotimes_{\ell=1}^{d} \boldsymbol{u}_q^{(\ell)}$$

with $\sigma_q > 0$, $\|\boldsymbol{u}_q^{(\ell)}\| = 1$ and R minimal, in the sense that it is not possible to express $\mathcal{T}$ with less rank-one components. This decomposition is known as the *canonical polyadic decomposition* (CPD), CANDECOMP or PARAFAC and dates back to Hitchcock (1927). It is unique (up to permutations the terms) under mild conditions (see, e.g., Sidiropoulos et al., 2017, Section IV) and defines R as the *rank* of the tensor $\mathcal{T}$.

However, we have lost a very important property in trying to generalize the matrix SVD in this way: the vectors $\{\boldsymbol{u}_q^{(\ell)}\}_{q \in [R]}$ are no longer orthogonal in general, $\langle \boldsymbol{u}_q^{(\ell)}, \boldsymbol{u}_{q'}^{(\ell)} \rangle \neq 0$ for all $q, q' \in [R]$. Furthermore, the set $\{\mathcal{X} \in \mathbb{R}^{n_1 \times \ldots \times n_d} \mid \operatorname{Rank} \mathcal{X} \leqslant R\}$ of tensors with rank at most R is not closed as soon as $R \geqslant 2$. This observation is disturbing in a context where we seek a low-rank approximation of a tensor — the best rank-R approximation may not exist!

Multilinear Singular Value Decomposition (MLSVD)

Another way to extend the matrix SVD is, instead of keeping the *diagonality* and irretrievably loosing the *orthogonality*, to keep the orthogonality, but this also implies loosing the diagonality. Specifically, the *multilinear singular value decomposition* (MLSVD) of $\mathcal{T}$ is

$$\mathcal{T} = \sum_{q_1}^{r_1} \cdots \sum_{q_d=1}^{r_d} \mathcal{G}_{q_1,\ldots,q_d} \bigotimes_{\ell=1}^{d} \boldsymbol{u}_{q_\ell}^{(\ell)} \overset{\text{def}}{=} \left[\!\!\left[\mathcal{G}; \boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(d)} \right]\!\!\right]$$

where $\mathcal{G}$ is an $r_1 \times \ldots \times r_d$ *core tensor* and $\boldsymbol{U}^{(\ell)} = \begin{bmatrix} \boldsymbol{u}_1^{(\ell)} & \ldots & \boldsymbol{u}_{r_\ell}^{(\ell)} \end{bmatrix}$ is an $n_\ell \times r_\ell$ matrix with orthonormal columns ($\boldsymbol{U}^{(\ell)\top}\boldsymbol{U}^{(\ell)} = \boldsymbol{I}_{r_\ell}$), i.e., it belongs to the *Stiefel manifold* $V_{r_\ell}(\mathbb{R}^{n_\ell})$ (Chikuse, 2003; Absil et al., 2009). $(\boldsymbol{u}_1^{(\ell)}, \ldots, \boldsymbol{u}_{r_\ell}^{(\ell)})$ forms an orthonormal basis of the subspace of $\mathbb{R}^{n_\ell}$ spanned by the ℓ-fibers of $\mathcal{T}$, that is, the left singular subspace of the unfolding $\boldsymbol{T}^{(\ell)}$. Moreover, although $\mathcal{G}$ is not diagonal, it keeps a "sub-diagonality" property in the sense that, if the columns of $\boldsymbol{U}^{(\ell)}$ are the left singular vectors of $\boldsymbol{T}^{(\ell)}$, then two distinct slices of $\mathcal{G}$ along a mode $\ell \in [d]$ are orthogonal (for the Frobenius inner product) and the Frobenius norm of each slice along this mode is a singular value of $\boldsymbol{T}^{(\ell)}$. This decomposition, introduced by Tucker (1966), is also called higher-order SVD (HOSVD, De Lathauwer et al., 2000b) and defines $(r_1, \ldots, r_d)$ as the *multilinear rank* of $\mathcal{T}$.

This second notion of rank seem better suited to low-rank approximation problems since the set $\{\mathcal{X} \in \mathbb{R}^{n_1 \times \ldots \times n_d} \mid \operatorname{Rank}\mathcal{X} \leqslant (r_1, \ldots, r_d)\}$ is closed whatever the value of $(r_1, \ldots, r_d)$. Hence, the best low-multilinear-rank approximation is well-defined. Note that, in the matrix case, both notions of rank coincide: $(r_1, r_2) = (R, R)$.

Remark 2.42 (Uniqueness of the MLSVD up to isometries). For all orthogonal matrices $\boldsymbol{O}^{(\ell)} \in \mathcal{O}_{r_\ell}(\mathbb{R})$, $\ell \in [d]$, we have the equivalent decomposition

$$\left[\!\left[\mathcal{G}; \boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(d)} \right]\!\right] = \left[\!\left[\check{\mathcal{G}}; \boldsymbol{U}^{(1)}\boldsymbol{O}^{(1)}, \ldots, \boldsymbol{U}^{(d)}\boldsymbol{O}^{(d)} \right]\!\right] \qquad \text{with} \qquad \check{\mathcal{G}} = \mathcal{G}(\boldsymbol{O}^{(1)}, \ldots, \boldsymbol{O}^{(d)}).$$

Nevertheless, up to isometries, the multilinear singular value decomposition is unique (De Lathauwer et al., 2000b, Property 4).

We call a *Tucker decomposition* the decomposition of $\mathcal{T}$ in the general form $[\![\mathcal{G}; \boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(d)}]\!]$ where $\mathcal{G}$ is a tensor and $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(d)}$ are matrices. The multilinear singular value decomposition (MLSVD) thus corresponds to a particular case of Tucker decomposition where $\boldsymbol{U}^{(\ell)}$ is the matrix of left singular vectors of $\boldsymbol{T}^{(\ell)}$, for all $\ell \in [d]$.

Remark 2.43 (Contraction and Tucker decomposition). If $\mathcal{T} = [\![\mathcal{G}; \boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(d)}]\!]$ then

$$\mathcal{T}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)}) = \left[\!\left[\mathcal{T}; \boldsymbol{A}^{(1)\top}, \ldots, \boldsymbol{A}^{(d)\top} \right]\!\right] = \left[\!\left[\mathcal{G}; \boldsymbol{A}^{(1)\top}\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{A}^{(d)\top}\boldsymbol{U}^{(d)} \right]\!\right].$$

Remark 2.44 (CPD and Tucker decomposition). The canonical polyadic decomposition is also a particular case of Tucker decomposition where $\mathcal{G}$ is diagonal.

The unfoldings of $\mathcal{T}$ can be expressed with matrix and Kronecker products between the factor matrices and the unfoldings of the core tensor $\mathcal{G}$. For instance, if $\mathcal{T} = [\![\mathcal{G}; \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}]\!]$,

$$\boldsymbol{T}^{(1)} = \boldsymbol{U}\boldsymbol{G}^{(1)}(\boldsymbol{V} \boxtimes \boldsymbol{W})^\top, \quad \boldsymbol{T}^{(2)} = \boldsymbol{V}\boldsymbol{G}^{(2)}(\boldsymbol{U} \boxtimes \boldsymbol{W})^\top, \quad \boldsymbol{T}^{(3)} = \boldsymbol{W}\boldsymbol{G}^{(3)}(\boldsymbol{U} \boxtimes \boldsymbol{V})^\top.$$

Note how these expressions generalize (2.11).

Signal-plus-Noise Models

Similarly to signal-plus-noise matrix models, a signal-plus-noise *tensor* model has the form $\mathcal{T} = \mathcal{P} + \mathcal{N}$ where $\mathcal{N}$ is a random tensor (the noise) and $\mathcal{P}$ is a finite-rank perturbation (the signal). Again, we say that $\mathcal{P}$ is a low-rank perturbation since its rank remains finite while the rank of $\mathcal{N}$ diverges almost surely as the tensor grows.

The CP-rank R of the CPD and the multilinear rank $(r_1, \ldots, r_d)$ of the MLSVD are constrained by the following inequality,

$$\max_{1 \leqslant \ell \leqslant d} r_\ell \leqslant R \leqslant \min_{1 \leqslant \ell \leqslant d} \prod_{\ell' \neq \ell} r_{\ell'}. \tag{2.12}$$

Hence, low CP-rank implies low multilinear rank and vice versa. Thus, we can generally speak of a *low-rank* perturbation $\mathcal{P}$ without mentioning which rank it actually refers to.

Yet, in order to estimate $\mathcal{P}$ from $\mathcal{T} = \mathcal{P} + \mathcal{N}$ with a *low-rank* approximation, it is much more convenient to consider a *multilinear* rank to avoid ill-posedness of the problem. Problem (1.9) mentioned in the introduction is well-posed if the minimization is over a set of low-*multilinear-rank* tensors $\mathscr{M} = \{\mathcal{X} \in \mathbb{R}^{n_1 \times \ldots \times n_d} \mid \text{Rank}\mathcal{X} \leqslant (r_1, \ldots, r_d)\}$.

Part I

Random Matrix Models for Spectral Clustering

Chapter 3

Asymptotic Gaussian Fluctuations of Eigenvectors in Spectral Clustering

Nota Bene. *Although this chapter arrives directly after the introduction of the technical tools, it does not represent the heart of this thesis and is meant to prove a fundamental CLT-like result on which we can rely afterwards.*

SPECTRAL clustering is a popular unsupervised classification technique which finds applications in many domains, such as image segmentation (Shi and Malik, 2000), text mining (Brew and im Walde, 2002), and as a general purpose method for data analysis (Ng et al., 2002; von Luxburg, 2007; Ding et al., 2001). It relies on the spectrum of a suitably chosen similarity matrix to perform dimensionality reduction before applying a standard clustering algorithm such as K-means. Consider, e.g., the following toy example where n vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ are separated in two clusters $\mathscr{C}^+$, $\mathscr{C}^-$ centered around $+\boldsymbol{\mu}$, $-\boldsymbol{\mu}$ respectively, i.e., $\boldsymbol{x}_i = \pm\boldsymbol{\mu} + \boldsymbol{n}_i$ where $\boldsymbol{n}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Then, the dominant eigenvector $\boldsymbol{u}_1(\boldsymbol{K})$ of the Gram kernel matrix $\boldsymbol{K} = \frac{1}{p}[\boldsymbol{x}_i^\top \boldsymbol{x}_j]_{1\leqslant i,j\leqslant n}$ is an information-theoretically optimal estimator (Onatski et al., 2013; Löffler et al., 2021) of the vector $\frac{1}{\sqrt{n}}\boldsymbol{j}$ such that $j_i = \pm 1$ if $\boldsymbol{x}_i \in \mathscr{C}^\pm$. In this case, clustering is achieved with the trivial decision rule $\boldsymbol{x}_i \to \mathscr{C}^\pm$ if $[\boldsymbol{u}_1(\boldsymbol{K})]_i \gtrless 0$.

The achievable performances of spectral clustering can be theoretically predicted thanks to the study of random matrix models corresponding to similarity matrices. For this purpose, we have at hand the powerful tools presented in Chapter 2. In particular, they allow to derive the limiting spectral distribution of the kernel matrix and to predict the position of isolated eigenvalues in *spiked* random matrix models (Baik and Silverstein, 2006; Benaych-Georges and Nadakuditi, 2011; Couillet and Benaych-Georges, 2016). The latter are of particular importance as, in a wide range of problems, the information of interest can be modeled as a low-rank signal corrupted with noise. In our previous toy example, the data matrix $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n\end{bmatrix}$ is a rank-one perturbation $\boldsymbol{\mu}\boldsymbol{j}^\top$ of a noise matrix $\boldsymbol{N}$ with i.i.d. $\mathcal{N}(0,1)$ entries. In order to theoretically predict the error rate of spectral clustering for a given signal-to-noise ratio, one must therefore study the behavior of the dominant eigenvectors of the similarity matrix. Tools such as the ones used in Section 2.3 or in Couillet and Benaych-Georges (2016); Couillet et al. (2021) allow to express the quality of their alignment with the true underlying signal, i.e., $|\langle \frac{1}{\sqrt{n}}\boldsymbol{j}, \boldsymbol{u}_1(\boldsymbol{K})\rangle|$. Although this tells us when an estimation of the signal is possible (depending on the signal-to-noise ratio) and its efficiency, a precise characterization of the fluctuations of the entries of spiked eigenvectors still lacks to rigorously predict the error rate of spectral clustering. Indeed, in our toy example, the expected error rate $\mathbb{P}(j_i[\boldsymbol{u}_1(\boldsymbol{K})]_i < 0)$ cannot be expressed unless the law of $\boldsymbol{u}_1(\boldsymbol{K})$ is known.

Yet, it is often stated that the entries of $\boldsymbol{u}_1(\boldsymbol{K})$ have Gaussian fluctuations in the large-dimensional regime, so that $\mathbb{P}(j_i[\boldsymbol{u}_1(\boldsymbol{K})]_i < 0)$ is a Gaussian integral. In Kadavankandy and Couillet (2019), this result is formally stated but no proof is given. Hence, we fill this missing gap with a rigorous proof of this phenomenon for a general spiked random matrix model. Although we stick to a simple signal-plus-noise model here, the proposed proof is not restricted to Gaussian noise (in fact, the noise only needs to be rotationally invariant) and can easily be adapted to most standard spiked models (such as, notably, the general model considered in Benaych-Georges and Nadakuditi, 2011). Our result and its proof thus support a wide range of previous works studying the performance of spectral algorithms. The demonstration can be summarized in two simple facts

(i) an eigenvector of the kernel matrix can be decomposed into the sum of a deterministic signal part and a random noise part

(ii) the random part is uniformly distributed on a certain sphere, hence any finite subset of its entries tends to a centered Gaussian vector in the large-dimensional limit.

Random matrix results on the performance of spectral clustering often (reasonably) assume a Gaussian distribution of the entries of spike eigenvectors without ever providing a proof. The purpose of this chapter is thus to do it once and for all. We consider a general signal-plus-noise random matrix model and briefly recall known results regarding its limiting spectral distribution and the behavior of its dominant eigenvalues and eigenvectors. Then, we show that the entries of the kernel eigenvectors indeed have Gaussian fluctuations in the large-dimensional regime. We present a short, self-contained and general proof which is our main contribution. Finally, we illustrate this result with numerical experiments on synthetic and real data.

Simulations. Python codes to reproduce simulations are available in the following GitHub repository `https://github.com/HugoLebeau/asymptotic_fluctuations_spectral_clustering`.

3.1 Model and Main Result

3.1.1 Spiked Matrix Model

Consider the following statistical model introduced in Section 2.2.3,

$$\boldsymbol{X} = \boldsymbol{P} + \boldsymbol{N} \in \mathbb{R}^{p\times n}, \qquad \boldsymbol{P} = \boldsymbol{L}\boldsymbol{V}^\top \tag{3.1}$$

with $\boldsymbol{L} \in \mathbb{R}^{p\times K}$ and $\boldsymbol{V} = \begin{bmatrix}\boldsymbol{v}_1 & \dots & \boldsymbol{v}_K\end{bmatrix} \in \mathbb{R}^{n\times K}$ such that $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_K$. It models a *low-rank* signal $\boldsymbol{P}$ corrupted by additive Gaussian noise $N_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. In a spectral clustering perspective, K represents the number of classes and $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{J}^\top$ where $\boldsymbol{M} = \begin{bmatrix}\boldsymbol{\mu}_1 & \dots & \boldsymbol{\mu}_K\end{bmatrix}$ is a matrix gathering the K cluster means and $J_{i,k} = 1$ if $\boldsymbol{x}_i$ is in the k-th cluster (i.e., $\boldsymbol{x}_i = \boldsymbol{\mu}_k + \boldsymbol{n}_i$) and 0 otherwise. This is congruent with model (3.1): define the $K\times K$ diagonal matrix $\boldsymbol{D} = \mathrm{Diag}(n_1,\dots,n_K)$ where n_k is the number of samples belonging to the k-th cluster, then $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{J}^\top = \boldsymbol{L}\boldsymbol{V}^\top$ with $\boldsymbol{L} = \boldsymbol{M}\boldsymbol{D}^{1/2}$ and $\boldsymbol{V} = \boldsymbol{J}\boldsymbol{D}^{-1/2}$. This is precisely the K-class generalization of Model (1.6) presented in the introduction.

Given model (3.1), we are interested in the reconstruction of $\boldsymbol{V}$ from the dominant eigenvectors of the Gram kernel matrix $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$. We study this problem in the asymptotic regime where $p,n \to +\infty$ at the same rate, i.e., $0 < c \overset{\text{def}}{=} p/n < +\infty$. This models the fact that, in practice, the number of samples n is comparable to the number of features p and they are both large. Moreover, we make the following assumptions.

Assumption 3.1. All classes are of comparable size, i.e., $\liminf n_k/n > 0$ as $p,n \to +\infty$ for all $k \in [K]$.

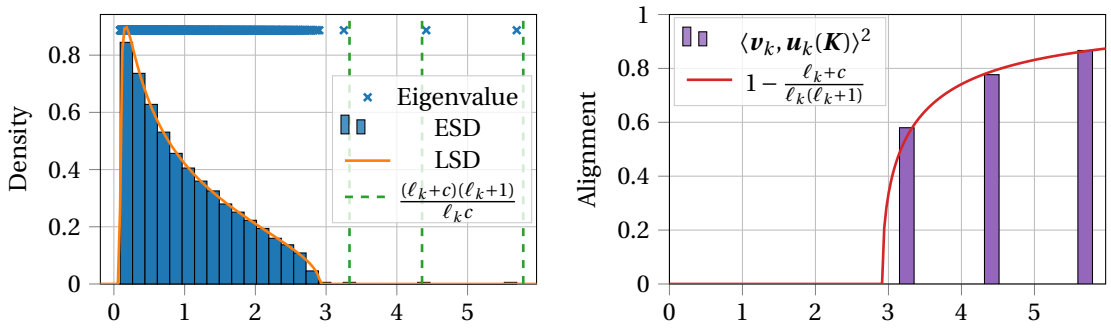


Figure 3.1: **Left:** Empirical spectral distribution (ESD) and limiting spectral distribution (LSD) of $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$ with the almost sure limits of $\lambda_k(\boldsymbol{K})$, $k \in [K]$. **Right:** Asymptotic alignments between $\boldsymbol{v}_k$ and $\boldsymbol{u}_k(\boldsymbol{K})$ as a function of the position of the corresponding eigenvalue of $\boldsymbol{K}$. **Experimental setting:** $p = 2000$, $n = 1000$, $K = 3$, $(n_1, n_2, n_3) = (333, 334, 333)$ and $(\|\boldsymbol{\mu}_1\|, \|\boldsymbol{\mu}_2\|, \|\boldsymbol{\mu}_3\|) = (3, 4, 5)$.

Assumption 3.2. $\lim_{p,n\to+\infty} \max_{\substack{1\leqslant i\leqslant n\\ 1\leqslant k\leqslant K}} \sqrt{n} V_{i,k}^2 = 0.$

Assumption 3.3. There exist $\lambda_1 > \ldots > \lambda_K > 0$ such that, with $\boldsymbol{\Lambda} \overset{\text{def}}{=} \mathrm{Diag}(\lambda_1, \ldots, \lambda_K)$, $\|\frac{1}{n}\boldsymbol{L}^\top\boldsymbol{L} - \boldsymbol{\Lambda}\| \to 0$ as $p, n \to +\infty$ and the columns of $\boldsymbol{V}$ are ordered accordingly (i.e., $\boldsymbol{v}_k$ is associated with λ_k).

The fact that $\lambda_k \neq \lambda_{k'}$ if $k \neq k'$ in Assumption 3.3 is only to simplify the presentation of the results so it is not necessary, but often verified in practice. However, Assumption 3.2 states that $\boldsymbol{V}$ must be *delocalized*, i.e., not sparse. It is naturally verified for spectral clustering as a result of Assumption 3.1 since $\boldsymbol{V} = \boldsymbol{J}\boldsymbol{D}^{-1/2}$, but the results presented below concern the statistical model (3.1), which is more general and also encompasses PCA for example (Couillet et al., 2021). In a spectral clustering perspective, since $\boldsymbol{L} = \boldsymbol{M}\boldsymbol{D}^{1/2}$, Assumption 3.3 implies that, as $p, n \to +\infty$, $\|\boldsymbol{\mu}_k\| = \Theta(1)$ for all $k \in [K]$ and $\langle \boldsymbol{\mu}_k, \boldsymbol{\mu}_{k'}\rangle \to 0$ if $k \neq k'$.

3.1.2 Eigenvalue Distribution and Spiked Eigenvalues

We briefly recall known results on model (3.1) in order to set the ground for our main result in Theorem 3.4.

Firstly, the limiting behavior of the empirical spectral distribution of $\boldsymbol{K}$, that is $\frac{1}{n}\sum_{\lambda\in\mathrm{Sp}\,\boldsymbol{K}}\delta_\lambda$ is given by Theorem 2.33: as $p, n \to +\infty$, the histogram of eigenvalues of $\boldsymbol{K}$ converges weakly to the Marčenko-Pastur distribution μ_{MP}, as depicted in the left panel of Figure 3.1.

Then, Theorem 2.34 specifies the limiting behavior of the K dominant eigenvalues and eigenvectors of $\boldsymbol{K}$. Due to the low-rank perturbation $\boldsymbol{P}$, the K dominant eigenvalues of $\boldsymbol{K}$ may isolate themselves from the bulk of Marčenko-Pastur if their corresponding signal-to-noise ratios (the eigenvalues of $\boldsymbol{\Lambda}$) are large enough — they are then called *spikes*. Theorem 2.34 states that $\lambda_k(\boldsymbol{K})$ leaves the bulk as soon as $\lambda_k(\boldsymbol{\Lambda}) > \sqrt{c}$ (this is a well-known phase transition phenomenon (Baik et al., 2005)) and further gives its almost sure asymptotic position $\xi_k \overset{\text{def}}{=} \frac{(\ell_k+c)(\ell_k+1)}{\ell_k c}$ where $\ell_k \overset{\text{def}}{=} \max(\sqrt{c}, \lambda_k(\boldsymbol{\Lambda}))$. This is illustrated in Figure 3.1 as well. Moreover, Theorem 2.34 indicates the almost sure asymptotic alignment $\zeta_k \overset{\text{def}}{=} 1 - \frac{\ell_k+c}{\ell_k(\ell_k+1)}$ of the corresponding eigenvector $\boldsymbol{u}_k(\boldsymbol{K})$ with the underlying signal $\boldsymbol{v}_k$. This is depicted in the right panel of Figure 3.1 and the top row of Figure 3.2.

As mentioned in the discussion in Section 2.2.3, these results are not restricted to Gaussian noise: up to a control on the moments of the distribution, they can be generalized thanks to an "interpolation

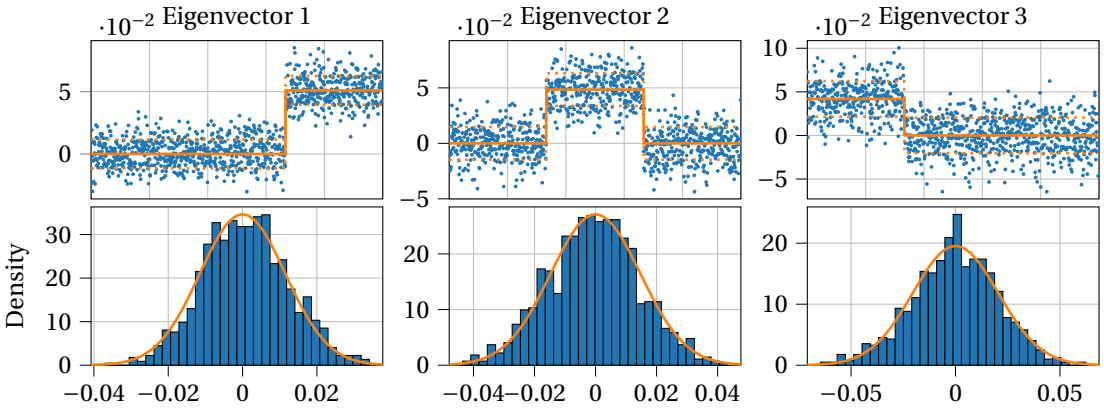


Figure 3.2: Dominant eigenvectors of $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$. **Top:** Coordinates of $\boldsymbol{u}_k(\boldsymbol{K})$ (blue) and the underlying signal $\sqrt{\zeta_k}\boldsymbol{v}_k$ (orange) with $\zeta_k = 1 - \frac{\ell_k + c}{\ell_k(\ell_k+1)}$. The dotted orange lines are the $\pm 1\sigma$-error curves deduced from Theorem 3.4. **Bottom:** Histogram of the entries of $\boldsymbol{u}_k(\boldsymbol{K}) - \sqrt{\zeta_k}\boldsymbol{v}_k$ (blue) and probability density function of $\mathcal{N}(0, \frac{1-\zeta_k}{n})$ (orange). **Experimental setting:** like in Figure 3.1.

trick" (Lytova and Pastur, 2009, Corollary 3.1). In addition, a similar spectral behavior is observed with non-i.i.d. noise following the realistic assumption that it is *concentrated* (El Karoui, 2009; Louart and Couillet, 2021).

3.1.3 Fluctuations of Spiked Eigenvectors Entries

The almost sure convergence of $\langle \boldsymbol{v}_k, \boldsymbol{u}_k(\boldsymbol{K})\rangle^2$ to $\zeta_k \overset{\text{def}}{=} 1 - \frac{\ell_k + c}{\ell_k(\ell_k+1)}$ stated in Theorem 2.34 is an important result which justifies the use of the dominant eigenvectors of $\boldsymbol{K}$ as estimators of the underlying signal $\boldsymbol{V}$. Yet, it is not enough to characterize their reconstruction performance. Indeed, the fluctuations of the entries of $\boldsymbol{u}_k(\boldsymbol{K})$ must be known to fully characterize *how* it is aligned with $\boldsymbol{v}_k$.

Consider, e.g., the multi-class spectral clustering problem with $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{J}^\top$. Here, $[\boldsymbol{v}_k]_i = J_{i,k}/\sqrt{n_k}$. Hence, $\boldsymbol{x}_i$ is classified in the k-th class if $[\boldsymbol{u}_k(\boldsymbol{K})]_i > [\boldsymbol{u}_{k'}(\boldsymbol{K})]_i$ for all $k' \neq k$. The reconstruction performance thus depends on the probability of correct classification $\mathbb{P}([\boldsymbol{u}_k(\boldsymbol{K})]_i > [\boldsymbol{u}_{k'}(\boldsymbol{K})]_i \mid J_{i,k} = 1)$. In the theorem below, we show that the entries of $\boldsymbol{u}_k(\boldsymbol{K})$ asymptotically have Gaussian fluctuations around those of $\boldsymbol{v}_k$ with variance $(1-\zeta_k)/n$, as illustrated in the bottom row of Figure 3.2.

Theorem 3.4. *For all finite ordered set of indices $\mathscr{I} = (i_1, \ldots, i_{|\mathscr{I}|}) \subset [n]$ and $k \in [K]$,*

$$\frac{\sqrt{n}}{\sqrt{1-\zeta_k}}\Big[\boldsymbol{u}_k(\boldsymbol{K}) - \sqrt{\zeta_k}\boldsymbol{v}_k\Big]_{\mathscr{I}} \xrightarrow[p,n\to+\infty]{\mathscr{D}} \mathcal{N}(\boldsymbol{0}_{|\mathscr{I}|}, \boldsymbol{I}_{|\mathscr{I}|}) \tag{3.2}$$

with $\boldsymbol{u}_k(\boldsymbol{K})$ such that $\langle \boldsymbol{v}_k, \boldsymbol{u}_k(\boldsymbol{K})\rangle \geqslant 0$ (otherwise, consider $-\boldsymbol{u}_k(\boldsymbol{K})$). Furthermore, the $|\mathscr{I}|$-dimensional vectors $\sqrt{n}[\boldsymbol{u}_1(\boldsymbol{K})]_{\mathscr{I}}, \ldots, \sqrt{n}[\boldsymbol{u}_K(\boldsymbol{K})]_{\mathscr{I}}$ are asymptotically mutually independent.

This result invokes the quantity ζ_k, which quantifies the alignment of $\boldsymbol{u}_k(\boldsymbol{K})$ with $\boldsymbol{v}_k$ (see Theorem 2.34). Theorem 3.4 specifies that $[\boldsymbol{u}_k(\boldsymbol{K})]_{\mathscr{I}}$ behaves like $\mathcal{N}(\sqrt{\zeta_k}[\boldsymbol{v}_k]_{\mathscr{I}}, \frac{1-\zeta_k}{n}\boldsymbol{I}_{|\mathscr{I}|})$ in the large-dimensional regime. That is, the more $\boldsymbol{u}_k(\boldsymbol{K})$ is aligned with $\boldsymbol{v}_k$ (i.e., the closer ζ_k is to 1), the more its entries concentrate around those of $\sqrt{\zeta_k}\boldsymbol{v}_k$, since the variance is $(1-\zeta_k)/n$. Moreover, the entries of $\sqrt{n}[\boldsymbol{u}_k(\boldsymbol{K})]_{\mathscr{I}}$ are *asymptotically independent* for any *finite* ordered set of indices $\mathscr{I}$. In the multi-class

spectral clustering problem considered above, since $\sqrt{n}[\boldsymbol{u}_k(\boldsymbol{K})]$ and $\sqrt{n}[\boldsymbol{u}_{k'}(\boldsymbol{K})]$ are asymptotically independent if $k' \neq k$, Theorem 3.4 yields

$$\mathbb{P}([\boldsymbol{u}_k(\boldsymbol{K})]_i > [\boldsymbol{u}_{k'}(\boldsymbol{K})]_i \mid J_{i,k} = 1) = \Phi\left(\sqrt{\frac{n}{n_k}\frac{\zeta_k}{2-(\zeta_k+\zeta_{k'})}}\right) + \mathrm{o}(1)$$

where $\Phi : x \mapsto \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,\mathrm{d}t$ is the Gaussian cumulative distribution function.

We prove Theorem 3.4 in Section 3.2 below. The proof hinges on the rotational invariance of the noise (Lemma 3.5). In fact, it does not need the entries of $\boldsymbol{N}$ to be distributed according to the Gaussian law, but only that its distribution be invariant under isometries. This makes it a very general proof, which can easily be adapted to most standard spiked models as those discussed, e.g., in Couillet and Liao (2022, Section 2.5.4).

3.2 Proof of Theorem 3.4

Consider the tangent-normal decomposition

$$\boldsymbol{u}_k(\boldsymbol{K}) = \sum_{\kappa=1}^{K} \tau_\kappa \boldsymbol{v}_\kappa + \sqrt{1-\|\boldsymbol{\tau}\|^2}\,\boldsymbol{u}_k^\sharp \tag{3.3}$$

where $\boldsymbol{u}_k^\sharp = (\boldsymbol{I}_n - \boldsymbol{V}\boldsymbol{V}^\top)\frac{\boldsymbol{u}_k(\boldsymbol{K})}{\sqrt{1-\|\boldsymbol{\tau}\|^2}}$ is a unit-norm vector orthogonal to the span of $\boldsymbol{V}$ and $\boldsymbol{\tau} \in [-1,1]^K$ with $\tau_\kappa = \langle \boldsymbol{v}_\kappa, \boldsymbol{u}_k(\boldsymbol{K})\rangle$, $\kappa \in [K]$, measuring the cosine between $\boldsymbol{v}_\kappa$ and $\boldsymbol{u}_k(\boldsymbol{K})$. Let $\boldsymbol{O} \in \mathscr{O}_n(\mathbb{R})$ be an $n \times n$ orthogonal matrix such that $\boldsymbol{O}\boldsymbol{V} = \boldsymbol{V}$ — i.e., a rotational symmetry about the span of $\boldsymbol{V}$ — and $\widetilde{\boldsymbol{K}} \overset{\text{def}}{=} \boldsymbol{O}\boldsymbol{K}\boldsymbol{O}^\top$. Then, since $\frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{V}^\top + \boldsymbol{N}$,

$$\widetilde{\boldsymbol{K}} = \frac{1}{p}\Big([\boldsymbol{O}\boldsymbol{V}]\boldsymbol{L}^\top\boldsymbol{L}[\boldsymbol{O}\boldsymbol{V}]^\top + [\boldsymbol{O}\boldsymbol{V}]\boldsymbol{L}^\top\big[\boldsymbol{N}\boldsymbol{O}^\top\big] + \big[\boldsymbol{N}\boldsymbol{O}^\top\big]^\top\boldsymbol{L}[\boldsymbol{O}\boldsymbol{V}]^\top + \big[\boldsymbol{N}\boldsymbol{O}^\top\big]^\top\big[\boldsymbol{N}\boldsymbol{O}^\top\big]\Big).$$

Lemma 3.5. *$\boldsymbol{N}$ and $\boldsymbol{N}\boldsymbol{O}^\top$ are identically distributed.*

Proof. The entries of $\boldsymbol{N}$ form a Gaussian random vector (by the independence of the entries). Thus, the distribution of $[\boldsymbol{N}\boldsymbol{O}^\top]_{i,j} = \sum_{k=1}^n N_{i,k}O_{j,k}$ is $\mathscr{N}(0,1)$ and $\mathrm{Cov}([\boldsymbol{N}\boldsymbol{O}^\top]_{i,j}, [\boldsymbol{N}\boldsymbol{O}^\top]_{i',j'}) = \delta_{i,i'}\delta_{j,j'}$ is 1 if $(i,j) = (i',j')$ and 0 otherwise. Hence $[\boldsymbol{N}\boldsymbol{O}^\top]_{i,j} \overset{\text{i.i.d.}}{\sim} \mathscr{N}(0,1)$ and $\boldsymbol{N}\boldsymbol{O}^\top$ is identically distributed to $\boldsymbol{N}$. □

According to the previous lemma, $\widetilde{\boldsymbol{K}}$ follows the same model as $\boldsymbol{K}$ since $\boldsymbol{O}\boldsymbol{V} = \boldsymbol{V}$. Therefore, its k-th dominant eigenvector can likewise be decomposed as

$$\boldsymbol{u}_k(\widetilde{\boldsymbol{K}}) = \sum_{\kappa=1}^{K} \widetilde{\tau}_\kappa \boldsymbol{v}_\kappa + \sqrt{1-\|\widetilde{\boldsymbol{\tau}}\|^2}\,\widetilde{\boldsymbol{u}}_k^\sharp$$

with $\widetilde{\tau}_\kappa = \langle \boldsymbol{v}_\kappa, \boldsymbol{u}_k(\widetilde{\boldsymbol{K}})\rangle$ and $\widetilde{\boldsymbol{u}}_k^\sharp = (\boldsymbol{I}_n - \boldsymbol{V}\boldsymbol{V}^\top)\frac{\boldsymbol{u}_k(\widetilde{\boldsymbol{K}})}{\sqrt{1-\|\widetilde{\boldsymbol{\tau}}\|^2}}$ *identically distributed to* $\boldsymbol{u}_k^\sharp$. Yet, $\boldsymbol{u}_k(\widetilde{\boldsymbol{K}}) = \boldsymbol{O}\boldsymbol{u}_k(\boldsymbol{K})$. Thus, $\boldsymbol{u}_k^\sharp$ and $\boldsymbol{O}\boldsymbol{u}_k^\sharp$ are identically distributed for all $\boldsymbol{O} \in \mathscr{O}_n(\mathbb{R})$ such that $\boldsymbol{O}\boldsymbol{V} = \boldsymbol{V}$. Let η denote the probability distribution of $\boldsymbol{u}_k^\sharp$ and $\boldsymbol{V}^\perp = \{\boldsymbol{a} \in \mathbb{R}^n \mid \boldsymbol{V}^\top\boldsymbol{a} = \boldsymbol{0}_K\}$. We claim that η is the uniform distribution on $\mathbb{S}^{n-1} \cap \boldsymbol{V}^\perp$. Indeed, given $\boldsymbol{x} \in \mathbb{S}^{n-1} \cap \boldsymbol{V}^\perp$, we have shown that $\mathrm{d}\eta(\boldsymbol{x}) = \mathrm{d}\eta(\boldsymbol{O}\boldsymbol{x})$ for all $\boldsymbol{O} \in \mathscr{O}_n(\mathbb{R})$ such that $\boldsymbol{O}\boldsymbol{V} = \boldsymbol{V}$. Then, given any $\boldsymbol{y} \in \mathbb{S}^{n-1} \cap \boldsymbol{V}^\perp$, the orthogonal matrix $\boldsymbol{O} = \boldsymbol{I}_n - \frac{(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^\top}{1-\boldsymbol{x}^\top\boldsymbol{y}}$ satisfies $\boldsymbol{O}\boldsymbol{V} = \boldsymbol{V}$ and $\boldsymbol{O}\boldsymbol{x} = \boldsymbol{y}$ (it is the reflection with respect to the hyperplane orthogonal to $\boldsymbol{x}-\boldsymbol{y}$). Hence $\mathrm{d}\eta(\boldsymbol{x}) = \mathrm{d}\eta(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{n-1} \cap \boldsymbol{V}^\perp$, i.e., $\boldsymbol{u}_k^\sharp$ is uniformly distributed on $\mathbb{S}^{n-1} \cap \boldsymbol{V}^\perp$.

Remark 3.6. Notice that $\mathbb{S}^{n-1} \cap \boldsymbol{V}^\top$ is isomorphic to $\mathbb{S}^{n-1-K}$. In the particular case where $n = 3$ and $K = 1$, this is simply the intersection between the 3-dimensional sphere and a 2-dimensional plane containing the origin — it is thus the unit circle.

Then, $\boldsymbol{u}_k^\sharp$ can be written as $\boldsymbol{U}\boldsymbol{w}_k$ where $\boldsymbol{w}_k$ is uniformly distributed on $\mathbb{S}^{n-K-1} \subset \mathbb{R}^{n-K}$ and $\boldsymbol{U} \in \mathbb{R}^{n\times(n-K)}$ is such that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{n-K}$ and $\boldsymbol{U}^\top\boldsymbol{V} = \boldsymbol{0}_{(n-K)\times K}$ (the columns of $\boldsymbol{U}$ form an orthonormal basis of $\boldsymbol{V}^\perp$ in $\mathbb{R}^n$). We use the following theorem to identify the asymptotic distribution of $\sqrt{n}[\boldsymbol{u}_k^\sharp]_{\mathscr{I}}$.

Theorem 3.7 (Schoenberg, 1938; Steerneman and van Perlo-ten Kleij, 2005)**.** *The characteristic function of a vector $\boldsymbol{w}$ uniformly distributed on $\mathbb{S}^{n-1}$ is given by $\varphi_{\boldsymbol{w}}(\boldsymbol{t}) \overset{def}{=} \mathbb{E}[e^{\mathrm{i}\boldsymbol{t}^\top\boldsymbol{w}}] = \Omega_n(\|\boldsymbol{t}\|)$ with*

$$\Omega_n : r \geqslant 0 \mapsto \frac{\Gamma(\frac{n}{2})}{\sqrt{\pi}\Gamma(\frac{n-1}{2})}\int_{-1}^{1} e^{\mathrm{i}rt}\left(1-t^2\right)^{\frac{n-3}{2}}\,\mathrm{d}t.$$

Moreover, $r \mapsto \Omega_n(r\sqrt{n})$ converges uniformly in $r \geqslant 0$ to $r \mapsto e^{-\frac{r^2}{2}}$ as $n \to +\infty$, i.e.,

$$\lim_{n\to+\infty}\sup_{r\geqslant 0}\left|\Omega_n(r\sqrt{n}) - e^{-\frac{r^2}{2}}\right| = 0.$$

Let $\boldsymbol{t} \in \mathbb{R}^n$ be such that $t_i = 0$ if $i \notin \mathscr{I}$. The characteristic function of $\sqrt{n}[\boldsymbol{u}_k^\sharp]_{\mathscr{I}}$ is

$$\begin{aligned}\varphi_{\sqrt{n}[\boldsymbol{u}_k^\sharp]_{\mathscr{I}}}(t_{i_1},\ldots,t_{i_{|\mathscr{I}|}}) &= \mathbb{E}\left[e^{\mathrm{i}\sqrt{n}\boldsymbol{t}^\top\boldsymbol{U}\boldsymbol{w}_k}\right]\\ &= \Omega_{n-K}\left(\sqrt{n}\left\|\boldsymbol{U}^\top\boldsymbol{t}\right\|\right)\end{aligned}$$

and $\|\boldsymbol{U}^\top\boldsymbol{t}\| = \sqrt{\|\boldsymbol{t}\|^2 - \|\boldsymbol{V}^\top\boldsymbol{t}\|^2} = \|\boldsymbol{t}\| + \mathscr{O}(\|\boldsymbol{V}^\top\boldsymbol{t}\|^2)$. According to Assumption 3.2, $\sqrt{n}\|\boldsymbol{V}^\top\boldsymbol{t}\|^2 \to 0$ as $p, n \to +\infty$, thus $\Omega_{n-K}(\sqrt{n}\|\boldsymbol{U}^\top\boldsymbol{t}\|) = \Omega_{n-K}(\sqrt{n-K}\|\boldsymbol{t}\| + \epsilon_n)$ with $\epsilon_n \to 0$ as $p, n \to +\infty$ and

$$\begin{aligned}&\left|\Omega_{n-K}(\sqrt{n-K}\|\boldsymbol{t}\| + \epsilon_n) - e^{-\frac{1}{2}\|\boldsymbol{t}\|^2}\right| \leqslant\\ &\qquad\left|\Omega_{n-K}(\sqrt{n-K}\|\boldsymbol{t}\| + \epsilon_n) - e^{-\frac{1}{2}\left[\|\boldsymbol{t}\| + \epsilon_n/\sqrt{n-K}\right]^2}\right| + \left|e^{-\frac{1}{2}\left[\|\boldsymbol{t}\| + \epsilon_n/\sqrt{n-K}\right]^2} - e^{-\frac{1}{2}\|\boldsymbol{t}\|^2}\right|.\end{aligned}$$

As $p, n \to +\infty$, the first term vanishes from the uniform convergence given in Theorem 3.7 and the second term vanishes by continuity. Therefore, $\varphi_{\sqrt{n}[\boldsymbol{u}_k^\sharp]_{\mathscr{I}}}(t_{i_1},\ldots,t_{i_{|\mathscr{I}|}}) \to e^{-\frac{\|\boldsymbol{t}\|^2}{2}}$ and, by Lévy's continuity theorem (Billingsley, 2012, Theorem 26.3), we can conclude that

$$\sqrt{n}[\boldsymbol{u}_k^\sharp]_{\mathscr{I}} \xrightarrow[p,n\to+\infty]{\mathscr{D}} \mathcal{N}(\boldsymbol{0}_{|\mathscr{I}|}, \boldsymbol{I}_{|\mathscr{I}|}).$$

Finally, the decomposition given by Equation (3.3) yields

$$\boldsymbol{u}_k(\boldsymbol{K}) = \sqrt{\zeta_k}\boldsymbol{v}_k + \sqrt{1-\zeta_k}\boldsymbol{u}_k^\sharp + \boldsymbol{\varepsilon} \quad \text{with} \quad \boldsymbol{\varepsilon} = \sum_{\kappa=1}^{K} r_\kappa \boldsymbol{v}_\kappa + \mathscr{O}(r_k) \times \boldsymbol{u}_k^\sharp$$

where $r_\kappa = \tau_\kappa$ if $\kappa \neq k$ and $r_k = \tau_k - \sqrt{\zeta_k}$. Thus, $r_\kappa \to 0$ almost surely as $p, n \to +\infty$ for all $\kappa \in [K]$ from Theorem 2.34. In order to conclude our proof of the convergence (3.2), we just need to show that $\sqrt{n}[\boldsymbol{\varepsilon}]_{\mathscr{I}} \xrightarrow[p,n\to+\infty]{\mathscr{D}} 0$. It is already clear that $\mathscr{O}(r_k) \times \sqrt{n}[\boldsymbol{u}_k^\sharp]_{\mathscr{I}} \xrightarrow[p,n\to+\infty]{\mathscr{D}} 0$ so it remains to show that $\sqrt{n}r_\kappa[\boldsymbol{v}_\kappa]_{\mathscr{I}} \xrightarrow[p,n\to+\infty]{\mathscr{D}} 0$ for all $\kappa \in [K]$. We have[1] $\operatorname{Var} r_\kappa^2 = \mathscr{O}(n^{-1})$ and, by Assumption 3.2, $\sqrt{n}[\boldsymbol{v}_\kappa]_i^2 =$

[1] As shown in Section 2.3.3, $\langle \boldsymbol{v}_\kappa, \boldsymbol{u}_k(\boldsymbol{K})\rangle^2 = -\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_\kappa} \boldsymbol{v}_\kappa^\top\boldsymbol{Q}(z)\boldsymbol{v}_\kappa\,\mathrm{d}z$ thus $\operatorname{Var}(r_\kappa^2) = \mathbb{E}[|-\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_\kappa} \boldsymbol{v}_\kappa^\top(\boldsymbol{Q}(z) - \mathbb{E}[\boldsymbol{Q}(z)])\boldsymbol{v}_\kappa\,\mathrm{d}z|^2] \leqslant \frac{1}{4\pi^2}\oint_{\gamma_\kappa}\mathbb{E}[|\boldsymbol{v}_\kappa^\top(\boldsymbol{Q}(z) - \mathbb{E}[\boldsymbol{Q}(z)])\boldsymbol{v}_\kappa|^2]\,\mathrm{d}z$ where the interchange between $\mathbb{E}$ and $\oint_{\gamma_\kappa}$ is allowed since the integrand is bounded. In Section 2.3.1, we have shown that $\mathbb{E}[|\boldsymbol{v}_\kappa^\top(\boldsymbol{Q}(z) - \mathbb{E}[\boldsymbol{Q}(z)])\boldsymbol{v}_\kappa|^2] = \mathscr{O}_z(n^{-1})$ therefore $\operatorname{Var}(r_\kappa^2) = \mathscr{O}(n^{-1})$.

$o(1)$ for all $i \in \mathscr{I}$. Therefore $(\sqrt{n} r_\kappa [\boldsymbol{v}_\kappa]_i)^2$ vanishes in quadratic mean as $p, n \to +\infty$ and this implies $\sqrt{n} r_\kappa [\boldsymbol{v}_\kappa]_{\mathscr{I}} \xrightarrow[p,n\to+\infty]{\mathscr{D}} 0$ by the continuity of $x \mapsto \sqrt{x}$ and the continuous mapping theorem (van der Vaart, 1998, Theorem 2.3).

In order to show the asymptotic mutual independence of $\sqrt{n}[\boldsymbol{u}_1(\boldsymbol{K})]_{\mathscr{I}}, \ldots, \sqrt{n}[\boldsymbol{u}_K(\boldsymbol{K})]_{\mathscr{I}}$, consider the tangent-normal decomposition (3.3) for each $k \in [K]$ and, with the same arguments, construct $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1 & \ldots & \boldsymbol{w}_K \end{bmatrix}$ where the $\boldsymbol{w}_k$'s are such that $\boldsymbol{u}_k^\sharp = \boldsymbol{U}\boldsymbol{w}_k$ for all $k \in [K]$. Let $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{W})^{-1/2}$. For all orthogonal matrix $\boldsymbol{O} \in \mathcal{O}_{n-K}(\mathbb{R})$, $\boldsymbol{O}\boldsymbol{W}$ is identically distributed to $\boldsymbol{W}$ and therefore $\boldsymbol{O}\widehat{\boldsymbol{W}}$ is identically distributed to $\widehat{\boldsymbol{W}}$ as well. Moreover, $\widehat{\boldsymbol{W}}^\top \widehat{\boldsymbol{W}} = \boldsymbol{I}_K$ so $\widehat{\boldsymbol{W}}$ is, in fact, uniformly distributed on the Stiefel manifold $V_K(\mathbb{R}^{n-K}) \overset{\text{def}}{=} \{\boldsymbol{A} \in \mathbb{R}^{(n-K)\times K} \mid \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_K\}$. From Theorem 2.2.1 of Chikuse (2003), $\widehat{\boldsymbol{W}}$ is identically distributed to $\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1/2}$ where $\boldsymbol{Z}$ is an $(n-K) \times K$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Thus $\sqrt{n}\widehat{\boldsymbol{W}}$ is identically distributed to $\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z}/n)^{-1/2}$ and, by the strong law of large numbers (Billingsley, 2012, Theorem 6.1), $\frac{1}{n}\boldsymbol{Z}^\top \boldsymbol{Z} \to \boldsymbol{I}_K$ almost surely. Hence, by Slutsky's theorem (Casella and Berger, 2001, Theorem 5.5.17), any finite set of the entries of $\sqrt{n}\widehat{\boldsymbol{W}}$ converges to a set of i.i.d. $\mathcal{N}(0,1)$ variables in the limit $p, n \to +\infty$. Furthermore, since $\boldsymbol{W} = \widehat{\boldsymbol{W}}(\boldsymbol{W}^\top \boldsymbol{W})^{1/2}$ and $\boldsymbol{W}^\top \boldsymbol{W} \to \boldsymbol{I}_K$ almost surely by Theorem 2.34, the same reasoning shows that any finite set of the entries of $\sqrt{n}\boldsymbol{W}$ converges to a set of i.i.d. $\mathcal{N}(0,1)$ variables in the limit $p, n \to +\infty$.

Then, let $\boldsymbol{U}^\sharp \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{u}_1^\sharp & \ldots & \boldsymbol{u}_K^\sharp \end{bmatrix}$ and $\boldsymbol{U}_{\mathscr{I}}^\sharp = \begin{bmatrix} [\boldsymbol{u}_1^\sharp]_{\mathscr{I}} & \ldots & [\boldsymbol{u}_K^\sharp]_{\mathscr{I}} \end{bmatrix}$. Consider the characteristic function of $\sqrt{n}\boldsymbol{U}_{\mathscr{I}}^\sharp$, that is, $\varphi_{\sqrt{n}\boldsymbol{U}_{\mathscr{I}}^\sharp}(\boldsymbol{T}) = \mathbb{E}[\exp(\mathrm{i}\,\mathrm{Tr}(\sqrt{n}\boldsymbol{T}^\top \boldsymbol{U}^\sharp))]$ where $\boldsymbol{T}$ is an $n \times K$ matrix whose entry $T_{i,k}$ is 0 if $i \notin \mathscr{I}$. Since $\boldsymbol{U}^\sharp = \boldsymbol{U}\boldsymbol{W}$, we have $\varphi_{\sqrt{n}\boldsymbol{U}_{\mathscr{I}}^\sharp}(\boldsymbol{T}) = \varphi_{\boldsymbol{W}}(\sqrt{n}\boldsymbol{U}^\top \boldsymbol{T})$ where $\varphi_{\boldsymbol{W}}$ is the characteristic function of $\boldsymbol{W}$, whose distribution is left-spherical ($\boldsymbol{O}\boldsymbol{W}$ is identically distributed to $\boldsymbol{W}$) therefore there exists a function $f_{n-K} : \mathbb{R}^{K\times K} \to \mathbb{R}$ such that $\varphi_{\boldsymbol{W}}(\boldsymbol{S}) = f_{n-K}(\boldsymbol{S}^\top \boldsymbol{S})$ for all $\boldsymbol{S} \in \mathbb{R}^{(n-K)\times K}$ (Li, 1993). As a result, $\varphi_{\sqrt{n}\boldsymbol{U}_{\mathscr{I}}^\sharp}(\boldsymbol{T}) = f_{n-K}(n\boldsymbol{T}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{T})$. Because any finite set of the entries of $\sqrt{n}\boldsymbol{W}$ converges to a set of i.i.d. $\mathcal{N}(0,1)$ variables in the limit $p, n \to +\infty$, we have $f_{n-K}(n\boldsymbol{T}^\top \boldsymbol{T}) \to \exp(-\frac{1}{2}\boldsymbol{T}^\top \boldsymbol{T})$ by Lévy's continuity theorem (Billingsley, 2012, Theorem 26.3). To conclude, recall that $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_n - \boldsymbol{V}\boldsymbol{V}^\top$ thus $n\boldsymbol{T}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{T} = n\boldsymbol{T}^\top \boldsymbol{T} + \sqrt{n}\boldsymbol{\epsilon}_n$ where $\boldsymbol{\epsilon}_n$ is such that $\|\boldsymbol{\epsilon}_n\| \to 0$ as per Assumption 3.2 and the fact that $\boldsymbol{T}$ has a *finite* number of non-zero entries. Hence, $\varphi_{\sqrt{n}\boldsymbol{U}_{\mathscr{I}}^\sharp}(\boldsymbol{T}) = f_{n-K}(n(\boldsymbol{T}^\top \boldsymbol{T} + \frac{\boldsymbol{\epsilon}_n}{\sqrt{n}})) \to \exp(-\frac{1}{2}\boldsymbol{T}^\top \boldsymbol{T})$. This shows that the entries of $\sqrt{n}\boldsymbol{U}_{\mathscr{I}}^\sharp$ are i.i.d. $\mathcal{N}(0,1)$ variables and concludes the proof of Theorem 3.4.

3.3 Numerical Experiments

To illustrate this result, we conduct a first experiment on synthetic data following model (3.1) with $K = 3$ classes of equal size and $(\|\boldsymbol{\mu}_1\|, \|\boldsymbol{\mu}_2\|, \|\boldsymbol{\mu}_3\|) = (3, 4, 5)$. The $\boldsymbol{x}_i$'s are ordered by class. The spectral distribution of $\boldsymbol{K}$ and the alignments with $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ of its dominant eigenvectors are plotted in Figure 3.1. Figure 3.2 shows the dominant eigenvectors with the histograms of residuals $\boldsymbol{u}_k(\boldsymbol{K}) - \sqrt{\zeta_k}\boldsymbol{v}_k$. We observe a very good fit of the latter to the probability density function of $\mathcal{N}(0, \frac{1-\zeta_k}{n})$ — the $\boldsymbol{u}_k(\boldsymbol{K})$'s exactly correspond to a deterministic signal $\sqrt{\zeta_k}\boldsymbol{v}_k$ corrupted by additive centered Gaussian noise. The signal-plus-noise structure of model (3.1) has been transferred to the spectral estimator of $\boldsymbol{V}$.

Then, we conduct a second experiment on the Fashion-MNIST dataset (Xiao et al., 2017) consisting of 28×28 images of clothes separated in 10 classes of size 7 000 each. We select two classes k_1, k_2 and perform binary spectral clustering using the dominant eigenvector of $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$ where the columns of $\boldsymbol{X}$ are the 784 pixels of the images from classes k_1 and k_2. The dimension of $\boldsymbol{X}$ is thus $784 \times 14\,000$. Here, we assume a similar model as our toy example in the introduction of this chapter: $\boldsymbol{X} = \boldsymbol{\mu}\boldsymbol{j}^\top + \boldsymbol{N}$ where $j_i = \pm 1$ depending on the class of the i-th image. Thus, according to Theorem 3.4, the i-th entry of the dominant eigenvector $\boldsymbol{u}_1(\boldsymbol{K})$ asymptotically follows $\mathcal{N}\left(\sqrt{\zeta}\frac{j_i}{\sqrt{n}}, \frac{1-\zeta}{n}\right)$ and, given $\boldsymbol{j}$, ζ can be

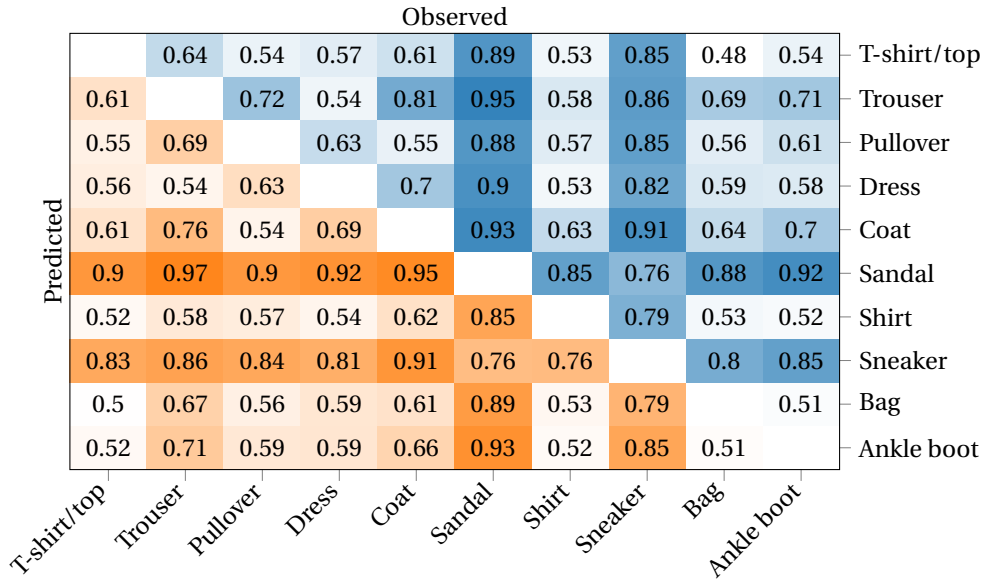


Figure 3.3: Observed (upper right, blue) and predicted (lower left, orange) classification accuracies of binary spectral clustering on the Fashion-MNIST dataset (Xiao et al., 2017).

estimated with the squared empirical mean[2] $(\sum_{i=1}^n \frac{j_i}{\sqrt{n}}[\boldsymbol{u}_1(\boldsymbol{K})]_i)^2$. We can then compare the observed accuracy $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{j_i[\boldsymbol{u}_1(\boldsymbol{K})]_i>0}$ to the one expected from Theorem 3.4, $\mathbb{P}(j_i[\boldsymbol{u}_1(\boldsymbol{K})]_i > 0) = \Phi\left(\sqrt{\frac{\zeta}{1-\zeta}}\right) + o(1)$. The results are presented in Figure 3.3 for each pair of classes k_1, k_2.

We find a very good agreement between the observed and predicted accuracies, regardless of whether the problem is easy (e.g., Trouser vs Sandal) or hard (e.g., Bag vs Ankle Boot). This observation confirms the general scope of Theorem 3.4: starting from real data $\boldsymbol{X}$ which is clearly *not* Gaussian, the normal distribution naturally emerges in the fluctuations of the entries of the large-dimensional eigenvector $\boldsymbol{u}_1(\boldsymbol{K})$.

3.4 Conclusion

After recalling known results on spectral clustering under a general signal-plus-noise random matrix model, we have shown that the entries of spiked eigenvectors have Gaussian fluctuations in the large-dimensional regime. This formalizes and clearly states a result which is often implicitly assumed in many problems, without ever being actually proven. The proposed proof relies solely on the rotational invariance of the noise. It is thus very general and can easily be extended to most standard spike models. Numerical experiments have demonstrated the universality of this phenomenon: the Gaussian behavior of the entries of spike eigenvectors can even be observed on real unprocessed data. This allows to accurately predict the classification performance of spectral clustering.

In the next chapter, we consider a more exotic spiked model which brings a powerful approach to spectral clustering under limited memory constraint. Our characterization of the fluctuations of the entries of spike eigenvectors (Theorem 3.4), combined with the random matrix tools introduced in Chapter 2, will be essential to thoroughly describe the performances of the proposed method.

[2] In practice, $\boldsymbol{j}$ is unknown and ζ can be estimated by maximum likelihood.

Chapter 4

A Random Matrix Analysis of Data Stream Clustering: Coping With Limited Memory Resources

THE ever-increasing amount of data coupled with the need for a more sober use of computational power puts online learning in the spotlight, as a way to deal with numerous and very large data with low memory resources. Be it because the volume of data is too high to be stored or because one is restricted to the sole use of a regular laptop, online learning appears as a handy and frugal way to process information. As data arrives in the learning pipeline, it is processed at a low computational cost before being discarded altogether, thus inducing a limited memory footprint.

Numerous works have proposed various algorithms to cluster data streams in an unsupervised manner (see, e.g., Ghesmoune et al., 2016; Zubaroğlu and Atalay, 2021 and references therein). Among standard methods are the construction of a graph (Fritzke, 1995) or a tree of clusters (Zhang et al., 1996) which is updated as new data arrives, or else, the formation of clusters using a distance function, as in K-means, (Aggarwal et al., 2003) or a density-based method (Ester et al., 1996). Such algorithms are often adaptations of existing offline algorithms, like OpticsStream (Tasoulis et al., 2007), StreamKM++ (Ackermann et al., 2012), online K-means (Liberty et al., 2015; Cohen-Addad et al., 2021), etc. These techniques operate on the entire feature space and their performance deteriorate as the dimension of the data increases. Therefore, Aggarwal et al. (2004) propose to cluster data streams after a projection on a lower-dimensional space. Sketching methods (Keriven et al., 2018; Gribonval et al., 2021) are also convenient to perform large-scale learning on data streams with a limited memory budget; the idea being to summarize the dataset into a single vector computed in one pass over the data.

Adapted from the standard spectral clustering algorithm (von Luxburg, 2007), techniques like incremental spectral clustering (Ning et al., 2010; Dhanjal et al., 2014) have been proposed to handle evolving data. Yet, they become quite memory-demanding when the number of samples grows large. Better suited to streaming applications, the spectral clustering algorithm of Yoo et al. (2016) constructs a spectral embedding of the stream in one pass by adapting ideas from matrix sketching (Liberty, 2013).

Spectral clustering has indeed remarkably good performances on high-dimensional data as it manages to greatly reduce the dimensionality by keeping just a few leading spectral components. It is therefore computationally less demanding than many other classical clustering algorithms. Moreover, it reaches the optimal phase transition threshold (i.e., it performs better than random guess as soon as theoretically possible) (Onatski et al., 2013) and achieves the optimal clustering error rate in

the Gaussian mixture model (Löffler et al., 2021).

From a random matrix theory perspective, spectral clustering is also of particular interest. Following the works of El Karoui (2010) and Cheng and Singer (2013) on the spectrum of kernel random matrices, Couillet and Benaych-Georges (2016) propose an analysis of kernel spectral clustering with numerous high-dimensional data. Then, Mai and Couillet (2018) demonstrate that many standard machine learning algorithms in fact suffer from being ill-used when dealing with such data. Besides, given some data matrix $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{p \times n}$, Couillet et al. (2021) show that it is possible to get huge reductions in computational and storage costs with almost no performance loss by puncturing the data, i.e., keeping only a few elements of $\boldsymbol{X}$ and computing only a few elements of the Gram kernel matrix $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$. In addition, Liao et al. (2021) demonstrate that, when carefully employed, sparsification and quantization of $\boldsymbol{K}$ incur negligible performance loss, while providing a great computational gain.

In light of these numerous benefits of spectral clustering when dealing with high-dimensional data, of the practicality of online learning to handle large data streams with limited memory, and of the promising path shown by random matrix theory towards resource-efficient learning with performance guarantees, this chapter introduces an "online spectral learning" algorithm to which we attach a rigorous performance analysis using random matrix theory.

The algorithm goes as follows: supposing that, due to memory limitations, only a small number L of data points can be kept in the pipeline, the computation of the $n \times n$ Gram kernel matrix is limited to the elements which are in a radius L around the diagonal of $\boldsymbol{K}$. This results in the following punctured kernel matrix model

$$\boldsymbol{K}_L = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \boldsymbol{T}$$

where $\odot$ denotes the Hadamard product and $\boldsymbol{T} \in \{0,1\}^{n \times n}$ is a Toeplitz mask: $T_{i,j} = \mathbf{1}_{|i-j|<L}$. A careful adaptation of spectral clustering is then performed on $\boldsymbol{K}_L$ to retrieve the class information.

In technical terms, the present analysis derives the limiting spectral distribution of $\boldsymbol{K}_L$ and analyzes the behavior of a few isolated eigenvalues (*spikes*) which carry information (that is, indicators for the data classes) in their associated eigenvectors. Two new interesting behaviors are observed: unlike classical spectral clustering, due to the Toeplitz filter, the number of informative spikes can potentially grow very large even in the case of binary classification. In addition, the eigenvectors are strongly tainted (in a way "convolved") by the eigenvectors of the Toeplitz mask, which then requires some careful post-processing for classification. Our results particularly shed light on how the learning performance is altered by the dimension of the data and the size of the pipeline, thus providing an analysis of the performance versus cost trade-off of online learning.

In a nutshell, our main contributions may be listed as follows

- we derive the limiting eigenvalue distribution of $\boldsymbol{K}_L$ as $p, n, L \to +\infty$ for data arising from a Gaussian mixture model: $\boldsymbol{x}_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I}_p)$;
- for centered data drawn from a two-class mixture $\boldsymbol{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \boldsymbol{I}_p)$, we show that a phase transition phenomenon occurs: depending on the signal power $\|\boldsymbol{\mu}\|$, some eigenvalues of $\boldsymbol{K}_L$ isolate themselves and their eigenvectors carry information about the classes;
- we propose an algorithm to retrieve information from isolated eigenvectors, thus performing high-dimensional "online spectral clustering";
- simulations of online spectral clustering on Fashion-MNIST and BigGAN-generated images confirm the predicted good behavior of the algorithm and support our theoretical findings.

The remainder of this chapter is organized as follows. Section 4.1 introduces the model and a circulant approximation of the Toeplitz mask $\boldsymbol{T}$, which will be used to derive our main results, presented

in Section 4.2. The limiting spectral distribution of the kernel matrix is studied first (Theorem 4.6) and a closer look is then given to the behavior of its isolated eigenvalues and associated eigenvectors (Theorem 4.8). Based on the previous results, Section 4.3 presents some theoretical considerations on the classification performance achievable on a data stream and proposes an online kernel spectral clustering algorithm, which is tested on image clustering tasks. Section 4.4 gives some concluding remarks.

Proofs and simulations. All proofs are deferred to the appendix. Python codes to reproduce simulations are available in the following GitHub repository `https://github.com/HugoLebeau/online_learning/`.

4.1 Online Learning Model and Problem Setting

4.1.1 General Framework

Let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{p\times n}$ be a collection of n data points of dimension p. They are noisy observations of K unknown classes whose means are $\begin{bmatrix} \boldsymbol{\mu}_1 & \dots & \boldsymbol{\mu}_K \end{bmatrix} \overset{\text{def}}{=} \boldsymbol{M} \in \mathbb{R}^{p\times K}$. Also define the $n\times K$ binary matrix $\boldsymbol{J}$ such that $J_{i,k} = 1$ if $\boldsymbol{x}_i$ belongs to class k and 0 otherwise. We make the following assumptions.

Assumption 4.1. The rows of $\boldsymbol{J}$ are independent realizations of a multinomial distribution with one trial and K outcomes, i.e., the class of $\boldsymbol{x}_i$ does not depend on the class of $\{\boldsymbol{x}_j\}_{j\neq i}$.

Assumption 4.2 (Non-triviality condition)**.** $\boldsymbol{M}$ is uniformly bounded in spectral norm and does not vanish asymptotically: $0 < \liminf_{p\to+\infty} \|\boldsymbol{M}\| \leqslant \limsup_{p\to+\infty} \|\boldsymbol{M}\| < +\infty$.

Assumption 4.3 (Additive noise model)**.** $\boldsymbol{X} = \boldsymbol{P} + \boldsymbol{N}$ where $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{J}^\top$ is a deterministic *signal* matrix and $\boldsymbol{N}$ is a random standard Gaussian *noise* matrix with independent entries[1].

Remark 4.4. The non-triviality condition (Assumption 4.2) places the work under scenarios of practical relevance, in the sense that the problem is asymptotically (as $p, n, L \to +\infty$) neither too easy ($\|\boldsymbol{M}\| \to +\infty$) nor too hard ($\|\boldsymbol{M}\| \to 0$). The classification error rate is therefore *not* expected to vanish asymptotically.

In the considered online setting, only the L previously seen data points are kept in memory. Thus, the element $K_{i,j} = \frac{1}{p}\boldsymbol{x}_i^\top \boldsymbol{x}_j$ of the Gram kernel matrix can only be computed if $|i-j| < L$. This is represented by the entrywise application of a Toeplitz mask $\boldsymbol{T} = [\mathbf{1}_{|i-j|<L}]_{1\leqslant i,j\leqslant n}$ resulting in

$$\boldsymbol{K}_L = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \boldsymbol{T} \qquad \text{with} \qquad \boldsymbol{T} = \begin{bmatrix} 1 & \dots & 1 & & 0 \\ \vdots & \ddots & & \ddots & \\ 1 & & \ddots & & 1 \\ & \ddots & & \ddots & \vdots \\ 0 & & 1 & \dots & 1 \end{bmatrix}.$$

As standard (offline) spectral clustering is "optimal"[2], we argue that spectral clustering on $\boldsymbol{K}_L$ ought to achieve good performance at least for not too small L/n ratios. Our technical goal is thus

[1] Refer to Section 2.2.3 for a discussion on this assumption.

[2] In that it performs better than random guess as soon as theoretically possible (Onatski et al., 2013).

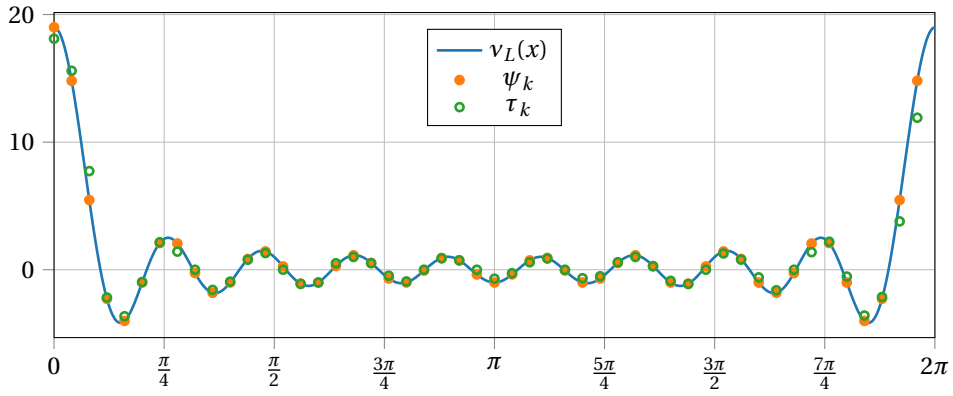


Figure 4.1: Graph of ν_L on $[0, 2\pi[$ (one period) with a plot of $\psi_k = \nu_L(\frac{2k\pi}{n})$ and τ_k for $0 \leqslant k < n$ (respectively, the eigenvalues of $\boldsymbol{C}$ and $\boldsymbol{T}$). **Experimental setting:** $n = 50$, $L = 10$.

to first provide a description of the spectral behavior of $\boldsymbol{K}_L$ as p, n and L are large. To this end, we place ourselves under the regime $p, n, L \to +\infty$ with $p/n \overset{\text{def}}{=} c \in]0, +\infty[$ and $(2L-1)/n \overset{\text{def}}{=} \varepsilon \in]0, +\infty[$.[3]

4.1.2 The Circulant Approximation

An important trick to derive our main results lies in the fact that the Toeplitz matrix $\boldsymbol{T}$ can be approximated to some extent by its circulant "version" $\boldsymbol{C} = [\mathbf{1}_{|i-j|<L} + \mathbf{1}_{|i-j|>n-L}]_{1 \leqslant i,j \leqslant n}$ (Gray, 2006). Indeed, denoting $\{\tau_k\}_{0 \leqslant k < n}$ and $\{\psi_k\}_{0 \leqslant k < n}$ their respective eigenvalues (which depend on n and L), then for *fixed* L and any continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} \left| f(\psi_k) - f(\tau_k) \right| = 0.$$

Remark 4.5. Keep in mind that, in our case, n and L grow together at the same rate. Therefore, approximating $\boldsymbol{T}$ by $\boldsymbol{C}$ is reasonable only if ε is sufficiently small.

The core advantage of $\boldsymbol{C}$ is that, unlike $\boldsymbol{T}$, its eigendecomposition is well-known:

$$\boldsymbol{C} = \widetilde{\boldsymbol{F}} \boldsymbol{\Psi} \widetilde{\boldsymbol{F}}^* \qquad \text{with} \qquad \widetilde{F}_{i,j} = \frac{1}{\sqrt{n}} e^{-\frac{2i\pi}{n}(i-1)(j-1)},$$

i.e., $\widetilde{\boldsymbol{F}}$ is the $n \times n$ Fourier matrix, $\widetilde{\boldsymbol{F}}^*$ is its Hermitian conjugate and $\boldsymbol{\Psi} = \mathrm{Diag}(\psi_k)_{0 \leqslant k < n}$ is the diagonal matrix of eigenvalues. The latter are a sampling of the Dirichlet kernel:

$$\psi_k = \nu_L\left(\frac{2k\pi}{n}\right) \qquad \text{with} \qquad \nu_L(x) = \frac{\sin((2L-1)\frac{x}{2})}{\sin(\frac{x}{2})}.$$

As we will often prefer a real eigendecomposition of $\boldsymbol{C}$, we introduce the matrix $\boldsymbol{F} = \Re\widetilde{\boldsymbol{F}} + \Im\widetilde{\boldsymbol{F}}$. It is a real orthogonal matrix ($\boldsymbol{F}^\top \boldsymbol{F} = \boldsymbol{F}\boldsymbol{F}^\top = \boldsymbol{I}_n$) and $\boldsymbol{C} = \boldsymbol{F}\boldsymbol{\Psi}\boldsymbol{F}^\top$.

In Figure 4.1 are superimposed to the graph of ν_L the eigenvalues of $\boldsymbol{C}$ and $\boldsymbol{T}$.[4] The τ_k's roughly follow the graph of ν_L, as if they were noisy versions of the ψ_k's.

[3] The provided results are asymptotic for theoretical convenience, modeling the fact that p, n and L are large. The convergence rates being at least $\mathcal{O}(\log n/\sqrt{n})$ as $p, n, L \to +\infty$, they remain valid for a large but finite horizon.

[4] Although there is a natural order for the eigenvalues of $\boldsymbol{C}$ given by $\psi_k = \nu_L(\frac{2k\pi}{n})$, we use a small trick to get the corresponding order for the eigenvalues of $\boldsymbol{T}$: after numerically computing them in ascending order, we apply the same permutation that maps the eigenvalues of $\boldsymbol{C}$ in ascending order to $(\psi_0, \ldots, \psi_{n-1})$. This yields the corresponding $(\tau_0, \ldots, \tau_{n-1})$.

4.2 Main Results

Following the random matrix approach presented in details in Chapter 2, the large dimensional spectral behavior of $\boldsymbol{K}_L$ is accessible through an analysis of the resolvent matrix

$$\boldsymbol{Q}(z) = (\boldsymbol{K}_L - z\boldsymbol{I}_n)^{-1}$$

defined for all $z \in \mathbb{C} \setminus \mathrm{Sp}(\boldsymbol{K}_L)$, where $\mathrm{Sp}(\boldsymbol{K}_L)$ is the spectrum of $\boldsymbol{K}_L$ (the set of it eigenvalues). Notably, the Stieltjes transform of the empirical spectral measure $\mu_n = \frac{1}{n}\sum_{\lambda\in\mathrm{Sp}(\boldsymbol{K}_L)} \delta_\lambda$ of $\boldsymbol{K}_L$ is the normalized trace of its resolvent (Proposition 2.6):

$$m_n(z) \overset{\text{def}}{=} \int_{\mathbb{R}} \frac{\mathrm{d}\mu_n(t)}{t-z} = \frac{1}{n}\operatorname{Tr}\boldsymbol{Q}(z).$$

The resolvent also encapsulates information about the eigenvectors of $\boldsymbol{K}_L$: given a simple closed positively-oriented complex contour γ circling around an eigenvalue λ of $\boldsymbol{K}_L$ and leaving all the other eigenvalues outside, $-\frac{1}{2\mathrm{i}\pi}\oint_\gamma \boldsymbol{Q}(z)\,\mathrm{d}z = \boldsymbol{u}\boldsymbol{u}^\top$, where $\boldsymbol{u}$ is a unit eigenvector associated to λ.[5]

4.2.1 Large Dimensional Spectral Behavior

Our main theorem provides a deterministic equivalent of the resolvent when the Toeplitz mask $\boldsymbol{T}$ is approximated by its circulant version $\boldsymbol{C}$, i.e., $\widetilde{\boldsymbol{Q}}(z) = (\widetilde{\boldsymbol{K}}_L - z\boldsymbol{I}_n)^{-1}$ with $\widetilde{\boldsymbol{K}}_L = \frac{\boldsymbol{X}^\top\boldsymbol{X}}{p} \odot \boldsymbol{C}$.

Theorem 4.6 (Deterministic equivalent of $\widetilde{\boldsymbol{Q}}(z)$)**.** *Let $\widetilde{\mu}_n$ be the empirical spectral distribution of $\widetilde{\boldsymbol{K}}_L = \frac{\boldsymbol{X}^\top\boldsymbol{X}}{p} \odot \boldsymbol{C}$ and $\widetilde{m}_n$ be its Stieltjes transform. Under Assumptions 4.1 to 4.3, there exists a unique (deterministic) Stieltjes transform (of a real probability measure) $\bar{m}_n$ satisfying*

$$1 + z\bar{m}_n(z) - c\sum_{k=0}^{n-1} \frac{\bar{m}_n(z)\frac{\psi_k}{p}}{1+\bar{m}_n(z)\frac{\psi_k}{p}} = 0, \qquad z \in \mathbb{C}\setminus\mathbb{R}, \tag{4.1}$$

and such that $|\widetilde{m}_n(z) - \bar{m}_n(z)| \to 0$ almost surely as $p,n,L \to +\infty$ with $p/n \overset{def}{=} c \in]0,+\infty[$ and $(2L-1)/n \overset{def}{=} \varepsilon \in]0,+\infty[$. Moreover, we have the following deterministic equivalent (Definition 2.17) of the resolvent $\widetilde{\boldsymbol{Q}}(z)$,

$$\widetilde{\boldsymbol{Q}}(z) \leftrightarrow \bar{\boldsymbol{Q}}(z) \overset{def}{=} \left[\frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n + \frac{\boldsymbol{P}^\top\boldsymbol{P}}{p} \odot \boldsymbol{C}\left(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^{-1}\right]^{-1}.$$

Proof. See Appendix 4.A. □

A first observation from Theorem 4.6 is that $\bar{\boldsymbol{Q}}(z)$ is the inverse of a perturbation of the identity which is *not* low-rank. This strikingly differs from standard spiked random matrix models (such as the ones presented in Section 2.2 or in Baik and Silverstein, 2006; Benaych-Georges and Nadakuditi, 2011) where a low-rank perturbation of the identity in the "population" matrix (here $\boldsymbol{P}$) usually results in the presence of only a few isolated eigenvalues in the "sample" matrix (here $\widetilde{\boldsymbol{K}}_L$). Yet, in most common settings, a majority of eigenvalues of $\frac{\boldsymbol{P}^\top\boldsymbol{P}}{p} \odot \boldsymbol{C}(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{C}}{p})^{-1}$ are too small and only a small number of them causes the appearance of isolated eigenvalues in the spectrum of $\widetilde{\boldsymbol{K}}_L$.

[5] In fact, this is only true if λ has multiplicity one. In the general case, the integral equals the projection matrix on the eigenspace associated to λ (see Section 2.1.2).

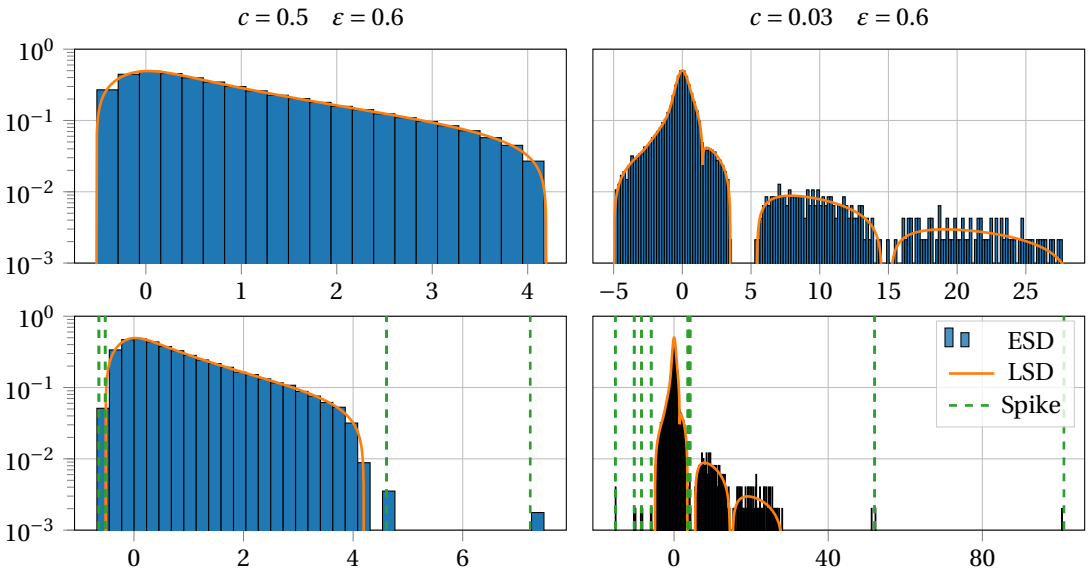


Figure 4.2: Empirical spectral distribution (ESD) and limiting spectral distribution (LSD) of $\widetilde{\boldsymbol{K}}_L$. **The y-axis is in log scale**. **Top:** noise only, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$. **Bottom:** two-class mixture, $\boldsymbol{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \boldsymbol{I}_p)$ with $\|\boldsymbol{\mu}\| = 2$. The green dashed lines are the asymptotic positions of the spikes ξ_k given by Theorem 4.8. **Experimental setting:** $n = 2\,500$, $L = 750$ and $p = 1\,250$ (**left**) or $p = 75$ (**right**).

Remark 4.7 (Link with Marčenko and Pastur (1967)). In the particular case $L \geqslant \frac{n}{2}$, the mask becomes $\boldsymbol{C} = \boldsymbol{1}_{n\times n}$ and $\widetilde{\boldsymbol{K}}_L = \boldsymbol{K}$. Thus, since $\psi_0 = n$ and $\psi_k = 0$ for $1 \leqslant k < n$, Equation (4.1) becomes

$$zc^{-1}\bar{m}_n^2(z) - \big(1 - c^{-1} - z\big)\bar{m}_n(z) + 1 = 0$$

which is the canonical equation defining the Stieltjes transform of the Marčenko-Pastur distribution (Equation (2.2)). In other words, the closer ε is to 1, the closer to the Marčenko-Pastur distribution is the limiting spectral distribution of $\widetilde{\boldsymbol{K}}_L$.

In practice, rather than computing $\bar{m}(z)$ directly from Equation (4.1), it is easier to solve numerically the following fixed-point equation in $\eta_0(z)$,

$$\eta_0(z) = \frac{p}{n}\sum_{k=0}^{n-1} \frac{\psi_k^2/p^2}{\big(1 - z - \eta_0(z)\big) + \frac{\psi_k}{p}}$$

and deduce $\bar{m}_n(z) = \frac{1}{1-z-\eta_0(z)}$.

Figure 4.2 displays, in log scale, the empirical spectral distribution of $\widetilde{\boldsymbol{K}}_L$ under two different settings (recall that $c = p/n$ and $\varepsilon = (2L-1)/n$) with its limiting spectral distribution computed by inverting the Stieltjes transform $\bar{m}_n$ given by Theorem 4.6. Two kinds of data are presented: noise-only, $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{I}_p)$, (top row) and a two-class mixture, $\boldsymbol{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \boldsymbol{I}_p)$, (bottom row). Notice how the shape of the distribution on the left column resembles the Marčenko-Pastur one (yet, some eigenvalues are negative here) while the second distribution has a completely different shape (there even are multiple bulks) for the same value of ε. This reveals that the parameter c also affects the closeness of the limiting spectral measure to the Marčenko-Pastur distribution. Also note that, under the two-class mixture setting, more than one isolated eigenvalue pops out of the limiting support. It now remains to give a close look to their associated eigenvectors to understand how to exploit the latter in a spectral clustering perspective.

4.2.2 Phase Transition and Spike Behavior

In this section, we focus back on our original clustering objective. We consider two classes $\mathscr{C}^+, \mathscr{C}^-$ whose means are $\pm\boldsymbol{\mu}$, i.e., $\boldsymbol{P} = \boldsymbol{\mu}\boldsymbol{j}^\top$ with $j_i = +1$ if $\boldsymbol{x}_i \in \mathscr{C}^+$ and $j_i = -1$ if $\boldsymbol{x}_i \in \mathscr{C}^-$. This corresponds to a two-class mixture with globally empirically centered data. Consistently with the previous setting, $\boldsymbol{M} = \begin{bmatrix} +\boldsymbol{\mu} & -\boldsymbol{\mu} \end{bmatrix}$ and $\boldsymbol{J}_{i,\cdot} = \begin{bmatrix} \mathbf{1}_{\{\boldsymbol{x}_i \in \mathscr{C}^+\}} & \mathbf{1}_{\{\boldsymbol{x}_i \in \mathscr{C}^-\}} \end{bmatrix}$.

Because of the rank-one structure, using the relation $\boldsymbol{M} \odot \boldsymbol{a}\boldsymbol{b}^\top = [\operatorname{Diag} \boldsymbol{a}]\boldsymbol{M}[\operatorname{Diag} \boldsymbol{b}]$, the deterministic equivalent of the resolvent given in Theorem 4.6 has a much simpler expression:

$$\bar{\boldsymbol{Q}}(z) = \begin{bmatrix} \boldsymbol{D_j F} \end{bmatrix} \left[\frac{1}{\bar{m}_n(z)} \boldsymbol{I}_n + \left\| \boldsymbol{\mu} \right\|^2 \frac{\boldsymbol{\Psi}}{p} \left(\boldsymbol{I}_n + \bar{m}_n(z) \frac{\boldsymbol{\Psi}}{p} \right)^{-1} \right]^{-1} \begin{bmatrix} \boldsymbol{D_j F} \end{bmatrix}^\top$$

where $\boldsymbol{D_j} = \operatorname{Diag} \boldsymbol{j}$ is the diagonal matrix induced by the vector $\boldsymbol{j}$. Now, $\bar{\boldsymbol{Q}}(z)$ no longer involves a Hadamard product and we already have its eigendecomposition since $\frac{\boldsymbol{\Psi}}{p}(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{\Psi}}{p})^{-1}$ is diagonal and $\boldsymbol{D_j F}$ is orthogonal. Note that the columns of $\boldsymbol{D_j F}$ are simply the columns of $\boldsymbol{F}$ with their signs switched at coordinates i such that $\boldsymbol{x}_i \in \mathscr{C}^-$.

With a deeper analysis of the resolvent $\bar{\boldsymbol{Q}}(z)$, the following theorem provides the position of the isolated eigenvalues and the shape of their associated eigenvectors.

Theorem 4.8 (Spike eigenvalues and eigenvectors)**.** *For all integer $k \in \{0, \ldots, n-1\}$, define*

$$\xi_k \overset{def}{=} \left(\left\| \boldsymbol{\mu} \right\|^2 + 1 \right) \frac{\psi_k}{p} \left(1 + c \sum_{l=0}^{n-1} \left[\left(\left\| \boldsymbol{\mu} \right\|^2 + 1 \right) \frac{\psi_k}{\psi_l} - 1 \right]^{-1} \right)$$

and

$$\zeta_k \overset{def}{=} \frac{\left\| \boldsymbol{\mu} \right\|^2}{\left\| \boldsymbol{\mu} \right\|^2 + 1} \left(1 - c \sum_{l=0}^{n-1} \left[\left(\left\| \boldsymbol{\mu} \right\|^2 + 1 \right) \frac{\psi_k}{\psi_l} - 1 \right]^{-2} \right).$$

Assuming Conjecture 4.32 on the global behavior of the spectrum presented and discussed in Appendix 4.B.1 is verified, if $\lim_{y \downarrow 0} \Im \bar{m}_n(\xi_k + \mathrm{i}y) = 0$ *(for p, n, L large enough) and $\psi_k \neq 0$ then*

- *almost surely, ξ_k is the asymptotic location of an eigenvalue of $\widetilde{\boldsymbol{K}}_L$;*
- *the matrix $\boldsymbol{U}_k$ whose columns gather all the eigenvectors of $\widetilde{\boldsymbol{K}}_L$ whose associated eigenvalues converge almost surely to ξ_k satisfies*

$$\boldsymbol{U}_k \boldsymbol{U}_k^\top \leftrightarrow \zeta_k \begin{bmatrix} \boldsymbol{D_j F} \end{bmatrix} \boldsymbol{\Delta}_k \begin{bmatrix} \boldsymbol{D_j F} \end{bmatrix}^\top$$

where $\boldsymbol{D_j} = \operatorname{Diag} \boldsymbol{j}$ and $\boldsymbol{\Delta}_k = \operatorname{Diag}\left(\mathbf{1}_{\{\psi_k = \psi_l\}} \right)_{0 \leqslant l < n}$.

Proof. See Appendix 4.B. □

To better understand this theorem, recall that, after Theorem 4.6, we predicted the presence of a few isolated eigenvalues in the spectrum of $\widetilde{\boldsymbol{K}}_L$. Theorem 4.8 details this assertion by specifying the positions ξ_k of the spikes and their alignments ζ_k with the signal. The quantity ζ_k can really be seen as an "indicator of spike" as it tells whether an isolated eigenvalue exists for index k (if $\zeta_k > 0$) and, if it does, the closer ζ_k is to 1, the better is the "quality" of the information carried in the corresponding eigenvector, i.e., the greater is the signal-to-noise ratio (see Figure 4.3).

Another difference with classical spiked random matrix models is that each asymptotic spike ξ_k, which has the same multiplicity as the population spike ψ_k, is rarely simple. In fact, the only simple eigenvalues of $\boldsymbol{C}$ are ψ_0, and $\psi_{n/2}$ if n is even. However, for finite values of p, n and L, the corresponding eigenvalues of $\widetilde{\boldsymbol{K}}_L$ are not necessarily degenerate (with probability one, they are not), but they have the same limit: $\xi_k = \xi_l$ for all l such that $\psi_k = \psi_l$. The number of columns of $\boldsymbol{U}_k$ is equal to the cardinality of $\{l \mid \psi_k = \psi_l\}$, which is the dimension of the limiting eigenspace associated to ξ_k.

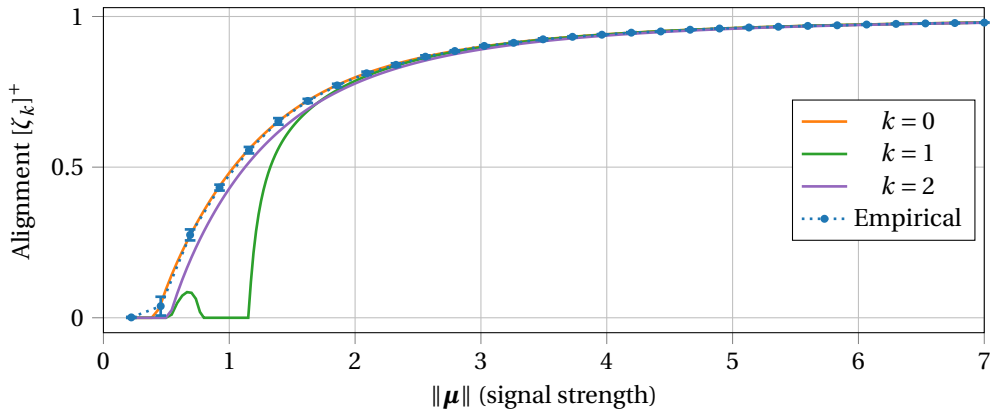


Figure 4.3: Asymptotic alignment $[\zeta_k]^+$ versus $\|\boldsymbol{\mu}\|$ for three values of k. The empirical alignment is computed as the mean of $\langle \boldsymbol{v}_0, \boldsymbol{u}_0 \rangle^2$ on 10 realizations (error bars indicate the standard deviation). **Experimental setting:** $n = 2\,500$, $p = 75$, $L = 750$.

One also notices from Theorem 4.8 that the number of isolated eigenvalues could potentially grow very large as $\|\boldsymbol{\mu}\|$ increases. Indeed, the value of $\|\boldsymbol{\mu}\|$ at which ζ_k changes sign (i.e, when one or more eigenvalues isolate themselves from the bulk around ξ_k at the *phase transition*) is given by

$$1 - c \sum_{l=0}^{n-1} \left[\left(\|\boldsymbol{\mu}\|^2 + 1 \right) \frac{\psi_k}{\psi_l} - 1 \right]^{-2} = 0. \tag{4.2}$$

Therefore, potentially any eigenvalue could leave the bulk, but this is prevented by the non-triviality condition (Assumption 4.2): $\|\boldsymbol{\mu}\| = \mathcal{O}(1)$. Moreover, since most ψ_k's are small (see Figure 4.1), the corresponding ξ_k's fall into the bulk and there are only a few spikes visible in practice. Yet, it is common to see negative isolated eigenvalues (see Figure 4.2). Indeed, since ψ_k can be negative, there can be spikes on *both sides* of the spectrum.

When positive, the quantity ζ_k is the asymptotic alignment between the empirical eigenvector[6] $\boldsymbol{u}_k$ and the corresponding eigenspace associated with ξ_k, i.e.,

$$\boldsymbol{u}_k^\top \left[\boldsymbol{D}_j \boldsymbol{F} \right] \boldsymbol{\Delta}_k \left[\boldsymbol{D}_j \boldsymbol{F} \right]^\top \boldsymbol{u}_k \xrightarrow[p,n,L \to +\infty]{\text{a.s.}} \zeta_k.$$

Thus, ζ_k measures the quality of the empirical eigenvector $\boldsymbol{u}_k$. In particular, with $k = 0$, this simply becomes $\langle \boldsymbol{v}_0, \boldsymbol{u}_0 \rangle^2 \to \zeta_0$ almost surely as $p, n, L \to +\infty$. Indeed, ξ_0 has multiplicity one and its eigenspace is spanned by $\boldsymbol{v}_0 = [\boldsymbol{D}_j \boldsymbol{F}]_{\cdot,0} = \frac{1}{\sqrt{n}} \boldsymbol{j}$. Said differently, $\boldsymbol{u}_0$ is a noisy version of the vector $\boldsymbol{v}_0$ and the noise level is indicated by $0 \leqslant 1 - \zeta_0 \leqslant 1$. In fact, each $\boldsymbol{v}_k = [\boldsymbol{D}_j \boldsymbol{F}]_{\cdot,k}$ is the vector $\boldsymbol{j}$ — the information sought — modulated by the k-th Fourier mode (recall that Fourier modes are the eigenvectors of $\boldsymbol{C}$).

Figure 4.3 displays the value of $[\zeta_k]^+ = \max(\zeta_k, 0)$ as a function of $\|\boldsymbol{\mu}\|$ for the setting corresponding to the bottom right part of Figure 4.2. The empirical alignment of the dominant eigenvector $\boldsymbol{u}_0$ with $\boldsymbol{v}_0 = \frac{1}{\sqrt{n}} \boldsymbol{j}$ fits perfectly with the curve of $[\zeta_0]^+$ predicted by Theorem 4.8. Moreover, notice the interesting fact that ξ_1 has several phase transitions: as $\|\boldsymbol{\mu}\|$ grows, it appears once, then disappears and appears once again! This is due to the limiting spectral distribution having several bulks under

[6] Note that, here, we use a different ordering of the eigenvalues and eigenvectors than the one used in the previous chapters. Previously, we considered the non-increasing order of eigenvalues ($\lambda_1(\boldsymbol{K}) \geqslant \lambda_2(\boldsymbol{K}) \geqslant \ldots$). In this chapter, it is instead more "natural" to consider the order induced by the ψ_k's. That is, $\boldsymbol{u}_k$ is the eigenvector corresponding to the k-th spike, induced by the k-th Fourier mode.

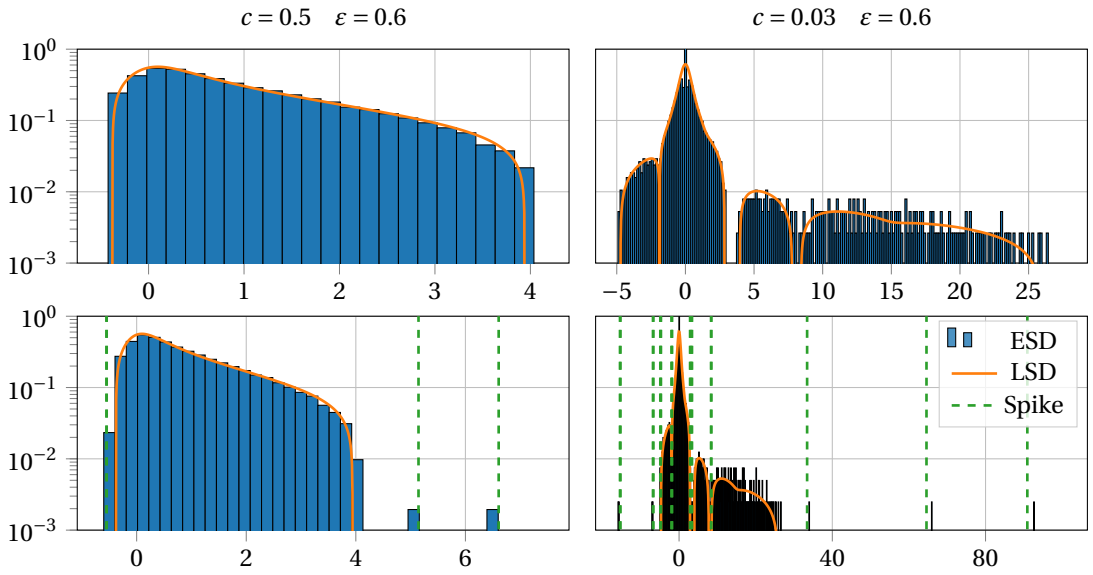


Figure 4.4: Same as Figure 4.2 but with a Toeplitz mask $\boldsymbol{T}$ instead of a circulant mask $\boldsymbol{C}$. Predictions (LSD and spike positions) are made with ψ_k replaced by τ_k and Fourier modes $\boldsymbol{F}$ replaced by $\boldsymbol{G}$, an eigenbasis of $\boldsymbol{T}$.

this setting (see Figure 4.2). The first time this spike appears, it is located between two bulks. It then goes through the rightmost bulk (so it is no longer an isolated eigenvalue thus $\zeta_1 \leqslant 0$), and finally goes out on the right edge of the distribution.

This last result may sound awkward and possibly testify of the suboptimality of our approach (when the signal-to-noise ratio increases, the information attached to some eigenvectors vanishes). This conclusion is not so immediate though, as the classification information is still contained within other eigenvectors which, as $\|\boldsymbol{\mu}\|$ increases, *do* carry increasingly clearer information.

4.2.3 Discussion on the Circulant Approximation

The approximation of the Toeplitz mask $\boldsymbol{T}$ by the circulant mask $\boldsymbol{C}$ used in Theorem 4.6 and Theorem 4.8 can be seen as a way to remove undesired edge effects, whose size is governed by L (notice that removing the first and last $L-1$ rows and columns of $\boldsymbol{C}$ and $\boldsymbol{T}$ yields the same two Toeplitz matrices). If L is chosen small compared to n (i.e., ε is small), edge effects are expected to be negligible and the previous results can plausibly be extended to the original setting, as observed empirically.

To adapt the previous results from $\boldsymbol{C}$ to $\boldsymbol{T}$, one only needs to change the eigenvalues and eigenvectors, i.e., replace ψ_k by τ_k — the eigenvalues of $\boldsymbol{T}$ — and replace $\boldsymbol{F}$ by $\boldsymbol{G} \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{g}_0 & \dots & \boldsymbol{g}_{n-1} \end{bmatrix}$, an eigenbasis of $\boldsymbol{T}$.

Very precise predictions on the original model can be made with these simple changes. Figures 4.4 and 4.5 compare simulations with a Toeplitz mask and the predictions of Theorem 4.6 and Theorem 4.8 with the ψ_k's replaced by the τ_k's and $\boldsymbol{F}$ replaced by $\boldsymbol{G}$.

Apart from extra mass around 0 in the second setting ($c = 0.03$ and $\varepsilon = 0.6$), the shape of the limiting spectral distribution is very well predicted, as well as the position of the isolated eigenvalues. Empirical alignments $\langle \boldsymbol{v}_0, \boldsymbol{u}_0 \rangle^2$ also fit well the predicted curve. Note that, contrary to the circulant mask, the eigenvalues of $\boldsymbol{T}$ are mostly simple (see Theorem 5 of Trench, 1994). Thus, we also repre-

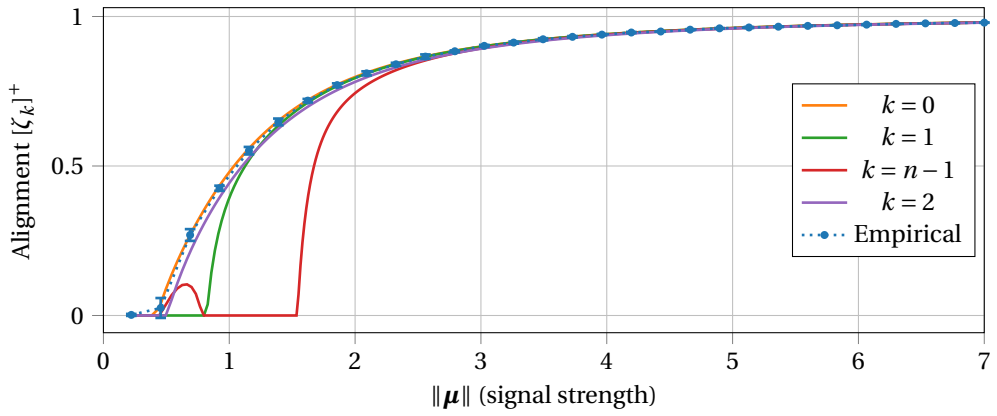


Figure 4.5: Same as Figure 4.3 but with a Toeplitz mask $\boldsymbol{T}$ instead of a circulant mask $\boldsymbol{C}$. Alignment curves are computed with ψ_k replaced by τ_k and Fourier modes $\boldsymbol{F}$ replaced by $\boldsymbol{G}$, an eigenbasis of $\boldsymbol{T}$.

sent $[\zeta_{n-1}]^+$ in Figure 4.5, which was confounded with $[\zeta_1]^+$ in Figure 4.3 ($\psi_1 = \psi_{n-1}$ but $\tau_1 \neq \tau_{n-1}$).

4.3 Online Spectral Clustering of Large Data

The previous results find direct applications to the online clustering of high-dimensional data streams.

4.3.1 Online Clustering Algorithm

Let us detail a clustering algorithm based on our previous results. We now use the banded version of the kernel matrix: $\boldsymbol{K}_L = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \boldsymbol{T}$ (the circulant mask is only useful for theoretical considerations) and recall the notation of the eigenbasis of $\boldsymbol{T}$: $\begin{bmatrix}\boldsymbol{g}_0 & \dots & \boldsymbol{g}_{n-1}\end{bmatrix} \overset{\text{def}}{=} \boldsymbol{G}$.

We consider a data stream of length T (possibly infinite). At each time step, a new vector $\boldsymbol{x}_t$ arrives in the pipeline while $\boldsymbol{x}_{t-L}$ is discarded (only the last L points are kept). The kernel matrix is then updated:

$$\left[\boldsymbol{K}_L^{(t)}\right]_{i,j} = \frac{1}{p}\boldsymbol{x}_{t-n+i}^\top \boldsymbol{x}_{t-n+j} \mathbf{1}_{|i-j|<L}.$$

Remark 4.9 (Memory management policy). A different memory management policy — not restricted to only choosing the previous L points to keep in memory — could be considered. However, we found that having points spread over a greater period of time (i.e., discarding newer ones to keep older ones) does not bring more information. To get a grasp, remark that the mean leaving time of the pipeline cannot be different than L, whatever the policy.

Remark 4.10 (Choice of n, L and eigenvector localization). It is important to emphasize that n is *not* the length of the data stream (given by the newly-introduced parameter $T \geqslant n$). As $\boldsymbol{K}_L$ has size $n \times n$, one can "only" classify the last n points of the stream, even when discarded from the length-L memory (older points are no longer classified though). The parameters n and L can be chosen by the user, accounting for memory limitations ($\mathcal{O}(Lp + Ln - \frac{1}{2}L(L-1))$ space is needed to store the data *and* the kernel matrix) and performance considerations, which we discuss in Section 4.3.2 below. Moreover, as the graph associated to $\boldsymbol{K}_L$ becomes sparser ($\varepsilon \to 0$), its eigenvectors tend to localize (Hata and Nakao, 2017), making classification more challenging. Therefore, $\varepsilon = (2L-1)/n$ should be neither too large (for the circulant approximation to hold) nor too small (to avoid eigenvector localization).

As per standard kernel spectral clustering, we use the dominant eigenvectors of $\boldsymbol{K}_L^{(t)}$ to estimate the classes. The last n points of the stream are classified at each time step so each point is classified n times. Then, the final class estimate can be chosen by a majority vote. However, standard clustering algorithms such as K-means — which are usually employed on spectral embeddings — perform poorly here, because of the particular shape of the eigenvectors caused by the Toeplitz mask[7] (see Figure 4.7).

Remark 4.11. The eigenvectors of $\boldsymbol{K}_L^{(t)}$ can be quickly computed at a low cost with a warm start of the power iteration algorithm from the previously computed eigenvectors of $\boldsymbol{K}_L^{(t-1)}$.

In a binary setting with globally centered data, classification can be performed using only the dominant eigenvector $\boldsymbol{u}_0^{(t)}$ of $\boldsymbol{K}_L^{(t)}$. Relying on the alignment of $\boldsymbol{u}_0^{(t)}$ with $\boldsymbol{v}_0^{(t)} = \boldsymbol{g}_0 \odot \boldsymbol{j}^{(t)}$ (Theorem 4.8) and the fact that the coordinates of $\boldsymbol{g}_0$ have constant sign[8], the class of $\boldsymbol{x}_{t-n+i}$ can be estimated from the sign of $\left[\boldsymbol{u}_0^{(t)}\right]_i$. This online clustering procedure is summarized in Algorithm 1.

Algorithm 1: Online Kernel Spectral Clustering (binary)

Output: class estimators $\{\hat{\mathcal{C}}_+^{(t)}, \hat{\mathcal{C}}_-^{(t)}\}_{n \leqslant t \leqslant T}$
for $t = 1$ **to** T **do**
  Get a new point $\boldsymbol{x}_t$ into the pipepline
  Update $\boldsymbol{K}_L^{(t-1)}$ into $\boldsymbol{K}_L^{(t)}$
  if $t \geqslant n$ **then**
    $\boldsymbol{u}_0^{(t)} \leftarrow \texttt{PowerIteration}(\boldsymbol{K}_L^{(t)}, \boldsymbol{u}_0^{(t-1)})$
    $\hat{\mathcal{C}}_\pm^{(t)} \leftarrow \{\boldsymbol{x}_{t-n+i} \mid [\boldsymbol{u}_0^{(t)}]_i \gtrless 0\}$
  end
end

The careful reader may wonder here whether the performance of the algorithm could be improved by using eigenvectors other than just the dominant one. In fact, the dominant eigenvector already contains all the information that can be retrieved. Since the classification is performed very easily with the signs of the coordinates in the binary setting $\boldsymbol{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \boldsymbol{I}_p)$, the use of other spike eigenvectors does not bring more information. However, in a general setting $\boldsymbol{x}_i \sim \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I}_p)$, we no longer have an alignment result such as Theorem 4.8 and it can become much harder to distinguish the classes from just the dominant eigenvector. In this case, the combination of several spike eigenvectors can make the classification easier. The interested reader is referred to Appendix 4.C, where we propose a — more complex and heuristic — online spectral clustering algorithm capable of handling K-class mixtures and test it on Fashion-MNIST images.

Note that these algorithms can easily be adapted to a setting where more than one vector $\boldsymbol{x}_t$ arrives at each time step (and this quantity does not need to be constant in time). This will nonetheless modify the structure of the kernel matrix $\boldsymbol{K}_L$ and additional work may be necessary to recover theoretical grounds.

4.3.2 Performance Versus Cost Trade-Off in Online Learning

One important question remains open in what we have presented so far: how to chose n and L? While the dimension p of the data is fixed by the problem considered, users are free to chose n and L as they wish. However, since we are concerned with memory-constrained settings, we introduce the *memory*

[7] The dominant eigenvector of $\boldsymbol{T}$, for example, is not constant, contrary to the first Fourier mode with the circulant mask.
[8] This is a consequence of the Perron-Frobenius theorem since the entries of $\boldsymbol{T}^k$ are all positive for $k \geqslant 1 + \lceil \frac{n-L}{L-1} \rceil$.

scale $M \overset{\text{def}}{=} Lp + Ln - \frac{1}{2}L(L-1)$, which decomposes into the memory allocated to the L data points kept in the pipeline (Lp) plus the memory allocated to the banded kernel matrix $\boldsymbol{K}_L^{(t)}$ ($Ln - \frac{1}{2}L(L-1)$ due to symmetry). Given that M and p are fixed, we seek the optimal way to chose n and L.

Recall Equation (4.2) defining the minimal value of $\|\boldsymbol{\mu}\|^2$ for a positive alignment between $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$. Taking $k = 0$, notice that,

$$\frac{\psi_0}{\psi_l} = \frac{2L-1}{\left[\frac{\sin((2L-1)\frac{l\pi}{n})}{\sin(\frac{l\pi}{n})}\right]} = \frac{\operatorname{sinc}(\frac{l\pi}{n})}{\operatorname{sinc}((2L-1)\frac{l\pi}{n})} \qquad \text{with} \qquad \operatorname{sinc} x = \frac{\sin x}{x}$$

and we have the following result.

Proposition 4.12. *As* $p, n, L \to +\infty$, $\sum_{l=1}^{\lfloor n/2 \rfloor} \left[\left(\|\boldsymbol{\mu}\|^2 + 1 \right) \frac{\psi_0}{\psi_l} - 1 \right]^{-2} = \sum_{l=1}^{\lfloor n/2 \rfloor} \left[\frac{\|\boldsymbol{\mu}\|^2 + 1}{\operatorname{sinc}((2L-1)\frac{l\pi}{n})} - 1 \right]^{-2} + \mathrm{o}(1)$.

Proof. See Appendix 4.E. □

Hence, as $p, n, L \to +\infty$, the position of the first phase transition is solution to

$$1 - c\left(\frac{1}{\|\boldsymbol{\mu}\|^4} + 2\sum_{l=1}^{+\infty} \left[\frac{\|\boldsymbol{\mu}\|^2 + 1}{\operatorname{sinc}(\varepsilon l \pi)} - 1 \right]^{-2} \right) = 0$$

where the term $\frac{1}{\|\boldsymbol{\mu}\|^4}$ stands for $l = 0$ and the sum $\sum_{l=1}^{+\infty}$ must appear twice to account for "l small" and "$n - l$ small" (this can be understood by the fact that $\frac{\psi_l}{\psi_0}$ behaves like $\operatorname{sinc}(\varepsilon l \pi)$ for $l \leqslant \frac{n}{2}$ and like $\operatorname{sinc}(\varepsilon(n-l)\pi)$ for $l \geqslant \frac{n}{2}$, see Figure 4.1). This shows that, asymptotically, the position of the phase transition depends only on the two parameters c and ε.

The constraint $M = Lp + Ln - \frac{1}{2}L(L-1)$ can be translated into $Lp = \alpha M$ and $Ln - \frac{1}{2}L(L-1) = (1-\alpha)M$ where $\alpha \in]0,1[$ represents the portion of memory dedicated to the pipeline while $1-\alpha$ is the portion of memory dedicated to the kernel matrix. The parameters c and ε can then be expressed as $c = [\frac{1-\alpha}{\alpha} + \frac{1}{2}\alpha r]^{-1}$ and $\varepsilon = 2\alpha r [\frac{1-\alpha}{\alpha} + \frac{1}{2}\alpha r]^{-1}$ where we have introduced $r \overset{\text{def}}{=} M/p^2$ representing the amount of memory "normalized" by the dimensionality of the data p^2.

Therefore, the position of the phase transition can be written as a function of α and r as well. Let $\phi(\alpha, r)$ be this function. Since the value of r is already fixed by the problem considered, α is the only free parameter. The best way to chose it is so that $\phi(\alpha, r)$ is minimal. Indeed, the earlier the phase transition, the better the alignment between $\boldsymbol{u}_0$ and $\boldsymbol{v}_0$ at a given signal-to-noise ratio $\|\boldsymbol{\mu}\|^2$. Thus, we set $\alpha_\star = \min_{0<\alpha\leqslant\bar{\alpha}} \phi(\alpha, r)$ where $\bar{\alpha} \overset{\text{def}}{=} \frac{-1+\sqrt{2r+1}}{r} < 1$. The constraint $\alpha \leqslant \bar{\alpha}$ is the translation of $L \leqslant n$. The value of $\alpha_\star$ can be numerically computed and the corresponding optimal values of n and L can be recovered: $n_\star = \frac{1-\alpha_\star}{\alpha_\star} p + \frac{1}{2}(\frac{\alpha_\star M}{p} - 1)$ and $L_\star = \frac{\alpha_\star M}{p}$.

The top panel of Figure 4.6 shows the value of $\alpha_\star$ as a function of r (in log scale). It is instructing to see that $\alpha_\star$ is monotonically decreasing from 1 to 0 as r grows. This means that, under very restrictive memory settings, a larger part of it should be dedicated to the storage of data while, as memory limitations become less constraining, a larger portion should be devoted to the storage of $\boldsymbol{K}_L$. As r becomes large, most of the memory should be occupied by the kernel matrix rather than the data pipeline ("small" L but big n).

The bottom panel displays the position of the phase transition as a function of r when n and L are chosen in the previously-introduced optimal way. It is compared with the phase transitions of *batch clustering* and *offline clustering*.

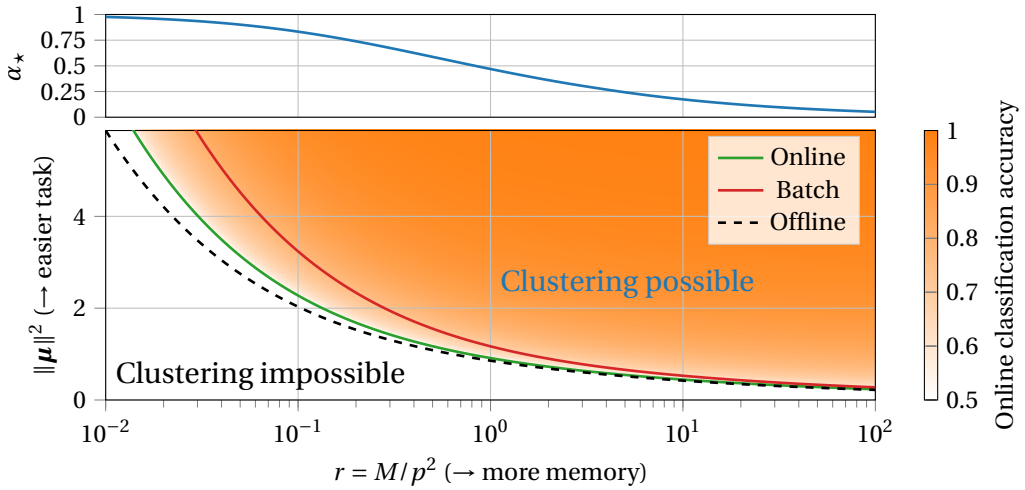


Figure 4.6: **Top:** $\alpha_\star$ against $r = M/p^2$. This represents the amount of memory $L_\star p$ dedicated to the pipeline depending on the total available memory. **Bottom:** Phase transition position of the dominant eigenvector of the banded Gram kernel matrix $\boldsymbol{K}_{L_\star}$ when $(n, L) = (n_\star, L_\star)$ (green curve) and of the standard $n' \times n'$ kernel matrix when performing batch spectral clustering with the corresponding available memory (red curve). Non-trivial clustering is only possible above the phase transition. The orange density map gives the predicted online clustering accuracy (with $\boldsymbol{K}_{L_\star}$) depending on the values of r and $\|\boldsymbol{\mu}\|^2$. The black dashed line shows the phase transition corresponding to the $n_\star \times n_\star$ *full* Gram kernel matrix (not achievable in an online setting).

- "Batch clustering" is the naive approach consisting in standard spectral clustering with an $n' \times n'$ kernel matrix such that $n'p + \frac{n'(n'+1)}{2} = M$. Recalling Theorem 2.34, the phase transition happens when $\|\boldsymbol{\mu}\|^2 = \sqrt{p/n'}$ and, given the memory constraint, this is $\frac{1}{\sqrt{-1+\sqrt{2r+1}}} + o(1)$.
- "Offline clustering" corresponds to the best achievable performance with the same ratio $p/n_\star$ but *without* memory constraint, i.e., with a full $n_\star \times n_\star$ kernel matrix ($L = n_\star$). Hence, the position of the phase transition is simply $\sqrt{p/n_\star} = [\frac{1-\alpha_\star}{\alpha_\star} + \frac{1}{2}\alpha_\star r]^{-1/2} + o(1)$.

Below the green curve, $\zeta_0 \leqslant 0$, no eigenvalue escapes the bulk and non-trivial clustering is impossible. After the phase transition, $\zeta_0 > 0$ and the closer it is to 1 the closer $\boldsymbol{u}_0$ is to $\boldsymbol{v}_0 = \frac{1}{\sqrt{n}}\boldsymbol{j}$. Following Theorem 3.4, the fluctuations of the entries of $\boldsymbol{u}_0$ are asymptotically Gaussian and pairwise independent with mean $\pm\sqrt{[\zeta_0]^+/n}$ and variance $(1-[\zeta_0]^+)/n$. Thus, the asymptotic classification accuracy is given by $\Phi\left(\sqrt{\frac{[\zeta_0]^+}{1-[\zeta_0]^+}}\right)$, where Φ is the Gaussian cumulative distribution function: $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,\mathrm{d}t$. This is represented by the orange density map.

From Figure 4.6, we see that, as r grows, i.e., as memory becomes less restrictive, the phase transition position of online spectral clustering reaches the optimal threshold $\|\boldsymbol{\mu}\|^2 = \sqrt{c}$ under which no information can be recovered (regardless of the method used and the data available). This is expected, since increasing the memory size allows to encapsulate more information. Still, the green curve specifies *how* memory limitations impair performance. Although we lack some information-theoretic result, the distance between the green curve and the black dashed line yields an upper bound on the difference between the performances of our method and an optimistic optimum (which, as wee see, can get very close to 0). Moreover, with r fixed (fixed memory size), the classification error vanishes as $\|\boldsymbol{\mu}\|$ increases (the signal becomes more powerful).

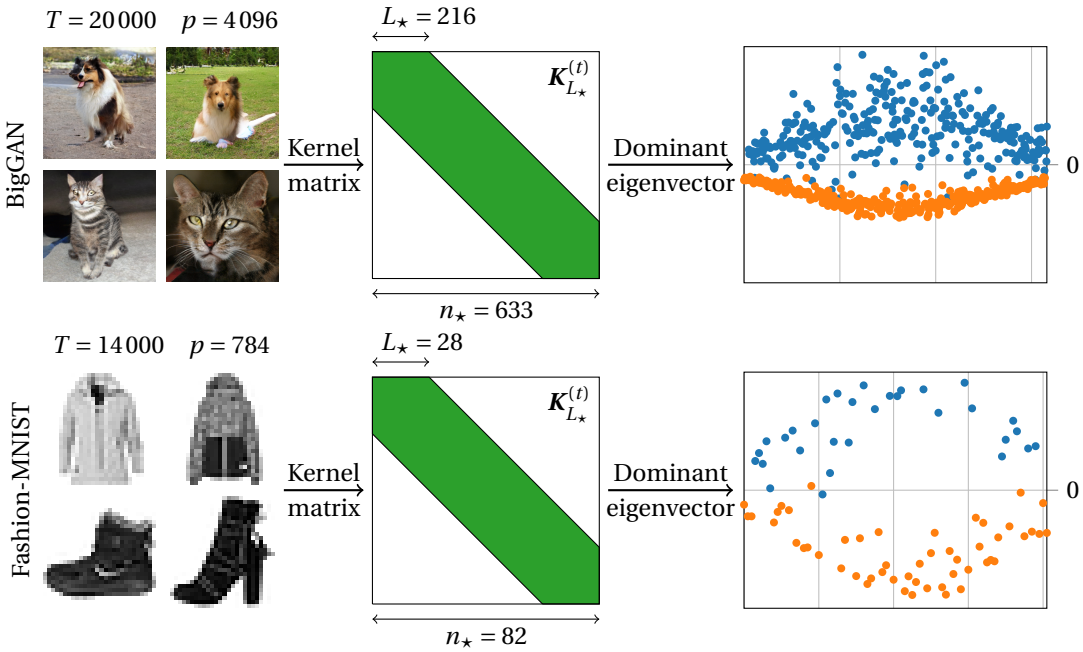


Figure 4.7: **Left:** Examples of BigGAN-generated images of `collie` dogs and `tabby` cats (**top**) and Fashion-MNIST images from the classes `Coat` and `Ankle Boot` (**bottom**). **Middle:** Illustration of the banded Gram kernel matrix $\boldsymbol{K}^{(t)}_{L_\star}$ with the corresponding values of $n_\star$ and $L_\star$. **Right:** Entries of the dominant eigenvector $\boldsymbol{u}^{(t)}_0$ of $\boldsymbol{K}^{(t)}_{L_\star}$ with each entry colored according to its class (ground truth).

Our method performs better (i.e., the phase transition occurs earlier) than a naive approach based on standard spectral clustering performed by batches, in particular under very restrictive memory settings. It is also able to classify the $n_\star$ previous points at any time, *although* $n_\star - L_\star$ *of them have already left memory!*

4.3.3 Simulations on Real-World Images

We illustrate our findings with two applications on image clustering tasks. Firstly, we apply Algorithm 1 on globally centered and scaled VGG-features (Simonyan and Zisserman, 2015) of randomly BigGAN-generated images (Brock et al., 2019) of `tabby` cats and `collie` dogs (see Figure 4.7). The vectors thus generated have dimension $p = 4\,096$ and simulate a stream of length $T = 20\,000$ with evenly likely cats and dogs. This experiment is performed twice: once on the original computed VGG-features and once on the same features perturbed with additive Gaussian noise. In addition, our algorithm is applied to a stream made of $T = 14\,000$ centered *raw* images from the Fashion-MNIST dataset (Xiao et al., 2017). Their dimension is $p = 784$ and we want to discriminate `coat` versus `ankle boot` in an online fashion. In both cases, n and L are chosen in order to minimize the position of the phase transition given a fixed amount of available memory M. For the BigGAN dataset, we take $M = 1\,000\,000$ (4 000 kB[9]) and, for the Fashion-MNIST data, we take $M = 24\,000$ (96 kB). This is a realistic choice of parameters which can easily be run on most standard laptops. At each time step, $L_\star$ images are kept in memory and, from the $n_\star \times n_\star$ kernel matrix $\boldsymbol{K}^{(t)}_{L_\star}$, we are able to classify the previous $n_\star$ images.

[9] This assumes that each floating-point number is stored on 32 bits.

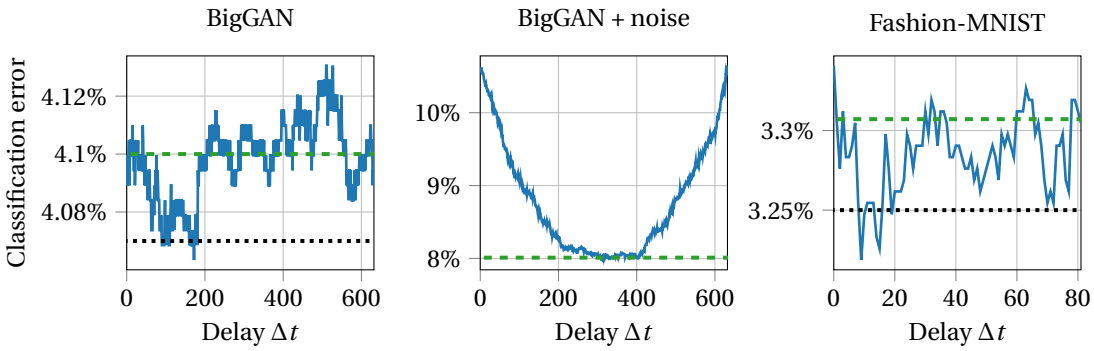


Figure 4.8: Classification error against delay Δt on BigGAN-generated images (**left** and **middle**) and Fashion-MNIST images (**right**). This is the mean classification error at time $t_0 + \Delta t$ of a point arrived at t_0. The green dashed line indicates the overall classification error when the class is chosen by a majority vote. The black dotted line is the classification error obtained with a $T \times T$ offline kernel spectral clustering. **Experimental setting:** See Figure 4.7.

		Limited memory			No memory constraint	
Stream	Memory	Online	Batch	Sketching	KSC	Sketching
BigGAN	4 000 kB	4.10	4.07	36.14	4.07	–
BigGAN + noise	4 000 kB	8.01	10.49	49.68	5.37	–
Fashion-MNIST	96 kB	3.31	3.38	14.29	3.25	5.36

Table 4.1: Classification error (%) on three image clustering tasks. KSC stands for kernel spectral clustering. Sketching on the whole BigGAN-generated dataset is too memory-demanding for a regular laptop.

The shape of the dominant eigenvector $\boldsymbol{u}_0^{(t)}$ at a certain time t during the execution of the algorithm is depicted in Figure 4.7. We clearly see a separation between the classes. For both settings, Figure 4.8 plots the mean classification error at $t_0 + \Delta t$ of a data point seen at t_0 (recall that a data point arriving at t_0 is classified at each time step between t_0 and $t_0 + n - 1$), as well as the overall classification error obtained after a majority vote (green dashed line), to be compared with the classification error obtained with a standard $T \times T$ *offline* kernel spectral clustering for which optimality results are known (black dotted line). Except in the setting where noise is added to the VGG-features of BigGAN-generated images, the mean classification error remains constant with Δt, thus showing that our algorithm does not lose any discriminative power between t_0 and $t_0 + n - 1$. Moreover, the classification performances of our algorithm are very close to those of the standard (offline and costly) spectral clustering but require much less memory resources: $\mathcal{O}(Lp + Ln - \frac{1}{2}L(L-1))$ against $\mathcal{O}(Tp + \frac{1}{2}T(T+1))$ space for the storage of the data and the kernel matrix. The "BigGAN + noise" experiment shows an interesting behavior: in a difficult (very noisy) setting, the lowest clustering error of a data point arrived at t_0 is achieved at $t \approx t_0 + \frac{n_\star}{2}$. This can be understood as a side effect of the U-shaped eigenvector used for the clustering (see Figure 4.7). Yet, this does not impair the overall performance when the clustering is performed by a majority vote over each time step (the green dashed line is near the bottom of the U in Figure 4.8).

In Table 4.1, we compare the performance of our algorithm (Online), batch clustering and sketching (Gribonval et al., 2021) with limited memory and these three clustering tasks. As a reference, we also give the performance of standard spectral clustering with the Gram kernel matrix (KSC) and

sketching achieved at once on the whole length-T dataset. Sketching is a general procedure which consists in computing a *sketch* $\boldsymbol{z} = \frac{1}{T}\sum_{t=1}^{T} \Phi(\boldsymbol{x}_i) \in \mathbb{C}^m$ with a transformation $\Phi : \boldsymbol{x} \mapsto \exp(-\mathrm{i}\boldsymbol{\Omega}^\top \boldsymbol{x})$ where $\boldsymbol{\Omega}$ is a $p \times m$ matrix of random frequencies and exp is applied entrywise (this is not the matrix exponential). The sketch can easily be computed in an online fashion and the clustering performed directly from it, see Gribonval et al. (2021) for details. However, this approach requires the storage of the large matrix $\boldsymbol{\Omega}$ and the size m of the sketch is therefore constrained by $M = m + pm \iff m = M/(p+1)$ (this is $m = 244$ for BigGAN images and $m = 30$ for Fashion-MNIST images). Hence, most of the memory is dedicated to the storage of a non-informative matrix, which explains the relatively poor results of this approach in our experiments[10].

On the BigGAN-generated stream of images, our online spectral clustering algorithm has the same performance as batch clustering, which is also the same performance as spectral clustering on the entire dataset. This problem is in fact "too easy" and the addition of noise exposes real differences in the performance (it places us closer to the phase transition by reducing the signal-to-noise ratio). Similar differences can be observed on the raw Fashion-MNIST dataset. These observations are in line with our predictions illustrated in Figure 4.6.

4.4 Concluding Remarks

Learning on large amounts of high-dimensional data is a challenging task requiring either large computational resources or efficient data-processing algorithms. This chapter is concerned with the second option: leveraging tools from random matrix theory and the circulant approximation of Toeplitz matrices, the analysis of the "banded kernel matrix" $\boldsymbol{K}_L = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \boldsymbol{T}$ has revealed interesting behaviors of its limiting spectrum. Notably, it can exhibit multiple bulks, with isolated eigenvalues on both sides, and even between the bulks. The characterization of the alignments of the corresponding isolated eigenvectors with the (right) singular subspace of the perturbation $\boldsymbol{P}$ displays a phase transition phenomenon: after a threshold value on the strength of the perturbation, isolated eigenvalues appear in the spectrum of $\boldsymbol{K}_L$ and their eigenvectors carry the sought information "convolved" with Fourier modes. This random matrix analysis finds direct applications to the study of online data stream clustering. Under limited memory resources, we have shown how to design the optimal banded kernel matrix and specified the accuracy of our approach. Near-optimal performances on high-dimensional data can be achieved with memory-limited systems using our online kernel spectral clustering algorithm and simulations on image classification tasks support our findings.

Besides introducing a new algorithm for online clustering, these results also pave the way towards large-dimensional learning on data streams with theoretical guarantees. This also raises several interesting questions to further strengthen these theoretical grounds. Firstly, our results on the performance of online kernel spectral clustering rely on the dominant eigenvector of $\boldsymbol{K}_L$ which, under the circulant approximation, is known to have asymptotic Gaussian fluctuations as per our results in Chapter 3. However, similar guarantees on other eigenvectors than the dominant one are not straightforward because of their particular intrication with Fourier modes. Then, we still miss an information-theoretic result of optimality for the proposed approach (which exists in the standard unbanded case), this key direction requires further investigation — a conceivable approach is through the techniques deployed by Nguyen and Couillet (2023) in the context of multitask learning. Finally, our analysis is performed under an i.i.d. assumption on the data (Assumption 4.1) which rules out the trickier but realistic situation where several points of the same class could arrive together in a batch and the data distribution could change over time. It would be interesting to know how our random matrix tools adapt to this kind of situation.

[10] Still, numerical methods exist to reduce this memory cost (Chatalic et al., 2018) and the purpose of sketching is not restricted to clustering so it should not be too severely judged in the light of these results!

4.A Proof of Theorem 4.6

4.A.1 Preliminary Results

Firstly, let us state a few useful results for the upcoming proof.

Singular Value Inequalities

Proposition 4.13 (Bai and Silverstein, 2010, Corollary A.12)**.** *For all $n \times n$ real (or complex) matrix $\boldsymbol{A}$,* $|\operatorname{Tr} \boldsymbol{A}| \leqslant \sum_{i=1}^{n} s_i(\boldsymbol{A})$.

Proposition 4.14 (Bai and Silverstein, 2010, Theorem A.14)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two real (or complex) matrices of size $p \times q$ and $q \times r$ respectively. For any integer $k \in \{1, \dots, \min(p,q,r)\}$,*

$$\sum_{i=1}^{k} s_i(\boldsymbol{AB}) \leqslant \sum_{i=1}^{k} s_i(\boldsymbol{A}) s_i(\boldsymbol{B}).$$

Hadamard Product Inequalities

Nota Bene. The (standard) matrix product has priority over the Hadamard product, i.e., $\boldsymbol{A} \odot \boldsymbol{BC} = \boldsymbol{A} \odot (\boldsymbol{BC})$.

We will often need to bound from above the spectral norm of Hadamard products. Thus, we state the following submultiplicativity property.

Proposition 4.15 (Bai and Silverstein, 2010, Theorem A.19)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two real (or complex) matrices of size $p \times n$. For any integer $k \in \{1, \dots, \min(p,n)\}$,*

$$\sum_{i=1}^{k} s_i(\boldsymbol{A} \odot \boldsymbol{C}) \leqslant \sum_{i=1}^{k} s_i(\boldsymbol{A}) s_i(\boldsymbol{C}).$$

Corollary 4.16. $\|\boldsymbol{A} \odot \boldsymbol{B}\| \leqslant \|\boldsymbol{A}\| \|\boldsymbol{B}\|$.

As we will be mostly interested in the case where $\boldsymbol{B} = \boldsymbol{C}$ (the circulant mask) and $\|\boldsymbol{C}\| = 2L - 1$, the upper bound $\|\boldsymbol{A} \odot \boldsymbol{C}\| \leqslant (2L-1)\|\boldsymbol{A}\|$ is not satisfying (because L grows as fast as n). Instead, we have the following upper bound.

Proposition 4.17. *There exists a constant $\alpha > 0$ such that, for all real $n \times n$ matrix $\boldsymbol{A}$, $\|\boldsymbol{A} \odot \boldsymbol{C}\| \leqslant \alpha \log n \|\boldsymbol{A}\|$.*

Proof. We use the complex eigendecomposition $\boldsymbol{C} = \widetilde{\boldsymbol{F}} \boldsymbol{\Psi} \widetilde{\boldsymbol{F}}^*$ where $\widetilde{F}_{i,j} = \frac{1}{\sqrt{n}} e^{-\frac{2\mathrm{i}\pi}{n}(i-1)(j-1)}$ and $\boldsymbol{\Psi} = \operatorname{Diag}(\psi_0, \dots, \psi_{n-1})$ with $\psi_k = \nu_L(\frac{2k\pi}{n})$ (see Section 4.1.2). We can write $\boldsymbol{A} \odot \boldsymbol{C}$ as

$$\boldsymbol{A} \odot \boldsymbol{C} = \sum_{k=0}^{n-1} \psi_k \boldsymbol{A} \odot \widetilde{\boldsymbol{F}}_{\cdot,k} \widetilde{\boldsymbol{F}}_{\cdot,k}^* = \sum_{k=0}^{n-1} \psi_k \operatorname{Diag}(\widetilde{\boldsymbol{F}}_{\cdot,k}) \boldsymbol{A} \operatorname{Diag}(\widetilde{\boldsymbol{F}}_{\cdot,k})^*$$

using the useful relation $\boldsymbol{A} \odot \boldsymbol{ab}^* = \operatorname{Diag}(\boldsymbol{a}) \boldsymbol{A} \operatorname{Diag}(\boldsymbol{b})^*$. Notice that $\|\operatorname{Diag}(\widetilde{\boldsymbol{F}}_{\cdot,k})\|^2 = \max_{i \in [n]} |\widetilde{F}_{i,k}|^2 = \frac{1}{n}$ therefore

$$\|\boldsymbol{A} \odot \boldsymbol{C}\| \leqslant \left(\frac{1}{n} \sum_{k=0}^{n-1} |\psi_k| \right) \|\boldsymbol{A}\|.$$

The rest of the proof consists in showing that $\frac{1}{n}\sum_{k=0}^{n-1}|\psi_k| = \mathcal{O}(\log n)$. From the symmetry of ν_L, $\frac{1}{n}\sum_{k=0}^{n-1}|\psi_k| = \frac{2}{n}\sum_{k=1}^{\lfloor n/2\rfloor}|\psi_k| + \mathcal{O}(1)$. Moreover, notice that

$$\nu_L(x) - \cos(Lx) = \frac{\sin((2L-1)\frac{x}{2})}{\sin(\frac{x}{2})} - \cos(Lx) = \sin(Lx)\cot\left(\frac{x}{2}\right)$$

and $0 < \frac{2}{x} - \cot(\frac{x}{2}) < \frac{2}{\pi}$ for all $x \in]0,\pi[$, therefore $\nu_L(x) = 2\frac{\sin(Lx)}{x} + \mathcal{O}(1)$ for all $x \in]0,\pi[$. Hence,

$$\frac{1}{n}\sum_{k=0}^{n-1}\left|\psi_k\right| = \frac{2}{n}\sum_{k=1}^{\lfloor n/2\rfloor}\left|\psi_k\right| + \mathcal{O}(1) = \frac{4}{n}\sum_{k=1}^{\lfloor n/2\rfloor}\left|\frac{\sin(L\frac{2k\pi}{n})}{\frac{2k\pi}{n}}\right| + \mathcal{O}(1) \leqslant \frac{2}{\pi}\sum_{k=1}^{\lfloor n/2\rfloor}\frac{1}{k} + \mathcal{O}(1) = \mathcal{O}(\log n).$$

□

In fact, we can state the following more general statement.

Proposition 4.18. *$n^{-q}\|\boldsymbol{A}\odot\boldsymbol{C}^{q+1}\| \leqslant (\alpha\log n)^{q+1}\|\boldsymbol{A}\|$ for all integer $q \geqslant 0$.*

Proof. We proceed similarly to the proof of Proposition 4.17.

$$\frac{1}{n^q}\left\|\boldsymbol{A}\odot\boldsymbol{C}^{q+1}\right\| = \frac{1}{n^q}\left\|\sum_{k=0}^{n-1}\psi_k^{q+1}\operatorname{Diag}(\widetilde{\boldsymbol{F}}_{\cdot,k})\boldsymbol{A}\operatorname{Diag}(\widetilde{\boldsymbol{F}}_{\cdot,k})^*\right\|$$
$$\leqslant \left(\frac{1}{n^{q+1}}\sum_{k=0}^{n-1}\left|\psi_k\right|^{q+1}\right)\|\boldsymbol{A}\| \leqslant \left(\frac{1}{n}\sum_{k=0}^{n-1}\left|\psi_k\right|\right)^{q+1}\|\boldsymbol{A}\| \leqslant \left(\alpha\log n\right)^{q+1}\|\boldsymbol{A}\|.$$

□

Finally, this last inequality will be useful as well.

Proposition 4.19. *$\|\boldsymbol{A}\odot\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}\| \leqslant \|\boldsymbol{A}\|$ for all $i,j \in [n]$.*

Proof. This is an elementary consequence of the relation $\boldsymbol{A}\odot\boldsymbol{a}\boldsymbol{b}^\top = \operatorname{Diag}(\boldsymbol{a})\boldsymbol{A}\operatorname{Diag}(\boldsymbol{b})$,

$$\left\|\boldsymbol{A}\odot\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}\right\| = \left\|\operatorname{Diag}(\boldsymbol{C}_{\cdot,i})\boldsymbol{A}\operatorname{Diag}(\boldsymbol{C}_{j,\cdot})\right\| \leqslant \|\boldsymbol{A}\|$$

since $\|\operatorname{Diag}(\boldsymbol{C}_{\cdot,i})\| = \|\operatorname{Diag}(\boldsymbol{C}_{j,\cdot})\| = 1$. □

4.A.2 Derivations with Stein's Lemma

Let us define the resolvent of the kernel matrix with the *circulant* mask,

$$\widetilde{\boldsymbol{Q}}(z) = \left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot\boldsymbol{C} - z\boldsymbol{I}_n\right)^{-1} \qquad \text{for all } z \in \mathbb{C}\setminus\operatorname{Sp}\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot\boldsymbol{C}\right).$$

As usual, we will often drop the dependence in z and simply write $\widetilde{\boldsymbol{Q}}$ instead of $\widetilde{\boldsymbol{Q}}(z)$ to ease the notation.

In this section, we prove the two results gathered in the following proposition.

Proposition 4.20. *For all $\boldsymbol{A} \in \mathbb{R}^{n\times n}$,*

$$\mathbb{E}\left[\left(\frac{\boldsymbol{N}^\top\boldsymbol{X}}{p}\odot\boldsymbol{A}\right)\widetilde{\boldsymbol{Q}}\right] = \mathbb{E}\left[(\boldsymbol{A}\odot\boldsymbol{I}_n)\widetilde{\boldsymbol{Q}}\right] - \mathbb{E}\left[\left(\frac{1}{p}\boldsymbol{A}\left(\widetilde{\boldsymbol{Q}}\odot\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\right)\boldsymbol{C}\odot\boldsymbol{I}_n\right)\widetilde{\boldsymbol{Q}}\right] - \mathbb{E}\left[\frac{1}{p}\left((\widetilde{\boldsymbol{Q}}\odot\boldsymbol{A})\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot\boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right], \quad (4.3)$$

$$\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top\boldsymbol{N}}{p}\odot\boldsymbol{A}\right)\widetilde{\boldsymbol{Q}}\right] = -\frac{1}{p}\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top\boldsymbol{X}}{p}\odot\boldsymbol{A}(\widetilde{\boldsymbol{Q}}\odot\boldsymbol{I}_n)\boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right] - \frac{1}{p}\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top\boldsymbol{X}}{p}(\widetilde{\boldsymbol{Q}}\odot\boldsymbol{C})\odot\boldsymbol{A}\right)\widetilde{\boldsymbol{Q}}\right]. \quad (4.4)$$

The proofs of Equations (4.3) and (4.4) rely on Stein's lemma (Lemma 2.18). Hence, we need the following expression for the derivatives of the resolvent $\widetilde{\boldsymbol{Q}}$.

Lemma 4.21 (Derivatives of the resolvent)**.**

$$\frac{\partial \widetilde{Q}_{a,b}}{\partial N_{c,d}} = -\frac{1}{p}\sum_{e=1}^{n}\big(\widetilde{Q}_{a,d}X_{c,e}C_{d,e}\widetilde{Q}_{e,b} + \widetilde{Q}_{a,e}X_{c,e}C_{e,d}\widetilde{Q}_{d,b}\big).$$

Proof. We have $(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot\boldsymbol{C} - z\boldsymbol{I}_n)\widetilde{\boldsymbol{Q}} = \boldsymbol{I}_n$ therefore $\partial\widetilde{\boldsymbol{Q}} = -\widetilde{\boldsymbol{Q}}\partial(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot\boldsymbol{C})\widetilde{\boldsymbol{Q}}$.

$$\begin{aligned}
\frac{\partial \widetilde{Q}_{a,b}}{\partial N_{c,d}} &= -\frac{1}{p}\sum_{e=1}^{n}\sum_{f=1}^{n}\widetilde{Q}_{a,e}\frac{\partial}{\partial N_{c,d}}\big[\boldsymbol{X}^\top\boldsymbol{X}\odot\boldsymbol{C}\big]_{e,f}\widetilde{Q}_{f,b}\\
&= -\frac{1}{p}\sum_{e=1}^{n}\sum_{f=1}^{n}\widetilde{Q}_{a,e}\frac{\partial}{\partial N_{c,d}}\left[\sum_{g=1}^{p}X_{g,e}X_{g,f}C_{e,f}\right]\widetilde{Q}_{f,b}\\
&= -\frac{1}{p}\sum_{e=1}^{n}\sum_{f=1}^{n}\sum_{g=1}^{p}\widetilde{Q}_{a,e}\big(\delta_{g,c}\delta_{e,d}X_{g,f}C_{e,f} + X_{g,e}\delta_{g,c}\delta_{f,d}C_{e,f}\big)\widetilde{Q}_{f,b}\\
&= -\frac{1}{p}\left(\sum_{f=1}^{n}\widetilde{Q}_{a,d}X_{c,f}C_{d,f}\widetilde{Q}_{f,b} + \sum_{e=1}^{n}\widetilde{Q}_{a,e}X_{c,e}C_{e,d}\widetilde{Q}_{d,b}\right)\\
&= -\frac{1}{p}\sum_{e=1}^{n}\big(\widetilde{Q}_{a,d}X_{c,e}C_{d,e}\widetilde{Q}_{e,b} + \widetilde{Q}_{a,e}X_{c,e}C_{e,d}\widetilde{Q}_{d,b}\big).
\end{aligned}$$

□

Proof of Equation (4.3)

We use successively Stein's lemma (Lemma 2.18) and Lemma 4.21.

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\boldsymbol{N}^\top\boldsymbol{X}}{p}\odot\boldsymbol{A}\right)\widetilde{\boldsymbol{Q}}\right]_{i,j} &= \frac{1}{p}\sum_{r=1}^{p}\sum_{s=1}^{n}\mathbb{E}\big[N_{r,i}X_{r,s}A_{i,s}\widetilde{Q}_{s,j}\big]\\
&= \frac{1}{p}\sum_{r=1}^{p}\sum_{s=1}^{n}\mathbb{E}\left[\frac{\partial X_{r,s}}{\partial N_{r,i}}A_{i,s}\widetilde{Q}_{s,j} + X_{r,s}A_{i,s}\frac{\partial\widetilde{Q}_{s,j}}{\partial N_{r,i}}\right]\\
&= \mathbb{E}\big[A_{i,i}\widetilde{Q}_{i,j}\big] - \frac{1}{p^2}\sum_{r=1}^{p}\sum_{s=1}^{n}\mathbb{E}\left[X_{r,s}A_{i,s}\sum_{t=1}^{n}\big(\widetilde{Q}_{s,i}X_{r,t}C_{i,t}\widetilde{Q}_{t,j} + \widetilde{Q}_{s,t}X_{r,t}C_{t,i}\widetilde{Q}_{i,j}\big)\right]\\
&= \mathbb{E}\big[A_{i,i}\widetilde{Q}_{i,j}\big] - \frac{1}{p^2}\mathbb{E}\big[((\widetilde{\boldsymbol{Q}}\odot\boldsymbol{A})\boldsymbol{X}^\top\boldsymbol{X}\odot\boldsymbol{C})\widetilde{\boldsymbol{Q}}\big]_{i,j} - \frac{1}{p^2}\mathbb{E}\Big[\big[\boldsymbol{A}(\widetilde{\boldsymbol{Q}}\odot\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{C}\big]_{i,i}\widetilde{Q}_{i,j}\Big].
\end{aligned}$$

Proof of Equation (4.4)

Similarly,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top\boldsymbol{N}}{p}\odot\boldsymbol{A}\right)\widetilde{\boldsymbol{Q}}\right]_{i,j} &= \frac{1}{p}\sum_{r=1}^{p}\sum_{s=1}^{n}\mathbb{E}\big[P_{r,i}N_{r,s}A_{i,s}\widetilde{Q}_{s,j}\big]\\
&\overset{\text{(Lemma 2.18)}}{=} \frac{1}{p}\sum_{r=1}^{p}\sum_{s=1}^{n}\mathbb{E}\left[P_{r,i}A_{i,s}\frac{\partial\widetilde{Q}_{s,j}}{\partial N_{r,s}}\right]\\
&\overset{\text{(Lemma 4.21)}}{=} -\frac{1}{p^2}\sum_{r=1}^{p}\sum_{s=1}^{n}\mathbb{E}\left[P_{r,i}A_{i,s}\sum_{t=1}^{n}\big(\widetilde{Q}_{s,s}X_{r,t}C_{s,t}\widetilde{Q}_{t,j} + \widetilde{Q}_{s,t}X_{r,t}C_{t,s}\widetilde{Q}_{s,j}\big)\right]
\end{aligned}$$

$$= -\frac{1}{p^2}\mathbb{E}\big[(\boldsymbol{P}^\top \boldsymbol{X} \odot \boldsymbol{A}(\widetilde{\boldsymbol{Q}} \odot \boldsymbol{I}_n)\boldsymbol{C})\widetilde{\boldsymbol{Q}}\big]_{i,j} - \frac{1}{p^2}\mathbb{E}\big[(\boldsymbol{P}^\top \boldsymbol{X}(\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C}) \odot \boldsymbol{A})\widetilde{\boldsymbol{Q}}\big]_{i,j}.$$

4.A.3 Concentration Results

Concentration of Bilinear Forms and Traces of the Resolvent

Here, we show the following important result.

Proposition 4.22. *$\mathbb{E}[\widetilde{\boldsymbol{Q}}]$ is a deterministic equivalent (Definition 2.17) of $\widetilde{\boldsymbol{Q}}$, i.e.,*

$$\boldsymbol{a}^\top\big(\widetilde{\boldsymbol{Q}} - \mathbb{E}[\widetilde{\boldsymbol{Q}}]\big)\boldsymbol{b} \xrightarrow[p,n,L\to+\infty]{a.s.} 0 \qquad \textit{and} \qquad \frac{1}{n}\operatorname{Tr}\boldsymbol{A}\big(\widetilde{\boldsymbol{Q}} - \mathbb{E}[\widetilde{\boldsymbol{Q}}]\big) \xrightarrow[p,n,L\to+\infty]{a.s.} 0.$$

for all bounded (sequences of) vectors $\boldsymbol{a},\boldsymbol{b} \in \mathbb{R}^n$ and matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

Proof. Let us start this proof with a "two-in-one" computation. Consider a matrix $\boldsymbol{B}$ which will either be $\boldsymbol{b}\boldsymbol{a}^\top$ or $\frac{1}{n}\boldsymbol{A}$. We can bound from above the variance of $\operatorname{Tr}\boldsymbol{B}\widetilde{\boldsymbol{Q}}$ using the Poincaré-Nash inequality (Lemma 2.19), Stein's lemma (Lemma 2.18) and Lemma 4.21.

$$\begin{aligned}
\operatorname{Var}\big(\operatorname{Tr}\boldsymbol{B}\widetilde{\boldsymbol{Q}}\big) &\overset{\text{(Lemma 2.19)}}{\leqslant} \sum_{i=1}^{p}\sum_{j=1}^{n}\mathbb{E}\left[\left|\frac{\partial \operatorname{Tr}\boldsymbol{B}\boldsymbol{Q}}{\partial N_{i,j}}\right|^2\right] \\
&\overset{\text{(Lemma 2.18)}}{=} \sum_{i=1}^{p}\sum_{j=1}^{n}\mathbb{E}\left[\left|\sum_{r=1}^{n}\sum_{s=1}^{n} B_{r,s}\frac{\partial \widetilde{Q}_{s,r}}{\partial N_{i,j}}\right|^2\right] \\
&\overset{\text{(Lemma 4.21)}}{=} \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{n}\mathbb{E}\left[\left|\sum_{r=1}^{n}\sum_{s=1}^{n} B_{r,s}\sum_{t=1}^{n}\big(\widetilde{Q}_{s,j}X_{i,t}C_{j,t}\widetilde{Q}_{t,r} + \widetilde{Q}_{s,t}X_{i,t}C_{t,j}\widetilde{Q}_{j,r}\big)\right|^2\right] \\
&= \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{n}\mathbb{E}\left[\left|\big[\boldsymbol{X}(\widetilde{\boldsymbol{Q}}(\boldsymbol{B}+\boldsymbol{B}^\top)\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C})\big]_{i,j}\right|^2\right] \\
&= \frac{1}{p^2}\mathbb{E}\Big[\big\|\boldsymbol{X}(\widetilde{\boldsymbol{Q}}(\boldsymbol{B}+\boldsymbol{B}^\top)\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C})\big\|_{\mathrm{F}}^2\Big].
\end{aligned}$$

Then, we use the inequalities $(a+b)^2 \leqslant 2(a^2+b^2)$ and $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}} \leqslant \|\boldsymbol{A}\|\|\boldsymbol{B}\|_{\mathrm{F}}$,

$$\begin{aligned}
\operatorname{Var}\big(\operatorname{Tr}\boldsymbol{B}\widetilde{\boldsymbol{Q}}\big) &\leqslant \frac{2}{p^2}\mathbb{E}\Big[\big\|\boldsymbol{X}(\widetilde{\boldsymbol{Q}}\boldsymbol{B}\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C})\big\|_{\mathrm{F}}^2 + \big\|\boldsymbol{X}(\widetilde{\boldsymbol{Q}}\boldsymbol{B}^\top\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C})\big\|_{\mathrm{F}}^2\Big] \\
&\leqslant \frac{2}{p^2}\mathbb{E}\Big[\|\boldsymbol{X}\|^2\big\|\widetilde{\boldsymbol{Q}}\boldsymbol{B}\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C}\big\|_{\mathrm{F}}^2 + \|\boldsymbol{X}\|^2\big\|\widetilde{\boldsymbol{Q}}\boldsymbol{B}^\top\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C}\big\|_{\mathrm{F}}^2\Big] \\
&\leqslant \frac{2}{p^2}\mathbb{E}\Big[\|\boldsymbol{X}\|^2\big\|\widetilde{\boldsymbol{Q}}\boldsymbol{B}\widetilde{\boldsymbol{Q}}\big\|_{\mathrm{F}}^2 + \|\boldsymbol{X}\|^2\big\|\widetilde{\boldsymbol{Q}}\boldsymbol{B}^\top\widetilde{\boldsymbol{Q}}\big\|_{\mathrm{F}}^2\Big] \leqslant \frac{4}{p^2}\|\boldsymbol{B}\|_{\mathrm{F}}^2\mathbb{E}\Big[\|\boldsymbol{X}\|^2\big\|\widetilde{\boldsymbol{Q}}\big\|^4\Big].
\end{aligned}$$

Since $\|\boldsymbol{A}\| \leqslant \sqrt{n}\|\boldsymbol{A}\|_{\mathrm{F}}$ and $\|\boldsymbol{b}\boldsymbol{a}^\top\|_{\mathrm{F}} = \|\boldsymbol{a}\|\|\boldsymbol{b}\|$, we directly see that

$$\operatorname{Var}\left(\frac{1}{n}\operatorname{Tr}\boldsymbol{A}\widetilde{\boldsymbol{Q}}\right) \leqslant \frac{4}{np^2}\|\boldsymbol{A}\|^2\mathbb{E}\Big[\|\boldsymbol{X}\|^2\big\|\widetilde{\boldsymbol{Q}}\big\|^4\Big] = \mathcal{O}_z(n^{-2})$$

$$\text{and} \qquad \operatorname{Var}\big(\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{b}\big) \leqslant \frac{4}{p^2}\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\Big[\|\boldsymbol{X}\|^2\big\|\widetilde{\boldsymbol{Q}}\big\|^4\Big] = \mathcal{O}_z(n^{-1}).$$

So we can already state that $\frac{1}{n}\operatorname{Tr}\boldsymbol{A}(\widetilde{\boldsymbol{Q}} - \mathbb{E}[\widetilde{\boldsymbol{Q}}]) \to 0$ almost surely as per Lemma 2.20. In order to show the concentration of bilinear forms, observe that

$$\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\big|^4\Big] = \mathrm{Var}\Big(\big(\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\big)^2\Big) + \Big|\mathbb{E}\Big[\big(\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\big)^2\Big]\Big|^2$$
$$\leqslant \mathrm{Var}\Big(\big(\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\big)^2\Big) + \underbrace{\mathrm{Var}\big(\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{b}\big)^2}_{=\mathscr{O}_z(n^{-2})}$$

and we can likewise bound from above the variance of $(\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b})^2$ with the Poincaré-Nash inequality (Lemma 2.19),

$$\begin{aligned}
\mathrm{Var}\Big(\big(\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\big)^2\Big) &\leqslant \sum_{i=1}^p\sum_{i=1}^n \mathbb{E}\left[\left|2\big(\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\big)\frac{\partial \boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{b}}{\partial N_{i,j}}\right|^2\right] \\
&= \frac{4}{p^2}\sum_{i=1}^p\sum_{i=1}^n \mathbb{E}\Big[\big|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\big|^2\big|\big[\boldsymbol{X}(\widetilde{\boldsymbol{Q}}(\boldsymbol{a}\boldsymbol{b}^\top+\boldsymbol{b}\boldsymbol{a}^\top)\widetilde{\boldsymbol{Q}}\odot\boldsymbol{C})\big]_{i,j}\big|^2\Big] \\
&= \frac{4}{p^2}\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\big|^2\big\|\boldsymbol{X}(\widetilde{\boldsymbol{Q}}(\boldsymbol{a}\boldsymbol{b}^\top+\boldsymbol{b}\boldsymbol{a}^\top)\widetilde{\boldsymbol{Q}}\odot\boldsymbol{C})\big\|_{\mathrm{F}}^2\Big] \\
&\leqslant \frac{16}{p^2}\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\big|^2\|\boldsymbol{X}\|^2\big\|\widetilde{\boldsymbol{Q}}\big\|^4\Big] = \mathscr{O}_z(n^{-2}).
\end{aligned}$$

Hence, $\mathbb{E}[|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}|^4] = \mathscr{O}_z(n^{-2})$ and the almost sure convergence $\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}-\mathbb{E}[\widetilde{\boldsymbol{Q}}])\boldsymbol{b}\to 0$ is given by Lemma 2.20. □

Remark 4.23. In the proof of Proposition 4.22, we have shown that

$$\mathrm{Var}\big(\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{b}\big) = \mathscr{O}_z(n^{-1}) \qquad \text{and} \qquad \mathrm{Var}\left(\frac{1}{n}\operatorname{Tr}\boldsymbol{A}\widetilde{\boldsymbol{Q}}\right) = \mathscr{O}_z(n^{-2}).$$

Variance of a Particular Quantity

We also state here a result on the variance of a specific quantity which will be useful in the upcoming analysis.

Proposition 4.24. $\mathrm{Var}\Big(\Big[\frac{1}{p}\boldsymbol{C}\Big(\widetilde{\boldsymbol{Q}}\odot\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\Big)\boldsymbol{C}\Big]_{r,s}\Big) = \mathscr{O}_z(n^{-2})$ *for all* $r,s\in[n]$.

Proof. We use the same technique as in the proof of Proposition 4.22 with the Poincaré-Nash inequality (Lemma 2.19), Stein's lemma (Lemma 2.18) and Lemma 4.21.

$$\begin{aligned}
&\mathrm{Var}\left(\left[\frac{1}{p}\boldsymbol{C}\Big(\widetilde{\boldsymbol{Q}}\odot\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\Big)\boldsymbol{C}\right]_{r,s}\right) \\
&\leqslant \sum_{i=1}^p\sum_{j=1}^n \mathbb{E}\left[\left|\frac{\partial}{\partial N_{i,j}}\left[\frac{1}{p}\boldsymbol{C}\Big(\widetilde{\boldsymbol{Q}}\odot\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\Big)\boldsymbol{C}\right]_{r,s}\right|^2\right] \\
&= \frac{1}{p^4}\sum_{i=1}^p\sum_{j=1}^n \mathbb{E}\left[\left|\frac{\partial}{\partial N_{i,j}}\sum_{t=1}^n\sum_{u=1}^n\sum_{v=1}^p C_{r,t}\widetilde{Q}_{t,u}X_{v,t}X_{v,u}C_{u,s}\right|^2\right] \\
&= \frac{1}{p^4}\sum_{i=1}^p\sum_{j=1}^n \mathbb{E}\left[\left|\sum_{t=1}^n\sum_{u=1}^n\sum_{v=1}^p C_{r,t}\frac{\partial\widetilde{Q}_{t,u}}{\partial N_{i,j}}X_{v,t}X_{v,u}C_{u,s} + \sum_{u=1}^n C_{r,j}\widetilde{Q}_{j,u}X_{i,u}C_{u,s} + \sum_{t=1}^n C_{r,t}\widetilde{Q}_{t,j}X_{i,t}C_{j,s}\right|^2\right].
\end{aligned}$$

Let us deal with each term inside $\mathbb{E}[|\cdot|^2]$ separately.

$$\begin{aligned}
&\sum_{t=1}^{n}\sum_{u=1}^{n}\sum_{v=1}^{p} C_{r,t}\frac{\partial \widetilde{Q}_{t,u}}{\partial N_{i,j}} X_{v,t}X_{v,u}C_{u,s} \\
&\qquad = -\frac{1}{p}\sum_{t=1}^{n}\sum_{u=1}^{n}\sum_{v=1}^{p} C_{r,t}\sum_{w=1}^{n}\left(\widetilde{Q}_{t,j}X_{i,w}C_{j,w}\widetilde{Q}_{w,u} + \widetilde{Q}_{t,w}X_{i,w}C_{w,j}\widetilde{Q}_{j,u}\right)X_{v,t}X_{v,u}C_{u,s} \\
&\qquad = -\frac{1}{p}\left[\boldsymbol{X}\left(\widetilde{\boldsymbol{Q}}\left(\boldsymbol{X}^\top\boldsymbol{X}\odot \boldsymbol{C}_{\cdot,s}\boldsymbol{C}_{r,\cdot} + \boldsymbol{X}^\top\boldsymbol{X}\odot \boldsymbol{C}_{\cdot,r}\boldsymbol{C}_{s,\cdot}\right)\widetilde{\boldsymbol{Q}}\odot \boldsymbol{C}\right)\right]_{i,j} \\
\text{and}\qquad &\sum_{u=1}^{n} C_{r,j}\widetilde{Q}_{j,u}X_{i,u}C_{u,s} + \sum_{t=1}^{n} C_{r,t}\widetilde{Q}_{t,j}X_{i,t}C_{j,s} = \left[\boldsymbol{X}(\widetilde{\boldsymbol{Q}}\odot \boldsymbol{C}_{\cdot,s}\boldsymbol{C}_{r,\cdot}) + \boldsymbol{X}(\widetilde{\boldsymbol{Q}}\odot \boldsymbol{C}_{\cdot,r}\boldsymbol{C}_{s,\cdot})\right]_{i,j}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathrm{Var}\left(\left[\frac{1}{p}\boldsymbol{C}\left(\widetilde{\boldsymbol{Q}}\odot\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\right)\boldsymbol{C}\right]_{r,s}\right) \\
&\leqslant \frac{1}{p^4}\mathbb{E}\left[\left\|-\boldsymbol{X}\left(\widetilde{\boldsymbol{Q}}\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot \boldsymbol{C}_{\cdot,s}\boldsymbol{C}_{r,\cdot} + \frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot \boldsymbol{C}_{\cdot,r}\boldsymbol{C}_{s,\cdot}\right)\widetilde{\boldsymbol{Q}}\odot \boldsymbol{C}\right) + \boldsymbol{X}(\widetilde{\boldsymbol{Q}}\odot \boldsymbol{C}_{\cdot,s}\boldsymbol{C}_{r,\cdot}) + \boldsymbol{X}(\widetilde{\boldsymbol{Q}}\odot \boldsymbol{C}_{\cdot,r}\boldsymbol{C}_{s,\cdot})\right\|_{\mathrm{F}}^2\right] \\
&\leqslant \frac{2}{p^4}\mathbb{E}\left[\|\boldsymbol{X}\|^2\left(\left\|\widetilde{\boldsymbol{Q}}\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot \boldsymbol{C}_{\cdot,s}\boldsymbol{C}_{r,\cdot} + \frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot \boldsymbol{C}_{\cdot,r}\boldsymbol{C}_{s,\cdot}\right)\widetilde{\boldsymbol{Q}}\odot \boldsymbol{C}\right\|_{\mathrm{F}}^2 + \left\|\widetilde{\boldsymbol{Q}}\odot \boldsymbol{C}_{\cdot,s}\boldsymbol{C}_{r,\cdot} + \widetilde{\boldsymbol{Q}}\odot \boldsymbol{C}_{\cdot,r}\boldsymbol{C}_{s,\cdot}\right\|_{\mathrm{F}}^2\right)\right] \\
&\leqslant \frac{2}{p^4}\mathbb{E}\left[\|\boldsymbol{X}\|^2\left(\left\|\widetilde{\boldsymbol{Q}}\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot \boldsymbol{C}_{\cdot,s}\boldsymbol{C}_{r,\cdot} + \frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot \boldsymbol{C}_{\cdot,r}\boldsymbol{C}_{s,\cdot}\right)\widetilde{\boldsymbol{Q}}\right\|_{\mathrm{F}}^2 + 2\left(\left\|\widetilde{\boldsymbol{Q}}\right\|_{\mathrm{F}}^2 + \left\|\widetilde{\boldsymbol{Q}}\right\|_{\mathrm{F}}^2\right)\right)\right] \\
&\leqslant \frac{2}{p^4}\mathbb{E}\left[\|\boldsymbol{X}\|^2\left(2\left\|\widetilde{\boldsymbol{Q}}\right\|^4\left(\left\|\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\right\|_{\mathrm{F}}^2 + \left\|\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\right\|_{\mathrm{F}}^2\right) + 4n\left\|\widetilde{\boldsymbol{Q}}\right\|^2\right)\right] \\
&\leqslant \frac{8n}{p^4}\mathbb{E}\left[\|\boldsymbol{X}\|^2\left(\left\|\widetilde{\boldsymbol{Q}}\right\|^4\left\|\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\right\|^2 + \left\|\widetilde{\boldsymbol{Q}}\right\|^2\right)\right] = \mathscr{O}_z(n^{-2}).
\end{aligned}$$

□

4.A.4 Limiting Spectral Distribution

In this section, we study the limiting behavior of the empirical spectral distribution of $\frac{\boldsymbol{X}^\top\boldsymbol{X}}{p}\odot\boldsymbol{C}$ via its Stieltjes transform, which is the normalized trace of its resolvent (Proposition 2.6).

Let us define the resolvent of the "noise part", $\widetilde{\boldsymbol{Q}}_0(z) \overset{\text{def}}{=} (\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\odot\boldsymbol{C} - z\boldsymbol{I}_n)^{-1}$ for all $z\in\mathbb{C}\setminus\mathrm{Sp}(\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\odot\boldsymbol{C})$. The following proposition shows that, in order to derive the limiting spectral distribution, we can study the simpler resolvent $\widetilde{\boldsymbol{Q}}_0$. In other words, the addition of a signal $\boldsymbol{P}$ does not change the limiting spectral distribution and so we can neglect it.

Proposition 4.25. $\frac{1}{n}\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0 - \frac{1}{n}\operatorname{Tr}\widetilde{\boldsymbol{Q}} \to 0$ *almost surely as* $p,n,L\to+\infty$.

Proof. Let $\boldsymbol{S} = \frac{\boldsymbol{P}^\top\boldsymbol{P}}{p} + \frac{\boldsymbol{P}^\top\boldsymbol{N}}{p} + \frac{\boldsymbol{N}^\top\boldsymbol{P}}{p}$. Note that the rank of $\boldsymbol{S}$ is at most $3K = \mathscr{O}(1)$.

$$\begin{aligned}
\left|\frac{1}{n}\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0 - \frac{1}{n}\operatorname{Tr}\widetilde{\boldsymbol{Q}}\right| &= \frac{1}{n}\left|\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0(\boldsymbol{S}\odot\boldsymbol{C})\widetilde{\boldsymbol{Q}}\right| \qquad \text{from the resolvent identity (Proposition 2.21)} \\
&\leqslant \frac{1}{n}\sum_{i=1}^{n} s_i(\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{Q}}_0)\,s_i(\boldsymbol{S}\odot\boldsymbol{C}) \qquad \text{from Propositions 4.13 and 4.14}
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \frac{1}{n}\|\widetilde{\boldsymbol{Q}}\|\|\widetilde{\boldsymbol{Q}}_0\|\sum_{i=1}^{n} s_i(\boldsymbol{S}\odot\boldsymbol{C}) \\
&\leqslant \frac{1}{n}\|\widetilde{\boldsymbol{Q}}\|\|\widetilde{\boldsymbol{Q}}_0\|\sqrt{n}\sqrt{\sum_{i=1}^{n} s_i^2(\boldsymbol{S}\odot\boldsymbol{C})} \qquad \text{from the Cauchy-Schwarz inequality} \\
&\leqslant \frac{1}{\sqrt{n}}\|\widetilde{\boldsymbol{Q}}\|\|\widetilde{\boldsymbol{Q}}_0\|\sqrt{\sum_{i=1}^{n} s_i^2(\boldsymbol{S})} \qquad \text{since } \|\boldsymbol{S}\odot\boldsymbol{C}\|_{\mathrm{F}} \leqslant \|\boldsymbol{S}\|_{\mathrm{F}} \\
&\leqslant \sqrt{\frac{3K}{n}}\|\widetilde{\boldsymbol{Q}}\|\|\widetilde{\boldsymbol{Q}}_0\|\|\boldsymbol{S}\| \qquad \text{since } \operatorname{Rank}\boldsymbol{S} \leqslant 3K.
\end{aligned}
$$

Hence, $|\frac{1}{n}\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0 - \frac{1}{n}\operatorname{Tr}\widetilde{\boldsymbol{Q}}| = \mathcal{O}_z(n^{-1/2})$. □

Since $\widetilde{\boldsymbol{Q}}_0^{-1}\widetilde{\boldsymbol{Q}}_0 = \boldsymbol{I}_n$, we have

$$
\mathbb{E}\left[\left(\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\odot\boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}_0\right] = z\mathbb{E}\big[\widetilde{\boldsymbol{Q}}_0\big] + \boldsymbol{I}_n
$$

and, by Equation (4.3) when $\boldsymbol{A} = \boldsymbol{C}$ and $\boldsymbol{P} = \boldsymbol{0}_{p\times n}$, this becomes

$$
\mathbb{E}\big[(\boldsymbol{C}\odot\boldsymbol{I}_n)\widetilde{\boldsymbol{Q}}_0\big] - \mathbb{E}\big[(\boldsymbol{H}_0\odot\boldsymbol{I}_n)\widetilde{\boldsymbol{Q}}_0\big] - \mathbb{E}\left[\frac{1}{p}\left((\widetilde{\boldsymbol{Q}}_0\odot\boldsymbol{C})\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\odot\boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}_0\right] = z\mathbb{E}\big[\widetilde{\boldsymbol{Q}}_0\big] + \boldsymbol{I}_n
$$

where $\boldsymbol{H}_0 \overset{\text{def}}{=} \frac{1}{p}\boldsymbol{C}(\widetilde{\boldsymbol{Q}}_0\odot\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p})\boldsymbol{C}$. In the left-hand side, $\boldsymbol{C}\odot\boldsymbol{I}_n = \boldsymbol{I}_n$ and $\|\mathbb{E}[\frac{1}{p}((\widetilde{\boldsymbol{Q}}_0\odot\boldsymbol{C})\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\odot\boldsymbol{C})\widetilde{\boldsymbol{Q}}_0]\| = \mathcal{O}_z(n^{-1}(\log n)^2)$ by Proposition 4.17. Moreover, $\mathbb{E}[(\boldsymbol{H}_0\odot\boldsymbol{I}_n)\widetilde{\boldsymbol{Q}}_0]_{i,j} = \mathbb{E}[[\boldsymbol{H}_0]_{i,i}[\widetilde{\boldsymbol{Q}}_0]_{i,j}] = \mathbb{E}[\boldsymbol{H}_0]_{i,i}\mathbb{E}[\widetilde{\boldsymbol{Q}}_0]_{i,j} + \operatorname{Cov}([\boldsymbol{H}_0]_{i,i},[\widetilde{\boldsymbol{Q}}_0]_{i,j})$ and $|\operatorname{Cov}([\boldsymbol{H}_0]_{i,i},[\widetilde{\boldsymbol{Q}}_0]_{i,j})| \leqslant \sqrt{\operatorname{Var}([\boldsymbol{H}_0]_{i,i})\operatorname{Var}([\widetilde{\boldsymbol{Q}}_0]_{i,j})} = \mathcal{O}_z(n^{-3/2})$ by the Cauchy-Schwarz inequality and Proposition 4.24 (see also Remark 4.23 for the variance of $[\widetilde{\boldsymbol{Q}}_0]_{i,j}$). Hence $\mathbb{E}[(\boldsymbol{H}_0\odot\boldsymbol{I}_n)\widetilde{\boldsymbol{Q}}_0]_{i,j} = \mathbb{E}[\boldsymbol{H}_0\odot\boldsymbol{I}_n]\mathbb{E}[\widetilde{\boldsymbol{Q}}_0] + \mathcal{O}_z^{\|\cdot\|_{\max}}(n^{-3/2})$ where $\mathcal{O}_z^{\|\cdot\|_{\max}}(n^{-3/2})$ is a matrix whose entries are $\mathcal{O}_z(n^{-3/2})$. Consequently,

$$
\mathbb{E}[(\boldsymbol{H}_0\odot\boldsymbol{I}_n)]\mathbb{E}\big[\widetilde{\boldsymbol{Q}}_0\big] + (z-1)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}_0\big] + \boldsymbol{I}_n = \mathcal{O}_z^{\|\cdot\|_{\max}}(n^{-3/2}) + \mathcal{O}_z^{\|\cdot\|}(n^{-1}(\log n)^2)
$$

where $\mathcal{O}_z^{\|\cdot\|}(n^{-1}(\log n)^2)$ is a matrix whose spectral norm $\|\cdot\|$ is $\mathcal{O}_z(n^{-1}(\log n)^2)$. In particular,

$$
\big(\mathbb{E}[\boldsymbol{H}_0]_{i,i} + z - 1\big)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}_0\big]_{i,i} + 1 = \mathcal{O}_z(n^{-1}(\log n)^2) \qquad \text{for all } i\in[n]. \tag{4.5}
$$

Proposition 4.26. *The diagonal entries of $\widetilde{\boldsymbol{Q}}_0$ are identically distributed.*

Proof. Consider the permutation matrix

$$
\boldsymbol{R} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 \\ 1 & 0 & \dots & \dots & 0 \end{bmatrix}.
$$

Notice that $\boldsymbol{R}(\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\odot\boldsymbol{C})\boldsymbol{R}^\top = \boldsymbol{R}\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\boldsymbol{R}^\top\odot\boldsymbol{C}$ since the action of $\boldsymbol{R}[\cdot]\boldsymbol{R}^\top$ is to shift circularly all the entries of the matrix towards the upper left therefore it preserves the diagonal structure of $\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\odot\boldsymbol{C}$. Moreover, $\left[\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\right]_{i,j} = \frac{1}{p}\sum_{k=1}^{p} N_{k,i}N_{k,j}$ is identically distributed to $\left[\frac{\boldsymbol{N}^\top\boldsymbol{N}}{p}\right]_{[(i-d-1)\bmod n]+1,[(j-d-1)\bmod n]+1} =$

$\left[\boldsymbol{R}^d \frac{\boldsymbol{N}^\top \boldsymbol{N}}{p} \boldsymbol{R}^{\top d}\right]_{i,j}$. Thus, $\widetilde{\boldsymbol{Q}}_0$ is identically distributed to $\boldsymbol{R}^d \widetilde{\boldsymbol{Q}}_0 \boldsymbol{R}^{\top d}$ and, in particular, $[\widetilde{\boldsymbol{Q}}_0]_{1,1}$ is identically distributed to $[\widetilde{\boldsymbol{Q}}_0]_{d,d}$ for all $d \in [n]$. □

According to Proposition 4.26, $\mathbb{E}[\widetilde{\boldsymbol{Q}}_{i,i}] = \frac{1}{n}\operatorname{Tr}\mathbb{E}[\widetilde{\boldsymbol{Q}}_0] = \mathbb{E}[\widetilde{m}_n(z)]$ for all $i \in [n]$, where $\widetilde{m}_n(z) \overset{\text{def}}{=} \frac{1}{n}\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0$ is the Stieltjes transform of the empirical spectral distribution of $\frac{\boldsymbol{N}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C}$. Recall that $\mathbb{E}[\widetilde{m}_n(z)]$ is also a Stieltjes transform according to Corollary 2.9. Hence, Equation (4.5) becomes

$$\big(\eta_0(z) + z - 1\big)\mathbb{E}[\widetilde{m}_n(z)] + 1 = 0 \tag{4.6}$$

where $\eta_0(z)$ is such that $\mathbb{E}[\boldsymbol{H}_0]_{i,i} = \eta_0(z) + \mathcal{O}_z(n^{-1}(\log n)^2)$ for all $i \in [n]$.

We will now study the entries of $\boldsymbol{H}$ to identify the quantity $\eta_0(z)$. Thanks to the relation $[\frac{1}{p}\boldsymbol{A}(\widetilde{\boldsymbol{Q}} \odot \frac{\boldsymbol{N}^\top \boldsymbol{N}}{p})\boldsymbol{C}]_{i,j} = \frac{1}{p}\operatorname{Tr}(\widetilde{\boldsymbol{Q}}(\frac{\boldsymbol{N}^\top \boldsymbol{N}}{p} \odot [\boldsymbol{A}_{i,\cdot}^\top \boldsymbol{C}_{j,\cdot}]))$, we can use Equation (4.3) in the expression of $\mathbb{E}[\boldsymbol{H}_0]_{i,j}$.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{H}_0]_{i,j} &= \frac{1}{p}\operatorname{Tr}\mathbb{E}\left[\left(\frac{\boldsymbol{N}^\top \boldsymbol{N}}{p} \odot [\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}]\right)\widetilde{\boldsymbol{Q}}_0\right] \\
&= \frac{1}{p}\sum_{k=1}^n \mathbb{E}\left[\left(\frac{\boldsymbol{N}^\top \boldsymbol{N}}{p} \odot [\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}]\right)\widetilde{\boldsymbol{Q}}_0\right]_{k,k} \\
&= \frac{1}{p}\sum_{k=1}^n \mathbb{E}\left[\left([\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}] \odot \boldsymbol{I}_n - \frac{1}{p}[\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}]\left(\widetilde{\boldsymbol{Q}}_0 \odot \frac{\boldsymbol{N}^\top \boldsymbol{N}}{p}\right)\boldsymbol{C} \odot \boldsymbol{I}_n - \frac{1}{p}(\widetilde{\boldsymbol{Q}}_0 \odot [\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}])\frac{\boldsymbol{N}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}_0\right]_{k,k} \\
&= \frac{1}{p}\sum_{k=1}^n \mathbb{E}\left[C_{k,i}C_{j,k}[\widetilde{\boldsymbol{Q}}_0]_{k,k} - \frac{1}{p}\operatorname{Tr}\left(\widetilde{\boldsymbol{Q}}_0\left(\frac{\boldsymbol{N}^\top \boldsymbol{N}}{p} \odot [[\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}]_{k,\cdot}^\top \boldsymbol{C}_{k,\cdot}]\right)\right)[\widetilde{\boldsymbol{Q}}_0]_{k,k}\right] + \mathcal{O}_z(n^{-1}\log n) \\
&= \frac{1}{p}\sum_{k=1}^n (C_{k,i} - [\boldsymbol{H}_0]_{i,k})\mathbb{E}\big[C_{j,k}[\widetilde{\boldsymbol{Q}}_0]_{k,k}\big] + \mathcal{O}_z(n^{-1}\log n) \\
&= \frac{1}{p}\sum_{k=1}^n (C_{k,i} - \mathbb{E}[\boldsymbol{H}_0]_{i,k})C_{j,k}\mathbb{E}\big[\widetilde{\boldsymbol{Q}}_0\big]_{k,k} + \mathcal{O}_z(n^{-1}\log n)
\end{aligned}$$

where the $\mathcal{O}_z(n^{-1}\log n)$ stems from the inequalities given in Propositions 4.17 and 4.19 and, in the last equality, we have used the fact that $\mathbb{E}[[\boldsymbol{H}_0]_{i,k}[\widetilde{\boldsymbol{Q}}_0]_{k,k}] = \mathbb{E}[\boldsymbol{H}_0]_{i,k}\mathbb{E}[\widetilde{\boldsymbol{Q}}_0]_{k,k} + \mathcal{O}_z(n^{-3/2})$ as proven above. Since $\mathbb{E}[\widetilde{\boldsymbol{Q}}_0]_{k,k} = \mathbb{E}[\widetilde{m}_n(z)]$ for all $k \in [n]$, we have $\mathbb{E}[\boldsymbol{H}_0] = \mathbb{E}[\widetilde{m}_n(z)](\boldsymbol{C} - \mathbb{E}[\boldsymbol{H}_0])\frac{\boldsymbol{C}}{p} + \boldsymbol{E}$ where $\boldsymbol{E} = \mathcal{O}_z^{\|\cdot\|_{\max}}(n^{-1}\log n)$. Let $\bar{\boldsymbol{H}}$ be the matrix such that

$$\bar{\boldsymbol{H}} = \mathbb{E}[\widetilde{m}_n(z)](\boldsymbol{C} - \bar{\boldsymbol{H}})\frac{\boldsymbol{C}}{p} \quad \Longleftrightarrow \quad \bar{\boldsymbol{H}}\left(\boldsymbol{I}_n + \mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right) = \mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}^2}{p}.$$

The existence and uniqueness of such a $\bar{\boldsymbol{H}}$ is guaranteed if $\boldsymbol{I}_n + \mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}$ is invertible, which is always the case if $z \in \mathbb{C} \setminus \mathbb{R}$ since $\Im z \neq 0 \implies 1 + \mathbb{E}[\widetilde{m}_n(z)]\frac{\psi_k}{p} \neq 0$. Then,

$$\begin{aligned}
\bar{\boldsymbol{H}} &= \mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}^2}{p}\left(\boldsymbol{I}_n + \mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right)^{-1} \\
&= \boldsymbol{C}\left(\boldsymbol{I}_n + \mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p} - \boldsymbol{I}_n\right)\left(\boldsymbol{I}_n + \mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right)^{-1} \\
&= \boldsymbol{C} - \boldsymbol{C}\left(\boldsymbol{I}_n + \mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right)^{-1}.
\end{aligned}$$

As $\mathbb{E}[\boldsymbol{H}_0] - \bar{\boldsymbol{H}} = -\mathbb{E}[\widetilde{m}_n(z)](\mathbb{E}[\boldsymbol{H}_0] - \bar{\boldsymbol{H}})\frac{\boldsymbol{C}}{p} + \boldsymbol{E}$, we have

$$\left|\left[\mathbb{E}[\boldsymbol{H}_0]-\bar{\boldsymbol{H}}\right]_{i,j}\right| = \left|\left[\boldsymbol{E}\left(\boldsymbol{I}_n+\mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right)^{-1}\right]_{i,j}\right| = \left|\sum_{k=1}^{n} E_{i,k}\left[\boldsymbol{I}_n+\mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right]^{-1}_{k,j}\right|$$
$$\leqslant \kappa\frac{\log n}{n}\left\|\left(\boldsymbol{I}_n+\mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right)^{-1}\right\|_1 = \kappa\frac{\log n}{n}\left\|\left(\boldsymbol{I}_n+\mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right)^{-1}\right\|_\infty$$

where $\kappa>0$ is a constant and $\|\cdot\|_1, \|\cdot\|_\infty$ are the induced matrix 1-norm and ∞-norm. The last equality stems from the fact that $\boldsymbol{I}_n+\mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}$ is symmetric. Moreover, from the inequality $\|\cdot\|_\infty \leqslant \sqrt{n}\|\cdot\|$ between ∞-norm and spectral norm,

$$\left\|\left(\boldsymbol{I}_n+\mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right)^{-1}\right\|_\infty \leqslant \sqrt{n}\left\|\left(\boldsymbol{I}_n+\mathbb{E}[\widetilde{m}_n(z)]\frac{\boldsymbol{C}}{p}\right)^{-1}\right\| = \sqrt{n}\max_{0\leqslant k\leqslant n-1}\left|1+\mathbb{E}[\widetilde{m}_n(z)]\frac{\psi_k}{p}\right|^{-1} = \mathcal{O}_z(\sqrt{n})$$

thus $\mathbb{E}[\boldsymbol{H}_0]_{i,j} = \bar{H}_{i,j}+\mathcal{O}_z(n^{-1/2}\log n)$. Hence,

$$\eta_0(z) = \bar{H}_{i,i}+\mathcal{O}_z(n^{-1/2}\log n) \quad \text{with} \quad \bar{H}_{i,i} = \frac{1}{n}\operatorname{Tr}\bar{\boldsymbol{H}} = 1-\sum_{k=0}^{n-1}\frac{\psi_k}{1+\mathbb{E}[\widetilde{m}_n(z)]\frac{\psi_k}{p}}$$

and Equation (4.6) becomes

$$1+z\mathbb{E}[\widetilde{m}_n(z)]-\frac{p}{n}\sum_{k=0}^{n-1}\frac{\mathbb{E}[\widetilde{m}_n(z)]\frac{\psi_k}{p}}{1+\mathbb{E}[\widetilde{m}_n(z)]\frac{\psi_k}{p}} = \mathcal{O}_z(n^{-1/2}\log n). \tag{4.7}$$

We will now show that there is a *deterministic* Stieltjes transform $\bar{m}_n$ such that $|\widetilde{m}_n(z)-\bar{m}_n(z)|\to 0$ almost surely as $p,n,L\to+\infty$. Let T_n be the operator acting on the set of functions of a complex variable such that

$$T_n[m](z) = \frac{1}{\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+m(z)\frac{\psi_k}{p}}-z}.$$

Proposition 4.27. *If m is the Stieltjes transform of a finite real measure, then $T_n[m]$ is the Stieltjes transform of a real probability measure.*

Proof. We prove this result with Theorem 2.8. As the composition of analytic functions, $T_n[m]$ is analytic on $\mathbb{C}^+$. Since $m(\bar{z}) = \overline{m(z)}$, we also have $T_n[m](\bar{z}) = \overline{T_n[m](z)}$. Moreover, if $\Im z>0$, then

$$\Im T_n[m] = \Im\left[\frac{\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k\left(1+m(z)\frac{\psi_k}{p}\right)}{\left|1+m(z)\frac{\psi_k}{p}\right|^2}-\bar{z}}{\left|\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+m(z)\frac{\psi_k}{p}}-z\right|^2}\right] = \frac{\frac{1}{n}\sum_{k=0}^{n-1}\frac{\Im[m(z)]\frac{\psi_k^2}{p}}{\left|1+m(z)\frac{\psi_k}{p}\right|^2}+\Im[z]}{\left|\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+m(z)\frac{\psi_k}{p}}-z\right|^2} > 0$$

and

$$|T_n[m](z)| = \frac{1}{\left|\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+m(z)\frac{\psi_k}{p}}-z\right|} \leqslant \frac{1}{\left|\Im\left[\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+m(z)\frac{\psi_k}{p}}-z\right]\right|} \leqslant \frac{1}{|\Im z|}$$

since $\Im\left[\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+m(z)\frac{\psi_k}{p}}\right] = \frac{1}{n}\sum_{k=0}^{n-1}\frac{\Im[\overline{m(z)}]\frac{\psi_k^2}{p}}{\left|1+m(z)\frac{\psi_k}{p}\right|^2} < 0$. Finally,

$$-\mathrm{i}yT_n[m](\mathrm{i}y) = \frac{-\mathrm{i}y}{\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+m(\mathrm{i}y)\frac{\psi_k}{p}}-\mathrm{i}y} \xrightarrow[y\to+\infty]{} 1.$$

because $m(\mathrm{i}y)\to 0$ as $y\to+\infty$ and $\frac{1}{n}\sum_{k=0}^{n}\psi_k = 1$. The assumptions of Theorem 2.8 are thus verified and $T_n[m]$ is the Stieltjes transform of unique probability measure on $\mathbb{R}$. □

As a consequence to Proposition 4.27, Equation (4.7) is equivalent to $\mathbb{E}[\tilde{m}_n](z) = T_n[\mathbb{E}[\tilde{m}_n]](z) + \mathcal{O}_z(n^{-1/2}\log n)$. Moreover, we have the following result.

Proposition 4.28. *There exists a unique real probability measure $\bar{\mu}_n$ whose Stieltjes transform $\bar{m}_n$ satisfies $T_n[\bar{m}_n] = \bar{m}_n$.*

Proof. This demonstration is in three steps: (i) we show the existence of a function $\bar{m}_n$ such that $\bar{m}_n = T_n[\bar{m}_n]$, (ii) we establish that $\bar{m}_n$ is a Stieltjes transform of a real probability measure, (iii) we prove its uniqueness.

(i) Existence. For any given positive integer $q \geqslant 1$, define the alternative mask $\mathfrak{C}_q \overset{\text{def}}{=} \mathbf{1}_{q\times q} \boxtimes \boldsymbol{C}$ where $\boxtimes$ denotes the Kronecker product, i.e., $\mathfrak{C}_q = \begin{bmatrix} \boldsymbol{C} & \dots & \boldsymbol{C} \\ \vdots & \ddots & \vdots \\ \boldsymbol{C} & \dots & \boldsymbol{C} \end{bmatrix} \in \mathbb{R}^{qn\times qn}$. It is also a circulant matrix. Notice that the line of reasoning followed so far is unchanged if we consider instead a $qp \times qn$ random matrix $\boldsymbol{N}$ and replace $\boldsymbol{C}$ by $\mathfrak{C}_q$. In particular, we can verify the following points.

- The eigenvalues of $\mathfrak{C}_q$ are $q\psi_0, \dots, q\psi_{n-1}$ and $(q-1)n$ zeros.
- We must only be careful that Proposition 4.17 becomes $\|\boldsymbol{A}\odot\mathfrak{C}_q\| \leqslant \alpha \log n \|\boldsymbol{A}\|$ (and not $\log(qn)$!). This result comes with the same proof and the eigendecomposition $(\frac{1}{\sqrt{q}}\mathbf{1}_q \boxtimes \widetilde{\boldsymbol{F}})(q\boldsymbol{\Psi})(\frac{1}{\sqrt{q}}\mathbf{1}_q \boxtimes \widetilde{\boldsymbol{F}})^*$.
- The result of Proposition 4.19 remains true if we replace $\boldsymbol{C}$ by $\mathfrak{C}_q$.
- The equations derived in Section 4.A.2 remain true because we only use the fact that the mask in symmetric.
- The concentration results in Section 4.A.3 remain true as well because the upper bound $\|\boldsymbol{A} \odot \mathfrak{C}_q\|_{\mathrm{F}} \leqslant \|\boldsymbol{A}\|_{\mathrm{F}}$ is still true.
- In the proof of Proposition 4.26, $\boldsymbol{R}$ can be replaced by its $qn \times qn$ version (or $\mathbf{1}_{q\times q} \boxtimes \boldsymbol{R}$ as well).

Therefore, we can simply replace $p, n, \boldsymbol{C}$ by $qp, qn, \mathfrak{C}_q$ everywhere with the only exception that $\log n$ does not become $\log(qn)$ but remains $\log n$. Notably, denoting $\widetilde{m}_{n,q}$ the Stieltjes transform of the empirical spectral distribution of $\frac{\boldsymbol{N}^\top \boldsymbol{N}}{qp} \odot \mathfrak{C}_q$ (recall that $\boldsymbol{N}$ is now of size $qp \times qn$), Equation (4.7) becomes

$$1 + z\mathbb{E}\big[\widetilde{m}_{n,q}(z)\big] - \frac{qp}{qn}\sum_{k=0}^{n-1} \frac{\mathbb{E}\big[\widetilde{m}_{n,q}(z)\big]\frac{q\psi_k}{qp}}{1+\mathbb{E}\big[\widetilde{m}_{n,q}(z)\big]\frac{q\psi_k}{qp}} = \mathcal{O}_z((qn)^{-1/2}\log n).$$

After simplification, we see that $\mathbb{E}[\widetilde{m}_{n,q}(z)]$ satisfies the same equation as $\mathbb{E}[\widetilde{m}_n(z)]$ with an error term $\mathcal{O}_z((qn)^{-1/2}\log n)$ instead of $\mathcal{O}_z(n^{-1/2}\log n)$, i.e, $\mathbb{E}[\widetilde{m}_{n,q}](z) = T_n[\mathbb{E}[\widetilde{m}_{n,q}]](z) + \mathcal{O}_z((qn)^{-1/2}\log n)$. Fix $z \in \mathbb{C}\setminus\mathbb{R}$. Since $\mathbb{E}[\widetilde{m}_{n,q}]$ is a Stieltjes transform (Corollary 2.9), $|\mathbb{E}[\widetilde{m}_{n,q}(z)]| \leqslant \frac{1}{|\Im z|}$ thus the sequence of complex numbers $(\mathbb{E}[\widetilde{m}_{n,q}(z)])_{q\geqslant 1}$ is bounded. Consider a converging subsequence (as $q \to +\infty$ but n is fixed) and denote its limit $\bar{m}_n(z)$. Thus, $\bar{m}_n$ is defined on $\mathbb{C}\setminus\mathbb{R}$ and satisfies $\bar{m}_n = T_n[\bar{m}_n]$.

(ii) Stieltjes transform. As per Proposition 2.10, $\bar{m}_n$ is the Stieltjes transform of a finite real measure and, by Proposition 4.27, since $\bar{m}_n = T_n[\bar{m}_n]$, it is the Stieltjes transform of a real *probability* measure.

(iii) Uniqueness. Assume there exists another Stieltjes transform (of a real probability measure) $\check{m}_n$ such that $\check{m}_n = T_n[\check{m}_n]$. Then,

$$\bar{m}_n(z) - \check{m}_n(z) = \frac{1}{\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+\bar{m}_n(z)\frac{\psi_k}{p}} - z} - \frac{1}{\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+\check{m}_n(z)\frac{\psi_k}{p}} - z} = \gamma(z)(\bar{m}_n(z) - \check{m}_n(z))$$

where $\gamma(z) = \dfrac{\frac{1}{n}\sum_{k=0}^{n-1} \frac{\psi_k^2/p}{\left(1+\bar{m}_n(z)\frac{\psi_k}{p}\right)\left(1+\check{m}_n(z)\frac{\psi_k}{p}\right)}}{\left(\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+\bar{m}_n(z)\frac{\psi_k}{p}} - z\right)\left(\frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+\check{m}_n(z)\frac{\psi_k}{p}} - z\right)} = \dfrac{p}{n}\sum_{k=0}^{n-1}\dfrac{\bar{m}_n(z)\frac{\psi_k}{p}}{1+\bar{m}_n(z)\frac{\psi_k}{p}}\dfrac{\check{m}_n(z)\frac{\psi_k}{p}}{1+\check{m}_n(z)\frac{\psi_k}{p}}.$

Since $\bar{m}_n(z), \check{m}_n z \to 0$ as $|z| \to +\infty$, $\gamma \neq 1$ and we conclude that $\bar{m}_n = \check{m}_n$. □

Following Proposition 4.28, $\bar{m}_n(z)$ satisfies

$$1 + z\bar{m}_n(z) - \frac{p}{n}\sum_{k=0}^{n-1}\frac{\bar{m}_n(z)\frac{\psi_k}{p}}{1+\bar{m}_n(z)\frac{\psi_k}{p}} = 0$$

and we can subtract this relation to Equation (4.7),

$$(\mathbb{E}[\tilde{m}_n(z)] - \bar{m}_n(z))\left[z - \frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{\left(1+\mathbb{E}[\tilde{m}_n(z)]\frac{\psi_k}{p}\right)\left(1+\bar{m}_n(z)\frac{\psi_k}{p}\right)}\right] = \mathcal{O}_z(n^{-1/2}\log n) \tag{4.8}$$

Let us state here the following lemma.

Lemma 4.29. *Let $z_1, z_2 \in \mathbb{C}$ such that $|z_1|, |z_2| < 1$. If the signs of $\Im z_1$ and $\Im z_2$ are equal then the sign of $\Im\left[\frac{-1}{(1+z_1)(1+z_2)}\right]$ is the same.*

Proof. Let $a_1, a_2, b_1, b_2 \in \mathbb{R}$ such that $z_1 = a_1 + \mathrm{i}b_1$ and $z_2 = a_2 + \mathrm{i}b_2$.

$$\Im\left[\frac{-1}{(1+z_1)(1+z_2)}\right] = -\Im\left[\frac{(1+a_1-\mathrm{i}b_1)(1+a_2-\mathrm{i}b_2)}{|1+z_1|^2|1+z_2|^2}\right] = \frac{(1+a_1)b_2 + (1+a_2)b_1}{|1+z_1|^2|1+z_2|^2}.$$

Since $|z_1|, |z_2| < 1$, the real numbers $1+a_1$ and $1+a_2$ are strictly positive. Hence, as the signs of b_1 and b_2 are the same, it is also the sign of $(1+a_1)b_2 + (1+a_2)b_1$. □

Assume, for the moment, that $|\mathbb{E}[\tilde{m}_n(z)]|\frac{2L-1}{p} < 1$ and $|\bar{m}_n(z)|\frac{2L-1}{p} < 1$. Since $\mathbb{E}[\tilde{m}_n]$ and $\bar{m}_n$ are Stieltjes transforms of real probability measures, this is true at least if $|\Im z| > \frac{2L-1}{p} = \frac{\varepsilon}{c}$. Moreover, if $\Im z > 0$, then $\mathbb{E}[\tilde{m}_n(z)](z), \bar{m}_n(z) > 0$ and the signs of $\Im[\mathbb{E}[\tilde{m}_n(z)]\frac{\psi_k}{p}]$ and $\Im[\bar{m}_n(z)\frac{\psi_k}{p}]$ are that of ψ_k. Then, according to Lemma 4.29, the sign of $\Im\left[\frac{-1}{\left(1+\mathbb{E}[\tilde{m}_n(z)]\frac{\psi_k}{p}\right)\left(1+\bar{m}_n(z)\frac{\psi_k}{p}\right)}\right]$ is also that of ψ_k and $\Im\left[\frac{-\psi_k}{\left(1+\mathbb{E}[\tilde{m}_n(z)]\frac{\psi_k}{p}\right)\left(1+\bar{m}_n(z)\frac{\psi_k}{p}\right)}\right]$ is positive. Similarly, if $\Im z < 0$ then $\Im\left[\frac{-\psi_k}{\left(1+\mathbb{E}[\tilde{m}_n(z)]\frac{\psi_k}{p}\right)\left(1+\bar{m}_n(z)\frac{\psi_k}{p}\right)}\right]$ is negative. Therefore,

$$\frac{1}{\left|z - \frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{\left(1+\mathbb{E}[\tilde{m}_n(z)]\frac{\psi_k}{p}\right)\left(1+\bar{m}_n(z)\frac{\psi_k}{p}\right)}\right|} \leqslant \frac{1}{\left|\Im z + \frac{1}{n}\sum_{k=0}^{n-1}\Im\frac{-\psi_k}{\left(1+\mathbb{E}[\tilde{m}_n(z)]\frac{\psi_k}{p}\right)\left(1+\bar{m}_n(z)\frac{\psi_k}{p}\right)}\right|} \leqslant \frac{1}{|\Im z|}$$

and Equation (4.8) simplifies into

$$\mathbb{E}[\tilde{m}_n(z)] - \bar{m}_n(z) = \mathcal{O}_z(n^{-1/2}\log n).$$

In summary, then sequence of *analytic* functions $z \mapsto \mathbb{E}[\tilde{m}_n(z)] - \bar{m}_n(z)$ converges pointwise to 0 (at least) on $\{z \in \mathbb{C} \mid |\Im z| > \frac{\varepsilon}{c}\}$. But since they are analytic on $\mathbb{C} \setminus \mathbb{R}$, the pointwise convergence to 0 holds on $\mathbb{C} \setminus \mathbb{R}$ as well, as per Vitali's convergence theorem (Theorem 2.24).

Finally, $|\tilde{m}_n(z) - \bar{m}_n(z)| \to 0$ almost surely as $p, n, L \to +\infty$ by the concentration of $\tilde{m}_n(z)$ around its expectation proven in Proposition 4.22.

4.A.5 Deterministic Equivalent of the Resolvent

In this section, we consider the "full" signal-plus-noise model $\boldsymbol{X} = \boldsymbol{P} + \boldsymbol{N}$ and derive a deterministic equivalent of the resolvent $\widetilde{\boldsymbol{Q}}(z) = (\frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C} - z\boldsymbol{I}_n)^{-1}$. As shown with Proposition 4.25, the asymptotic behavior of the limiting spectral distribution of $\frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}$ is characterized by the deterministic Stieltjes transform $\bar{m}_n$.

Since $\widetilde{\boldsymbol{Q}}^{-1}\widetilde{\boldsymbol{Q}} = \boldsymbol{I}_n$, we have

$$\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right] + \mathbb{E}\left[\left(\frac{\boldsymbol{N}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right] = z\mathbb{E}\big[\widetilde{\boldsymbol{Q}}\big] + \boldsymbol{I}_n$$

and, by Equation (4.3), this becomes

$$\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right] + \mathbb{E}\big[\widetilde{\boldsymbol{Q}}\big] - \mathbb{E}[(\boldsymbol{H} \odot \boldsymbol{I}_n)]\mathbb{E}\big[\widetilde{\boldsymbol{Q}}\big] + \mathscr{O}_z^{\|\cdot\|_{\max}}(n^{-3/2}) + \mathscr{O}_{z,\|\cdot\|}(n^{-1}(log n)^2) = z\mathbb{E}\big[\widetilde{\boldsymbol{Q}}\big] + \boldsymbol{I}_n \quad (4.9)$$

where $\boldsymbol{H} = \frac{1}{p}\boldsymbol{C}(\widetilde{\boldsymbol{Q}} \odot \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p})\boldsymbol{C}$ and $\mathbb{E}[(\boldsymbol{H} \odot \boldsymbol{I}_n)\widetilde{\boldsymbol{Q}}] = \mathbb{E}[\boldsymbol{H} \odot \boldsymbol{I}_n]\mathbb{E}[\widetilde{\boldsymbol{Q}}] + \mathscr{O}_z^{\|\cdot\|_{\max}}(n^{-3/2})$ as for the model without signal.

Proposition 4.30. *$\boldsymbol{H} = \boldsymbol{H}_0 + \mathscr{O}_z^{\|\cdot\|_{\max}}(n^{-1})$.*

Proof. Recall the relation $[\frac{1}{p}\boldsymbol{A}(\widetilde{\boldsymbol{Q}} \odot \frac{\boldsymbol{X}^\top \boldsymbol{X}}{p})\boldsymbol{C}]_{i,j} = \frac{1}{p}\operatorname{Tr}(\widetilde{\boldsymbol{Q}}(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot [\boldsymbol{A}_{i,\cdot}^\top \boldsymbol{C}_{j,\cdot}]))$. Then,

$$\begin{aligned}
H_{i,j} &= \frac{1}{p}\operatorname{Tr}\left(\widetilde{\boldsymbol{Q}}\left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot [\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}]\right)\right) \\
&= \frac{1}{p}\operatorname{Tr}\left(\widetilde{\boldsymbol{Q}}\left(\left[\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} + \frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} + \frac{\boldsymbol{N}^\top \boldsymbol{P}}{p}\right] \odot [\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}]\right)\right) + \frac{1}{p}\operatorname{Tr}\left(\widetilde{\boldsymbol{Q}}\left(\frac{\boldsymbol{N}^\top \boldsymbol{N}}{p} \odot [\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}]\right)\right) \\
&= \frac{1}{p}\operatorname{Tr}\big(\widetilde{\boldsymbol{Q}}\big(\boldsymbol{S} \odot [\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}]\big)\big) + [\boldsymbol{H}_0]_{i,j}
\end{aligned}$$

where we have introduced $\boldsymbol{S} \overset{\text{def}}{=} \frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} + \frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} + \frac{\boldsymbol{N}^\top \boldsymbol{P}}{p}$, whose rank is at most $3K = \mathscr{O}(1)$.

$$\begin{aligned}
\left|\frac{1}{p}\operatorname{Tr}\big(\widetilde{\boldsymbol{Q}}\big(\boldsymbol{S} \odot [\boldsymbol{C}_{\cdot,i}\boldsymbol{C}_{j,\cdot}]\big)\big)\right| &= \frac{1}{p}\left|\operatorname{Tr}\big(\widetilde{\boldsymbol{Q}}\operatorname{Diag}(\boldsymbol{C}_{\cdot,i})\boldsymbol{S}\operatorname{Diag}(\boldsymbol{C}_{j,\cdot})\big)\right| \\
&\leqslant \frac{1}{p}\sum_{k=1}^{n} s_k(\widetilde{\boldsymbol{Q}}\operatorname{Diag}(\boldsymbol{C}_{\cdot,i})\boldsymbol{S}\operatorname{Diag}(\boldsymbol{C}_{j,\cdot})) \qquad \text{from Proposition 4.13} \\
&\leqslant \frac{1}{p}\sum_{k=1}^{n} s_k(\widetilde{\boldsymbol{Q}})\, s_k(\operatorname{Diag}(\boldsymbol{C}_{\cdot,i})\boldsymbol{S}\operatorname{Diag}(\boldsymbol{C}_{j,\cdot})) \qquad \text{from Proposition 4.14} \\
&\leqslant \frac{1}{p}\big\|\widetilde{\boldsymbol{Q}}\big\| \sum_{k=1}^{n} s_k(\operatorname{Diag}(\boldsymbol{C}_{\cdot,i})\boldsymbol{S}\operatorname{Diag}(\boldsymbol{C}_{j,\cdot})) \\
&\leqslant \frac{3K}{p}\big\|\widetilde{\boldsymbol{Q}}\big\|\big\|\operatorname{Diag}(\boldsymbol{C}_{\cdot,i})\boldsymbol{S}\operatorname{Diag}(\boldsymbol{C}_{j,\cdot})\big\| \qquad \text{since } \operatorname{Rank}\boldsymbol{S} \leqslant 3K \\
&\leqslant \frac{3K}{p}\big\|\widetilde{\boldsymbol{Q}}\big\|\|\boldsymbol{S}\| = \mathscr{O}_z(n^{-1}) \qquad \text{since } \big\|\operatorname{Diag}(\boldsymbol{C}_{\cdot,i})\big\| = \big\|\operatorname{Diag}(\boldsymbol{C}_{j,\cdot})\big\| = 1.
\end{aligned}$$

□

Combining Proposition 4.30 and the fact (proven above) that $\mathbb{E}[\boldsymbol{H}_0]_{i,j} = \bar{H}_{i,j} + \mathcal{O}_z(n^{-1/2}\log n)$, we have $\mathbb{E}[H_{i,i}] = \eta_0(z) + \mathcal{O}_z(n^{-1/2}\log n)$ for all $i \in [n]$. Thus, using Equation (4.6) in Equation (4.9) yields

$$\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right] + \frac{1}{\mathbb{E}[\widetilde{m}_n(z)]}\mathbb{E}\big[\widetilde{\boldsymbol{Q}}\big] = \boldsymbol{I}_n + \mathcal{O}_z^{\|\cdot\|}(n^{-1/2}\log n) \tag{4.10}$$

where the $\mathcal{O}_z^{\|\cdot\|_{\max}}(n^{-3/2})$ in Equation (4.9) is absorbed in the $\mathcal{O}_z^{\|\cdot\|}(n^{-1/2}\log n)$ since $\mathcal{O}_z^{\|\cdot\|_{\max}}(n^{-3/2}) = \mathcal{O}_z^{\|\cdot\|}(n^{-1/2})$ (in accordance with the inequalities $\|\cdot\| \leqslant \|\cdot\|_{\mathrm{F}} \leqslant n \times \|\cdot\|_{\max}$). Moreover, with the decomposition $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{J}^\top$ (Assumption 4.3), we have

$$\begin{aligned}
\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right]_{i,i} &= \sum_{k=1}^{n}\left[\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p}\right]_{i,k} C_{i,k}\widetilde{Q}_{k,i} \\
&= \frac{1}{\sqrt{p}}\sum_{k=1}^{n}\sum_{l=1}^{K} J_{i,l}\left[\boldsymbol{M}^\top \frac{\boldsymbol{X}}{\sqrt{p}}\right]_{l,k} C_{i,k}\widetilde{Q}_{k,i} \\
&= \frac{1}{\sqrt{p}}\sum_{l=1}^{K} J_{i,l}\left[\boldsymbol{M}^\top \frac{\boldsymbol{X}}{\sqrt{p}}(\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C})\right]_{l,i}
\end{aligned}$$

where $J_{i,l} \in \{0,1\}$ and $\left|\left[\boldsymbol{M}^\top \frac{\boldsymbol{X}}{\sqrt{p}}(\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C})\right]_{l,i}\right| \leqslant \left\|\boldsymbol{M}^\top \frac{\boldsymbol{X}}{\sqrt{p}}(\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C})\right\| = \mathcal{O}_z(\log n)$ almost surely. Therefore, $\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right]_{i,i} = \mathcal{O}_z(n^{-1/2}\log n)$ and, by Equation (4.10),

$$\begin{aligned}
\frac{1}{\mathbb{E}[\widetilde{m}_n(z)]}\mathbb{E}\big[\widetilde{Q}_{i,i}\big] = 1 + \mathcal{O}_z(n^{-1/2}\log n) \qquad \Longleftrightarrow \qquad \mathbb{E}\big[\widetilde{Q}_{i,i}\big] &= \mathbb{E}[\widetilde{m}_n(z)] + \mathcal{O}_z(n^{-1/2}\log n) \\
&= \bar{m}_n(z) + \mathcal{O}_z(n^{-1/2}\log n).
\end{aligned}$$

In the proof of Proposition 4.22, we have shown that $\mathbb{E}[|\widetilde{Q}_{i,i} - \mathbb{E}[\widetilde{Q}_{i,i}]|^4] = \mathcal{O}_z(n^{-2})$. Thus, by Markov's inequality (Billingsley, 2012, Equation 5.31),

$$\mathbb{P}\Big(\big|\widetilde{Q}_{i,i} - \mathbb{E}\big[\widetilde{Q}_{i,i}\big]\big| \geqslant n^{-\delta}\Big) \leqslant n^{4\delta} \times \mathcal{O}_z(n^{-2}) = \mathcal{O}_z(n^{4\delta-2}).$$

Then, according to the first Borel-Cantelli lemma (Billingsley, 2012, Theorem 4.3), $\widetilde{Q}_{i,i} = \mathbb{E}[\widetilde{Q}_{i,i}] + \mathcal{O}_z(n^{-\delta})$ almost surely for all $\delta < \frac{1}{4}$. Consequently, $\widetilde{Q}_{i,i} = \bar{m}_n(z) + \mathcal{O}_z(n^{-\delta})$ almost surely as well. This observation is useful to handle the term $\mathbb{E}[(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C})\widetilde{\boldsymbol{Q}}]$ in Equation (4.10). Indeed, $\mathbb{E}[(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C})\widetilde{\boldsymbol{Q}}] = \mathbb{E}[(\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C})\widetilde{\boldsymbol{Q}}] + \mathbb{E}[(\frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C})\widetilde{\boldsymbol{Q}}]$ and we have the following result.

Proposition 4.31. *For all integer $q \geqslant 0$,*

$$\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C}\left(-\bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^q\right)\widetilde{\boldsymbol{Q}}\right] = \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\left(-\bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^{q+1}\right)\widetilde{\boldsymbol{Q}}\right] + \mathcal{O}_z^{\|\cdot\|}(n^{-\delta}).$$

Proof. We use Equation (4.4) with $\boldsymbol{A} = \boldsymbol{C}(-\bar{m}_n(z)\frac{\boldsymbol{C}}{p})^q$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C}\left(-\bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^q\right)\widetilde{\boldsymbol{Q}}\right] \\
&\qquad = -\frac{1}{p}\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\left(-\bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^q(\widetilde{\boldsymbol{Q}} \odot \boldsymbol{I}_n)\boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right] - \frac{1}{p}\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p}(\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C}) \odot \boldsymbol{C}\left(-\bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^q\right)\widetilde{\boldsymbol{Q}}\right] \\
&\qquad = \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\left(-\bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^{q+1}\right)\widetilde{\boldsymbol{Q}}\right] - (-\bar{m}_n(z))^q\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \frac{\boldsymbol{C}^{q+1}}{p^q}\operatorname{Diag}(\boldsymbol{d})\frac{\boldsymbol{C}}{p}\right)\widetilde{\boldsymbol{Q}}\right] \\
&\qquad\qquad - \frac{(-\bar{m}_n(z))^q}{p}\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p}(\widetilde{\boldsymbol{Q}} \odot \boldsymbol{C}) \odot \frac{\boldsymbol{C}^{q+1}}{p^q}\right)\widetilde{\boldsymbol{Q}}\right]
\end{aligned}$$

where $\boldsymbol{d} \in \mathbb{C}^n$ is such that $d_i = \widetilde{Q}_{i,i} - \bar{m}_n(z)$, i.e., $d_i = \mathcal{O}_z(n^{-\delta})$ almost surely. According to Proposition 4.18, the last term in the previous equation is $\mathcal{O}_z^{\|\cdot\|}(n^{-1}(\log n)^{q+2})$. Evaluating the middle one is less straightforward but we can proceed similarly to the proofs of Propositions 4.17 and 4.18, using the eigendecomposition $\boldsymbol{C} = \widetilde{\boldsymbol{F}}\boldsymbol{\Psi}\widetilde{\boldsymbol{F}}^*$.

$$\begin{aligned}
\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \frac{\boldsymbol{C}^{q+1}}{p^q} \operatorname{Diag}(\boldsymbol{d}) \frac{\boldsymbol{C}}{p} &= \frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \left(\frac{1}{p^q} \sum_{k=0}^{n-1} \psi_k^{q+1} \widetilde{\boldsymbol{F}}_{\cdot,k} \widetilde{\boldsymbol{F}}_{\cdot,k}^* \right) \operatorname{Diag}(\boldsymbol{d}) \left(\frac{1}{p} \sum_{l=0}^{n-1} \psi_l \widetilde{\boldsymbol{F}}_{\cdot,l} \widetilde{\boldsymbol{F}}_{\cdot,l}^* \right) \\
&= \frac{1}{p^{q+1}} \sum_{k=0}^{n-1} \sum_{l=0}^{n-1} \psi_k^{q+1} \psi_l \left[\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \left(\widetilde{\boldsymbol{F}}_{\cdot,k} \widetilde{\boldsymbol{F}}_{\cdot,k}^* \operatorname{Diag}(\boldsymbol{d}) \widetilde{\boldsymbol{F}}_{\cdot,l} \widetilde{\boldsymbol{F}}_{\cdot,l}^* \right) \right] \\
&= \frac{1}{p^{q+1}} \sum_{k=0}^{n-1} \sum_{l=0}^{n-1} \psi_k^{q+1} \psi_l \left(\widetilde{\boldsymbol{F}}_{\cdot,k}^* \operatorname{Diag}(\boldsymbol{d}) \widetilde{\boldsymbol{F}}_{\cdot,l} \right) \left[\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \widetilde{\boldsymbol{F}}_{\cdot,k} \widetilde{\boldsymbol{F}}_{\cdot,l}^* \right] \\
&= \frac{1}{p^{q+1}} \sum_{k=0}^{n-1} \sum_{l=0}^{n-1} \psi_k^{q+1} \psi_l \left(\widetilde{\boldsymbol{F}}_{\cdot,k}^* \operatorname{Diag}(\boldsymbol{d}) \widetilde{\boldsymbol{F}}_{\cdot,l} \right) \left[\operatorname{Diag}(\widetilde{\boldsymbol{F}}_{\cdot,k}) \frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \operatorname{Diag}(\widetilde{\boldsymbol{F}}_{\cdot,l})^* \right].
\end{aligned}$$

Since $\|\operatorname{Diag}(\widetilde{\boldsymbol{F}}_{\cdot,k})\|^2 = \max_{i \in [n]} |\widetilde{F}_{i,k}|^2 = \frac{1}{n}$,

$$\begin{aligned}
\left\| \frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \frac{\boldsymbol{C}^{q+1}}{p^q} \operatorname{Diag}(\boldsymbol{d}) \frac{\boldsymbol{C}}{p} \right\| &\leqslant \frac{1}{p^{q+1}} \frac{1}{n} \sum_{k=0}^{n-1} \sum_{l=0}^{n-1} |\psi_k|^{q+1} |\psi_l| \left| \widetilde{\boldsymbol{F}}_{\cdot,k}^* \operatorname{Diag}(\boldsymbol{d}) \widetilde{\boldsymbol{F}}_{\cdot,l} \right| \left\| \frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \right\| \\
&\leqslant \left(\frac{1}{p^{q+1}} \sum_{k=0}^{n-1} |\psi_k|^{q+1} \right) \left(\frac{1}{n} \sum_{l=0}^{n-1} |\psi_l| \right) \|\operatorname{Diag}(\boldsymbol{d})\| \left\| \frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \right\| \\
&\leqslant \left(\frac{1}{p} \sum_{k=0}^{n-1} |\psi_k| \right)^{q+1} \left(\frac{1}{n} \sum_{l=0}^{n-1} |\psi_l| \right) \|\operatorname{Diag}(\boldsymbol{d})\| \left\| \frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \right\|.
\end{aligned}$$

Hence, $\left\| \frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \frac{\boldsymbol{C}^{q+1}}{p^q} \operatorname{Diag}(\boldsymbol{d}) \frac{\boldsymbol{C}}{p} \right\| = \mathcal{O}_z(n^{-\delta}(\log n)^{q+2})$ and

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C} \left(-\bar{m}_n(z) \frac{\boldsymbol{C}}{p} \right)^q \right) \widetilde{\boldsymbol{Q}} \right] \\
&\qquad = \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C} \left(-\bar{m}_n(z) \frac{\boldsymbol{C}}{p} \right)^{q+1} \right) \widetilde{\boldsymbol{Q}} \right] + \mathcal{O}_z^{\|\cdot\|}(n^{-\delta}(\log n)^{q+2}) + \mathcal{O}_z^{\|\cdot\|}(n^{-1}(\log n)^{q+2}).
\end{aligned}$$

This is true for all $\delta < \frac{1}{4}$ and $\mathcal{O}_z^{\|\cdot\|}(n^{-\delta}(\log n)^{q+2}) = \mathcal{O}_z^{\|\cdot\|}(n^{-\delta'})$ for all $\delta' < \delta$, therefore the rest is $\mathcal{O}_z^{\|\cdot\|}(n^{-\delta})$ for all $\delta < \frac{1}{4}$. □

Following Proposition 4.31,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C} \right) \widetilde{\boldsymbol{Q}} \right] &= \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C} \right) \widetilde{\boldsymbol{Q}} \right] + \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C} \right) \widetilde{\boldsymbol{Q}} \right] \\
&= \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C} \right) \widetilde{\boldsymbol{Q}} \right] + \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C} \left(-\bar{m}_n(z) \frac{\boldsymbol{C}}{p} \right) \right) \widetilde{\boldsymbol{Q}} \right] + \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C} \left(-\bar{m}_n(z) \frac{\boldsymbol{C}}{p} \right) \right) \widetilde{\boldsymbol{Q}} \right] \\
&\qquad + \mathcal{O}_z^{\|\cdot\|}(n^{-\delta}) \\
&= \dots \\
&= \sum_{q=0}^{\mathfrak{q}} \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C} \left(-\bar{m}_n(z) \frac{\boldsymbol{C}}{p} \right)^q \right) \widetilde{\boldsymbol{Q}} \right] + \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C} \left(-\bar{m}_n(z) \frac{\boldsymbol{C}}{p} \right)^{\mathfrak{q}} \right) \widetilde{\boldsymbol{Q}} \right] + \mathcal{O}_z^{\|\cdot\|}(n^{-\delta}) \\
&= \left(\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C} \sum_{q=0}^{\mathfrak{q}} \left(-\bar{m}_n(z) \frac{\boldsymbol{C}}{p} \right)^q \right) \mathbb{E}[\widetilde{\boldsymbol{Q}}] + \mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C} \left(-\bar{m}_n(z) \frac{\boldsymbol{C}}{p} \right)^{\mathfrak{q}} \right) \widetilde{\boldsymbol{Q}} \right] + \mathcal{O}_z^{\|\cdot\|}(n^{-\delta})
\end{aligned}$$

for all integer $\mathfrak{q} \geqslant 0$. If $\|\bar{m}_n(z)\frac{\boldsymbol{C}}{p}\| = |\bar{m}_n(z)|\frac{2L-1}{p} < 1$ then we can make $\mathfrak{q}$ grow to $+\infty$, which yields

$$\mathbb{E}\left[\left(\frac{\boldsymbol{P}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}\right)\widetilde{\boldsymbol{Q}}\right] = \left(\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C}\left(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^{-1}\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}\big] + \mathcal{O}_z^{\|\cdot\|}(n^{-\delta}).$$

Finally, Equation (4.10) becomes

$$\left[\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C}\left(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^{-1} + \frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n\right]\mathbb{E}\big[\widetilde{\boldsymbol{Q}}\big] = \boldsymbol{I}_n + \mathcal{O}_z^{\|\cdot\|}(n^{-\delta})$$

so we have the following deterministic equivalent (Definition 2.17) of $\widetilde{\boldsymbol{Q}}$,

$$\bar{\boldsymbol{Q}}(z) \overset{\text{def}}{=} \left[\frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n + \frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C}\left(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^{-1}\right]^{-1}.$$

Notice that, although it is derived under the assumption $|\bar{m}_n(z)|\frac{2L-1}{p} < 1$, the matrix $\bar{\boldsymbol{Q}}$ is still well-defined if this condition is not satisfied (as long as $z \in \mathbb{C} \setminus \mathbb{R}$). Furthermore, for all bounded (sequences of) vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the functions $z \mapsto \boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}(z) - \bar{\boldsymbol{Q}}(z))\boldsymbol{b}$ and $z \mapsto \frac{1}{n}\operatorname{Tr}\boldsymbol{A}(\widetilde{\boldsymbol{Q}} - \bar{\boldsymbol{Q}})$ are analytic and converge pointwise to 0 almost surely as $p, n, L \to +\infty$ on $\{z \in \mathbb{C} \mid |\Im z| > \frac{\varepsilon}{c}\}$ (which is a subset of $\{z \in \mathbb{C} \mid |\bar{m}_n(z)|\frac{2L-1}{p} < 1\}$ for all p, n, L). Therefore, by Vitali's convergence theorem (Theorem 2.24), the almost sure pointwise convergence to 0 holds on $\mathbb{C} \setminus \mathbb{R}$ and $\bar{\boldsymbol{Q}}(z)$ is a deterministic equivalent of $\widetilde{\boldsymbol{Q}}(z)$ for all $z \in \mathbb{C} \setminus \mathbb{R}$.

4.B Proof of Theorem 4.8

4.B.1 Conjecture on the Global Behavior of the Spectrum

Before delving into mathematical considerations, we must state that the study of isolated eigenvalues and eigenvectors of $\frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}$ raises a few technical issues. The model considered differs notably from standard spiked random matrix models in the fact that the rank of the perturbation $\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C}$ scales with n. Yet, it can still be seen as if it had a low rank because the majority of eigenvalues of $\frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C}$ are negligible and only a small number (not scaling with n) of them are large enough to actually exceed the expected threshold value of signal-to-noise ratio causing the presence of isolated eigenvalues outside the spectrum of $\frac{\boldsymbol{N}^\top \boldsymbol{N}}{p} \odot \boldsymbol{C}$, just as in standard spiked random matrix models (such as the one studied in Section 2.3).

Nevertheless this "exoticism" comes with a cost: the tools presented in Chapter 2 (and, in particular, Section 2.3.3) are no longer quite suited to the study of the present model. Notably, the "trick" consisting in computing the position of isolated eigenvalues outside the bulk by means of a *small* determinant given by Sylvester's identity (Proposition 2.22) cannot help us here. Therefore, we must go through a simplifying assumption to work this out.

Precisely, we know that the resolvent $\widetilde{\boldsymbol{Q}}(z)$ is defined for all z which is not an eigenvalue of $\frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}$. Therefore, we make the following conjecture where, by *isolated eigenvalue*, we mean an eigenvalue lying outside the distribution characterized by $\bar{m}_n$.

Conjecture 4.32. *With probability* 1, *as* $p, n, L \to +\infty$, *the only isolated eigenvalues in the spectrum of* $\frac{\boldsymbol{X}^\top \boldsymbol{X}}{p} \odot \boldsymbol{C}$ *are the ones converging to the singular points of the deterministic equivalent* $\bar{\boldsymbol{Q}}$, *that is,* $\xi \in \mathbb{R}$ *such that* $\frac{1}{\bar{m}_n(\xi)}\boldsymbol{I}_n + \frac{\boldsymbol{P}^\top \boldsymbol{P}}{p} \odot \boldsymbol{C}\left(\boldsymbol{I}_n + \bar{m}_n(\xi)\frac{\boldsymbol{C}}{p}\right)^{-1}$ *is not invertible (with* $\bar{m}_n(\xi) \overset{\text{def}}{=} \lim_{y\downarrow 0} \bar{m}_n(\xi + \mathrm{i}y)$*).*

This property is verified for the random matrix model $\frac{\boldsymbol{X}^\top \boldsymbol{X}}{p}$ as shown in Section 2.3 and we argue that adding a mask $\boldsymbol{C}$ — i.e., simply zeroing the entries which are far from the diagonal — is a soft enough transformation and so it keeps these "nice properties" of the global behavior of the spectrum. Furthermore, a large number of numerical simulations make us confident with this statement. The reader is kindly invited to reproduce these simulations with the Python notebooks given in the GitHub repository `https://github.com/HugoLebeau/online_learning/`.

Consequently, in the following sections, we exhibit the singular points of the deterministic equivalent $\bar{\boldsymbol{Q}}$ in the two-class setting ($\boldsymbol{P} = \boldsymbol{\mu}\boldsymbol{j}^\top$). This guarantees that one or more (depending on the multiplicity) eigenvalues converge to this location if it is outside the support of the distribution characterized by $\bar{m}_n$ *and we conjecture that these are the only isolated eigenvalues.* Then, we study the alignments of the corresponding eigenvectors with the signal $\boldsymbol{j}$ through the standard approach relying on contour integrals presented in Section 2.3.3.

4.B.2 Deterministic Equivalent

Here $\boldsymbol{P} = \boldsymbol{\mu}\boldsymbol{j}^\top$ therefore we can simplify the expression of the deterministic equivalent $\bar{\boldsymbol{Q}}$ given in Theorem 4.6. We use the relation $\boldsymbol{M} \odot \boldsymbol{a}\boldsymbol{b}^\top = [\mathrm{Diag}\,\boldsymbol{a}]\boldsymbol{M}[\mathrm{Diag}\,\boldsymbol{b}]$ and the eigendecomposition $\boldsymbol{C} = \boldsymbol{F}\boldsymbol{\Psi}\boldsymbol{F}^\top$ where $\boldsymbol{F} = \Re\widetilde{\boldsymbol{F}} + \Im\widetilde{\boldsymbol{F}}$ (see Section 4.1.2).

$$\begin{aligned}
\bar{\boldsymbol{Q}}(z) &= \left[\frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n + \frac{\boldsymbol{P}^\top\boldsymbol{P}}{p} \odot \boldsymbol{C}\left(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^{-1}\right]^{-1} \\
&= \left[\frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n + \|\boldsymbol{\mu}\|^2\frac{\boldsymbol{j}\boldsymbol{j}^\top}{p} \odot \boldsymbol{C}\left(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^{-1}\right]^{-1} \\
&= \left[\frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n + \|\boldsymbol{\mu}\|^2\boldsymbol{D}_{\boldsymbol{j}}\frac{\boldsymbol{C}}{p}\left(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{C}}{p}\right)^{-1}\boldsymbol{D}_{\boldsymbol{j}}\right]^{-1} \\
&= \left[\frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n + \|\boldsymbol{\mu}\|^2\boldsymbol{D}_{\boldsymbol{j}}\boldsymbol{F}\frac{\boldsymbol{\Psi}}{p}\left(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{\Psi}}{p}\right)^{-1}\boldsymbol{F}^\top\boldsymbol{D}_{\boldsymbol{j}}\right]^{-1} \\
&= [\boldsymbol{D}_{\boldsymbol{j}}\boldsymbol{F}]\left[\frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n + \|\boldsymbol{\mu}\|^2\frac{\boldsymbol{\Psi}}{p}\left(\boldsymbol{I}_n + \bar{m}_n(z)\frac{\boldsymbol{\Psi}}{p}\right)^{-1}\right]^{-1}[\boldsymbol{D}_{\boldsymbol{j}}\boldsymbol{F}]^\top.
\end{aligned}$$

Note that $\boldsymbol{D}_{\boldsymbol{j}}\boldsymbol{F}$ is also an orthogonal matrix ($[\boldsymbol{D}_{\boldsymbol{j}}\boldsymbol{F}]^{-1} = [\boldsymbol{D}_{\boldsymbol{j}}\boldsymbol{F}]^\top$), hence the last equality.

4.B.3 Isolated Eigenvalues

Given the expression of $\bar{\boldsymbol{Q}}$ in the two-class setting, its singular points ξ_k satisfy

$$\frac{1}{\bar{m}_n(\xi_k)} + \frac{\|\boldsymbol{\mu}\|^2\frac{\psi_k}{p}}{1+\bar{m}_n(\xi_k)\frac{\psi_k}{p}} = 0 \quad\Longleftrightarrow\quad 1 + \bar{m}_n(\xi_k)\frac{\psi_k}{p} + \|\boldsymbol{\mu}\|^2\bar{m}_n(\xi_k)\frac{\psi_k}{p} = 0.$$

If $\psi_k = 0$, there is no solution. If $\psi_k \neq 0$, then $\bar{m}_n(\xi_k) = \frac{-1}{(\|\boldsymbol{\mu}\|^2+1)\frac{\psi_k}{p}}$ and we can inject this expression into Equation (4.1),

$$1 + \frac{-1}{\left(\|\boldsymbol{\mu}\|^2+1\right)\frac{\psi_k}{p}}\left(\xi - \frac{1}{n}\sum_{l=0}^{n-1}\frac{\psi_l}{1+\frac{-1}{\left(\|\boldsymbol{\mu}\|^2+1\right)\frac{\psi_k}{p}}\frac{\psi_l}{p}}\right) = 0$$

$$\iff \quad \xi_k = \left(\|\boldsymbol{\mu}\|^2+1\right)\frac{\psi_k}{p} + \frac{1}{n}\sum_{l=0}^{n-1}\frac{\psi_l}{1-\frac{\psi_l/p}{\left(\|\boldsymbol{\mu}\|^2+1\right)\frac{\psi_k}{p}}}$$
$$= \left(\|\boldsymbol{\mu}\|^2+1\right)\frac{\psi_k}{p}\left(1+c\sum_{l=0}^{n-1}\left[\left(\|\boldsymbol{\mu}\|^2+1\right)\frac{\psi_k}{\psi_l}-1\right]^{-1}\right).$$

Since ξ_k must not lie inside the support of the distribution characterized by $\bar{m}_n$, it must also satisfy $\lim_{y\downarrow 0}\Im\bar{m}_n(\xi_k+\mathrm{i}y)=0$, at least for p,n,L large enough.

4.B.4 Eigenvector Alignments

Let us denote $\{(\lambda_k,\boldsymbol{u}_k)\}_{0\leqslant k<n}$ the eigenvalue-eigenvector pairs of $\widetilde{\boldsymbol{K}}_L$. From the definition of the resolvent, we known that

$$\widetilde{\boldsymbol{Q}}(z) = \sum_{l=0}^{n-1}\frac{\boldsymbol{u}_l\boldsymbol{u}_l^\top}{\lambda_l - z}.$$

Therefore, with Cauchy's integral formula (Proposition 2.15) and a positively-oriented simple closed contour γ_k circling around λ_k and leaving the other eigenvalues outside, we have access to the quantity

$$\sum_{\substack{0\leqslant l\leqslant n-1\\ \xi_l=\xi_k}}\boldsymbol{u}_l\boldsymbol{u}_l^\top = -\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_k}\widetilde{\boldsymbol{Q}}(z)\,\mathrm{d}z$$

which is simply $\boldsymbol{u}_k\boldsymbol{u}_k^\top$ when the associated eigenvalue has multiplicity one. Then, we can calculate the alignment of any vector $\boldsymbol{a}\in\mathbb{R}^n$ with the eigenspace associated to λ_k:

$$\sum_{\substack{0\leqslant l\leqslant n-1\\ \xi_l=\xi_k}}\langle\boldsymbol{a},\boldsymbol{u}_l\rangle^2 = -\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_k}\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}(z)\boldsymbol{a}\,\mathrm{d}z.$$

The asymptotic behavior of this quantity is therefore given by the deterministic equivalent $\bar{\boldsymbol{Q}}$. With residue calculus (Proposition 2.16), we compute

$$-\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_k}\bar{\boldsymbol{Q}}(z)\,\mathrm{d}z = -\lim_{z\to\xi_k}(z-\xi_k)[\boldsymbol{D_jF}]\left[\frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n+\|\boldsymbol{\mu}\|^2\frac{\boldsymbol{\Psi}}{p}\left(\boldsymbol{I}_n+\bar{m}_n(z)\frac{\boldsymbol{\Psi}}{p}\right)^{-1}\right]^{-1}[\boldsymbol{D_jF}]^\top$$
$$= -[\boldsymbol{D_jF}]\left(\lim_{z\to\xi_k}(z-\xi_k)\left[\frac{1}{\bar{m}_n(z)}\boldsymbol{I}_n+\|\boldsymbol{\mu}\|^2\frac{\boldsymbol{\Psi}}{p}\left(\boldsymbol{I}_n+\bar{m}_n(z)\frac{\boldsymbol{\Psi}}{p}\right)^{-1}\right]^{-1}\right)[\boldsymbol{D_jF}]^\top.$$

If $l\in\{0,\dots,n-1\}$ is such that $\psi_l\neq\psi_k$, then

$$\lim_{z\to\xi_k}\frac{z-\xi_k}{\frac{1}{\bar{m}_n(z)}+\frac{\|\boldsymbol{\mu}\|^2\frac{\psi_l}{p}}{1+\bar{m}_n(z)\frac{\psi_l}{p}}} = 0$$

whereas if $\psi_l=\psi_k$, L'Hôpital's rule yields

$$\lim_{z\to\xi_k}\frac{z-\xi_k}{\frac{1}{\bar{m}_n(z)}+\frac{\|\boldsymbol{\mu}\|^2\frac{\psi_k}{p}}{1+\bar{m}_n(z)\frac{\psi_k}{p}}} = \frac{\frac{\mathrm{d}}{\mathrm{d}z}[(z-\xi_k)\bar{m}_n(z)]_{z=\xi_k}}{\frac{\mathrm{d}}{\mathrm{d}z}\left[1+\|\boldsymbol{\mu}\|^2\frac{\bar{m}_n(z)\frac{\psi_k}{p}}{1+\bar{m}_n(z)\frac{\psi_k}{p}}\right]_{z=\xi_k}}$$

$$= \frac{\bar{m}_n(\xi_k)}{\|\boldsymbol{\mu}\|^2 \frac{\bar{m}_n'(\xi_k)\frac{\psi_k}{p}}{\left(1+\bar{m}_n(\xi_k)\frac{\psi_k}{p}\right)^2}}$$
$$= \frac{\bar{m}_n(\xi_k)\left(1+\bar{m}_n(\xi_k)\frac{\psi_k}{p}\right)^2}{\|\boldsymbol{\mu}\|^2 \bar{m}_n'(\xi_k)\frac{\psi_k}{p}}.$$

Recalling that $\bar{m}_n(\xi_k) = \frac{-1}{(\|\boldsymbol{\mu}\|^2+1)\frac{\psi_k}{p}}$, we have $1 + m(\xi_k)\frac{\psi_k}{p} = \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2+1}$. Hence,

$$\lim_{z\to\xi_k} \frac{z-\xi_k}{\frac{1}{\bar{m}_n(z)} + \frac{\|\boldsymbol{\mu}\|^2\frac{\psi_k}{p}}{1+\bar{m}_n(z)\frac{\psi_k}{p}}} = \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2+1} \frac{1}{\left(\|\boldsymbol{\mu}\|^2+1\right)\frac{\psi_k}{p}} \frac{\bar{m}_n(\xi_k)}{\bar{m}_n'(\xi_k)}$$
$$= -\frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2+1} \frac{\bar{m}_n^2(\xi_k)}{\bar{m}_n'(\xi_k)}.$$

Let us calculate an expression of $\frac{\bar{m}_n^2(\xi_k)}{\bar{m}_n'(\xi_k)}$. Differentiating in z Equation (4.1) yields

$$\bar{m}_n(z) + z\bar{m}_n'(z) = c\sum_{r=0}^{n-1} \frac{\bar{m}_n'(z)\frac{\psi_r}{p}}{\left(1+\bar{m}_n(z)\frac{\psi_r}{p}\right)^2}$$

thus,

$$\frac{\bar{m}_n^2(\xi_k)}{\bar{m}_n'(\xi_k)} = -\xi_k\bar{m}_n(\xi_k) + c\sum_{r=0}^{n-1} \frac{\bar{m}_n(\xi_k)\frac{\psi_r}{p}}{\left(1+\bar{m}_n(\xi_k)\frac{\psi_r}{p}\right)^2}$$
$$= 1 - c\sum_{r=0}^{n-1} \frac{\bar{m}_n(\xi_k)\frac{\psi_r}{p}}{1+\bar{m}_n(\xi_k)\frac{\psi_r}{p}} + c\sum_{r=0}^{n-1} \frac{\bar{m}_n(\xi_k)\frac{\psi_r}{p}}{\left(1+\bar{m}_n(\xi_k)\frac{\psi_r}{p}\right)^2} \qquad \text{from Equation (4.1)}$$
$$\frac{\bar{m}_n^2(\xi_k)}{\bar{m}_n'(\xi_k)} = 1 - c\sum_{r=0}^{n-1}\left[\frac{\bar{m}_n(\xi_k)\frac{\psi_r}{p}}{1+\bar{m}_n(\xi_k)\frac{\psi_r}{p}}\right]^2.$$

Finally, since $\bar{m}_n(\xi_k) = \frac{-1}{(\|\boldsymbol{\mu}\|^2+1)\frac{\psi_k}{p}}$, we obtain

$$\lim_{z\to\xi_k} \frac{z-\xi_k}{\frac{1}{\bar{m}_n(z)} + \frac{\|\boldsymbol{\mu}\|^2\frac{\psi_k}{p}}{1+\bar{m}_n(z)\frac{\psi_k}{p}}} = -\frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2+1}\left(1 - c\sum_{l=0}^{n-1}\left[\frac{\bar{m}_n(\xi_k)\frac{\psi_l}{p}}{1+\bar{m}_n(\xi_k)\frac{\psi_l}{p}}\right]^2\right)$$
$$= -\frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2+1}\left(1 - c\sum_{l=0}^{n-1}\left[\left(\|\boldsymbol{\mu}\|^2+1\right)\frac{\psi_k}{\psi_l} - 1\right]^{-2}\right).$$

Hence,

$$-\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_k} \bar{\boldsymbol{Q}}(z)\,\mathrm{d}z = [\boldsymbol{D_j F}]\,\mathrm{Diag}(\zeta_0,\ldots,\zeta_{n-1})\,[\boldsymbol{D_j F}]^\top$$

with $\zeta_k = \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2+1}\left(1 - c\sum_{l=0}^{n-1}\left[\left(\|\boldsymbol{\mu}\|^2+1\right)\frac{\psi_k}{\psi_l} - 1\right]^{-2}\right)$. This shows that, if ξ_k is an isolated eigenvalue, then ζ_k is the asymptotic value of the alignment $\langle \boldsymbol{u}_k, [\boldsymbol{D_j F}]_{\cdot,k}\rangle^2$.

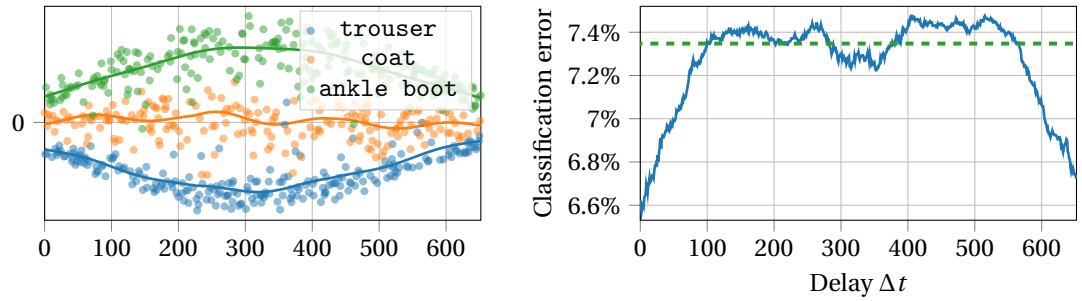


Figure 4.9: Clustering on Fashion-MNIST images (`trouser` vs. `coat` vs. `ankle boot`). **Left:** Dominant eigenvector of $\boldsymbol{K}_{L_\star}^{(t)}$. Solid curves are the estimated trend of each class $\boldsymbol{h}_{k,\mathscr{C}}^{(t)}$. **Right:** Classification error against delay Δt. This is the mean classification error at time $t_0 + \Delta t$ of a point arrived at t_0. The green dashed line indicate the overall classification error when the class is chosen by a majority vote. **Experimental setting:** $T = 21\,000$, $n_\star = 653$, $p = 784$, $L_\star = 226$.

4.C K-Classes Online Kernel Spectral Clustering Algorithm

4.C.1 General Presentation and Simulations

We use a set of spike eigenvectors $\{\boldsymbol{u}_k^{(t)}\}_{k\in\mathscr{K}}$ (with a set of indices $\mathscr{K}$) to estimate the $|\mathscr{K}|$-dimensional "trend" of each class. That is, denoting $\mathscr{C}[t]$ the class of $\boldsymbol{x}_t$, we consider the following model

$$\left[\boldsymbol{u}_k^{(t)}\right]_i = \left[\boldsymbol{h}_{k,\mathscr{C}[t-n+i]}^{(t)} + \boldsymbol{\epsilon}_k^{(t)}\right]_i$$

where, for $k \in \mathscr{K}$, $\boldsymbol{h}_{k,\mathscr{C}}^{(t)} \in \mathbb{R}^n$ is the "trend" of class $\mathscr{C}$ and $\boldsymbol{\epsilon}_k^{(t)}$ is a centered noise vector. A deeper analysis of the deterministic equivalent of Theorem 4.6 is needed to properly understand the behavior of the vectors $\boldsymbol{h}_{k,\mathscr{C}}^{(t)}$. From our general understanding so far, it is expected that they are linear combinations of a few dominant eigenvectors of $\boldsymbol{T}$. Using this approach, we are able to estimate the trends from $\{\boldsymbol{u}_k^{(t)}\}_{k\in\mathscr{K}}$ (see the left part of Figure 4.9). Each point is then associated to the class whose curve is the nearest in the $|\mathscr{K}|$-dimensional space. The details of this algorithm are given in the following subsection.

This algorithm is tested on a stream made of $T = 21\,000$ centered raw-images from the Fashion-MNIST dataset (Xiao et al., 2017). Their dimension is $p = 784$ and we want to discriminate between `trouser`, `coat` and `ankle boot` images in an online fashion. We choose $n = n_\star = 653$ and $L = L_\star = 226$ and we use the 5 dominant eigenvectors of $\boldsymbol{K}_{L_\star}^{(t)}$ for the estimation (thus $|\mathscr{K}| = 5$).

In Figure 4.9 are displayed the shape of the dominant eigenvector $\boldsymbol{u}_0^{(t)}$ at a given time during the execution of the algorithm with the estimated trends of each class (this is only the first dimension of a 5-dimensional trend) and the mean clustering error at $t_0 + \Delta t$ of a data point seen at t_0 with the overall classification error obtained after a majority vote. Surprisingly, the classification error curve has an inverse-U shape: it is easier to discriminate between the trends on the edges of the eigenvector than in the middle. The majority vote leads to an overall classification error which is close to the top of the inverted U. Unfortunately, in practice, it is hard to determine whether we are in a U-shaped / stable / inverted-U-shaped setting and, without this knowledge, a majority vote is the best decision policy.

Here, the overall classification error is 7.348% while a standard $T \times T$ *offline* kernel spectral clustering has only a 3.671% error rate.

4.C.2 Details of the Algorithm

We consider a set $\mathcal{K}$ of indices of spikes and the following model for $\boldsymbol{u}_k^{(t)}$, $k \in \mathcal{K}$,

$$\left[\boldsymbol{u}_k^{(t)}\right]_i = \left[\boldsymbol{h}_{k,\mathscr{C}[t-n+i]}^{(t)} + \boldsymbol{\epsilon}_k^{(t)}\right]_i, \qquad 1 \leqslant i \leqslant n$$

where $\boldsymbol{h}_{k,\mathscr{C}}^{(t)} \in \mathbb{R}^n$ is the trend of class $\mathscr{C}$ and $\boldsymbol{\epsilon}_k^{(t)}$ is a centered noise vector.

Our goal is to estimate the trend $\boldsymbol{h}_{k,\mathscr{C}}^{(t)}$ from the eigenvectors $\{\boldsymbol{u}_k^{(t)}\}_{k\in\mathcal{K}}$. Since we assume they are linear combinations of a few dominant eigenvectors of $\boldsymbol{T}$, we define a set of indices $\mathcal{K}_*$ specifying the eigenvectors $\{\boldsymbol{g}_k\}_{k\in\mathcal{K}_*}$ which we expect the $\boldsymbol{h}_{k,\mathscr{C}}^{(t)}$'s to be linear combinations of.

We denote $\hat{\mathscr{C}}^{(t)}[s]$ the class of $\boldsymbol{x}_{t-n+s}$ estimated at time t.

In order to compute an estimation $\{\hat{\mathscr{C}}^{(t)}[i]\}_{1\leqslant i\leqslant n}$ of the classes at a given time t, we propose a two-step algorithm. Firstly, we compute a rough estimation $\{\hat{\mathscr{C}}_0^{(t)}[i]\}_{1\leqslant i\leqslant n}$ of the classes by following the K paths with an exponential smoothing in the coordinates of the eigenvectors $\{\boldsymbol{u}_k^{(t)}\}_{k\in\mathcal{K}}$, this is called the *pre-clustering* step. Then, we refine this estimation with projections on $\mathrm{Span}\{\boldsymbol{g}_k\}_{k\in\mathcal{K}_*}$, this is the *clustering* step.

In the following, we drop the time dependency when it is not needed to ease notations.

Pre-Clustering Step

Given the number of classes K and the eigenvectors $\{\boldsymbol{u}_k^{(t)}\}_{k\in\mathcal{K}}$, we consider the set of n points in $\mathbb{R}^{|\mathcal{K}|}$ defined by the coordinates of each eigenvector: $[\boldsymbol{u}_{\mathcal{K}}]_i \overset{\text{def}}{=} \left([\boldsymbol{u}_k]_i\right)_{k\in\mathcal{K}}$ for $1 \leqslant i \leqslant n$. As i goes from 1 to n, these points draw K paths. The goal is to guess which path (and therefore which class) each point belongs to.

Iteration Let us suppose we have already estimated $\hat{\mathscr{C}}_0[1], \ldots, \hat{\mathscr{C}}_0[i-1]$ and the first $i-1$ coordinates of the vectors $\{\widetilde{\boldsymbol{h}}_k\}_{k\in\mathcal{K}}$ such that $\left[\widetilde{\boldsymbol{h}}_k\right]_j$ is an estimation of $\left[\boldsymbol{h}_{k,\hat{\mathscr{C}}_0[j]}\right]_j$ (initialization is discussed later). As for $\{\boldsymbol{u}_k\}_{k\in\mathcal{K}}$, we see $\{\widetilde{\boldsymbol{h}}_k\}_{k\in\mathcal{K}}$ as a set of n points in $\mathbb{R}^{|\mathcal{K}|}$, which have to be estimated. The estimation of the i-th point $\left[\widetilde{\boldsymbol{h}}_{\mathcal{K}}\right]_i$ is induced by the class estimate $\hat{\mathscr{C}}_0[i]$ — the corresponding path is updated with an exponential smoothing:

$$\left[\widetilde{\boldsymbol{h}}_{\mathcal{K}}\right]_i = \mathscr{E}_\alpha(i, \boldsymbol{u}_{\mathcal{K}}, \widetilde{\boldsymbol{h}}_{\mathcal{K}}, \hat{\mathscr{C}}_0[i]) \overset{\text{def}}{=} \frac{\alpha[\boldsymbol{u}_{\mathcal{K}}]_i + \theta\left[\widetilde{\boldsymbol{h}}_{\mathcal{K}}\right]_{I[\hat{\mathscr{C}}_0[i],i]}}{\alpha+\theta}$$

$$\text{where} \qquad \theta = \frac{1-\alpha}{i - I[\hat{\mathscr{C}}_0[i], i]}\left[1 + \frac{1-\alpha}{\alpha}\left(1 - (1-\alpha)^{i-I[\hat{\mathscr{C}}_0[i],i]-1}\right)\right],$$

$I[\hat{\mathscr{C}}_0[i], i] = \max\{1 \leqslant j \leqslant i-1 \mid \hat{\mathscr{C}}_0[j] = \hat{\mathscr{C}}_0[i]\}$ is the index of the last seen point in $\hat{\mathscr{C}}_0[i]$ and $\alpha \in [0,1]$ is the smoothing parameter. The reasons for such a formula are detailed in Appendix 4.D.

However, $\hat{\mathscr{C}}_0[i]$ is chosen as the class which minimizes the growth of the corresponding path:

$$\hat{\mathscr{C}}_0[i] = \underset{\hat{\mathscr{C}}\in\{\hat{\mathscr{C}}_1,\ldots,\hat{\mathscr{C}}_K\}}{\arg\min} \frac{\left\|\mathscr{E}_\alpha(i, \boldsymbol{u}_{\mathcal{K}}, \widetilde{\boldsymbol{h}}_{\mathcal{K}}, \hat{\mathscr{C}}) - \left[\widetilde{\boldsymbol{h}}_{\mathcal{K}}\right]_{I[\hat{\mathscr{C}},i]}\right\|}{i - I[\hat{\mathscr{C}}, i]}.$$

Indeed, by doing so, we minimize the Lipschitz constant of the estimated trend and ensure some regularity.

Initialization From the regularity of the true trend, $\boldsymbol{h}_{k,\mathscr{C}}$ is almost flat on its very first coordinates. Therefore, we can initialize the values $\hat{\mathscr{C}}_0[i]$ for $1 \leqslant i \leqslant h$ with a standard clustering algorithm applied to $\{[\boldsymbol{u}_{\mathcal{K}}]_i\}_{1\leqslant i\leqslant h}$. h is a parameter which should be taken as small as possible to stay in a domain where the trends are almost flat while still having a few representatives of each class. The computation of $\{\left[\widetilde{\boldsymbol{h}}_{\mathcal{K}}\right]_i\}_{1\leqslant i\leqslant h}$ follows from the class estimates, as presented above.

We found that a hierarchical clustering algorithm and $h \approx 10K$ worked well for the initialization. As for the smoothing parameter, a good value is $\alpha \approx 0.15$.

The pre-clustering step is summarized in Algorithm 2.

Algorithm 2: Pre-Clustering

Input: K, $\{\boldsymbol{u}_k\}_{k\in\mathcal{K}}$, h, α
Output: $\{\hat{\mathscr{C}}_0[i]\}_{1\leqslant i\leqslant n}$
Set $\hat{\mathscr{C}}_0[i]$ for $i = 1$ to h with agglomerative clustering
for $i = 1$ **to** h **do**
  $\left[\widetilde{\boldsymbol{h}}_{\mathcal{K}}\right]_i \leftarrow \mathcal{E}_\alpha(i, \boldsymbol{u}_{\mathcal{K}}, \widetilde{\boldsymbol{h}}_{\mathcal{K}}, \hat{\mathscr{C}}_0[i])$
end
for $i = h+1$ *to* n **do**
  $\hat{\mathscr{C}}_0[i] \leftarrow \arg\min_{\hat{\mathscr{C}}\in\{\hat{\mathscr{C}}_1,\dots,\hat{\mathscr{C}}_K\}} \frac{\left\|\mathcal{E}_\alpha(i,\boldsymbol{u}_{\mathcal{K}},\widetilde{\boldsymbol{h}}_{\mathcal{K}},\hat{\mathscr{C}}) - [\widetilde{\boldsymbol{h}}_{\mathcal{K}}]_{I[\hat{\mathscr{C}},i]}\right\|}{i - I[\hat{\mathscr{C}},i]}$
  $\left[\widetilde{\boldsymbol{h}}_{\mathcal{K}}\right]_i \leftarrow \mathcal{E}_\alpha(i, \boldsymbol{u}_{\mathcal{K}}, \widetilde{\boldsymbol{h}}_{\mathcal{K}}, \hat{\mathscr{C}}_0[i])$
end

Clustering Step

The class estimates obtained after the pre-clustering step are usually not very satisfying but still remain a good basis to estimate $\boldsymbol{h}_{k,\mathscr{C}}$ with regressions.

In the second step of the algorithm, we are given a set $\{\boldsymbol{g}_k\}_{k\in\mathcal{K}_*}$ of eigenvectors of $\boldsymbol{T}$. It is supposed that the trends $\{\boldsymbol{h}_{k,\mathscr{C}}\}_{k\in\mathcal{K}}$ are mixtures of these eigenvectors.

From the class estimates $\{\hat{\mathscr{C}}[i]\}_{1\leqslant i\leqslant n}$, we can compute an estimation $\hat{\boldsymbol{h}}_{\mathcal{K},\hat{\mathscr{C}}}$ of the trend of each estimated class $\hat{\mathscr{C}}$ with a linear regression

$$\hat{\boldsymbol{h}}_{k,\hat{\mathscr{C}}} = \boldsymbol{g}_{\mathcal{K}_*}\boldsymbol{\beta}_{k,\hat{\mathscr{C}}} \qquad \text{where} \qquad \boldsymbol{\beta}_{k,\hat{\mathscr{C}}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{|\mathcal{K}_*|}}{\arg\min} \left\|[\boldsymbol{u}_k]_{\hat{\mathscr{C}}} - \left[\boldsymbol{g}_{\mathcal{K}_*}\right]_{\hat{\mathscr{C}}}\boldsymbol{\beta}\right\|^2$$

where we use the notation $[\cdot]_{\hat{\mathscr{C}}}$ to represent the restriction to $\hat{\mathscr{C}}$.

Then, new class estimates can be computed by associating each point to the class whose trend is the closest:

$$\hat{\mathscr{C}}[i] = \underset{\hat{\mathscr{C}}\in\{\hat{\mathscr{C}}_1,\dots,\hat{\mathscr{C}}_K\}}{\arg\min} \left\|[\boldsymbol{u}_{\mathcal{K}}]_i - \left[\hat{\boldsymbol{h}}_{\mathcal{K},\hat{\mathscr{C}}}\right]_i\right\|.$$

We repeat this process until convergence of the class estimates. The classification step is summarized in Algorithm 3.

Final Algorithm

In an online fashion, pre-clustering can be performed as a warm-up during the first n time steps. Then, as $t \geqslant n$, only the clustering step is needed: the classes $\{\hat{\mathscr{C}}^{(t-1)}[s]\}_{1\leqslant s\leqslant n}$ estimated at $t-1$ (or during pre-clustering if $t = n$) serve as a good basis to estimate the classes at time t (both $\hat{\mathscr{C}}^{(t-1)}[s+1]$ and $\hat{\mathscr{C}}^{(t)}[s]$ are estimates of the class of $\boldsymbol{x}_{t-n+s}$). Moreover, the few interesting eigenvectors $\boldsymbol{u}_{\mathcal{K}}^{(t)}$ of $\boldsymbol{K}_L^{(t)}$

Algorithm 3: Clustering

Input: K, $\{\hat{\mathscr{C}}_0[i]\}_{1\leqslant i\leqslant n}$, $\{\boldsymbol{u}_k\}_{k\in\mathcal{K}}$, $\{\boldsymbol{v}_k\}_{k\in\mathcal{K}_*}$
Output: $\{\hat{\mathscr{C}}[i]\}_{1\leqslant i\leqslant n}$
for $i=1$ **to** n **do**
  $\hat{\mathscr{C}}[i]\leftarrow\hat{\mathscr{C}}_0[i]$
end
repeat
  for $\hat{\mathscr{C}}\in\{\hat{\mathscr{C}}_1,\ldots,\hat{\mathscr{C}}_K\}$ **do**
    $\hat{\boldsymbol{h}}_{\mathcal{K},\hat{\mathscr{C}}}\leftarrow\boldsymbol{g}_{\mathcal{K}_*}\left(\left[\boldsymbol{g}_{\mathcal{K}_*}\right]_{\hat{\mathscr{C}}}^\top\left[\boldsymbol{g}_{\mathcal{K}_*}\right]_{\hat{\mathscr{C}}}\right)^{-1}\left[\boldsymbol{g}_{\mathcal{K}_*}\right]_{\hat{\mathscr{C}}}^\top\left[\boldsymbol{u}_{\mathcal{K}}\right]_{\hat{\mathscr{C}}}$
  end
  for $i=1$ **to** n **do**
    $\hat{\mathscr{C}}[i]\leftarrow\arg\min_{\hat{\mathscr{C}}\in\{\hat{\mathscr{C}}_1,\ldots,\hat{\mathscr{C}}_K\}}\left\|\left[\boldsymbol{u}_{\mathcal{K}}\right]_i-\left[\hat{\boldsymbol{h}}_{\mathcal{K},\hat{\mathscr{C}}}\right]_i\right\|$
  end
until *convergence*

can be quickly computed with a power iteration algorithm starting at $\boldsymbol{u}_{\mathcal{K}}^{(t-1)}$ (they do not differ much from one time step to another). The final algorithm is presented in Algorithm 4.

Algorithm 4: Online Kernel Spectral Clustering

Input: K, $\mathcal{K}$, $\{\boldsymbol{g}_k\}_{k\in\mathcal{K}_*}$, h, α
Output: $\{\hat{\mathscr{C}}_t[s]\}_{\substack{1\leqslant s\leqslant n\\ n\leqslant t\leqslant T}}$
for $t=1$ **to** T **do**
  Get a new point $\boldsymbol{x}_t$ into the pipeline
  Update $\boldsymbol{K}_L^{(t-1)}$ into $\boldsymbol{K}_L^{(t)}$
  $\boldsymbol{u}_{\mathcal{K}}^{(t)}\leftarrow\texttt{PowerIteration}(\boldsymbol{K}_L^{(t)},\boldsymbol{u}_{\mathcal{K}}^{(t-1)})$
  if $1\leqslant t\leqslant n$ **then**
    Do an iteration as in Algorithm 2
  end
  if $t\geqslant n$ **then**
    Compute $\left\{\hat{\mathscr{C}}^{(t)}[s]\right\}_{1\leqslant s\leqslant n}$ according to Algorithm 3 with $\left\{\hat{\mathscr{C}}^{(t-1)}[s]\right\}_{1\leqslant s\leqslant n-1}$
  end
end

4.D Exponential Smoothing with Missing Data

Let $(\boldsymbol{y}_t)_{t\geqslant 0}$ be a time series. Assume we want to compute its trend $(\boldsymbol{s}_t)_{t\geqslant 0}$. A common technique is to perform an exponential smoothing:

$$\boldsymbol{s}_0=\boldsymbol{y}_0\qquad\text{and}\qquad\boldsymbol{s}_{t+1}=\alpha\boldsymbol{y}_{t+1}+(1-\alpha)\boldsymbol{s}_t\quad\text{for all } t\geqslant 0$$

where $\alpha\in[0,1]$ is the smoothing parameter. It acts as a low-pass filter which removes high-frequency noise.

Let us now assume that we do not have access to $(\boldsymbol{y}_t)_{t\geqslant 0}$ at each time step and we want to compute $\boldsymbol{s}_{t+h}$ ($h \geqslant 1$) with $\boldsymbol{y}_{t+h}$ and $\boldsymbol{s}_t$ only. Expanding the recurrence relation, we have

$$\boldsymbol{s}_{t+h} = \alpha \boldsymbol{y}_{t+h} + \alpha \sum_{k=1}^{h-1} (1-\alpha)^k \boldsymbol{y}_{t+h-k} + (1-\alpha)^h \boldsymbol{s}_t.$$

We propose to replace the unknown values $\boldsymbol{y}_{t+h-k}$ for $1 \leqslant k \leqslant h-1$ by the linear interpolation of the trend:

$$\begin{aligned}\boldsymbol{s}_{t+h} &= \alpha \boldsymbol{y}_{t+h} + \alpha \sum_{k=1}^{h-1} (1-\alpha)^k \left[\frac{k}{h} \boldsymbol{s}_t + \frac{h-k}{h} \boldsymbol{s}_{t+h} \right] + (1-\alpha)^h \boldsymbol{s}_t \\ &= \alpha \boldsymbol{y}_{t+h} + \frac{\alpha}{h} \left(\boldsymbol{s}_t \sum_{k=1}^{h-1} k(1-\alpha)^k + \boldsymbol{s}_{t+h} \sum_{k=1}^{h-1} k(1-\alpha)^{h-k} \right) + (1-\alpha)^h \boldsymbol{s}_t.\end{aligned}$$

Using the following formulae,

$$\begin{aligned}&\sum_{k=1}^{h-1} k(1-\alpha)^k = \frac{1-\alpha}{\alpha}\Big(1 - h(1-\alpha)^{h-1}\Big) + \left(\frac{1-\alpha}{\alpha}\right)^2 \Big(1-(1-\alpha)^{h-1}\Big) \\ \text{and} \quad &\sum_{k=1}^{h-1} k(1-\alpha)^{h-k} = \frac{1-\alpha}{\alpha}(h-1) - \left(\frac{1-\alpha}{\alpha}\right)^2 \Big(1-(1-\alpha)^{h-1}\Big),\end{aligned}$$

we have

$$\left(\alpha + \frac{1-\alpha}{h}\left[1 + \frac{1-\alpha}{\alpha}\Big(1-(1-\alpha)^{h-1}\Big)\right]\right)\boldsymbol{s}_{t+h} = \alpha \boldsymbol{y}_{t+h} + \frac{1-\alpha}{h}\left[1 + \frac{1-\alpha}{\alpha}\Big(1-(1-\alpha)^{h-1}\Big)\right]\boldsymbol{s}_t.$$

4.E Proof of Proposition 4.12

Let $f_L : x \in [0,\pi] \mapsto \left[\left(\|\boldsymbol{\mu}\|^2 + 1\right)\dfrac{\operatorname{sinc}(\frac{x}{2})}{\operatorname{sinc}((2L-1)\frac{x}{2})} - 1\right]^{-2}$. Then,

$$\sum_{l=1}^{\lfloor n/2 \rfloor} \left[\left(\|\boldsymbol{\mu}\|^2 + 1\right)\frac{\psi_0}{\psi_l} - 1\right]^{-2} = \sum_{l=1}^{\lfloor n/2 \rfloor} f_L\left(\frac{2l\pi}{n}\right).$$

For all $\delta \in]\frac{1}{2}, \frac{2}{3}[$, we can split this sum into

$$\begin{aligned}\sum_{l=1}^{\lfloor n/2 \rfloor} f_L\left(\frac{2l\pi}{n}\right) &= \sum_{l=1}^{\lfloor n^\delta/2 \rfloor} f_L\left(\frac{2l\pi}{n}\right) + \sum_{l=\lfloor n^\delta/2 \rfloor + 1}^{\lfloor n/2 \rfloor} f_L\left(\frac{2l\pi}{n}\right) \\ &= \sum_{l=1}^{\lfloor n^\delta/2 \rfloor} \left[\frac{\|\boldsymbol{\mu}\|^2 + 1}{\operatorname{sinc}((2L-1)\frac{l\pi}{n})}\left[1 + \mathcal{O}\left(\left(\frac{l\pi}{n}\right)^2\right)\right] - 1\right]^{-2} + \sum_{l=\lfloor n^\delta/2 \rfloor + 1}^{\lfloor n/2 \rfloor} f_L\left(\frac{2l\pi}{n}\right) \\ &= \sum_{l=1}^{\lfloor n^\delta/2 \rfloor} \left[\frac{\|\boldsymbol{\mu}\|^2 + 1}{\operatorname{sinc}((2L-1)\frac{l\pi}{n})} - 1\right]^{-2} + \sum_{l=1}^{\lfloor n^\delta/2 \rfloor} \mathcal{O}\left(\left(\frac{l\pi}{n}\right)^2\right) + \sum_{l=\lfloor n^\delta/2 \rfloor + 1}^{\lfloor n/2 \rfloor} f_L\left(\frac{2l\pi}{n}\right) \\ &= \sum_{l=1}^{\lfloor n^\delta/2 \rfloor} \left[\frac{\|\boldsymbol{\mu}\|^2 + 1}{\operatorname{sinc}((2L-1)\frac{l\pi}{n})} - 1\right]^{-2} + \mathcal{O}\left(n^{3\delta-2}\right) + \sum_{l=\lfloor n^\delta/2 \rfloor + 1}^{\lfloor n/2 \rfloor} f_L\left(\frac{2l\pi}{n}\right)\end{aligned}$$

where we have used the Taylor expansions $\operatorname{sinc} x = 1 + \mathcal{O}(x^2)$ and $(1+x)^{-2} = 1 + \mathcal{O}(x)$ as $x \to 0$. In order to control the last sum, notice that, for all $x \in]0, \pi]$, we have,

$$f_L(x) = \left[\left(\|\boldsymbol{\mu}\|^2 + 1 \right) - \frac{\operatorname{sinc}((2L-1)\frac{x}{2})}{\operatorname{sinc}(\frac{x}{2})} \right]^{-2} \left[\frac{\operatorname{sinc}(\frac{x}{2})}{\operatorname{sinc}((2L-1)\frac{x}{2})} \right]^{-2}$$

and, since $\frac{\operatorname{sinc}((2L-1)\frac{x}{2})}{\operatorname{sinc}(\frac{x}{2})} = \frac{\nu_L(x)}{2L-1} \leqslant 1$,

$$f_L(x) \leqslant \frac{1}{\|\boldsymbol{\mu}\|^4} \left[\frac{\sin((2L-1)\frac{x}{2})}{(2L-1)\sin(\frac{x}{2})} \right]^2 \leqslant \frac{1}{\|\boldsymbol{\mu}\|^4 (2L-1)^2 \sin^2(\frac{x}{2})}.$$

Hence,

$$\sum_{l=\lfloor n^\delta/2 \rfloor + 1}^{\lfloor n/2 \rfloor} f_L\left(\frac{2l\pi}{n}\right) \leqslant \frac{\lfloor n/2 \rfloor - \lfloor n^\delta/2 \rfloor}{\|\boldsymbol{\mu}\|^4 (2L-1)^2 \sin^2(\frac{n^\delta \pi}{2n})} = \frac{\lfloor n/2 \rfloor - \lfloor n^\delta/2 \rfloor}{\|\boldsymbol{\mu}\|^4 (2L-1)^2 \left(\frac{n^\delta \pi}{2n} + \mathcal{O}(n^{3(1-\delta)}) \right)^2} = \mathcal{O}\left(n^{1-2\delta} \right).$$

Finally, with $\delta = \frac{3}{5}$, we have,

$$\sum_{l=1}^{\lfloor n/2 \rfloor} \left[\left(\|\boldsymbol{\mu}\|^2 + 1 \right) \frac{\psi_0}{\psi_l} - 1 \right]^{-2} = \sum_{l=1}^{\lfloor n/2 \rfloor} \left[\frac{\|\boldsymbol{\mu}\|^2 + 1}{\operatorname{sinc}((2L-1)\frac{l\pi}{n})} - 1 \right]^{-2} + \mathcal{O}\left(n^{-1/5} \right).$$

Part II

Tensor Unfolding Approaches: Insights Into the Computational Limits

Chapter 5

A Random Matrix Approach to Low-Multilinear-Rank Tensor Approximation

INFORMATION retrieval from large amounts of data has become a common task of signal processing and machine learning in the past decades. Often, these data have several modes as they may come from various sources, modalities, domains, and so on. Tensors (multi-way arrays) are therefore a natural representation for such datasets — they appear in multiple areas such as brain imaging (Zhou et al., 2013), neurophysiological measurements (Seely et al., 2016), community detection (Anandkumar et al., 2013), compression of hyperspectral images (Li and Li, 2010), spatio-temporal gene expression (Liu et al., 2022), recommender systems (Karatzoglou et al., 2010; Rendle and Schmidt-Thieme, 2010; Frolov and Oseledets, 2017) and topic modeling (Anandkumar et al., 2014). Indeed, tensors as multi-way arrays provide a more detailed representation of data than mere matrices (two-way arrays) as they convey a structural information. For instance, the modes of a data tensor can represent *pixel* × *pixel* × *wavelength* × *sample* in hyperspectral imaging (Zhang et al., 2013; Kanatsoulis et al., 2018), *time* × *spatial scale* × *electrode* in the EEG analysis by Acar et al. (2007) or *neuron* × *time* × *stimuli* in the study of the visual cortex by Rabinowitz et al. (2015).

In an information retrieval context, it is common to make use of tensor decompositions in order to estimate a sought signal. In their fMRI study, Hunyadi et al. (2017) perform a blind source separation via a joint tensor decomposition on a *channel* × *time* × *patient* array, whereas Williams et al. (2018) use a low-rank tensor approximation on a *neuron* × *time* × *trial* array as a dimensionality reduction technique to study neural dynamics. In fact, supposing that the signal has a low-rank structure is a natural sparsity assumption (Kadmon and Ganguli, 2018; Anandkumar et al., 2014), and low-rank tensor approximations are key tools to extract information from multi-way data.

In the present chapter, we propose a random matrix analysis of a general *low-rank information* + *noise* tensor model and precisely quantify the amount of information which can be recovered with a low-rank tensor approximation depending on the signal-to-noise ratio (SNR). For a more general introduction to tensors than the elements we gave in Section 2.4, we refer the reader to Comon (2014, 2009); Landsberg (2011); Hackbusch (2012) and, for an emphasis on statistical learning applications, Bi et al. (2021); Sun et al. (2021). In the remainder of the introduction, the main concepts and challenges behind low-rank tensor estimation are presented in Section 5.0.1. Then, Section 5.0.2 introduces some important related works. Our main results are finally summarized in Section 5.0.3. Recall that tensor-related notations, operations and decompositions are properly defined in Section 2.4.

5.0.1 Low-Rank Tensor Estimation

What is meant by a low-rank approximation of a tensor? And how is the *rank* of a tensor actually defined? Let us start with a familiar matrix case: a matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ is a two-way array (or order-2 tensor). A singular value decomposition (SVD) allows us to write $\boldsymbol{M}$ in a compact way as the sum of R rank-1 terms, $\boldsymbol{M} = \sum_{i=1}^{R} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ where $\boldsymbol{U}$, $\boldsymbol{V}$ are respectively $n_1 \times R$ and $n_2 \times R$ matrices with *orthonormal* columns and $\boldsymbol{\Sigma}$ is the $R \times R$ *diagonal* matrix of singular values. The rank of $\boldsymbol{M}$ is here the minimal number of rank-1 terms in which it can be exactly decomposed. Extending this notion to tensors therefore seems straightforward: a tensor[1] $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ has rank R if it is the minimal number of rank-1 terms in which it can be exactly decomposed, $\mathcal{T} = \sum_{i=1}^{R} \sigma_i \boldsymbol{a}_i \otimes \boldsymbol{b}_i \otimes \boldsymbol{c}_i$. What we have just described is the canonical polyadic decomposition (CPD) of $\mathcal{T}$, it dates back to Hitchcock (1927) and is unique under very mild conditions (Kolda and Bader, 2009). However, we have lost an important property in this process: the unit vectors $\boldsymbol{a}_i$ (resp. $\boldsymbol{b}_i$, $\boldsymbol{c}_i$) are, in general, no longer orthonormal. Conversely, retaining the orthonormality property inevitably results in the loss of the diagonality property, $\mathcal{T} = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^{r_3} \mathcal{G}_{i,j,k} \boldsymbol{u}_i \otimes \boldsymbol{v}_j \otimes \boldsymbol{w}_k$. This latter decomposition is called a Tucker decomposition and dates back to Tucker (1966). In fact, the best way to represent $\mathcal{T}$ with a Tucker decomposition is to choose the $\boldsymbol{u}_i$ (resp. $\boldsymbol{v}_i$, $\boldsymbol{w}_i$) as the left singular vectors of the unfolding of $\mathcal{T}$ along mode 1 (resp. 2, 3)[2]. This is called the multilinear SVD (MLSVD, De Lathauwer et al., 2000b) and gives rise to a new definition of rank: the *multilinear*-rank (r_1, r_2, r_3). Note that, in the matrix case, $r_1 = r_2 = R$ since both the diagonality and orthonormality properties are verified. However, r_1, r_2, r_3 are, in general, not equal in the tensor case, but $\max(r_1, r_2, r_3) \leqslant R \leqslant \min(r_1 r_2, r_2 r_3, r_1 r_3)$ (Equation (2.12)). See, e.g., Sidiropoulos et al. (2017) for details. Other relevant references for the reader interested in tensor decompositions are Kolda and Bader (2009); Cichocki et al. (2015); Rabanser et al. (2017).

Given an order-d tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ of possibly very high rank, we are interested in finding a low-rank approximation, i.e., an $n_1 \times \ldots \times n_d$ tensor $\mathcal{X}$ which minimizes the distance $\|\mathcal{T} - \mathcal{X}\|_{\mathrm{F}}$ on a set of low-rank tensors. Yet, the problem of the best rank-R approximation of a tensor is ill-posed as soon as $R > 1$ because the set of rank-R tensors is not closed (Kolda and Bader, 2009). Instead, we shall consider the best low-*multilinear*-rank problem, which is always well-posed,

$$\min_{\mathrm{Rank}(\mathcal{X}) \leqslant (r_1, \ldots, r_d)} \|\mathcal{T} - \mathcal{X}\|_{\mathrm{F}}^2. \tag{5.1}$$

It is well known in the matrix case that the best rank-R approximation can be easily computed by truncating the SVD to its R most energetic terms (Eckart and Young, 1936; Mirsky, 1960). Could this also be true for the MLSVD? Unfortunately, counter-examples exist (Kolda, 2003), showing that there is no tensor equivalent of the Eckart-Young-Mirsky theorem. Worse still, Problem (5.1) is in fact NP-hard (Hillar and Lim, 2013). Nevertheless, despite not being the best low-multilinear-rank approximation, the truncated MLSVD $\hat{\mathcal{T}}$ remains a very good "first guess" as it verifies $\|\mathcal{T} - \mathcal{T}_\star\|_{\mathrm{F}} \leqslant \|\mathcal{T} - \hat{\mathcal{T}}\|_{\mathrm{F}} \leqslant \sqrt{d}\|\mathcal{T} - \mathcal{T}_\star\|_{\mathrm{F}}$ where $\mathcal{T}_\star$ denotes a solution to Problem (5.1) and d is the order of the tensor (Grasedyck et al., 2013; Hackbusch, 2012). It is a cheap (it consists only in d standard matrix SVDs) and quasi-optimal low-multilinear-rank approximation of $\mathcal{T}$. Moreover, it is often used as an initialization of numerical methods which estimate a solution to Problem (5.1), among which the most common is the higher-order orthogonal iteration (HOOI) algorithm (Kroonenberg and de Leeuw, 1980; Kapteyn et al., 1986; De Lathauwer et al., 2000a).

Another motivation for the analysis of the low-multilinear-rank approximation problem is that it has also a practical interest for the numerical computation of the canonical polyadic decomposition (CPD). Indeed, when dealing with large tensors, it is computationally more efficient to first compress the tensor with a low-multilinear-rank approximation and then compute the CPD on the smaller core

[1] It is chosen of order 3 for simplicity of exposure.

[2] This is properly defined in Section 2.4.2.

tensor rather than computing the CPD of the large tensor directly (Bro and Andersson, 1998). This is done, e.g., by the `cpd` function of the `MATLAB` toolbox `Tensorlab` (Vervliet et al., 2016).

5.0.2 Related Work

Multilinear SVD (MLSVD) has a wide range of applications and is often used to extract relevant information from multi-way arrays. For instance, it has been used in human motion recognition (Vasilescu, 2002), face recognition (Vasilescu and Terzopoulos, 2003), handwritten digit classification (Savas and Eldén, 2007) but also genomics (Omberg et al., 2007, 2009; Muralidhara et al., 2011) and syndromic surveillance (Fanaee-T and Gama, 2015).

The analysis of *spiked* tensor models — i.e., low-rank perturbations of large random tensors — has started with the introduction by Richard and Montanari (2014) of the rank-1 symmetric spiked tensor model, $\mathcal{T} = \beta \boldsymbol{x}^{\otimes d} + \mathcal{N}$ with $\|\boldsymbol{x}\| = 1$, $\mathcal{N}$ Gaussian noise and β a parameter controlling the signal-to-noise ratio (SNR). They show that estimation of $\boldsymbol{x}$ from $\mathcal{T}$ is *theoretically* possible as soon as β is above a certain threshold β_c behaving like $\sqrt{d \log d}$, which is reminiscent of the now well-known spiked *matrix* model where signal reconstruction is only possible above a critical threshold (Péché, 2006) — a phenomenon called the BBP phase transition (Baik et al., 2005). The behavior of singular values and singular vectors of spiked matrix models is comprehensively studied by Benaych-Georges and Nadakuditi (2012). Contrary to the matrix case however, Richard and Montanari (2014) make the disturbing observation that none of the polynomial-time estimation algorithms among tensor unfolding, power iteration and approximate message passing (AMP) succeed unless β diverges as the dimensions of the tensor grow large. The results of Hopkins et al. (2015, 2017) suggest that no polynomial-time algorithm can succeed unless $\beta \gtrsim N^{\frac{d-2}{4}}$, where N scales as the dimensions of the data tensor. While Perry et al. (2020) show that the information-theoretic threshold is of order 1, this indicates the existence of a *computational-to-statistical gap* in spiked tensor estimation, as in a myriad of other problems (Bandeira et al., 2018; Zdeborová and Krzakala, 2016).

The landscape of the rank-1 symmetric spiked tensor model is studied by Ben Arous et al. (2019), who show that the number of local optima to Problem (5.1) grows exponentially with the size of the tensor, but all lie close to a subspace orthogonal to the sought solution, except for one if β exceeds a critical threshold β_c. Completing this analysis, Jagannath et al. (2020) show[3] the existence of two close but different thresholds $\beta_s < \beta_c$ such that the solution aligned with the underlying signal is a local minimum of Problem (5.1) as soon as $\beta > \beta_s$ but becomes a *global* one only if $\beta > \beta_c$. Relying on the Kac-Rice method, Ros et al. (2019) thoroughly study such high-dimensional landscapes and classify the different behaviors and phase transitions which can occur.

So far, we have only referred to works dealing with the rank-1 symmetric case, but there are also some studies on higher-(low-)rank spiked models. Chevreuil and Loubaton (2018) give a sufficient (but not necessary) condition for the *non-detectability* of a rank-R asymmetric signal perturbed by an additive Gaussian noise. Chen et al. (2021) also discuss signal detectability in the rank-R symmetric case. The statistical inference of finite rank tensors is studied by Chen et al. (2022) who identify the limit free energy of the model in terms of a variational formula. Zhang and Xia (2018) consider a general *low-multilinear-rank signal* $\mathcal{P}$ + *Gaussian noise* $\mathcal{N}$ model and bring to light the same statistical-to-computational gap: if $\|\mathcal{P}\|_{\mathrm{F}}$ is above a statistical threshold of order 1 then Problem (5.1) has a solution which is aligned with the signal but is computationally intractable unless $\|\mathcal{P}\|_{\mathrm{F}}$ is above a computational threshold of order $N^{\frac{d-2}{4}}$. In this strong SNR regime, the higher-order orthogonal iteration (HOOI) algorithm (De Lathauwer et al., 2000a) is minimax-optimal. In fact, it is also proved by Ben Arous et al. (2020) that, with Langevin dynamics and gradient descent, the algorithmic threshold behaves like N^{α} with $\alpha > 0$. Other algorithmic thresholds have been shown as well for semi-

[3]The setting considered by Jagannath et al. (2020) is more general than the one of Ben Arous et al. (2019) because the noise in their model is not necessarily symmetric but the perturbation is still a rank-1 symmetric tensor.

definite and spectral relaxations of the maximum likelihood problem (Hopkins et al., 2015, 2016; Kim et al., 2017). AMP and tensor power iteration algorithms achieve $N^{\frac{d-1}{2}}$ (Lesieur et al., 2017; Huang et al., 2022) while tensor unfolding methods (truncated MLSVD and HOOI algorithm) achieve $N^{\frac{d-2}{4}}$ as already conjectured by Richard and Montanari (2014) and later proven by Hopkins et al. (2015); Ben Arous et al. (2023) in the rank-1 case. The convergence of the HOOI algorithm towards a local maximum for a *sufficiently close* initialization is proven by Xu (2018) and Feldman and Donoho (2023) show that it achieves exact recovery of a rank-1 perturbation in the large N regime when it is initialized with the dominant singular vectors of the unfoldings.

Recently, a new approach relying on tools from random matrix theory has broaden the understanding of spiked tensor models. In particular, Goulart et al. (2022) study the rank-1 symmetric case and are able to recover explicitly the same β_s threshold as Jagannath et al. (2020) as well as to precisely quantify the alignment between a solution to Problem (5.1) and the signal. A similar analysis is carried out by Seddik et al. (2022) for the more general asymmetric case, relying solely on classical techniques from random matrix theory. Such tools show promise for the theoretical understanding of learning from tensor data (Seddik et al., 2023b). In particular, the results of Ben Arous et al. (2023) and Feldman (2023) on *long* random matrices, similar to those considered in this work, provide instructive insight into the recovery performance of tensor unfolding methods.

5.0.3 Summary of Contributions

In low-rank tensor approximation, tensor unfolding methods achieve the best known performance among polynomial-time algorithms. Motivated by several works suggesting that such method could actually reach the computational threshold (Hopkins et al., 2015, 2017; Zhang and Xia, 2018; Wein et al., 2019), we propose a thorough random matrix analysis of the low-multilinear-rank tensor approximation problem.

Consider the general spiked tensor model,

$$\mathcal{T} = \mathcal{P} + \frac{1}{\sqrt{N}}\mathcal{N} \quad \in \mathbb{R}^{n_1 \times \ldots \times n_d}, \qquad \mathcal{N}_{i_1,\ldots,i_d} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{5.2}$$

where $d \geqslant 3$ is the order of the tensor, $\mathcal{N}$ is an additive Gaussian noise, N is a parameter controlling the size of the tensor such that the ratio n_ℓ / N is constant[4] (at least for N above a certain threshold value N_0) for all $\ell \in \{1,\ldots,d\}$ (for instance, $N = n_1$ or $N = \sum_{\ell=1}^{d} n_\ell$) and $\mathcal{P}$ is a low-multilinear-rank deterministic tensor, i.e., which can be decomposed as

$$\mathcal{P} = \sum_{q_1=1}^{r_1} \ldots \sum_{q_d=1}^{r_d} \mathcal{H}_{q_1,\ldots,q_d} [\boldsymbol{x}_{q_1}^{(1)} \otimes \ldots \otimes \boldsymbol{x}_{q_d}^{(d)}] \overset{\text{def}}{=} \left[\!\!\left[\mathcal{H}; \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(d)} \right]\!\!\right], \tag{5.3}$$

with $\mathcal{H} \in \mathbb{R}^{r_1 \times \ldots \times r_d}$ and $\boldsymbol{X}^{(\ell)}$ an $n_\ell \times r_\ell$ matrix with orthonormal columns $\boldsymbol{x}_{q_\ell}^{(\ell)}$. The range of $\boldsymbol{X}^{(\ell)}$ is the ℓ-th singular subspace of $\mathcal{P}$. This decomposition is illustrated for the case $d = 3$ in Figure 5.1. Model (5.2) with $\mathcal{P}$ as in Equation (5.3) is the most general spiked tensor model — i.e., low-rank perturbation of a large random tensor. Indeed, any of the models referred to in the previous Section 5.0.2 fall into this definition since decomposition (5.3) always exists and low CP-rank is equivalent to low multilinear rank thanks to the inequality $\max_\ell \{r_\ell\} \leqslant R \leqslant \min_\ell \{\prod_{\ell' \neq \ell} r_{\ell'}\}$ (Equation (2.12)).

In the regime where $N \to +\infty$ — representing the fact that, in practice, the dimensions of the tensor are large compared to its rank —, we study the estimation of $\mathcal{P}$ from $\mathcal{T}$ with a truncated MLSVD,

[4] This ensures that the spectral norm of $\frac{1}{\sqrt{N}}\mathcal{N}$ is of order 1 (Tomioka and Suzuki, 2014).

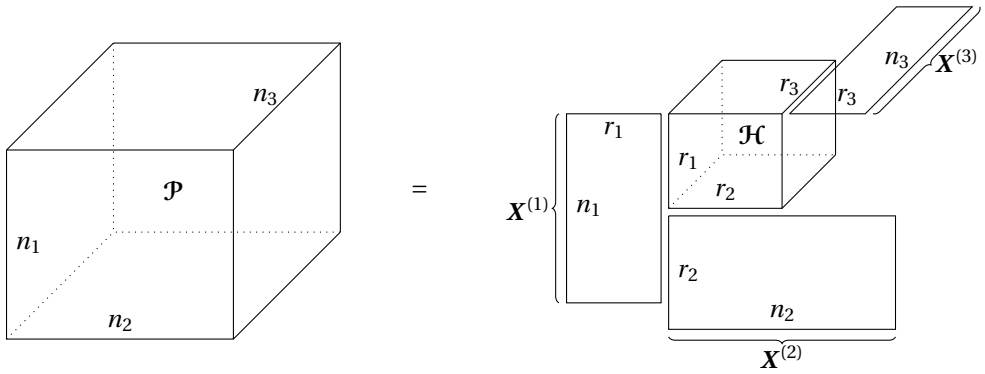


Figure 5.1: Illustration of the Tucker decomposition (5.3) of an $n_1 \times n_2 \times n_3$ tensor $\mathcal{P}$ with multilinear rank (r_1, r_2, r_3). $\mathcal{H}$ is the $r_1 \times r_2 \times r_3$ core tensor and $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \boldsymbol{X}^{(3)}$ are matrices with orthonormal columns spanning the singular subspaces of $\mathcal{P}$.

which serves as initialization of the HOOI algorithm. In particular, we reveal that the interesting *non-trivial* regime is characterized by the $\Theta(N^{\frac{d-2}{2}})$ quantity[5] $\sigma_N = \frac{1}{N}\prod_{\ell=1}^{d}\sqrt{n_\ell}$.

- If $\|\mathcal{P}\|_{\mathrm{F}}^2/\sigma_N \xrightarrow[N\to+\infty]{} 0$, then the noise completely masks the signal, and truncated MLSVD fails to recover $\mathcal{P}$.
- If $\|\mathcal{P}\|_{\mathrm{F}}^2/\sigma_N \xrightarrow[N\to+\infty]{} +\infty$, then the signal clearly stands out from the noise, and reconstruction of $\mathcal{P}$ with a truncated MLSVD is easy.
- If $\|\mathcal{P}\|_{\mathrm{F}}^2/\sigma_N = \Theta(1)$ as $N \to +\infty$, then we are precisely in the *non-trivial* regime between the two previous situations, and truncated MLSVD may partially recover $\mathcal{P}$.

It is the analysis of this last regime which is of practical interest. Given the low-multilinear-rank approximation $\hat{\mathcal{T}} = [\![\hat{\mathcal{G}}; \hat{\boldsymbol{U}}^{(1)}, \ldots, \hat{\boldsymbol{U}}^{(d)}]\!]$ obtained with a truncated MLSVD of $\mathcal{T}$, we quantify how well $\hat{\mathcal{T}}$ reconstructs $\mathcal{P}$ in this non-trivial regime. To do so, we study the spectral properties of the unfoldings (i.e., matricizations) of the tensor $\mathcal{T}$, i.e., the $n_\ell \times \prod_{\ell' \neq \ell} n_{\ell'}$ matrices $\boldsymbol{T}^{(\ell)}$ whose columns are mode-ℓ fibers of $\mathcal{T}$ and the columns of $\hat{\boldsymbol{U}}^{(\ell)}$ are its dominant left singular vectors. Such *long* matrices (the second dimension grows faster than the first one) have already been studied by Ben Arous et al. (2023) in order to analyze the properties of tensor-unfolding methods in the particular setting of a rank-1 spike. Here, we tackle this problem with a highly different approach relying solely on classical tools from random matrix theory (Couillet and Liao, 2022) and give very general results on the spiked tensor model that go beyond the specific rank-1 one case. Moreover, we justify the practical use of truncated MLSVD as an initialization of the HOOI algorithm by showing its optimality in the large N regime.

Although the spectrum of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$, $\ell \in \{1, \ldots, d\}$, diverges as $N \to +\infty$, we show that its eigenvalues (i.e., the squared singular values of $\boldsymbol{T}^{(\ell)}$) gather in an interval $[\mu_N^{(\ell)} \pm 2\sigma_N]$ with $\mu_N^{(\ell)} = \frac{1}{N}\prod_{\ell' \neq \ell} n_{\ell'} = \Theta(N^{d-2})$. More precisely, the empirical spectral distribution of the centered-and-scaled matrix $\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$ converges weakly to the semicircle distribution on $[-2, 2]$ (see Figure 5.3, Theorem 5.1 and Corollary 5.2). Furthermore, we show that a BBP phase transition phenomenon occurs: for each singular value of $\boldsymbol{P}^{(\ell)}$ (the unfolding of $\mathcal{P}$ along mode ℓ) which is above the threshold $\sqrt{\sigma_N}$, an

[5] Note that this corresponds to the regime of the algorithmic threshold but, here, our measure of the signal power is $\|\mathcal{P}\|_{\mathrm{F}}^2$ (and not $\|\mathcal{P}\|_{\mathrm{F}}$) hence the $\Theta(N^{\frac{d-2}{2}})$ instead of $\Theta(N^{\frac{d-2}{4}})$.

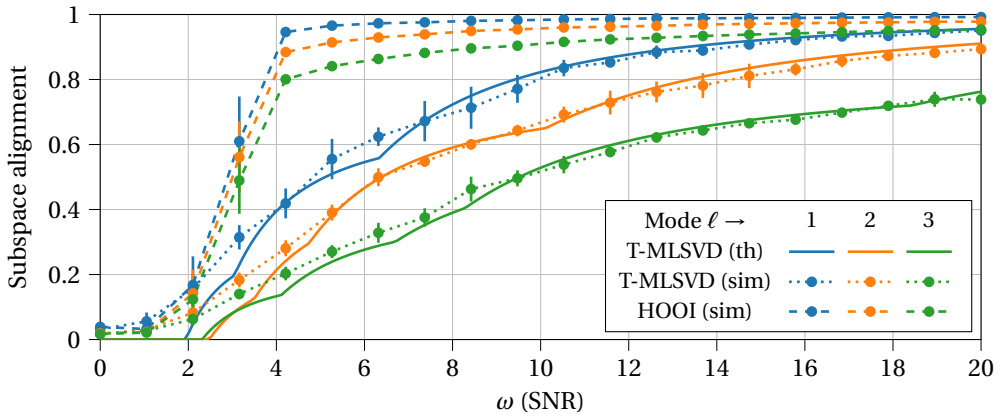


Figure 5.2: Alignments between singular subspaces (see Section 5.1.2) of the observation $\mathcal{T} = \sqrt{\omega}\mathcal{P}_\circ + \frac{1}{\sqrt{N}}\mathcal{N}$ and of the signal $\mathcal{P}_\circ$, with $\|\mathcal{P}_\circ\|_{\mathrm{F}}^2 = \frac{\sqrt{n_1 n_2 n_3}}{N}$, as a function of the signal-to-noise ratio ω. Theoretical alignments (Theorem 5.5) achieved with truncated MLSVD are compared with simulations and those achieved with the HOOI algorithm. Empirical results are averaged over 10 trials, with error bars representing standard deviation. **Experimental setting:** $d = 3$, $(n_1, n_2, n_3) = (100, 200, 300)$, $N = n_1 + n_2 + n_3$ and $(r_1, r_2, r_3) = (3, 4, 5)$.

eigenvalue of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ isolates itself on the right side of the *bulk* (see Figure 5.3) and its corresponding eigenvector (i.e., left singular vector of $\boldsymbol{T}^{(\ell)}$) is aligned with the corresponding singular subspace of $\mathcal{P}$. The position of the isolated eigenvalue and this alignment are efficiently predicted by Theorem 5.5 (see also Figure 5.3).

As a result, Figure 5.2 plots, for an order-3 tensor, as a function of the signal-to-noise ratio (SNR) $\omega = \|\mathcal{P}\|_{\mathrm{F}}^2/\sigma_N$, the alignments between the singular subspace of the signal $\mathcal{P}$ spanned by $\boldsymbol{X}^{(\ell)}$ and the dominant singular subspace of the observation $\mathcal{T}$ spanned by $\hat{\boldsymbol{U}}^{(\ell)}$. Solid curves are the alignments predicted by Theorem 5.5 while dotted curves are empirical alignments computed on a $100 \times 200 \times 300$ tensor with signal-rank $(3, 4, 5)$. If the SNR ω is too small, there is no alignment, meaning that truncated MLSVD fails to recover $\mathcal{P}$ — the signal is masked by the noise. When it exceeds a critical value (see Theorem 5.5 and Section 5.1.2 for details), a phase transition phenomenon occurs[6]: the alignment starts to grow — i.e., truncated MLSVD now partially recovers $\mathcal{P}$ — and converges to 1 as $\omega \to +\infty$.

Besides, Figure 5.2 also plots the empirical alignments between the singular subspaces of $\mathcal{P}$ and those estimated with the HOOI algorithm (De Lathauwer et al., 2000a) given in Algorithm 5, whose truncated MLSVD serves as initialization. This yields much better alignments, especially close to the phase transition. In fact, we show in Theorem 5.11 that the HOOI algorithm converges to a solution to Problem (5.2) as soon as its initialization sufficiently preserves the underlying signal. This provides new insight into the computational barrier: initialization is the limiting factor here. Had one prior information on the solution, one could initialize the HOOI algorithm in the right basin of attraction and still be able to perfectly (i.e., with alignment 1) reconstruct the signal in the regime $1 \ll \|\mathcal{P}\|_{\mathrm{F}} \ll N^{\frac{d-2}{4}}$, which is computationally hard but statistically easy (see details in Section 5.2.1 and discussion in Section 5.2.2).

In a nutshell, our contributions can be summarized as follows.

[6] In fact, we will see in Section 5.1.2 that there is one phase transition for each principal direction of the singular subspaces of $\mathcal{P}$, resulting in $\sum_{\ell=1}^{d} r_\ell$ phase transitions. Their positions corresponds to sudden changes of slope in the solid curves of Figure 5.2.

- We characterize, in the large N limit, the behavior of the singular values of the unfoldings of the tensor $\mathcal{T}$ — denoted $\boldsymbol{T}^{(\ell)}$, $\ell \in \{1,\dots,d\}$ — when it follows the general spiked tensor model (5.2) (Theorem 5.1 and Corollary 5.2). This is performed through the analysis of the limiting spectral distribution of the symmetric matrix $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ using standard tools from the theory of large random matrices.

- We give a precise condition, depending on a $\Theta(N^{\frac{d-2}{4}})$ threshold on the signal, for the detectability of a principal direction of the ℓ-th singular subspace of $\mathcal{P}$ from the unfolding $\boldsymbol{T}^{(\ell)}$. This corresponds to the presence of an isolated eigenvalue in the spectrum of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ with associated eigenvector aligned with the sought singular subspace. We find similar formulae as Feldman (2023)[7] for the asymptotic position of this isolated eigenvalue, as well as the quality of the alignment (Theorem 5.5).

- Relying on our random matrix analysis, we characterize the performance of truncated MLSVD in the reconstruction of the signal $\mathcal{P}$ from the observation $\mathcal{T}$ (Section 5.1.2).

- We show that exact reconstruction of $\mathcal{P}$ from $\mathcal{T}$ is possible in the large N regime with the HOOI algorithm (De Lathauwer et al., 2000a) as long as $\|\mathcal{P}\|_{\mathrm{F}} \gg 1$ and it is initialized in the right bassin of attraction (Theorem 5.11). Without prior information, this depends on the detectability of $\mathcal{P}$ in the truncated MLSVD of $\mathcal{T}$, which is only possible above the $\Theta(N^{\frac{d-2}{4}})$ computational threshold. Moreover, as $N \to +\infty$, the number of iterations needed for the convergence of the algorithm converges to 1.

In section 5.1, we present the random matrix analysis of long matrices emerging from the unfoldings of tensors following the general spiked tensor model (5.2). These results are presented in the context of truncated MLSVD and exploited to quantitatively explain its reconstruction performances. Then, relying on these results, Section 5.2 deals with the numerical estimation of a solution to Problem (5.1) with the HOOI algorithm. We show its asymptotic optimality and provide insight into the limiting factors for numerical estimation below the computational threshold. We conclude and discuss our results in Section 5.3. Most proofs are deferred to the appendix.

5.1 Analysis of Truncated MLSVD under the General Spiked Tensor Model

This section presents a random matrix analysis of the general spiked tensor model introduced in Equation (5.2) using tools presented in Section 2.1. We give precise results on the spectral behavior of the unfoldings of the observed tensor $\mathcal{T}$, and specify the achievable performance in the estimation of the underlying signal $\mathcal{P}$ with a truncated MLSVD. Although, as explained in Section 5.0.1, this approach is only *quasi-optimal*, it is very easy to implement and represents an excellent "first guess" to initialize a numerical scheme converging to a solution to Problem (5.1), which is discussed in Section 5.2.

5.1.1 Random Matrix Results on the Model

Under the general spiked tensor model (5.2), we consider an $n_1 \times \ldots \times n_d$ tensor $\mathcal{T} = \mathcal{P} + \frac{1}{\sqrt{N}}\mathcal{N}$ of order $d \geqslant 3$, modeling a low-rank deterministic signal $\mathcal{P}$ corrupted by an additive Gaussian noise tensor

[7] Feldman (2023) studies the spiked model associated with long random matrices within a similar framework as Benaych-Georges and Nadakuditi (2011). Our approach is different in that it mostly relies on a *deterministic equivalent* which is introduced in Theorem 5.1.

$\mathcal{N}$ whose entries are independent $\mathcal{N}(0,1)$ random variables[8]. We denote by $(r_1,\dots,r_d)$ the multilinear rank of $\mathcal{P}$ and study this model in the asymptotic regime where $N \to +\infty$ with $n_\ell = \Theta(N)$ and $r_\ell = \Theta(1)$, $\ell \in [d]$.

The estimation of $\mathcal{P}$ with a truncated MLSVD on $\mathcal{T}$ is simply the computation of the dominant singular subspaces of $\mathcal{T}$. Specifically, $\hat{\boldsymbol{U}}^{(\ell)} \in \mathbb{R}^{n_\ell \times r_\ell}$ gathers the r_ℓ dominant left singular vectors of $\boldsymbol{T}^{(\ell)}$ — and thus, $\hat{\boldsymbol{U}}^{(\ell)\top}\hat{\boldsymbol{U}}^{(\ell)} = \boldsymbol{I}_{r_\ell}$. Then, a low-multilinear-rank approximation of $\mathcal{T}$ is $\hat{\mathcal{T}} = [\![\hat{\mathcal{G}}; \hat{\boldsymbol{U}}^{(1)},\dots,\hat{\boldsymbol{U}}^{(d)}]\!]$ with an $r_1 \times \dots \times r_d$ core tensor $\hat{\mathcal{G}} = \mathcal{T}(\hat{\boldsymbol{U}}^{(1)},\dots,\hat{\boldsymbol{U}}^{(d)})$. An equivalent expression explicitly showing that $\hat{\mathcal{T}}$ is the projection of $\mathcal{T}$ on its dominant singular subspaces is[9] $\hat{\mathcal{T}} = [\![\mathcal{T}; \hat{\boldsymbol{U}}^{(1)}\hat{\boldsymbol{U}}^{(1)\top},\dots,\hat{\boldsymbol{U}}^{(d)}\hat{\boldsymbol{U}}^{(d)\top}]\!]$. Thus, the quality of this estimation hinges upon the alignments between the singular subspaces of $\mathcal{P}$ and the dominant singular subspaces of $\mathcal{T}$. Namely, denoting $\hat{\boldsymbol{u}}_{q_\ell}^{(\ell)}$, for $q_\ell \in [r_\ell]$, the columns of $\hat{\boldsymbol{U}}^{(\ell)} = \begin{bmatrix} \hat{\boldsymbol{u}}_1^{(\ell)} & \dots & \hat{\boldsymbol{u}}_{r_\ell}^{(\ell)} \end{bmatrix}$ and given that $\mathcal{P} = [\![\mathcal{H}; \boldsymbol{X}^{(1)},\dots,\boldsymbol{X}^{(d)}]\!]$, the quantities of interest are $\|\boldsymbol{X}^{(\ell)\top}\hat{\boldsymbol{u}}_{q_\ell}^{(\ell)}\|^2$ since they represent how much of $\hat{\boldsymbol{u}}_{q_\ell}^{(\ell)}$ is in the ℓ-th singular subspace of the signal $\mathcal{P}$.

In order to understand how the singular subspaces of $\mathcal{P}$ are perturbed by the addition of noise, we study the spectral properties of the unfoldings $\boldsymbol{T}^{(\ell)} = \boldsymbol{P}^{(\ell)} + \frac{1}{\sqrt{N}}\boldsymbol{N}^{(\ell)}$. In fact, since we are only interested in the *left* singular vectors of $\boldsymbol{T}^{(\ell)}$, it is more convenient to consider the $n_\ell \times n_\ell$ *symmetric* matrix $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$. Note that this is different from standard spiked matrix models (Benaych-Georges and Nadakuditi, 2011) because the second dimension of $\boldsymbol{T}^{(\ell)}$ grows at a faster polynomial rate than the first one ($\Theta(N^{d-1})$ versus $\Theta(N)$). Hence, it is easy to see that the spectrum of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ should diverge as $N \to +\infty$: set $\mathcal{P} = \boldsymbol{0}_{n_1\times\dots\times n_d}$ for simplicity and consider the expected mean of the eigenvalues,

$$\frac{1}{n_\ell}\mathbb{E}\left[\sum_{\lambda\in\mathrm{Sp}\,\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}}\lambda\right] = \frac{1}{n_\ell}\mathbb{E}\left[\mathrm{Tr}\left(\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}\right)\right] = \frac{1}{N}\prod_{\ell'\neq\ell} n_{\ell'} \xrightarrow[N\to+\infty]{} +\infty.$$

Hence, we need to consider instead a *centered-and-scaled* version of our random matrix $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ to properly study the behavior of its spectrum. The quantities $\mu_N^{(\ell)}$ and σ_N introduced in Theorem 5.1 below are such that the eigenvalues of $\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$ neither diverge nor vanish but stay at a $\Theta(1)$ scale as $N \to +\infty$.

Theorem 5.1 (Deterministic equivalent)**.** *For $\ell \in [d]$, define the following quantities,*

$$\mu_N^{(\ell)} = \frac{1}{N}\prod_{\ell'\neq\ell} n_{\ell'}, \qquad \sigma_N = \frac{1}{N}\sqrt{\prod_{\ell\in[d]} n_\ell}.$$

As $N \to +\infty$, if the ratio $\|\mathcal{P}\|_{\mathrm{F}}^2/\sigma_N$ is bounded, then the resolvent of the centered-and-scaled matrix $\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$ has the following deterministic equivalent (Definition 2.17), for all $\tilde{z} \in \mathbb{C}\setminus\mathbb{R}$,

$$\tilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) \overset{def}{=} \left(\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right] - \tilde{z}\boldsymbol{I}_{n_\ell}\right)^{-1} \longleftrightarrow \bar{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) \overset{def}{=} \left(\frac{1}{\sigma_N}\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top} + \frac{1}{\tilde{m}(\tilde{z})}\boldsymbol{I}_{n_\ell}\right)^{-1}$$

where $\tilde{m}(\tilde{z}) \overset{def}{=} \lim_{N\to+\infty}\frac{1}{n_\ell}\mathrm{Tr}\,\tilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z})$ does not depend on $\ell \in [d]$ and satisfies the following equation,

$$\tilde{m}^2(\tilde{z}) + \tilde{z}\tilde{m}(\tilde{z}) + 1 = 0. \tag{5.4}$$

Proof. See Appendix 5.A. □

[8] Regarding this *Gaussian* noise assumption, the universality result of Gurau (2014) shows that, as $N \to +\infty$, the distribution of a random tensor with i.i.d. entries has the same limit than that of a tensor with i.i.d. *Gaussian* entries. Moreover, recall our discussion in Section 2.2.3 on the generality of this assumption.

[9] In the more familiar matrix case ($d = 2$), the expression $[\![\boldsymbol{T}; \hat{\boldsymbol{U}}^{(1)}\hat{\boldsymbol{U}}^{(1)\top}, \hat{\boldsymbol{U}}^{(2)}\hat{\boldsymbol{U}}^{(2)\top}]\!]$ is equivalent to $\hat{\boldsymbol{U}}^{(1)}\hat{\boldsymbol{U}}^{(1)\top}\boldsymbol{T}\hat{\boldsymbol{U}}^{(2)}\hat{\boldsymbol{U}}^{(2)\top}$.

This theorem is fundamental. It gives a deterministic equivalent $\bar{\boldsymbol{Q}}^{(\ell)}(\tilde{z})$ of the resolvent of the centered-and-scaled matrix $\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$, which is our entry point into the precise characterization of its spectral behavior, following the approach presented in Section 2.3. First of all, notice that the "scaling parameters" $\mu_N^{(\ell)}$ and σ_N are respectively $\Theta(N^{d-2})$ and $\Theta(N^{\frac{d-2}{2}})$, meaning that the eigenvalues of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ grow at a speed N^{d-2} and spread over an interval whose length grows as $\sqrt{N^{d-2}}$ and does not depend on the mode ℓ.

Moreover, the relation given in Equation (5.4) characterizes $\tilde{m}$, the Stieltjes transform of the limiting spectral distribution (LSD) of $\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$. We see that this LSD is the same regardless of the low-rank perturbation $\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top}$, as it is expected that the perturbation should only cause the presence of a *finite* number of isolated eigenvalues in the empirical spectral distribution. Notice that, if $\boldsymbol{P}^{(\ell)} = \boldsymbol{0}_{n_\ell \times \prod_{\ell' \neq \ell} n_{\ell'}}$, we simply have $\bar{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) = \tilde{m}(\tilde{z})\boldsymbol{I}_{n_\ell}$. As a corollary, we recover the limiting spectral distribution of long random matrices, which was first characterized by Bai and Yin (1988).

Corollary 5.2 (Limiting spectral distribution)**.** *As $N \to +\infty$, the empirical spectral distribution of the centered-and-scaled matrix $\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$ converges weakly almost surely to μ_{SC}, the semicircle distribution on $[-2,2]$,*

$$\mathrm{d}\mu_{\mathrm{SC}}(x) \overset{def}{=} \frac{1}{2\pi}\sqrt{\left[4-x^2\right]^+}\,\mathrm{d}x.$$

Proof. Following Equation (5.4), $\tilde{m}(\tilde{z}) = \frac{1}{2}\left[-z \pm \sqrt{z^2-4}\right]$ where the $\pm$ sign is chosen so that $\tilde{m}$ satisfies the properties of a Stieltjes transform, in particular $\Im[\tilde{z}]\Im[\tilde{m}(\tilde{z})] > 0$ for all $\tilde{z} \in \mathbb{C}\setminus\mathbb{R}$. Then, the result follows from the Stieltjes transform inversion formula introduced in Proposition 2.11. □

This result states that the limiting spectral distribution of $\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$ is, in fact, the very-well-known semicircle distribution, first observed by Wigner (1955, 1958) in the study of certain special classes of random matrices arising in quantum mechanics. It indicates that, as $N \to +\infty$, the density of eigenvalues of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ is a stretched semicircle on $[\mu_N^{(\ell)} \pm 2\sigma_N]$. This phenomenon is illustrated in the first row of Figure 5.3, where the empirical spectral distribution (ESD) of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ is represented with the corresponding stretched semicircle for every mode ℓ of an order-3 tensor of size $300 \times 500 \times 700$ following the general spiked tensor model (5.2).

Remark 5.3 (From Marčenko-Pastur to Wigner). Given a random matrix $\boldsymbol{X} \in \mathbb{R}^{p_1 \times p_2}$ with i.i.d. $\mathcal{N}(0, \frac{1}{p_2})$ entries, it is well known that the empirical spectral distribution of $\boldsymbol{X}\boldsymbol{X}^\top$ converges weakly to the Marčenko-Pastur distribution as $p_1, p_2 \to +\infty$ with $p_1/p_2 = c > 0$ (Theorem 2.29). On the other hand, the standard semicircle distribution μ_{SC} is known to be the limiting spectral distribution of symmetric $p \times p$ random matrices with i.i.d. (up to symmetry) $\mathcal{N}(0, \frac{1}{p})$ entries (Theorem 2.25). Here, Corollary 5.2 shows that if p_2 grows at a faster polynomial rate than p_1, the matrix $\boldsymbol{X}\boldsymbol{X}^\top$ behaves asymptotically (up to a deterministic rescaling and shift) like a Wigner matrix, even if its entries are not independent. Experimentally, we observe that, if n_2 and n_3 are chosen small compared to n_1 (in contradiction with our assumption $n_1, n_2, n_3 = \Theta(N)$), e.g., $(n_1, n_2, n_3) = (1000, 40, 40)$, then the empirical spectral distribution of $\boldsymbol{T}^{(1)}\boldsymbol{T}^{(1)\top}$ is better modeled by a Marčenko-Pastur distribution than by a Wigner semicircle.

Remark 5.4 (Confinement of the spectrum). The weak convergence of the empirical spectral distribution to μ_{SC} stated in Corollary 5.2 could allow for a negligible amount of eigenvalues to stay outside the support of the limiting spectral distribution. In fact, in Appendix 5.A.5, we show an even more precise statement on the global behavior of the eigenvalues of $\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}$ (the model without signal): for all $\varepsilon > 0$, there exists an integer N_0 such that $\mathrm{Dist}(\frac{1}{\sigma_N}[\lambda - \mu_N^{(\ell)}], [-2,2]) \leqslant \varepsilon$ almost surely for all $\lambda \in \mathrm{Sp}\,\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}$ as soon as $N \geqslant N_0$. This means that no eigenvalue of $\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$ stays outside the support of the semicircle distribution $[-2,2]$ almost surely as $N \to +\infty$.

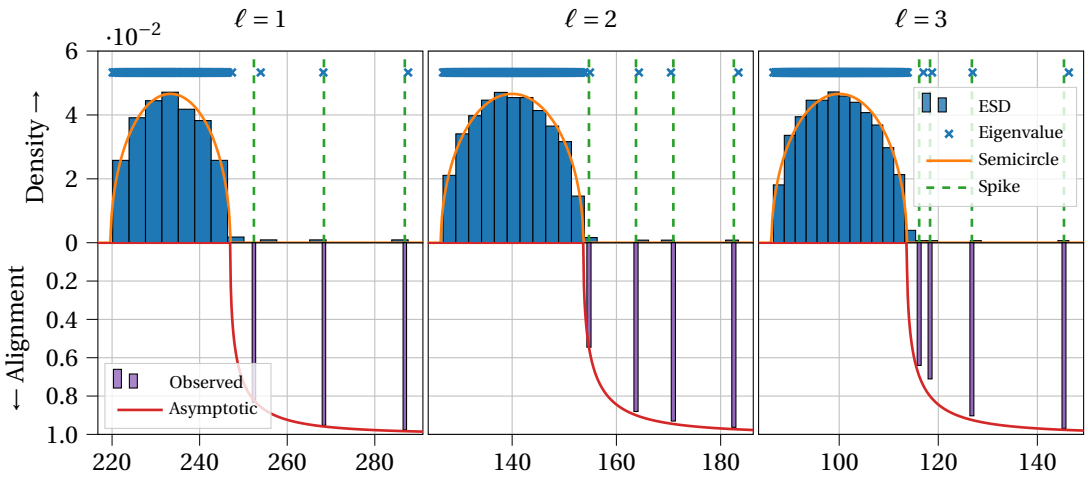


Figure 5.3: **Top:** empirical spectral distribution (ESD) of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$. The orange curve is the density of the stretched semicircle on $[\mu_N^{(\ell)} \pm 2\sigma_N]$ (Corollary 5.2). Green dashed lines represent asymptotic positions of spikes $\mu_N^{(\ell)} + \sigma_N \tilde{\xi}_{q_\ell}^{(\ell)}$ (Theorem 5.5). **Bottom:** Observed alignments between the dominant eigenvectors of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ and $\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top}$ (purple bars) with their predicted asymptotic values $[\zeta_{q_\ell}^{(\ell)}]^+$ (red curve, Theorem 5.5). **Experimental setting:** $d = 3$, $(n_1, n_2, n_3) = (300, 500, 700)$, $N = n_1 + n_2 + n_3$, $(r_1, r_2, r_3) = (3, 4, 5)$ and $\|\mathcal{P}\|_{\mathrm{F}}^2/\sigma_N = 15$.

The empirical spectral distributions of Figure 5.3 also show isolated eigenvalues on the right side of each semicircle. They are caused by the low-rank perturbation $\mathcal{P}$ which, in this setting, has multilinear-rank $(3, 4, 5)$. The estimate of $\mathcal{P}$ given by a truncated MLSVD on $\mathcal{T}$ has its singular subspaces spanned by the dominant eigenvectors of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$, i.e., precisely those associated with these isolated eigenvalues. Hence, a precise characterization of the behavior of these *spikes* is needed to plainly understand the recovery performance of this estimate. As explained in Section 2.1, this can be achieved with the deterministic equivalent given in Theorem 5.1.

Theorem 5.5 (Spike behavior)**.** *For $\ell \in [d]$ and $q_\ell \in [r_\ell]$, define the following quantities*[10]*,*

$$\rho_{q_\ell}^{(\ell)} = \frac{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})}{\sigma_N}, \qquad \tilde{\xi}_{q_\ell}^{(\ell)} = \rho_{q_\ell}^{(\ell)} + \frac{1}{\rho_{q_\ell}^{(\ell)}} \qquad \text{and} \qquad \zeta_{q_\ell}^{(\ell)} = 1 - \frac{1}{\left[\rho_{q_\ell}^{(\ell)}\right]^2}.$$

As $N \to +\infty$, if the ratio $\|\mathcal{P}\|_{\mathrm{F}}^2/\sigma_N$ is bounded and $\rho_{q_\ell}^{(\ell)} > 1$, then

$$\frac{1}{\sigma_N}\left[s_{q_\ell}^2(\boldsymbol{T}^{(\ell)}) - \mu_N^{(\ell)}\right] \xrightarrow{a.s.} \tilde{\xi}_{q_\ell}^{(\ell)} \qquad \text{and} \qquad \left\|\boldsymbol{X}^{(\ell)\top}\hat{\boldsymbol{u}}_{q_\ell}^{(\ell)}\right\|^2 \xrightarrow{a.s.} \zeta_{q_\ell}^{(\ell)}$$

where $\hat{\boldsymbol{u}}_{q_\ell}^{(\ell)}$ is the q_ℓ-th dominant left singular vector of $\boldsymbol{T}^{(\ell)}$.

Proof. See Appendix 5.B. □

The first quantity defined in this theorem, $\rho_{q_\ell}^{(\ell)}$, should be understood as a signal-to-noise ratio (SNR). Indeed, the squared q_ℓ-th singular value of $\boldsymbol{P}^{(\ell)}$ (i.e., the q_ℓ-th eigenvalue of $\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top}$),

[10] We recall that the notation $s_i(\boldsymbol{A})$ denotes the i-th singular value of $\boldsymbol{A}$ in non-increasing order.

$s^2_{q_\ell}(\boldsymbol{P}^{(\ell)})$, measures the "strength" of the signal in its q_ℓ-th principal direction, whereas σ_N measures the spread of the noise, as seen in Theorem 5.1. The two quantities $\tilde{\xi}^{(\ell)}_{q_\ell}$ and $\zeta^{(\ell)}_{q_\ell}$ depend only on the value of this SNR and indicate respectively the position of an isolated eigenvalue in the sprectrum of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ and the alignment of the corresponding eigenvector with the sought signal. In fact, we observe a *phase transition* phenomenon: if the SNR is large enough, i.e., if $\rho^{(\ell)}_{q_\ell} > 1$, an eigenvalue of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ isolates itself from the semicircle[11] around $\mu^{(\ell)}_N + \sigma_N\tilde{\xi}^{(\ell)}_{q_\ell}$. Moreover, recalling that $\mathcal{P} = [\![\mathcal{H}; \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(d)}]\!]$, the eigenvector associated with this isolated eigenvalue is aligned with the subspace spanned by $\boldsymbol{X}^{(\ell)}$, which is the ℓ-th singular subspace of $\mathcal{P}$. The quality of this alignment is given by $0 < \zeta^{(\ell)}_{q_\ell} \leqslant 1$.

Most importantly, this result reveals the *non-trivial regime* for the estimation of $\mathcal{P}$ with a truncated MLSVD. Since $\sigma_N = \Theta(N^{\frac{d-2}{2}})$, it shows that $\|\mathcal{P}\|^2_{\mathrm{F}} = \sum_{q_\ell=1}^{r_\ell} s^2_{q_\ell}(\boldsymbol{P}^{(\ell)})$ must also be of the same order. Indeed, if $\|\mathcal{P}\|^2_{\mathrm{F}} \ll N^{\frac{d-2}{2}}$, then $\rho^{(\ell)}_{q_\ell} \to 0$, the SNR is too small and no signal can be recovered, whereas if $\|\mathcal{P}\|^2_{\mathrm{F}} \gg N^{\frac{d-2}{2}}$, then $\rho^{(\ell)}_{q_\ell} \to +\infty$, the SNR is very high and recovery of $\mathcal{P}$ is easy. It is precisely between these two regimes, i.e., $\|\mathcal{P}\|_{\mathrm{F}} = \Theta(N^{\frac{d-2}{4}})$, that the recovery is non-trivial. Note that this observation is in line with the results of Ben Arous et al. (2023) and Zhang and Xia (2018). In this non-trivial regime, the quantities $\zeta^{(\ell)}_{q_\ell}$ given in Theorem 5.5 precisely quantify how well the dominant eigenvectors of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ are aligned with the sought signal, i.e., the singular subspaces of $\mathcal{P}$. In section 5.1.2 below, this result is used to study the reconstruction performance of truncated MLSVD.

Remark 5.6 (Link with the spiked Wigner model). The expressions of $\tilde{\xi}^{(\ell)}_{q_\ell}$ and $\zeta^{(\ell)}_{q_\ell}$ given in Theorem 5.5 are similar to those given in Corollary 2.28 for the spiked Wigner model. Indeed, given a symmetric $p \times p$ random matrix $\boldsymbol{W}$ with i.i.d. (up to symmetry) $\mathcal{N}(0, \frac{1}{p})$ entries, the spectrum of $[\rho\boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{W}]$ with $\|\boldsymbol{x}\| = 1$ follows a semicircle distribution as $p \to +\infty$ with an isolated eigenvalue at $\rho + \frac{1}{\rho}$ if, and only if, $\rho > 1$ (Féral and Péché, 2007; Edwards and Jones, 1976; Füredi and Komlós, 1981). Moreover, the corresponding eigenvector $\boldsymbol{u}$ is such that $\langle \boldsymbol{x}, \boldsymbol{u}\rangle^2 \to 1 - \frac{1}{\rho^2}$ almost surely as $p \to +\infty$ (Benaych-Georges and Nadakuditi, 2011). As discussed in Remark 5.3, up to a deterministic rescaling and shift, $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ asymptotically behaves like a spiked Wigner matrix.

Theorem 5.5 is illustrated in Figure 5.3. In the first row, asymptotic positions of isolated eigenvalues $\mu^{(\ell)}_N + \sigma_N\tilde{\xi}^{(\ell)}_{q_\ell}$ are represented by the green dashed lines. In our experiment, $\mathcal{P}$ has multilinear rank $(3,4,5)$. Hence 3, 4 and 5 isolated eigenvalues are expected in the spectrum of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ for $\ell = 1$, 2 and 3 respectively. This is indeed the case for $\ell = 1$ and $\ell = 2$ but not $\ell = 3$ where there are only 4 spike eigenvalues. In fact, $s_5(\boldsymbol{P}^{(3)})$ is not "energetic enough" to extricate itself from the bulk of eigenvalues, i.e, the SNR $\rho^{(3)}_5$ is below the phase transition threshold. Hence the fifth dominant left singular vector of $\boldsymbol{T}^{(3)}$ is not informative as it is not aligned with the third singular subspace of $\mathcal{P}$, spanned by $\boldsymbol{X}^{(3)}$. The second row of Figure 5.3 depicts the alignments of the spiked eigenvectors $\hat{\boldsymbol{u}}^{(\ell)}_{q_\ell}$ with the corresponding singular subspaces of $\mathcal{P}$ as well as the asymptotic alignment given by Theorem 5.5 as a function of the position of the associated eigenvalue. It appears that the higher is the SNR $\rho^{(\ell)}_{q_\ell}$, the farther is the isolated eigenvalue from the bulk and the more is the corresponding eigenvector aligned with the span of $\boldsymbol{X}^{(\ell)}$. This assertion can be intuitively understood in terms of "energy" $s^2_{q_\ell}(\boldsymbol{P}^{(\ell)})$ of a given principal direction. More energy pushes the eigenvalue farther from the bulk and aligns the corresponding eigenvector with the corresponding singular subspace of $\mathcal{P}$.

[11] Indeed, note that $\rho^{(\ell)}_{q_\ell} > 1 \implies \tilde{\xi}^{(\ell)}_{q_\ell} > 2$.

5.1.2 Reconstruction Performance of Truncated MLSVD

Our random matrix results allow to accurately study the reconstruction performance of truncated MLSVD. Given a data tensor $\mathcal{T}$ following the general spiked tensor model (5.2), we consider its low-rank approximation $\hat{\mathcal{T}} = [\![\hat{\mathcal{G}}; \hat{\boldsymbol{U}}^{(1)}, \ldots, \hat{\boldsymbol{U}}^{(d)}]\!]$ where $\hat{\boldsymbol{U}}^{(\ell)}$ is the $n_\ell \times r_\ell$ matrix whose columns are the r_ℓ dominant singular vectors of $\mathcal{T}$ and $\hat{\mathcal{G}} = \mathcal{T}(\hat{\boldsymbol{U}}^{(1)}, \ldots, \hat{\boldsymbol{U}}^{(d)})$. $\hat{\mathcal{T}}$ is the projection of $\mathcal{T}$ on its dominant singular subspaces, hence the name *truncated MLSVD* as it generalizes the truncated SVD of matrices. Given Model (5.2), the underlying signal estimated by $\hat{\mathcal{T}}$ is $\mathcal{P} = [\![\mathcal{H}; \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(d)}]\!]$. The reconstruction performance of $\hat{\mathcal{T}}$ hence depends on how well the *subspace*[12] spanned by $\hat{\boldsymbol{U}}^{(\ell)}$ estimates the one spanned by $\boldsymbol{X}^{(\ell)}$.

Metrics between singular subspaces are often expressed in terms of *principal angles* (Björck and Golub, 1973; Stewart and Sun, 1990, II.4), which generalize the concept of angle between lines. Given two subspaces (here, $\operatorname{Span} \boldsymbol{X}^{(\ell)}$ and $\operatorname{Span} \hat{\boldsymbol{U}}^{(\ell)}$), one can define a set of mutual angles which are invariant under isometric transformation.

Definition 5.7 (Principal angles)**.** The principal angles $\theta_{q_\ell}^{(\ell)} \in [0, \frac{\pi}{2}]$ between the subspaces spanned by $\boldsymbol{X}^{(\ell)}$ and $\hat{\boldsymbol{U}}^{(\ell)}$ are recursively defined for $q_\ell = 1, \ldots, r_\ell$ by

$$\cos\theta_{q_\ell}^{(\ell)} = \langle \boldsymbol{x}_{q_\ell}, \boldsymbol{u}_{q_\ell} \rangle \quad \text{with} \quad (\boldsymbol{x}_{q_\ell}, \boldsymbol{u}_{q_\ell}) \in \underset{\substack{(\boldsymbol{x},\boldsymbol{u}) \in \operatorname{Span} \boldsymbol{X}^{(\ell)} \times \operatorname{Span} \hat{\boldsymbol{U}}^{(\ell)} \\ \boldsymbol{x}^\top \boldsymbol{x}_{q'_\ell} = 0,\ \boldsymbol{u}^\top \boldsymbol{u}_{q'_\ell} = 0,\ 1 \leqslant q'_\ell < q_\ell}}{\arg\max} \langle \boldsymbol{x}, \boldsymbol{u} \rangle.$$

Moreover, we have the following useful property.

Proposition 5.8 (Björck and Golub, 1973)**.** *The q_ℓ-th singular value of $\boldsymbol{X}^{(\ell)\top} \hat{\boldsymbol{U}}^{(\ell)}$ in non-increasing order equals the cosine of the q_ℓ-th principal angle,*

$$s_{q_\ell}(\boldsymbol{X}^{(\ell)\top} \hat{\boldsymbol{U}}^{(\ell)}) = \cos\theta_{q_\ell}^{(\ell)}, \qquad \ell \in [d], \quad q_\ell \in [r_\ell].$$

In fact, information about the alignment between the subspaces induced by $\boldsymbol{X}^{(\ell)}$ and $\hat{\boldsymbol{U}}^{(\ell)}$ are contained entirely in the $r_\ell \times r_\ell$ matrix $\boldsymbol{X}^{(\ell)\top} \hat{\boldsymbol{U}}^{(\ell)}$. Following Definition 5.7, we know from Theorem 5.5 that, as $N \to +\infty$, $\cos^2\theta_{q_\ell}^{(\ell)} \to \left[\zeta_{q_\ell}^{(\ell)}\right]^+$ almost surely[13], since $\rho_{q_\ell}^{(\ell)} > 1 \iff \zeta_{q_\ell}^{(\ell)} > 0$. Hence, using Proposition 5.8,

$$\frac{1}{r_\ell} \left\| \boldsymbol{X}^{(\ell)\top} \hat{\boldsymbol{U}}^{(\ell)} \right\|_{\mathrm{F}}^2 = \frac{1}{r_\ell} \sum_{q_\ell=1}^{r_\ell} \cos^2\theta_{q_\ell}^{(\ell)} \xrightarrow[N \to +\infty]{\text{a.s.}} \frac{1}{r_\ell} \sum_{q_\ell=1}^{r_\ell} \left[\zeta_{q_\ell}^{(\ell)}\right]^+. \tag{5.5}$$

Therefore, the quantity $\frac{1}{r_\ell} \| \boldsymbol{X}^{(\ell)\top} \hat{\boldsymbol{U}}^{(\ell)} \|_{\mathrm{F}}^2 \in [0,1]$ appears as a relevant measure of alignment between the singular subspaces of $\mathcal{P}$ and $\hat{\mathcal{T}}$ and does not depend on the chosen orthonormal bases $\boldsymbol{X}^{(\ell)}$ and $\hat{\boldsymbol{U}}^{(\ell)}$. More details on metrics between subspaces can be found in Stewart and Sun (1990, II.4).

In Figure 5.2 are represented the alignments between the singular subspaces of $\hat{\mathcal{T}}$ and $\mathcal{P} = \sqrt{\omega}\mathcal{P}_\circ$ with $\|\mathcal{P}_\circ\|^2 = \sigma_N$ as a function of the signal-to-noise ratio ω. The fact that $\|\mathcal{P}_\circ\|^2 = \sigma_N$ ensures that the estimation problem is *non-trivial* (neither too easy nor too hard) as $\rho_{q_\ell}^{(\ell)} = \Theta(1)$. Plain curves are the alignments given by Theorem 5.5 as $N \to +\infty$ (right-hand side of Equation (5.5)) whereas dotted curves are simulation results at finite N (left-hand side of Equation (5.5)). In this setting, $\mathcal{P}_\circ$ has multilinear rank $(3,4,5)$. Hence, its "energy" $\|\mathcal{P}_\circ\|^2$ is spread among 3, 4 and 5 directions along modes 1, 2 and 3 respectively. Each break in the plain curves correspond to a value of ω such that $\rho_{q_\ell}^{(\ell)} = 1$,

[12] The object of importance is indeed the subspace and not the matrix $\hat{\boldsymbol{U}}^{(\ell)}$ since any other matrix $\hat{\boldsymbol{U}}^{(\ell)} \boldsymbol{O}^{(\ell)}$, with $\boldsymbol{O}^{(\ell)} \in \mathcal{O}_{r_\ell}(\mathbb{R})$, would span the same subspace and therefore give the same approximation.

[13] Indeed, since $\zeta_1^{(\ell)} \geqslant \ldots \geqslant \zeta_{r_\ell}^{(\ell)}$, observe that, in Definition 5.7, $\langle \boldsymbol{x}_{q_\ell}, \boldsymbol{u}_{q_\ell} \rangle^2$ is asymptotically bounded by $\left[\zeta_{q_\ell}^{(\ell)}\right]^+$.

that is, $\omega = \sigma_N / s^2_{q_\ell}(\boldsymbol{P}_\circ^{(\ell)})$. In other words, there are r_ℓ phase transitions along mode ℓ whose positions depend on the singular values of $\boldsymbol{P}_\circ^{(\ell)}$. If ω is too small (here, $\omega \lesssim 2$), truncated MLSVD fails to recover any direction of the singular subspaces of $\mathcal{P}_\circ$. As ω passes the first phase transition (here, at $\omega = \sigma_N / s^2_1(\boldsymbol{P}_\circ^{(1)}) \approx 2$), a first principal direction is partially reconstructed. Then, more and more phase transitions occur, corresponding to more and more principal directions being recovered as ω grows. Simultaneously, the reconstruction of previous directions keeps improving. Eventually, as ω is large, subspace alignments approach 1, indicating that truncated MLSVD accurately recovers the singular subspaces of $\mathcal{P}_\circ$.

The reader has not missed the dashed lines in Figure 5.2 showing much better subspace alignments than truncated MLSVD. They result from the numerical estimation of the *best* rank-$(3,4,5)$ approximation of $\mathcal{T}$ with the HOOI algorithm, which is discussed in the next section.

Remark 5.9 (Reconstruction of $\mathcal{H}$). Guarantees on the recovery of the core tensor $\mathcal{H}$ can be deduced from Theorem 5.5 as well. Without loss of generality, we can assume that $\boldsymbol{X}^{(\ell)\top}\hat{\boldsymbol{U}}^{(\ell)}$ is, up to an almost-surely vanishing additive term, a diagonal matrix with entries $\zeta_1^{(\ell)}, \dots, \zeta_{r_\ell}^{(\ell)}$ (otherwise replace $\hat{\boldsymbol{U}}^{(\ell)}$ by $\hat{\boldsymbol{U}}^{(\ell)}\boldsymbol{O}^{(\ell)}$ for a well-chosen orthogonal matrix $\boldsymbol{O}^{(\ell)}$). Then, $\hat{\mathcal{G}} = \mathcal{T}(\hat{\boldsymbol{U}}^{(1)}, \dots, \hat{\boldsymbol{U}}^{(d)})$ is the corresponding estimator of $\mathcal{H}$ and

$$\hat{\mathcal{G}} = \hat{\mathcal{H}} + \frac{1}{\sqrt{N}}\mathcal{N}(\hat{\boldsymbol{U}}^{(1)}, \dots, \hat{\boldsymbol{U}}^{(d)}) \quad \text{with} \quad \hat{\mathcal{H}} = \left[\!\!\left[\mathcal{H}; \hat{\boldsymbol{U}}^{(1)\top}\boldsymbol{X}^{(1)}, \dots, \hat{\boldsymbol{U}}^{(d)\top}\boldsymbol{X}^{(d)} \right]\!\!\right].$$

We will see in Lemma 5.10 below that $\|\frac{1}{\sqrt{N}}\mathcal{N}(\hat{\boldsymbol{U}}^{(1)}, \dots, \hat{\boldsymbol{U}}^{(d)})\|_{\mathrm{F}} = \mathcal{O}(1)$ almost surely as $N \to +\infty$. On the other hand, we know that $\|\mathcal{H}\|_{\mathrm{F}} = \|\mathcal{P}\|_{\mathrm{F}} = \Theta(N^{\frac{d-2}{4}}) \gg \mathcal{O}(1)$ as soon as $d \geqslant 3$ and the entries of $\hat{\mathcal{H}}$ are proportional to those of $\mathcal{H}$:

$$\hat{\mathcal{H}}_{q_1,\dots,q_d} = \mathcal{H}_{q_1,\dots,q_d} \prod_{\ell=1}^{d} \zeta_{q_\ell}^{(\ell)} + \epsilon_{q_1,\dots,q_d}, \qquad \ell \in [d], \quad q_\ell \in [r_\ell],$$

up to an almost-surely vanishing additive term $\epsilon_{q_1,\dots,q_d}$ as $N \to +\infty$.

Moreover, regarding the reconstruction of $\mathcal{T}$, we know from De Lathauwer et al. (2000b, Property 10) that

$$\left\| \mathcal{T} - \hat{\mathcal{T}} \right\|_{\mathrm{F}}^2 \leqslant \sum_{\ell=1}^{d} \sum_{i_\ell = r_\ell + 1}^{n_\ell} s^2_{i_\ell}(\boldsymbol{T}^{(\ell)}) = \sum_{\ell=1}^{d} \left(\|\mathcal{T}\|_{\mathrm{F}}^2 - \sum_{q_\ell=1}^{r_\ell} s^2_{q_\ell}(\boldsymbol{T}^{(\ell)}) \right)$$

and the asymptotic behavior of $s^2_{q_\ell}(\boldsymbol{T}^{(\ell)})$ is given by Theorem 5.5.

5.2 Numerical Estimation of the Best Low-Multilinear-Rank Approximation

In search of an efficient estimator of the planted signal $\mathcal{P}$, one naturally considers the best low-multilinear-rank approximation of $\mathcal{T}$, that is, a solution to Problem (5.1). As explained in the introduction, this is NP-hard in general but numerical schemes can compute it in polynomial time above the computational threshold (Richard and Montanari, 2014; Zhang and Xia, 2018). In this section, we examine the most standard of these numerical schemes, namely the Higher Order Orthogonal Iteration (HOOI) algorithm (De Lathauwer et al., 2000a; Kroonenberg and de Leeuw, 1980; Kapteyn et al., 1986), and discuss the numerical difficulties faced in the computation of a solution to Problem (5.1).

5.2.1 Higher-Order Orthogonal Iteration

Following De Lathauwer et al. (2000a, Theorem 4.2), the maximum likelihood estimation formulated in Problem (5.1) is equivalent to

$$\left(\boldsymbol{U}_\star^{(1)},\ldots,\boldsymbol{U}_\star^{(d)}\right) \in \underset{\boldsymbol{U}^{(\ell)}\in V_{r_\ell}(\mathbb{R}^{n_\ell}),\ \ell\in[d]}{\arg\max} \frac{1}{2}\left\|\mathcal{T}(\boldsymbol{U}^{(1)},\ldots,\boldsymbol{U}^{(d)})\right\|_{\mathrm{F}}^2 \tag{5.6}$$

where $V_{r_\ell}(\mathbb{R}^{n_\ell}) = \{\boldsymbol{U}^{(\ell)} \in \mathbb{R}^{n_\ell \times r_\ell} \mid \boldsymbol{U}^{(\ell)\top}\boldsymbol{U}^{(\ell)} = \boldsymbol{I}_{r_\ell}\}$ is the set of $r_\ell \times n_\ell$ matrices with orthonormal columns, known as the *Stiefel manifold* (Chikuse, 2003; Absil et al., 2009). Then, since the Frobenius norm of $\mathcal{T}(\boldsymbol{U}^{(1)},\ldots,\boldsymbol{U}^{(d)})$ is equal to the Frobenius norm of any of its unfoldings,

$$\left\|\mathcal{T}(\boldsymbol{U}^{(1)},\ldots,\boldsymbol{U}^{(d)})\right\|_{\mathrm{F}} = \left\|\boldsymbol{U}^{(\ell)\top}\boldsymbol{T}^{(\ell)}\bigboxtimes_{\ell'\neq\ell}\boldsymbol{U}^{(\ell')}\right\|_{\mathrm{F}}, \qquad \ell\in[d].$$

And we see from Problem (5.6) that $\boldsymbol{U}_\star^{(\ell)}$ is the matrix gathering the r_ℓ dominant left singular vectors of $\boldsymbol{T}^{(\ell)}\boxtimes_{\ell'\neq\ell}\boldsymbol{U}_\star^{(\ell')}$. This is precisely what motivates the HOOI algorithm presented in Algorithm 5. It performs fixed-point iterations to compute a solution $\boldsymbol{U}_\star^{(1)},\ldots,\boldsymbol{U}_\star^{(d)}$ satisfying the previous property[14].

Algorithm 5: Higher-Order Orthogonal Iteration (De Lathauwer et al., 2000a)

for $\ell = 1,\ldots,d$ **do** $\boldsymbol{U}_0^{(\ell)} \leftarrow r_\ell$ dominant left singular vectors of $\boldsymbol{T}^{(\ell)}$
repeat
  for $\ell = 1,\ldots,d$ **do** $\boldsymbol{U}_{t+1}^{(\ell)} \leftarrow r_\ell$ dominant left singular vectors of $\boldsymbol{T}^{(\ell)}\boxtimes_{\ell'\neq\ell}\boldsymbol{U}_t^{(\ell')}$
until *convergence at* $t = T$
$\mathcal{G}_{\mathrm{HOOI}} \leftarrow \mathcal{T}(\boldsymbol{U}_T^{(1)},\ldots,\boldsymbol{U}_T^{(d)})$

The HOOI algorithm is initialized with $\boldsymbol{U}_0^{(1)},\ldots,\boldsymbol{U}_0^{(d)}$, the truncated MLSVD[15] of $\mathcal{T}$. Given the results of Section 5.1, this is indeed a very good and easily computable first guess. Then, fixed-point iterations are repeated in order to find a solution $\boldsymbol{U}_\star^{(1)},\ldots,\boldsymbol{U}_\star^{(d)}$ such that $\boldsymbol{U}_\star^{(\ell)}$ spans the left r_ℓ-dimensional dominant singular subspace of $\boldsymbol{T}^{(\ell)}\boxtimes_{\ell'\neq\ell}\boldsymbol{U}_\star^{(\ell')}$ for all $\ell\in[d]$, which corresponds to a solution to Problem (5.6). In practice, the stopping criterion can be chosen as a negligible change in the norm of the estimated core tensor, $\|\mathcal{T}(\boldsymbol{U}_t^{(1)},\ldots,\boldsymbol{U}_t^{(d)})\|_{\mathrm{F}}$.

Xu (2018) showed that the convergence of this algorithm towards a local minimum of Problem (5.6) is guaranteed as long as its initialization is *sufficiently close* to this local minimum. In light of our previous results, we can provide further insight into this "sufficiently close" property. Indeed, in Theorem 5.11 below, we show that an initialization with non-zero alignment with the signal $\mathcal{P}$ is sufficient to ensure that the HOOI algorithm perfectly reconstructs it *after a single iteration*.

Before introducing Theorem 5.11, we formulate an important preliminary result which essentially states that, given $\boldsymbol{A}^{(\ell)} \in V_{r_\ell}(\mathbb{R}^{n_\ell})$, $\ell\in[d]$, the quantity $\frac{1}{\sqrt{N}}\|\mathcal{N}(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(d)})\|_{\mathrm{F}}$ is almost surely bounded as $N\to+\infty$.

Lemma 5.10. *With probability at least* $1-\delta$,

$$\sup_{\boldsymbol{A}^{(\ell)}\in V_{r_\ell}(\mathbb{R}^{n_\ell}),\ \ell\in[d]} \left\|\mathcal{N}(\boldsymbol{A}^{(1)},\ldots,\boldsymbol{A}^{(d)})\right\|_{\mathrm{F}}^2 \leqslant 16\left[\left(\sum_{\ell=1}^d r_\ell\left(n_\ell - \frac{r_\ell+1}{2}\right)\right)\log\frac{Cd}{\log\frac{3}{2}} + \log\left(\frac{1}{\delta}\max\left(1, e^{\frac{1}{2}\prod_{\ell=1}^d r_\ell - 1}\right)\right)\right]$$

where $C>0$ *is a universal constant.*

[14] In fact, this property corresponds to first-order optimality conditions of Problem (5.6), with the squared singular values as Lagrange multipliers.

[15] Consistently with the notations of Section 5.1, this is $\hat{\boldsymbol{U}}^{(1)},\ldots,\hat{\boldsymbol{U}}^{(d)}$.

Proof. See Appendix 5.C. □

This result is crucial to handle the behavior of the noise in our analysis of Algorithm 5 (Appendix 5.D), which leads to the following result on the alignment between the singular subspaces of the signal $\mathcal{P}$ (spanned by $\boldsymbol{X}^{(\ell)}$) and those estimated from the observation $\mathcal{T}$ after the first iteration of HOOI (spanned by $\boldsymbol{U}_1^{(\ell)}$).

Theorem 5.11 (Asymptotic optimality of HOOI)**.** *As $N \to +\infty$, if $\|\mathcal{P}\|_{\mathrm{F}} \gg 1$ and*

$$\min_{\ell \in [d],\ q_\ell \in [r_\ell]} \left\| \mathcal{P}(\boldsymbol{U}_0^{(1)}, \ldots, \boldsymbol{x}_{q_\ell}^{(\ell)}, \ldots, \boldsymbol{U}_0^{(d)}) \right\|_{\mathrm{F}} \overset{def}{=} L_N \gg \|\mathcal{P}\|_{\mathrm{F}}^{1/2}$$

where $\boldsymbol{x}_{q_\ell}^{(\ell)}$ is the q_ℓ-th column of $\boldsymbol{X}^{(\ell)}$, then,

$$\frac{1}{r_\ell} \left\| \boldsymbol{X}^{(\ell)\top} \boldsymbol{U}_1^{(\ell)} \right\|_{\mathrm{F}}^2 = 1 + \mathcal{O}\left(\frac{\|\mathcal{P}\|_{\mathrm{F}}}{L_N^2} \right) \qquad \textit{almost surely.}$$

Proof. See Appendix 5.D. □

It is important to carefully understand the assumptions of Theorem 5.11. Firstly, it assumes that $\|\mathcal{P}\|_{\mathrm{F}} \gg 1$ as $N \to +\infty$, that is, the signal is not necessarily in the non-trivial regime $\Theta(N^{\frac{d-2}{4}})$ but can be smaller or bigger as long as $\|\mathcal{P}\|_{\mathrm{F}} \to +\infty$, regardless its speed. Then, the second assumption $\|\mathcal{P}(\boldsymbol{U}_0^{(1)}, \ldots, \boldsymbol{x}_{q_\ell}^{(\ell)}, \ldots, \boldsymbol{U}_0^{(d)})\|_{\mathrm{F}} \gg \|\mathcal{P}\|_{\mathrm{F}}^{1/2}$ means that each principal directions of the ℓ-th singular subspace are sufficiently preserved after contraction on $\{\boldsymbol{U}_0^{(\ell')}\}_{\ell' \neq \ell}$. When these assumptions are verified, Theorem 5.11 states that the matrices $\boldsymbol{U}_1^{(1)}, \ldots, \boldsymbol{U}_1^{(d)}$ computed after the *first* iteration of HOOI *perfectly* reconstruct the singular subspaces of the sought signal $\mathcal{P}$ as the dimensions of the tensor, $n_1, \ldots, n_d$, grow large. More formally, as $N \to +\infty$, the alignment $\frac{1}{r_\ell} \|\boldsymbol{X}^{(\ell)\top} \boldsymbol{U}_1^{(\ell)}\|_{\mathrm{F}}^2$ approaches 1 almost surely. Furthermore, the speed of this convergence behaves like $\|\mathcal{P}\|_{\mathrm{F}} / L_N^2$.

Theorem 5.11 does not assume a particular choice of initialization $\boldsymbol{U}_0^{(1)}, \ldots, \boldsymbol{U}_0^{(d)}$ and gives a sufficient condition for it to ensure the convergence of the algorithm. Nevertheless, as it is presented in Algorithm 5, truncated MLSVD is a standard choice of initialization. In this case, the assumption $L_N = \Theta(\|\mathcal{P}\|)$ is verified as soon as enough principal directions are recovered. According to Theorem 5.5, this is only possible if $\|\mathcal{P}\|_{\mathrm{F}} \geqslant \sqrt{\sigma_N} = \Theta(N^{\frac{d-2}{4}})$ since a *necessary* condition is $\rho_1^{(\ell)} > 1$ for all $\ell \in [d]$, while a *sufficient* condition is $\rho_{q_\ell}^{(\ell)} > 1$ for all $\ell \in [d]$ and $q_\ell \in [d]$. In other words, convergence at speed $\|\mathcal{P}\|_{\mathrm{F}}^{-1}$ as assured above a critical signal-to-noise ratio lying between the first and the last phase transition of each mode, and which depends on the particular structure of the core tensor $\mathcal{H}$. Yet, in most cases, this happens quite early, right after the first phase transitions, see for example Figure 5.2.

We emphasize the fact that the assumption of Theorem 5.11 can already be verified as soon as $\min_\ell \rho_1^{(\ell)} > 1$. That is, there is no need for all the principal directions to be reconstructed at initialization. In fact, it could very well be that $\max_\ell \rho_2^{(\ell)} < 1$. If the singular subspaces of $\mathcal{P}$ are sufficiently preserved with the initialization $\boldsymbol{U}_0^{(1)}, \ldots, \boldsymbol{U}_0^{(d)}$ — i.e., if $L_N \gg \|\mathcal{P}\|^{1/2}$ —, then the other principal directions still emerge after the first iteration.

In the simpler rank-1 case, these technical considerations vanish and we recover the result of Feldman and Donoho (2023, Theorem 4.2): if $\mathcal{P} = \beta_N \bigotimes_{\ell \in [d]} \boldsymbol{x}^{(\ell)}$ then a *necessary and sufficient* condition for $L_N = \Theta(\|\mathcal{P}\|_{\mathrm{F}})$ is simply $\beta_N^2 > \sigma_N$. Indeed, $\rho_1^{(\ell)} = \beta_N^2 / \sigma_N$ for all $\ell \in [d]$. Moreover, when this assumption if verified, Theorem 5.11 ensures the asymptotic exact reconstruction of $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)}$ in a single iteration with a $\|\mathcal{P}\|_{\mathrm{F}}^{-1} = \beta_N^{-1}$ speed of convergence.

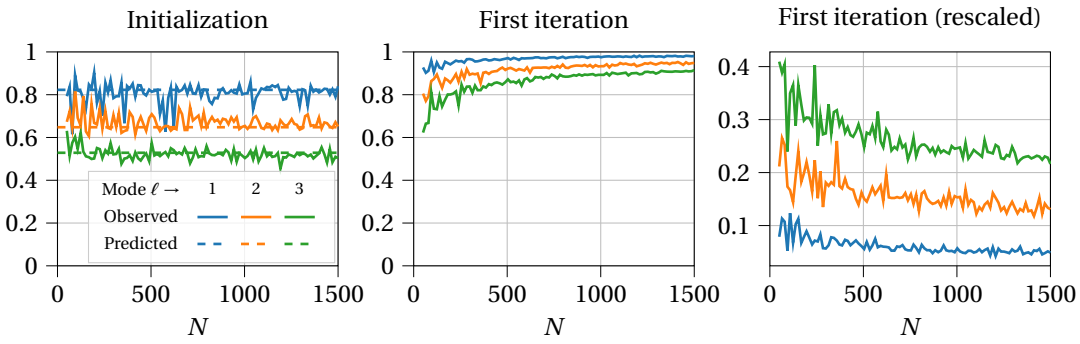


Figure 5.4: Alignments between singular subspaces of the observation $\mathcal{T} = \mathcal{P} + \frac{1}{\sqrt{N}}\mathcal{N}$ and of the signal $\mathcal{P}$, with $\|\mathcal{P}\|_{\mathrm{F}}^2/\sigma_N = 10$, at initialization of Algorithm 5 (i.e., truncated MLSVD) and after the first iteration, as a function of the size of the tensor given by the parameter N. **Left:** $\frac{1}{r_\ell}\|\boldsymbol{X}^{(\ell)\top}\boldsymbol{U}_0^{(\ell)}\|_{\mathrm{F}}^2$. **Middle:** $\frac{1}{r_\ell}\|\boldsymbol{X}^{(\ell)\top}\boldsymbol{U}_1^{(\ell)}\|_{\mathrm{F}}^2$. **Right:** $(1-\frac{1}{r_\ell}\|\boldsymbol{X}^{(\ell)\top}\boldsymbol{U}_1^{(\ell)}\|_{\mathrm{F}}^2)\times\sqrt{\sigma_N}$. **Experimental setting:** $d = 3$, $(\frac{n_1}{N}, \frac{n_2}{N}, \frac{n_3}{N}) = (\frac{1}{6}, \frac{2}{6}, \frac{3}{6})$, $N = n_1 + n_2 + n_3$ and $(r_1, r_2, r_3) = (3, 4, 5)$.

Remark 5.12 (Practical implications). In practice, one should still run several iterations of Algorithm 5 until a certain stopping criterion is verified as this effectively improves the final estimate and converges to a solution to the maximum likelihood estimation (5.1) (Xu, 2018). Theorem 5.11 states that the reconstruction performance of HOOI after the first iteration increases as we consider larger tensors, until it reaches perfect recovery in the large N limit. In other words, the number of iterations required to achieve a specific level of accuracy in maximum likelihood estimation tends to 1 as $N \to +\infty$.

Theorem 5.11 is illustrated in Figure 5.4. As a function of N — the size of the tensor — we represent the subspace alignments observed at initialization and after the first iteration for a fixed signal-to-noise ratio $\|\mathcal{P}\|_{\mathrm{F}}^2/\sigma_N = 10$. The left panel compares the observed alignments achieved with truncated MLSVD (initialization of Algorithm 5) with the asymptotic alignments predicted by Theorem 5.5. As N grows, the observed alignments remain around their asymptotic values, with only a decrease in variance. The middle panel presents the alignments after the first iteration. Here, as N increases, we observe an increase in the values of the alignments, which approach 1, consistently with Theorem 5.11. This is specified in the right panel where the value $(1 - \frac{1}{r_\ell}\|\boldsymbol{X}^{(\ell)\top}\boldsymbol{U}_1^{(\ell)}\|_{\mathrm{F}}^2) \times \sqrt{\sigma_N}$ is plotted. According to Theorem 5.11, this value should be $\mathcal{O}(1)$ since $L_N = \Theta(\|\mathcal{P}\|_{\mathrm{F}}) = \Theta(\sqrt{\sigma_N})$ here. The observed behavior confirms the $\|\mathcal{P}\|_{\mathrm{F}}^{-1}$ speed of convergence asserted in Theorem 5.11.

5.2.2 Discussion on Signal Reconstructibility

Our results on truncated MLSVD (Section 5.1) and HOOI (Section 5.2.1) bring insight into the computational-to-statistical gap observed in the low-multilinear-rank approximation problem. Truncated MLSVD can only work efficiently if $\|\mathcal{P}\|_{\mathrm{F}}$ is at least $\Theta(N^{\frac{d-2}{4}})$ and its reconstruction performance has been discussed in Section 5.1.2. However, Theorem 5.11 suggests that it is possible to perfectly reconstruct the signal $\mathcal{P}$ from the observation $\mathcal{T}$ as long as $\|\mathcal{P}\|_{\mathrm{F}} \gg 1$ and HOOI is accurately initialized. Yet, it is known that, without prior information on $\mathcal{P}$, maximum likelihood estimation (5.1) is NP-hard below the $\Theta(N^{\frac{d-2}{4}})$ computational threshold (Zhang and Xia, 2018), which lies precisely in the non-trivial regime of truncated MLSVD.

In fact, what can be understood from Theorem 5.11 is that it suffices to have an initialization $\boldsymbol{U}_0^{(\ell)}$, $\ell \in [d]$, slightly aligned with the underlying signal $\mathcal{P}$ to be in the right basin of attraction and allow

the convergence of Algorithm 5 towards a solution to Problem (5.1). This complements the results of Xu (2018) on the conditions of convergence of HOOI. Furthermore, a solution to Problem (5.1) is aligned with $\mathcal{P}$ as soon as $\|\mathcal{P}\|_{\mathrm{F}} = \Theta(1)$ (Ben Arous et al., 2019; Jagannath et al., 2020; Zhang and Xia, 2018). Hence, as the HOOI algorithm is meant to compute a maximum likelihood estimator, with the assumption $\|\mathcal{P}\|_{\mathrm{F}} \gg 1$ made in Theorem 5.11, it is expected that these iterations allow to perfectly recover the signal asymptotically. Maximum likelihood estimation is indeed (theoretically) trivial if $\|\mathcal{P}\|_{\mathrm{F}} \to +\infty$. It is more surprising however that this already happens at the first iteration.

As said previously, the choice of initialization $\boldsymbol{U}_0^{(\ell)}$, $\ell \in [d]$, does not matter in Theorem 5.11. In fact, without prior information, a truncated MLSVD is the best choice as it allows to partially reconstruct the signal at the $\Theta(N^{\frac{d-2}{4}})$ computational threshold. Nevertheless, had one prior information allowing such a reconstruction in the regime $1 \ll \mathcal{P} \ll N^{\frac{d-2}{4}}$ — where truncated MLSVD would not be fruitful —, HOOI would still be able to perfectly reconstruct the signal $\mathcal{P}$ given this initialization.

Hence, HOOI initialized with a truncated MLSVD, as it is presented in Algorithm 5, allows to numerically compute a maximum likelihood estimator (solution to Problem (5.1)) but only above the phase transition of truncated MLSVD. Indeed, its initialization plays a determining role: it must place $\boldsymbol{U}_0^{(\ell)}$, $\ell \in [d]$, in the right basin of attraction, which, without prior information, is only possible above the $\Theta(N^{\frac{d-2}{4}})$ computational threshold.

Finally, we highlight the fact that these results concern the large N limit. In practice, it makes no sense to talk about $\Theta(1)$ or $\Theta(N^{\frac{d-2}{4}})$ regimes at finite N. Figure 5.2 also depicts the subspace alignments achieved with the maximum likelihood estimator computed with Algorithm 5 on $\mathcal{T} = \sqrt{\omega}\mathcal{P}_\circ + \frac{1}{\sqrt{N}}\mathcal{N}$. Although $\|\mathcal{P}_\circ\|_{\mathrm{F}} = \sqrt{\sigma_N} = \Theta(N^{\frac{d-2}{4}})$, HOOI does not achieve perfect recovery of $\mathcal{P}$ as one might expect from Theorem 5.11 (even if several iterations were performed here). In fact, at finite N, perfect recovery is not feasible. But, as N grows, the dashed line would approach 1 above the (computational) phase transition determined by the truncated MLSVD and stay close to 0 below.

5.3 Concluding Remarks

The analysis presented in this work yields theoretical and practical insights into the estimation of a low-rank signal from an observation $\mathcal{T} = \mathcal{P} + \frac{1}{\sqrt{N}}\mathcal{N}$ following the most general spiked tensor model. While Zhang and Xia (2018) gave a general overview of the different regimes governing the estimation of $\mathcal{P}$ with a low-multilinear-rank approximation of $\mathcal{T}$ — thereby confirming the existence of a computational-to-statistical gap —, our results shed light on the non-trivial aspects at stake around the $\Theta(N^{\frac{d-2}{4}})$ computational threshold. This is of particular importance as practical applications lie in this non-trivial regime where signal and noise have the same magnitude and must be decoupled. In particular, truncated MLSVD and HOOI are very standard and efficient algorithms to compute low-multilinear-rank approximations. Performances of the latter rely strongly on the quality of its initialization, which is usually performed with a truncated MLSVD in the absence of prior information. This approach allows the detection of the underlying signal as early as the computational threshold contrary to other methods such as AMP or tensor power iteration, which are efficient above a $\Theta(N^{\frac{d-1}{2}})$ algorithmic threshold (Richard and Montanari, 2014).

Relying on standard tools and methods from the theory of large random matrices, we have characterized the spectral behavior of the unfoldings of $\mathcal{T}$ in the large N limit. Specifically, our first main result shows that, when properly centered and scaled, the eigenvalues of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ asymptotically follow a semicircle distribution. The rescaling exhibits their mean $\mu_N^{(\ell)} = \Theta(N^{d-2})$ and a quantity $\sigma_N = \Theta(N^{\frac{d-2}{2}})$ governing their spread. From our denoising perspective, σ_N indicates the *strength* of the noise. Indeed, while the global behavior of the eigenvalues is controlled by the noise, the addition

of a low-rank signal causes the presence of a finite number of eigenvalues outside the limiting semi-circle distribution with corresponding eigenvectors aligned with the singular subspaces of the sought signal $\mathcal{P}$. Yet, the existence of these outlier eigenvalues hinges on the values of the signal-to-noise ratios $\rho_{q_\ell}^{(\ell)} = s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})/\sigma_N$, manifesting a BBP phase transition phenomenon. When they exist, the positions of these isolated eigenvalues and the quality of the corresponding alignments are completely determined by $\rho_{q_\ell}^{(\ell)}$. These results justify the use of a truncated MLSVD to estimate $\mathcal{P}$ from the observation $\mathcal{T}$ and allow the precise characterization of the achievable reconstruction performances in the non-trivial regime, i.e., close to the computational threshold. In particular, we have seen that each singular value of $\boldsymbol{P}^{(\ell)}$ determines the position of a phase transition corresponding to the detectability of the corresponding principal direction.

Although truncated MLSVD does not yield the best low-multilinear-rank approximation — i.e., a maximum likelihood solution —, it serves as an excellent initialization to the HOOI algorithm, which converges to such an estimator if it is initialized *sufficiently close* to it (Xu, 2018). In fact, we precise this last assertion by showing that, as long as the initialization preserves the singular subspaces of $\mathcal{P}$ in a sense precised in Theorem 5.11, HOOI converges to a maximum likelihood solution in a number of iterations which tends to 1 as $N \to +\infty$. Hence, when it is initialized with a truncated MLSVD, it shares the same phase transition, whose position depends on the singular values $s_{q_\ell}^{(\ell)}(\boldsymbol{P}^{(\ell)})$ of the unfoldings of $\mathcal{P}$. Yet, given prior information, HOOI can still reconstruct the maximum likelihood solution below the computational threshold $\|\mathcal{P}\|_{\mathrm{F}} = \Theta(N^{\frac{d-2}{4}})$, where its success depends entirely on the quality of its initialization.

This work gives a comprehensive understanding of the low-multilinear-rank approximation problem near the computational threshold, which has both practical and theoretical implications. Besides, from a theoretical perspective, the behavior of the maximum likelihood estimator is still unclear near the statistical threshold — that is, in the regime where $\|\mathcal{P}\|_{\mathrm{F}} = \Theta(1)$. Several works have studied the rank-1 symmetric case (Ben Arous et al., 2019; Jagannath et al., 2020) and the approach developed by Seddik et al. (2022) in their analysis of the rank-1 *asymmetric* case may be an attractive direction to consider. Relying solely on classical tools from random matrix theory, they bring the study of the best rank-1 tensor approximation down to that of a structured matrix defined from the contractions of the data tensor on its dominant singular vectors. Extending this procedure to our general spiked tensor model (5.2) presents no conceptual difficulty, despite being computationally cumbersome due to the multiple dimensions of the singular subspaces. It is an interesting line of investigation to refine our understanding of the statistical limits to low-rank tensor estimation from spiked models.

5.A Proof of Theorem 5.1

The resolvent of $\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ is defined for all $z \in \mathbb{C} \setminus \operatorname{Sp} \boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top}$ as

$$\boldsymbol{Q}^{(\ell)}(z) \stackrel{\text{def}}{=} \left(\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - z\boldsymbol{I}_{n_\ell}\right)^{-1}.$$

We will often drop the dependence in z to simplify notations.

Since $\boldsymbol{Q}^{(\ell)-1}\boldsymbol{Q}^{(\ell)} = \boldsymbol{I}_{n_\ell}$ and $\boldsymbol{T}^{(\ell)} = \boldsymbol{P}^{(\ell)} + \frac{1}{\sqrt{N}}\boldsymbol{N}^{(\ell)}$, we have,

$$\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)} + \frac{1}{\sqrt{N}}\boldsymbol{N}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)} - z\boldsymbol{Q}^{(\ell)} = \boldsymbol{I}_{n_\ell}. \tag{5.7}$$

5.A.1 Expressions with Stein's Lemma

Using Stein's lemma (Lemma 2.18), we find the following expressions.

$$\mathbb{E}\left[\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}\right] = \mathbb{E}\left[\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}\right] - \frac{1}{N}\mathbb{E}\left[\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}\operatorname{Tr}\boldsymbol{Q}^{(\ell)} + \boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)2}\right], \tag{5.8}$$

$$\mathbb{E}\left[\boldsymbol{N}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}\right] = \frac{\prod_{\ell'\neq\ell} n_{\ell'}}{\sqrt{N}}\mathbb{E}\left[\boldsymbol{Q}^{(\ell)}\right] - \frac{1}{\sqrt{N}}\mathbb{E}\left[(n_\ell+1)\boldsymbol{Q}^{(\ell)} + z\left(\boldsymbol{Q}^{(\ell)2} + \boldsymbol{Q}^{(\ell)}\operatorname{Tr}\boldsymbol{Q}^{(\ell)}\right)\right]. \tag{5.9}$$

Derivatives of $\boldsymbol{Q}^{(\ell)}$ Firstly, we need to show that

$$\frac{\partial Q^{(\ell)}_{a,b}}{\partial N^{(\ell)}_{c,d}} = -\frac{1}{\sqrt{N}}\left(Q^{(\ell)}_{a,c}\left[\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}\right]_{d,b} + \left[\boldsymbol{Q}^{(\ell)}\boldsymbol{T}^{(\ell)}\right]_{a,d}Q^{(\ell)}_{c,b}\right). \tag{5.10}$$

Indeed, using the fact that $\partial\boldsymbol{Q}^{(\ell)} = -\boldsymbol{Q}^{(\ell)}\partial(\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top})\boldsymbol{Q}^{(\ell)}$, we have,

$$\frac{\partial Q^{(\ell)}_{a,b}}{\partial N^{(\ell)}_{c,d}} = -\left[\boldsymbol{Q}^{(\ell)}\frac{\partial\boldsymbol{T}^{(\ell)}}{\partial N^{(\ell)}_{c,d}}\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}\right]_{a,b} - \left[\boldsymbol{Q}^{(\ell)}\boldsymbol{T}^{(\ell)}\frac{\partial\boldsymbol{T}^{(\ell)\top}}{\partial N^{(\ell)}_{c,d}}\boldsymbol{Q}^{(\ell)}\right]_{a,b}$$

and, since $\boldsymbol{T}^{(\ell)} = \boldsymbol{P}^{(\ell)} + \frac{1}{\sqrt{N}}\boldsymbol{N}^{(\ell)}$,

$$\left[\boldsymbol{Q}^{(\ell)}\frac{\partial\boldsymbol{T}^{(\ell)}}{\partial N^{(\ell)}_{c,d}}\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}\right]_{a,b} = \frac{1}{\sqrt{N}}Q^{(\ell)}_{a,c}\left[\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}\right]_{d,b},$$

$$\left[\boldsymbol{Q}^{(\ell)}\boldsymbol{T}^{(\ell)}\frac{\partial\boldsymbol{T}^{(\ell)\top}}{\partial N^{(\ell)}_{c,d}}\boldsymbol{Q}^{(\ell)}\right]_{a,b} = \frac{1}{\sqrt{N}}\left[\boldsymbol{Q}^{(\ell)}\boldsymbol{T}^{(\ell)}\right]_{a,d}Q^{(\ell)}_{c,b}.$$

Proof of Equation (5.8) Since $\boldsymbol{T}^{(\ell)} = \boldsymbol{P}^{(\ell)} + \frac{1}{\sqrt{N}}\boldsymbol{N}^{(\ell)}$, we have,

$$\mathbb{E}[\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}] = \mathbb{E}[\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}] + \frac{1}{\sqrt{N}}\mathbb{E}[\boldsymbol{P}^{(\ell)}\boldsymbol{N}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}].$$

To deal with the rightmost term, we successively use Stein's lemma (Lemma 2.18) and Equation (5.10).

$$\mathbb{E}\left[\boldsymbol{P}^{(\ell)}\boldsymbol{N}^{(\ell)\top}\boldsymbol{Q}^{(\ell)}\right]_{i,j} = \sum_{k=1}^{\prod_{\ell'\neq\ell} n_{\ell'}}\sum_{l=1}^{n_\ell}\mathbb{E}\left[P^{(\ell)}_{i,k}N^{(\ell)}_{l,k}Q^{(\ell)}_{l,j}\right]$$

$$
\begin{aligned}
&= \sum_{k=1}^{\prod_{\ell'\neq\ell} n_{\ell'}} \sum_{l=1}^{n_\ell} \mathbb{E}\left[P^{(\ell)}_{i,k} \frac{\partial Q^{(\ell)}_{l,j}}{\partial N^{(\ell)}_{l,k}} \right] \\
&= -\frac{1}{\sqrt{N}} \sum_{k=1}^{\prod_{\ell'\neq\ell} n_{\ell'}} \sum_{l=1}^{n_\ell} \mathbb{E}\left[P^{(\ell)}_{i,k} Q^{(\ell)}_{l,l} \left[\boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)} \right]_{k,j} \right] \\
&\qquad -\frac{1}{\sqrt{N}} \sum_{k=1}^{\prod_{\ell'\neq\ell} n_{\ell'}} \sum_{l=1}^{n_\ell} \mathbb{E}\left[P^{(\ell)}_{i,k} \left[\boldsymbol{Q}^{(\ell)} \boldsymbol{T}^{(\ell)} \right]_{l,k} Q^{(\ell)}_{l,j} \right] \\
&= -\frac{1}{\sqrt{N}} \mathbb{E}\left[\boldsymbol{P}^{(\ell)} \boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)} \operatorname{Tr} \boldsymbol{Q}^{(\ell)} + \boldsymbol{P}^{(\ell)} \boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)2} \right]_{i,j}.
\end{aligned}
$$

Proof of Equation (5.9) We proceed similarly with Stein's lemma (Lemma 2.18) and Equation (5.10).

$$
\begin{aligned}
&\mathbb{E}\left[\boldsymbol{N}^{(\ell)} \boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)} \right]_{i,j} \\
&\quad = \sum_{k=1}^{\prod_{\ell'\neq\ell} n_{\ell'}} \sum_{l=1}^{n_\ell} \mathbb{E}\left[N^{(\ell)}_{i,k} T^{(\ell)}_{l,k} Q^{(\ell)}_{l,j} \right] \\
&\quad = \sum_{k=1}^{\prod_{\ell'\neq\ell} n_{\ell'}} \sum_{l=1}^{n_\ell} \mathbb{E}\left[\frac{\partial T^{(\ell)}_{l,k}}{\partial N^{(\ell)}_{i,k}} Q^{(\ell)}_{l,j} + T^{(\ell)}_{l,k} \frac{\partial Q^{(\ell)}_{l,j}}{\partial N^{(\ell)}_{i,k}} \right] \\
&\quad = \frac{\prod_{\ell'\neq\ell} n_{\ell'}}{\sqrt{N}} \mathbb{E}\left[\boldsymbol{Q}^{(\ell)} \right]_{i,j} - \frac{1}{\sqrt{N}} \sum_{k=1}^{\prod_{\ell'\neq\ell} n_{\ell'}} \sum_{l=1}^{n_\ell} \mathbb{E}\left[T^{(\ell)}_{l,k} Q^{(\ell)}_{l,i} \left[\boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)} \right]_{k,j} \right] \\
&\qquad\qquad -\frac{1}{\sqrt{N}} \sum_{k=1}^{\prod_{\ell'\neq\ell} n_{\ell'}} \sum_{l=1}^{n_\ell} \mathbb{E}\left[T^{(\ell)}_{l,k} \left[\boldsymbol{Q}^{(\ell)} \boldsymbol{T}^{(\ell)} \right]_{l,k} Q^{(\ell)}_{i,j} \right] \\
&\quad = \frac{\prod_{\ell'\neq\ell} n_{\ell'}}{\sqrt{N}} \mathbb{E}\left[\boldsymbol{Q}^{(\ell)} \right]_{i,j} - \frac{1}{\sqrt{N}} \mathbb{E}\left[\boldsymbol{Q}^{(\ell)} \boldsymbol{T}^{(\ell)} \boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)} + \boldsymbol{Q}^{(\ell)} \operatorname{Tr} \boldsymbol{T}^{(\ell)} \boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)} \right]_{i,j}.
\end{aligned}
$$

Since $\boldsymbol{T}^{(\ell)} \boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)} - z\boldsymbol{Q}^{(\ell)} = \boldsymbol{I}_{n_\ell}$, we find the result stated in Equation (5.9).

5.A.2 Asymptotic Behavior of the Resolvent

Taking the expectation of Equation (5.7) and injecting Equation (5.9) yields

$$
\mathbb{E}\left[\boldsymbol{P}^{(\ell)} \boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)} \right] + \frac{\prod_{\ell'\neq\ell} n_{\ell'}}{N} \mathbb{E}\left[\boldsymbol{Q}^{(\ell)} \right] - \frac{1}{N} \mathbb{E}\left[(n_\ell + 1) \boldsymbol{Q}^{(\ell)} + z\left(\boldsymbol{Q}^{(\ell)2} + \boldsymbol{Q}^{(\ell)} \operatorname{Tr} \boldsymbol{Q}^{(\ell)} \right) \right] - z\mathbb{E}\left[\boldsymbol{Q}^{(\ell)} \right] = \boldsymbol{I}_{n_\ell}.
$$

We rearrange this expression into the more convenient following form,

$$
z \frac{n_\ell}{N} \mathbb{E}\left[\boldsymbol{Q}^{(\ell)} \frac{\operatorname{Tr} \boldsymbol{Q}^{(\ell)}}{n_\ell} \right] + \left(z + \frac{n_\ell - \prod_{\ell'\neq\ell} n_{\ell'}}{N} \right) \mathbb{E}\left[\boldsymbol{Q}^{(\ell)} \right] + \boldsymbol{I}_{n_\ell} = -\frac{1}{N} \mathbb{E}\left[\boldsymbol{Q}^{(\ell)} + z\boldsymbol{Q}^{(\ell)2} \right] + \mathbb{E}\left[\boldsymbol{P}^{(\ell)} \boldsymbol{T}^{(\ell)\top} \boldsymbol{Q}^{(\ell)} \right].
$$

Here, the divergence of the spectrum of $\boldsymbol{T}^{(\ell)} \boldsymbol{T}^{(\ell)\top}$ becomes problematic: its resolvent $\boldsymbol{Q}^{(\ell)}$ vanishes asymptotically, allowing the presence of the diverging coefficient $\frac{1}{N} \prod_{\ell'\neq\ell} n_{\ell'}$ in the previous equation. To bypass this difficulty, we proceed to a rescaling,

$$
\tilde{z} \overset{\text{def}}{=} \frac{z - \mu^{(\ell)}_N}{\sigma_N}, \qquad \widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) \overset{\text{def}}{=} \left(\frac{\boldsymbol{T}^{(\ell)} \boldsymbol{T}^{(\ell)\top} - \mu^{(\ell)}_N \boldsymbol{I}_{n_\ell}}{\sigma_N} - \tilde{z} \boldsymbol{I}_{n_\ell} \right)^{-1} = \sigma_N \boldsymbol{Q}^{(\ell)}(z),
$$

$$\text{with} \qquad \mu_N^{(\ell)} = \frac{1}{N}\prod_{\ell'\neq\ell} n_{\ell'}, \qquad \sigma_N = \frac{1}{N}\sqrt{\prod_{\ell\in[d]} n_\ell}.$$

This changes our equation into

$$\frac{\mu_N^{(\ell)}+\sigma_N\tilde{z}}{\sigma_N^2}\frac{n_\ell}{N}\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(\ell)}}{n_\ell}\right] + \left(\frac{\mu_N^{(\ell)}}{\sigma_N}+\tilde{z}+\frac{n_\ell-\prod_{\ell'\neq\ell}n_{\ell'}}{\sigma_N N}\right)\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right] + \boldsymbol{I}_{n_\ell}$$
$$= -\frac{1}{\sigma_N N}\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}+\left(\frac{\mu_N^{(\ell)}}{\sigma_N}+\tilde{z}\right)\widetilde{\boldsymbol{Q}}^{(\ell)2}\right] + \frac{1}{\sigma_N}\mathbb{E}\left[\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\widetilde{\boldsymbol{Q}}^{(\ell)}\right].$$

Let us define $\widetilde{m}_N^{(\ell)} : \tilde{z} \mapsto \frac{1}{n_\ell}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z})$, the Stieltjes transform of the empirical spectral distribution of $\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$. With this definition, and using the fact that $\frac{\mu_N^{(\ell)}}{\sigma_N^2}\frac{n_\ell}{N} = 1$, we can rewrite the previous relation as

$$\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\widetilde{\boldsymbol{Q}}^{(\ell)}\right] + \tilde{z}\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right] + \boldsymbol{I}_{n_\ell} - \frac{1}{\sigma_N}\mathbb{E}\left[\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\widetilde{\boldsymbol{Q}}^{(\ell)}\right]$$
$$= -\frac{1}{\sigma_N}\frac{n_\ell}{N}\mathbb{E}\left[\left(\tilde{z}\widetilde{m}_N^{(\ell)}(\tilde{z})+1\right)\widetilde{\boldsymbol{Q}}^{(\ell)}\right] - \frac{1}{N}\left(\frac{\mu_N^{(\ell)}}{\sigma_N^2}+\frac{\tilde{z}}{\sigma_N}\right)\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)2}\right] - \frac{1}{\sigma_N}\frac{1}{N}\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right] \tag{5.11}$$

where we have kept only the non-vanishing terms on the left-hand side.

5.A.3 Concentration of Bilinear Forms and Traces

Let us show that $\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(\ell)}]$ is a deterministic equivalent of $\widetilde{\boldsymbol{Q}}^{(\ell)}$ (Definition 2.17). That is, for all bounded (sequences of) vectors $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^{n_\ell}$ and matrices $\boldsymbol{A}\in\mathbb{R}^{n_\ell\times n_\ell}$,

$$\boldsymbol{a}^\top\left(\widetilde{\boldsymbol{Q}}^{(\ell)} - \mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right]\right)\boldsymbol{b} \xrightarrow[N\to+\infty]{\text{a.s.}} 0 \qquad \text{and} \qquad \frac{1}{n_\ell}\operatorname{Tr}\boldsymbol{A}\left(\widetilde{\boldsymbol{Q}}^{(\ell)} - \mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right]\right) \xrightarrow[N\to+\infty]{\text{a.s.}} 0.$$

To that end, we use the Poincaré-Nash inequality (Lemma 2.19) to prove that

$$\operatorname{Var}\left(\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}^{(\ell)}\boldsymbol{b}\right) \leqslant \frac{4}{\sigma_N^2 N}\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\left[\left\|\widetilde{\boldsymbol{Q}}^{(\ell)}\right\|^4\left\|\boldsymbol{T}^{(\ell)}\right\|^2\right] = \mathcal{O}_{\tilde{z}}(N^{-1}),$$

$$\mathbb{E}\left[\left|\boldsymbol{a}^\top\left(\widetilde{\boldsymbol{Q}}^{(\ell)} - \mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right]\right)\boldsymbol{b}\right|^4\right]$$
$$\leqslant \frac{16}{\sigma_N^2 N}\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\left[\left|\boldsymbol{a}^\top\left(\widetilde{\boldsymbol{Q}}^{(\ell)} - \mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right]\right)\boldsymbol{b}\right|^2\left\|\widetilde{\boldsymbol{Q}}^{(\ell)}\right\|^4\left\|\boldsymbol{T}^{(\ell)}\right\|^2\right] + \operatorname{Var}\left(\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}^{(\ell)}\boldsymbol{b}\right)^2$$
$$= \mathcal{O}_{\tilde{z}}(N^{-2})$$

$$\text{and} \qquad \operatorname{Var}\left(\frac{1}{n_\ell}\operatorname{Tr}\boldsymbol{A}\widetilde{\boldsymbol{Q}}^{(\ell)}\right) \leqslant \frac{4}{N}\frac{1}{\sigma_N^2 N}\|\boldsymbol{A}\|^2\mathbb{E}\left[\left\|\widetilde{\boldsymbol{Q}}^{(\ell)}\right\|^4\left\|\boldsymbol{T}^{(\ell)}\right\|^2\right] = \mathcal{O}_{\tilde{z}}(N^{-2}).$$

Then, we conclude with Lemma 2.20. The computations are similar to those performed in Section 2.3.1 therefore we skip them for brevity.

5.A.4 Expansion of the Mean Empirical Stieltjes Transform

Let $\widetilde{\boldsymbol{Q}}_0^{(\ell)}$ denote the resolvent of $\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$. Using the resolvent identity (Proposition 2.21), we can see that $\frac{1}{n_\ell}\operatorname{Tr}(\widetilde{\boldsymbol{Q}}_0^{(\ell)} - \widetilde{\boldsymbol{Q}}^{(\ell)}) = \mathcal{O}_{\tilde{z}}(N^{-1})$ therefore the low-rank perturbation $\boldsymbol{P}^{(\ell)}$ does not change the limiting spectral distribution and we can consider $\boldsymbol{P}^{(\ell)} = \boldsymbol{0}_{n_\ell\times\prod_{\ell'\neq\ell}n_{\ell'}}$ from now on.

Applying $\frac{1}{n_\ell}\operatorname{Tr}$ to Equation (5.11) and using the fact that $\frac{1}{n_\ell}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(\ell)2} = \widetilde{m}_N^{(\ell)\prime}$ as well as $\mathbb{E}[\widetilde{m}_N^{(\ell)2}(\tilde{z})] = \mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})]^2 + \operatorname{Var}(\widetilde{m}_N^{(\ell)}(\tilde{z}))$ with $\operatorname{Var}(\widetilde{m}_N^{(\ell)}(\tilde{z})) = \mathcal{O}_{\tilde{z}}(N^{-2})$, we find

$$\begin{aligned}\mathbb{E}\Big[\widetilde{m}_N^{(\ell)}(\tilde{z})\Big]^2 + \tilde{z}\mathbb{E}\Big[\widetilde{m}_N^{(\ell)}(\tilde{z})\Big] + 1 \\ = -\frac{1}{\sigma_N}\frac{n_\ell}{N}\mathbb{E}\Big[\tilde{z}\widetilde{m}_N^{(\ell)2}(\tilde{z}) + \widetilde{m}_N^{(\ell)}(\tilde{z})\Big] - \frac{1}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\mathbb{E}\Big[\widetilde{m}_N^{(\ell)\prime}(\tilde{z})\Big] + \mathcal{O}_{\tilde{z}}(N^{-\min(\frac{d}{2},2)}) \quad (5.12)\end{aligned}$$

where the $\mathcal{O}_{\tilde{z}}(N^{-\min(\frac{d}{2},2)})$ stems from $\sigma_N = \Theta(N^{\frac{d-2}{2}})$. We know that the Stieltjes transform m_{SC} of the Wigner semicircle distribution on $[-2,2]$ satisfies $m_{\mathrm{SC}}^2(\tilde{z}) + \tilde{z}m_{\mathrm{SC}}(\tilde{z}) + 1 = 0$ for all $\tilde{z} \in \mathbb{C}\setminus[-2,2]$ (see Section 2.2.1). Let us subtract this relation to Equation (5.12) and factorize by $\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})] - m_{\mathrm{SC}}(\tilde{z})$ using the relation $a^2 - b^2 = (a-b)(a+b)$.

$$\begin{aligned}\Big(\mathbb{E}\Big[\widetilde{m}_N^{(\ell)}(\tilde{z})\Big] - m_{\mathrm{SC}}(\tilde{z})\Big)\Big(\mathbb{E}\Big[\widetilde{m}_N^{(\ell)}(\tilde{z})\Big] + m_{\mathrm{SC}}(\tilde{z}) + \tilde{z}\Big) \\ = -\frac{1}{\sigma_N}\frac{n_\ell}{N}\mathbb{E}\Big[\tilde{z}\widetilde{m}_N^{(\ell)2}(\tilde{z}) + \widetilde{m}_N^{(\ell)}(\tilde{z})\Big] - \frac{1}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\mathbb{E}\Big[\widetilde{m}_N^{(\ell)\prime}(\tilde{z})\Big] + \mathcal{O}_{\tilde{z}}(N^{-\min(\frac{d}{2},2)}).\end{aligned}$$

Let $g_N^{(\ell)} : \tilde{z} \mapsto \frac{-1}{\tilde{z}+m_{\mathrm{SC}}(\tilde{z})+\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})]}$. Since, m_{SC} and $\mathbb{E}[\widetilde{m}_N^{(\ell)}]$ are Stieltjes transforms (Corollary 2.9), we have $|g_N^{(\ell)}(\tilde{z})| \leqslant |\Im[\tilde{z} + m_{\mathrm{SC}}(\tilde{z}) + \mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})]]|^{-1} \leqslant |\Im\tilde{z}|^{-1}$ therefore

$$\begin{aligned}&\mathbb{E}\Big[\widetilde{m}_N^{(\ell)}(\tilde{z})\Big] - m_{\mathrm{SC}}(\tilde{z}) \\ &= g_N^{(\ell)}(\tilde{z})\left[\frac{1}{\sigma_N}\frac{n_\ell}{N}\mathbb{E}\Big[\tilde{z}\widetilde{m}_N^{(\ell)2}(\tilde{z}) + \widetilde{m}_N^{(\ell)}(\tilde{z})\Big] + \frac{1}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\mathbb{E}\Big[\widetilde{m}_N^{(\ell)\prime}(\tilde{z})\Big]\right] + \mathcal{O}_{\tilde{z}}(N^{-\min(\frac{d}{2},2)}) \quad (5.13)\\ &= g_N^{(\ell)}(\tilde{z})\frac{1}{\sigma_N}\frac{n_\ell}{N}\mathbb{E}\Big[\tilde{z}\widetilde{m}_N^{(\ell)2}(\tilde{z}) + \widetilde{m}_N^{(\ell)}(\tilde{z})\Big] + \mathcal{O}_{\tilde{z}}(N^{-1}) \quad (5.14)\\ &= \mathcal{O}_{\tilde{z}}(N^{-\min(\frac{d-2}{2},1)}). \quad (5.15)\end{aligned}$$

Notice that the dominant term in the difference $\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})] - m_{\mathrm{SC}}(\tilde{z})$ differs depending on whether $d = 3$, $d = 4$ or $d \geqslant 5$ because $\sigma_N = \Theta(N^{\frac{d-2}{2}})$. The higher d, the faster the convergence of the empirical spectral distribution to the semicircle distribution. Indeed, at this point, Equation (5.15) already shows the pointwise convergence of $\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})]$ to m_{SC} and therefore the weak convergence of the corresponding distributions (Proposition 2.10). Yet, in order to show the confinement of the spectrum below, we need an explicit expansion of $\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})]$ with all the terms not dominated by N^{-1}, which we state in the following lemma.

Lemma 5.13. *$\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})] = m_{\mathrm{SC}}(\tilde{z}) + h_N^{(\ell)}(\tilde{z}) + \mathcal{O}_{\tilde{z}}(N^{-\min(\frac{d}{2},2)})$ with*

$$\begin{aligned}h_N^{(\ell)}(\tilde{z}) = \tau_N^{(\ell)}(\tilde{z})\left(\frac{1}{\sigma_N}\frac{n_\ell}{N}\Big(\tilde{z}t_N^{(\ell)2}(\tilde{z}) + t_N^{(\ell)}(\tilde{z})\Big) - \frac{1}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\frac{m_{\mathrm{SC}}(\tilde{z})}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\right) \\ \textit{where}\quad \tau_N^{(\ell)}(\tilde{z}) = \frac{-1}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\left(1 + \frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{\tilde{z}m_{\mathrm{SC}}^2(\tilde{z}) + m_{\mathrm{SC}}(\tilde{z})}{(\tilde{z}+2m_{\mathrm{SC}}(\tilde{z}))^2}\right) \\ \textit{and}\quad t_N^{(\ell)}(\tilde{z}) = m_{\mathrm{SC}}(\tilde{z}) - \frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{\tilde{z}m_{\mathrm{SC}}^2(\tilde{z}) + m_{\mathrm{SC}}(\tilde{z})}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}.\end{aligned}$$

Proof. Firstly, with the Poincaré-Nash inequality (Lemma 2.19), we can show that

$$\mathrm{Var}(\widetilde{m}_N^{(\ell)\prime}(\tilde{z})) \leqslant \sum_{i=1}^{n_\ell} \sum_{j=1}^{\prod_{\ell'\neq\ell} n_{\ell'}} \mathbb{E}\left[\left|\frac{1}{n_\ell}\frac{\partial \mathrm{Tr}\,\widetilde{\boldsymbol{Q}}^{(\ell)2}}{\partial N_{i,j}^{(\ell)}}\right|^2\right] \leqslant \frac{16}{n_\ell \sigma_N^2 N^2}\mathbb{E}\left[\left\|\widetilde{\boldsymbol{Q}}^{(\ell)}\right\|^6 \left\|\boldsymbol{N}^{(\ell)}\right\|^2\right] = \mathscr{O}_{\tilde{z}}(N^{-2}).$$

Taking $\frac{1}{n_\ell}\mathrm{Tr}$ of Equation (5.11) with $\boldsymbol{P}^{(\ell)} = \boldsymbol{0}_{n_\ell \times \prod_{\ell'\neq\ell} n_{\ell'}}$ and differentiating with respect to the complex variable $\tilde{z}$ under $\mathbb{E}$ (which is possible because the integrand can be upper bounded on every compact subset of $\mathbb{C}\setminus\mathbb{R}$), we find

$$\begin{aligned}
&\mathbb{E}\left[\left(2\widetilde{m}_N^{(\ell)}(\tilde{z}) + \tilde{z}\right)\widetilde{m}_N^{(\ell)\prime}(\tilde{z}) + \widetilde{m}_N^{(\ell)}(\tilde{z})\right] \\
&= \mathbb{E}\left[-\frac{1}{\sigma_N}\frac{n_\ell}{N}\left(\widetilde{m}_N^{(\ell)2}(\tilde{z}) + 2\tilde{z}\widetilde{m}_N^{(\ell)}(\tilde{z})\widetilde{m}_N^{(\ell)\prime}(\tilde{z}) + \widetilde{m}_N^{(\ell)\prime}(\tilde{z})\right) - \frac{1}{N}\left(\frac{\mu_N^{(\ell)}}{\sigma_N^2} + \frac{\tilde{z}}{\sigma_N}\right)\widetilde{m}_N^{(\ell)\prime\prime}(\tilde{z}) - \frac{2\widetilde{m}^{(\ell)\prime}(\tilde{z})}{\sigma_N N}\right]. \quad (5.16)
\end{aligned}$$

Then, since $\mathrm{Var}\,\widetilde{m}_N^{(\ell)}(\tilde{z}) = \mathscr{O}_{\tilde{z}}(N^{-2})$ and $\mathrm{Var}\,\widetilde{m}_N^{(\ell)\prime}(\tilde{z}) = \mathscr{O}_{\tilde{z}}(N^{-2})$, we have $\mathbb{E}[\widetilde{m}_N^{(\ell)2}(\tilde{z})] = \mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})]^2 + \mathscr{O}_{\tilde{z}}(N^{-2})$ and $\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})\widetilde{m}_N^{(\ell)\prime}(\tilde{z})] = \mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})]\mathbb{E}[\widetilde{m}_N^{(\ell)\prime}(\tilde{z})] + \mathscr{O}_{\tilde{z}}(N^{-2})$ (using the Cauchy-Schwarz inequality to upper bound the covariance). And, with the fact that $\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})] = m_{\mathrm{SC}}(\tilde{z}) + \mathscr{O}_{\tilde{z}}(N^{-\min(\frac{d-2}{2},1)})$ (Equation (5.15)), we obtain from Equation (5.16) that

$$\mathbb{E}\left[\widetilde{m}_N^{(\ell)\prime}(\tilde{z})\right] = \frac{-m_{\mathrm{SC}}(\tilde{z})}{\tilde{z} + 2m_{\mathrm{SC}}(\tilde{z})} + \mathscr{O}_{\tilde{z}}(N^{-\min(\frac{d-2}{2},1)})$$

since $|\tilde{z} + 2m_{\mathrm{SC}}(\tilde{z})|^{-1} \leqslant |\tilde{z}|^{-1} \leqslant |\Im\tilde{z}|^{-1}$. Moreover, using Equation (5.15) in Equation (5.14), we find

$$\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right] = m_{\mathrm{SC}}(\tilde{z}) + g_N^{(\ell)}(\tilde{z})\frac{1}{\sigma_N}\frac{n_\ell}{N}\left(\tilde{z}m_{\mathrm{SC}}^2(\tilde{z}) + m_{\mathrm{SC}}(\tilde{z})\right) + \mathscr{O}_{\tilde{z}}(N^{-1}). \quad (5.17)$$

We can now inject the last two relations into Equation (5.13):

$$\begin{aligned}
\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right] = m_{\mathrm{SC}}(\tilde{z}) &+ g_N^{(\ell)}(\tilde{z})\frac{\tilde{z}}{\sigma_N}\frac{n_\ell}{N}\left(m_{\mathrm{SC}}(\tilde{z}) + g_N^{(\ell)}(\tilde{z})\frac{1}{\sigma_N}\frac{n_\ell}{N}\left(\tilde{z}m_{\mathrm{SC}}^2(\tilde{z}) + m_{\mathrm{SC}}(\tilde{z})\right)\right)^2 \\
&+ g_N^{(\ell)}(\tilde{z})\frac{1}{\sigma_N}\frac{n_\ell}{N}\left(m_{\mathrm{SC}}(\tilde{z}) + g_N^{(\ell)}(\tilde{z})\frac{1}{\sigma_N}\frac{n_\ell}{N}\left(\tilde{z}m_{\mathrm{SC}}^2(\tilde{z}) + m_{\mathrm{SC}}(\tilde{z})\right)\right) \\
&- g_N^{(\ell)}(\tilde{z})\frac{1}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\frac{m_{\mathrm{SC}}(\tilde{z})}{\tilde{z} + 2m_{\mathrm{SC}}(\tilde{z})} + \mathscr{O}_{\tilde{z}}(N^{-\min(\frac{d}{2},2)}). \quad (5.18)
\end{aligned}$$

We only need to handle the asymptotic behavior of $g_N^{(\ell)}(\tilde{z})$ to conclude the proof. With Equation (5.15) and the fact that $|\tilde{z} + 2m_{\mathrm{SC}}(\tilde{z})|^{-1} \leqslant |\Im\tilde{z}|^{-1}$, we have $g_N^{(\ell)}(\tilde{z}) = \frac{-1}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})} + \mathscr{O}_{\tilde{z}}(N^{-\min(\frac{d-2}{2},1)})$. We can then use this relation in Equation (5.17):

$$\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right] - m_{\mathrm{SC}}(\tilde{z}) = -\frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{\tilde{z}m_{\mathrm{SC}}^2(\tilde{z}) + m_{\mathrm{SC}}(\tilde{z})}{\tilde{z} + 2m_{\mathrm{SC}}(\tilde{z})} + \mathscr{O}_{\tilde{z}}(N^{-1}).$$

Therefore, we have

$$g_N^{(\ell)}(\tilde{z}) = \frac{-1}{\tilde{z} + m_{\mathrm{SC}}(\tilde{z}) + \mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right]} = \frac{-1}{\tilde{z} + 2m_{\mathrm{SC}}(\tilde{z})}\left(1 + \frac{\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right] - m_{\mathrm{SC}}(\tilde{z})}{\tilde{z} + 2m_{\mathrm{SC}}(\tilde{z})}\right)^{-1}$$

$$= \frac{-1}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\left(1-\frac{\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right]-m_{\mathrm{SC}}(\tilde{z})}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}+\mathscr{O}_{\tilde{z}}(N^{-\min(d-2,2)})\right)$$
$$= \frac{-1}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{\tilde{z}m_{\mathrm{SC}}^2(\tilde{z})+m_{\mathrm{SC}}(\tilde{z})}{(\tilde{z}+2m_{\mathrm{SC}}(\tilde{z}))^2}\right)+\mathscr{O}_{\tilde{z}}(N^{-1}).$$

Eventually, Equation (5.18) becomes

$$\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right] = m_{\mathrm{SC}}(\tilde{z})+\tau_N^{(\ell)}(\tilde{z})\frac{\tilde{z}}{\sigma_N}\frac{n_\ell}{N}\left(m_{\mathrm{SC}}(\tilde{z})-\frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{\tilde{z}m_{\mathrm{SC}}^2(\tilde{z})+m_{\mathrm{SC}}(\tilde{z})}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\right)^2$$
$$+\tau_N^{(\ell)}(\tilde{z})\frac{1}{\sigma_N}\frac{n_\ell}{N}\left(m_{\mathrm{SC}}(\tilde{z})-\frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{\tilde{z}m_{\mathrm{SC}}^2(\tilde{z})+m_{\mathrm{SC}}(\tilde{z})}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\right)$$
$$-\tau_N^{(\ell)}(\tilde{z})\frac{1}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\frac{m_{\mathrm{SC}}(\tilde{z})}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}+\mathscr{O}_{\tilde{z}}(N^{-\min(\frac{d}{2},2)}).$$

with $\tau_N^{(\ell)}:\tilde{z}\mapsto\frac{-1}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{\tilde{z}m_{\mathrm{SC}}^2(\tilde{z})+m_{\mathrm{SC}}(\tilde{z})}{(\tilde{z}+2m_{\mathrm{SC}}(\tilde{z}))^2}\right)$. □

Remark 5.14. Note that $\tilde{z}\mapsto\frac{-1}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}$ is the Stieltjes transform of a probability measure with the same support as μ_{SC}, i.e., $[-2,2]$ (Theorem 2.8). Thus, $h_N^{(\ell)}$ can be analytically extended on $\mathbb{C}\setminus[-2,2]$ and $h_N^{(\ell)}(x)\in\mathbb{R}$ for all $x\in\mathbb{R}\setminus[-2,2]$.

5.A.5 Confinement of the Spectrum

The confinement of the spectrum is proven in the same manner as in Section 2.3.2. Let $\varepsilon>0$, $\varphi:\mathbb{R}\mapsto[0,1]$ be an infinitely differentiable function which equals 1 on $[-2,2]$ and 0 on $\mathbb{R}\setminus[-2-\varepsilon,2+\varepsilon]$ and $\psi=1-\varphi$. We want to show that $\mathrm{Tr}\left(\psi\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right)\to 0$ almost surely as $N\to+\infty$.

First of all, we show the convergence in mean with the Helffer-Sjöstrand formula (Proposition 2.12).

$$\mathbb{E}\left[\frac{1}{n_\ell}\mathrm{Tr}\left(\varphi\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right)\right] = \frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar{\tilde{z}}}(\tilde{z})\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right]\mathrm{d}\tilde{z}$$
$$= \frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar{\tilde{z}}}(\tilde{z})m_{\mathrm{SC}}(\tilde{z})\,\mathrm{d}\tilde{z}+\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar{\tilde{z}}}(\tilde{z})h_N^{(\ell)}(\tilde{z})\,\mathrm{d}\tilde{z}$$
$$+\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar{\tilde{z}}}(\tilde{z})\times\mathscr{O}_{\tilde{z}}(N^{-\min(\frac{d}{2},2)})\,\mathrm{d}\tilde{z}.$$

where we have used the expression of $\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})]$ given by Lemma 5.13. The first integral is $\int_{\mathbb{R}}\varphi\,\mathrm{d}\mu_{\mathrm{SC}}=1$ while the last one is $\mathscr{O}(N^{-\min(\frac{d}{2},2)})$ with q chosen sufficiently large so that $\frac{\partial\Phi_q[\varphi]}{\partial\bar{\tilde{z}}}(\tilde{z})$ cancels the divergence of $\mathscr{O}_{\tilde{z}}(N^{-\min(\frac{d}{2},2)})$ near the real axis. In order to evaluate the second integral, we use Lemma 2.14. The function $h_N^{(\ell)}$ is analytic on $\mathbb{C}\setminus[-2,2]$, $\lim_{|\tilde{z}|\to+\infty}h(\tilde{z})=0$ and $h_N^{(\ell)}(\bar{\tilde{z}})=\overline{h(\tilde{z})}$ for all $\tilde{z}\in\mathbb{C}\setminus[-2,2]$. Thus, we just need to show that there exist an integer n_0 and a constant $C>0$ such that $|h_N^{(\ell)}(\tilde{z})|\leqslant C\max(\mathrm{Dist}(\tilde{z},[-2,2])^{-n_0},1)$ for all $\tilde{z}\in\mathbb{C}\setminus[-2,2]$. Firstly, we find an upper bound for $t_N^{(\ell)}(\tilde{z})$. Since $\tilde{z}\mapsto\frac{-1}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}$ is the Stieltjes transform of a probability measure on $[-2,2]$ and $-(\tilde{z}m_{\mathrm{SC}}(\tilde{z})+1)=m_{\mathrm{SC}}^2(\tilde{z})$, we have

$$\left|t_N^{(\ell)}(\tilde{z})\right| = |m_{\mathrm{SC}}(\tilde{z})|\left|1-\frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{\tilde{z}m_{\mathrm{SC}}(\tilde{z})+1}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\right|\leqslant\frac{1}{\mathrm{Dist}(\tilde{z},[-2,2])}\left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{1}{\mathrm{Dist}(\tilde{z},[-2,2])^3}\right)$$

$$\leqslant \left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\right)\max(\mathrm{Dist}(\tilde{z},[-2,2])^{-4},1).$$

Moreover, since $|\tilde{z}| \leqslant \mathrm{Dist}(\tilde{z},[-2,2])+2$, we also have

$$\left|\tilde{z}t_N^{(\ell)}(\tilde{z})\right| \leqslant \left(1+\frac{2}{\mathrm{Dist}(\tilde{z},[-2,2])}\right)\left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{1}{\mathrm{Dist}(\tilde{z},[-2,2])^3}\right) \leqslant 3\left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\right)\max(\mathrm{Dist}(\tilde{z},[-2,2])^{-4},1).$$

Similarly,

$$\left|\tau_N^{(\ell)}(\tilde{z})\right| = \left|\frac{-1}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\right|\left|1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\frac{m_{\mathrm{SC}}(\tilde{z})(\tilde{z}m_{\mathrm{SC}}(\tilde{z})+1)}{(\tilde{z}+2m_{\mathrm{SC}}(\tilde{z}))^2}\right| \leqslant \left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\right)\max(\mathrm{Dist}(\tilde{z},[-2,2])^{-6},1).$$

Hence, we can upper bound $|h_N^{(\ell)}(\tilde{z})|$:

$$\begin{aligned}\left|h_N^{(\ell)}(\tilde{z})\right| &\leqslant \left|\tau_N^{(\ell)}(\tilde{z})\right|\left(\frac{1}{\sigma_N}\frac{n_\ell}{N}\left|t_N^{(\ell)}(\tilde{z})\right|\left(\left|\tilde{z}t_N^{(\ell)}(\tilde{z})\right|+1\right)+\frac{1}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\left|\frac{m_{\mathrm{SC}}(\tilde{z})}{\tilde{z}+2m_{\mathrm{SC}}(\tilde{z})}\right|\right)\\ &\leqslant \left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\right)\left[\frac{1}{\sigma_N}\frac{n_\ell}{N}\left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\right)\left[3\left(1+\frac{1}{\sigma_N}\frac{n_\ell}{N}\right)+1\right]+\frac{1}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\right]\max(\mathrm{Dist}(\tilde{z},[-2,2])^{-14},1).\end{aligned}$$

Therefore, following Lemma 2.14, as soon as $q \geqslant 14$,

$$\begin{aligned}\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\bar{\tilde{z}}}(\tilde{z})h_N^{(\ell)}(\tilde{z})\,\mathrm{d}\tilde{z} &= \lim_{y\downarrow 0}\frac{1}{\pi}\int_{-2}^{2}\varphi(x)\Im[h_N^{(\ell)}(x+\mathrm{i}y)]\,\mathrm{d}x\\ &= \lim_{y\downarrow 0}\frac{1}{\pi}\int_{-2}^{2}\Im[h_N^{(\ell)}(x+\mathrm{i}y)]\,\mathrm{d}x\\ &= 0\end{aligned}$$

since $\varphi(x)=1$ for all $x\in[-2,2]$ and $\lim_{y\to+\infty}-\mathrm{i}yh_N^{(\ell)}(\mathrm{i}y)=0$. Thus we can conclude that

$$\mathbb{E}\left[\frac{1}{n_\ell}\mathrm{Tr}\left(\varphi\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right)\right] = 1+\mathcal{O}(N^{-\min(\frac{d}{2},2)}),$$

$$\text{i.e.,}\quad \mathbb{E}\left[\mathrm{Tr}\left(\psi\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right)\right] = \mathcal{O}(N^{-\min(\frac{d-2}{2},1)}).$$

Secondly, we prove the almost sure convergence of $\mathrm{Tr}\left(\psi\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right)$ to 0 by showing that its variance is $\mathcal{O}(N^{-\min(\frac{d}{2},2)})$ (and Lemma 2.20 implies the result). With the Poincaré-Nash inequality (Lemma 2.19), we have

$$\begin{aligned}&\mathrm{Var}\left(\mathrm{Tr}\left(\psi\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right)\right) = \mathrm{Var}\left(\mathrm{Tr}\left(\varphi\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right)\right)\\ &\leqslant \sum_{i=1}^{n_\ell}\sum_{j=1}^{\prod_{\ell'\neq\ell}n_{\ell'}}\mathbb{E}\left[\left|\frac{\partial\mathrm{Tr}\left(\varphi\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right)}{\partial N_{i,j}^{(\ell)}}\right|^2\right]\\ &= \sum_{i=1}^{n_\ell}\sum_{j=1}^{\prod_{\ell'\neq\ell}n_{\ell'}}\mathbb{E}\left[\left|\mathrm{Tr}\left(\varphi'\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\frac{\partial}{\partial N_{i,j}^{(\ell)}}\left[\frac{1}{\sigma_N}\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}\right]\right)\right|^2\right]\\ &= \frac{1}{\sigma_N^2}\frac{1}{N^2}\sum_{i=1}^{n_\ell}\sum_{j=1}^{\prod_{\ell'\neq\ell}n_{\ell'}}\mathbb{E}\left[\left|\left[2\varphi'\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top}-\mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\boldsymbol{N}^{(\ell)}\right]_{i,j}\right|^2\right]\end{aligned}$$

$$= \frac{1}{\sigma_N}\frac{4}{N}\mathbb{E}\left[\operatorname{Tr} u\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right] + \frac{4}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\mathbb{E}\left[\operatorname{Tr}\varphi'^2\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right]$$

where $u : x \mapsto x\varphi'^2(x)$ and φ'^2 are infinitely differentiable functions with compact support which equal 0 on $[-2,2]$. Hence, applying similarly the Helffer-Sjöstrand formula (Proposition 2.12), we find

$$\begin{aligned}\frac{1}{\sigma_N}\frac{4}{N}\mathbb{E}\left[\operatorname{Tr} u\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right] &= \frac{4}{\sigma_N}\frac{n_\ell}{N}\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[u]}{\partial\bar{\tilde{z}}}(\tilde{z})\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right]\mathrm{d}\tilde{z}\\ &= \mathcal{O}(N^{-\min(d-1,\frac{d+2}{2})})\end{aligned}$$

and

$$\begin{aligned}\frac{4}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\mathbb{E}\left[\operatorname{Tr}\varphi'^2\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right] &= 4\frac{n_\ell}{N}\frac{\mu_N^{(\ell)}}{\sigma_N^2}\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi']}{\partial\bar{\tilde{z}}}(\tilde{z})\mathbb{E}\left[\widetilde{m}_N^{(\ell)}(\tilde{z})\right]\mathrm{d}\tilde{z}\\ &= \mathcal{O}(N^{-\min(\frac{d}{2},2)})\end{aligned}$$

for q chosen sufficiently large. Thus, $\operatorname{Var}\left(\operatorname{Tr}\left(\psi\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]\right)\right)\right) = \mathcal{O}(N^{-\min(\frac{d}{2},2)})$.

5.A.6 Deterministic Equivalent

With the rescaling $(z, \boldsymbol{Q}^{(\ell)}(z)) \curvearrowright (\tilde{z}, \widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z}))$, Equation (5.8) becomes

$$\frac{1}{\sigma_N}\mathbb{E}\left[\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\widetilde{\boldsymbol{Q}}^{(\ell)}\right] = \frac{1}{\sigma_N}\mathbb{E}\left[\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top}\widetilde{\boldsymbol{Q}}^{(\ell)}\right] - \frac{1}{\sigma_N^2}\mathbb{E}\left[\frac{n_\ell}{N}\widetilde{m}_N^{(\ell)}(\tilde{z})\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\widetilde{\boldsymbol{Q}}^{(\ell)} + \frac{1}{N}\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\widetilde{\boldsymbol{Q}}^{(\ell)2}\right]$$

where $\left\|\frac{1}{\sigma_N^2}\mathbb{E}\left[\frac{n_\ell}{N}\widetilde{m}_N^{(\ell)}(\tilde{z})\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\widetilde{\boldsymbol{Q}}^{(\ell)} + \frac{1}{N}\boldsymbol{P}^{(\ell)}\boldsymbol{T}^{(\ell)\top}\widetilde{\boldsymbol{Q}}^{(\ell)2}\right]\right\| \to 0$ as $N \to +\infty$ since $\|\boldsymbol{P}^{(\ell)}\| = \mathcal{O}(N^{\frac{d-2}{4}})$ (this is true both in Frobenius norm and spectral norm as $\boldsymbol{P}^{(\ell)}$ has rank r_ℓ) and $\|\boldsymbol{T}^{(\ell)}\| = \mathcal{O}(N^{\frac{d-2}{2}})$. Hence, with Equation (5.11) and Equation (5.15), we have

$$\left\|m_{\mathrm{SC}}(\tilde{z})\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right] + \tilde{z}\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right] + \boldsymbol{I}_{n_\ell} - \frac{1}{\sigma_N}\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top}\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(\ell)}\right]\right\| \xrightarrow[N\to+\infty]{} 0$$

and, since $m_{\mathrm{SC}}(\tilde{z}) + \tilde{z} = \frac{-1}{m_{\mathrm{SC}}(\tilde{z})}$, we can define the following deterministic equivalent (Definition 2.17):

$$\widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) \longleftrightarrow \bar{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) \overset{\text{def}}{=} \left(\frac{1}{\sigma_N}\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top} + \frac{1}{m_{\mathrm{SC}}(\tilde{z})}\boldsymbol{I}_{n_\ell}\right)^{-1}.$$

5.B Proof of Theorem 5.5

Recall that $\widetilde{\boldsymbol{Q}}^{(\ell)}$ is the resolvent of $\frac{1}{\sigma_N}\left[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$ while $\widetilde{\boldsymbol{Q}}_0^{(\ell)}$ denotes the resolvent of the same model without signal, $\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right]$.

5.B.1 Convergence of Bilinear Forms

First of all, we must show the *almost sure* convergence $\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}_0^{(\ell)}\boldsymbol{b} - m_{\mathrm{SC}}(\tilde{z})\langle\boldsymbol{a},\boldsymbol{b}\rangle \to 0$ for all bounded (sequences of) vectors $\boldsymbol{a},\boldsymbol{b} \in \mathbb{R}^{n_\ell}$. Given the concentration result proven in Section 5.A.3, we just need to show that $\boldsymbol{a}^\top\mathbb{E}[\widetilde{\boldsymbol{Q}}_0^{(\ell)}]\boldsymbol{b} - m_{\mathrm{SC}}(\tilde{z})\langle\boldsymbol{a},\boldsymbol{b}\rangle \to 0$ as $N \to +\infty$.

Let us multiply Equation (5.11) when $\boldsymbol{P}^{(\ell)} = \boldsymbol{0}_{n_\ell \times \prod_{\ell' \neq \ell} n_{\ell'}}$ by $\boldsymbol{a}^\top$ on the left and $\boldsymbol{b}$ on the right.

$$\mathbb{E}\Big[\widetilde{m}_N^{(\ell)}(\tilde{z})\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(\ell)}\boldsymbol{b}\Big] + \tilde{z}\mathbb{E}\Big[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(\ell)}\boldsymbol{b}\Big] + \langle \boldsymbol{a}, \boldsymbol{b}\rangle = \mathcal{O}_z(N^{-\min(\frac{d-2}{2},1)}).$$

Then, using the fact that $\mathbb{E}[\widetilde{m}_N^{(\ell)}(\tilde{z})] = m_{\mathrm{SC}}(\tilde{z}) + \mathcal{O}_{\tilde{z}}(N^{-\min(\frac{d-2}{2},1)})$ and $m_{\mathrm{SC}}(\tilde{z}) + \tilde{z} = \frac{-1}{m_{\mathrm{SC}}(\tilde{z})}$, we obtain the desired result: $\mathbb{E}[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(\ell)}\boldsymbol{b}] = m_{\mathrm{SC}}(\tilde{z})\langle \boldsymbol{a}, \boldsymbol{b}\rangle + \mathcal{O}_{\tilde{z}}(N^{-\min(\frac{d-2}{2},1)})$.

5.B.2 Isolated Eigenvalues

We seek eigenvalues of $\frac{1}{\sigma_N}\Big[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\Big]$ which stay outside the support of the semicircle distribution $[-2,2]$. That is, we seek $\tilde{\xi} \in \mathbb{R}\setminus[-2,2]$ such that

$$\det\left(\frac{1}{\sigma_N}\Big[\boldsymbol{T}^{(\ell)}\boldsymbol{T}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\Big] - \tilde{\xi}\boldsymbol{I}_{n_\ell}\right) = 0.$$

Using the expansion $\boldsymbol{T}^{(\ell)} = \boldsymbol{P}^{(\ell)} + \frac{1}{\sqrt{N}}\boldsymbol{N}^{(\ell)}$, this is equivalent to

$$\det\left(\frac{1}{\sigma_N}\left(\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top} + \frac{1}{\sqrt{N}}\boldsymbol{P}^{(\ell)}\boldsymbol{N}^{(\ell)\top} + \frac{1}{\sqrt{N}}\boldsymbol{N}^{(\ell)}\boldsymbol{P}^{(\ell)\top}\right)\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi}) + \boldsymbol{I}_{n_\ell}\right)$$
$$\times \det\left(\frac{1}{\sigma_N}\left[\frac{1}{N}\boldsymbol{N}^{(\ell)}\boldsymbol{N}^{(\ell)\top} - \mu_N^{(\ell)}\boldsymbol{I}_{n_\ell}\right] - \tilde{\xi}\boldsymbol{I}_{n_\ell}\right) = 0$$

where the second determinant is non-zero for N large enough from the confinement of the spectrum proven in Section 5.A.5. Then, we know that $\mathcal{P} = [\![\mathcal{H}; \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(d)}]\!]$ therefore $\boldsymbol{P}^{(\ell)} = \boldsymbol{X}^{(\ell)}\boldsymbol{L}^{(\ell)}$ with $\boldsymbol{L}^{(\ell)} = \boldsymbol{H}^{(\ell)}\Big(\bigotimes_{\ell' \neq \ell}\boldsymbol{X}^{(\ell')\top}\Big)$ and we can write $\frac{1}{\sigma_N}\Big(\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top} + \frac{1}{\sqrt{N}}\boldsymbol{P}^{(\ell)}\boldsymbol{N}^{(\ell)\top} + \frac{1}{\sqrt{N}}\boldsymbol{N}^{(\ell)}\boldsymbol{P}^{(\ell)\top}\Big)$ as the matrix product $\begin{bmatrix}\boldsymbol{X}^{(\ell)} & \boldsymbol{X}^{(\ell)} & \frac{1}{\sigma_N\sqrt{N}}\boldsymbol{N}^{(\ell)}\boldsymbol{L}^{(\ell)\top}\end{bmatrix}\begin{bmatrix}\frac{1}{\sigma_N}\boldsymbol{H}^{(\ell)}\boldsymbol{H}^{(\ell)\top}\boldsymbol{X}^{(\ell)\top} \\ \frac{1}{\sigma_N\sqrt{N}}\boldsymbol{L}^{(\ell)}\boldsymbol{N}^{(\ell)\top} \\ \boldsymbol{X}^{(\ell)\top}\end{bmatrix}$. Thus, with Sylvester's identity (Proposition 2.22), we are left to evaluate a $3r_\ell \times 3r_\ell$ determinant:

$$\det\begin{bmatrix}
\frac{1}{\sigma_N}\boldsymbol{H}^{(\ell)}\boldsymbol{H}^{(\ell)\top}\boldsymbol{X}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{X}^{(\ell)} + \boldsymbol{I}_{r_\ell} & \frac{1}{\sigma_N}\boldsymbol{H}^{(\ell)}\boldsymbol{H}^{(\ell)\top}\boldsymbol{X}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{X}^{(\ell)} & \frac{1}{\sigma_N^2\sqrt{N}}\boldsymbol{H}^{(\ell)}\boldsymbol{H}^{(\ell)\top}\boldsymbol{X}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{N}^{(\ell)}\boldsymbol{L}^{(\ell)\top} \\
\frac{1}{\sigma_N\sqrt{N}}\boldsymbol{L}^{(\ell)}\boldsymbol{N}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{X}^{(\ell)} & \frac{1}{\sigma_N\sqrt{N}}\boldsymbol{L}^{(\ell)}\boldsymbol{N}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{X}^{(\ell)} + \boldsymbol{I}_{r_\ell} & \frac{1}{\sigma_N^2 N}\boldsymbol{L}^{(\ell)}\boldsymbol{N}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{N}^{(\ell)}\boldsymbol{L}^{(\ell)\top} \\
\boldsymbol{X}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{X}^{(\ell)} & \boldsymbol{X}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{X}^{(\ell)} & \frac{1}{\sigma_N\sqrt{N}}\boldsymbol{X}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{N}^{(\ell)}\boldsymbol{L}^{(\ell)\top} + \boldsymbol{I}_{r_\ell}
\end{bmatrix} = 0.$$

From the convergence of bilinear forms and the orthonormality of the columns of $\boldsymbol{X}^{(\ell)}$, we have $\boldsymbol{X}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{X}^{(\ell)} \to m_{\mathrm{SC}}(\tilde{\xi})\boldsymbol{I}_{r_\ell}$ almost surely. Moreover, we can see that $\frac{1}{\sigma_N\sqrt{N}}\boldsymbol{L}^{(\ell)}\boldsymbol{N}^{(\ell)\top}\widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi})\boldsymbol{X}^{(\ell)}$ vanishes almost surely as $N \to +\infty$ since $\|\boldsymbol{L}^{(\ell)}\| = \mathcal{O}(N^{\frac{d-2}{4}})$ and $\|\boldsymbol{N}^{(\ell)}\boldsymbol{a}\| = \mathcal{O}(\sqrt{N})$ almost surely[16] for all

[16] This fact is not so easy to see (note that it does not depend on d!). If we naively upper bound $\|\boldsymbol{N}^{(\ell)}\boldsymbol{a}\|^2$ by $\|\boldsymbol{N}^{(\ell)}\|^2\|\boldsymbol{a}\|^2 = \mathcal{O}(N^{\frac{d-1}{2}})$, we do not find the desired upper bound. Instead, we can remark that $\|\boldsymbol{N}^{(\ell)}\boldsymbol{a}\|^2 = \boldsymbol{a}^\top\boldsymbol{N}^{(\ell)\top}\boldsymbol{N}^{(\ell)}\boldsymbol{a} = \boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{a}$ where $\boldsymbol{D} = \mathrm{Diag}(\lambda_1(\boldsymbol{N}^{(\ell)\top}\boldsymbol{N}^{(\ell)}), \ldots, \lambda_{n_\ell}(\boldsymbol{N}^{(\ell)\top}\boldsymbol{N}^{(\ell)}))$ (we assume N large enough so that $\prod_{\ell' \neq \ell} n_{\ell'} > n_\ell$ thus $\lambda_i(\boldsymbol{N}^{(\ell)\top}\boldsymbol{N}^{(\ell)}) = 0$ for $i > n_\ell$) and $\boldsymbol{V}$ follows a uniform distribution on the Stiefel manifold $V_{n_\ell}(\mathbb{R}^{\prod_{\ell' \neq \ell} n_{\ell'}})$ (Chikuse, 2003, Theorem 2.2.1). Therefore $\frac{1}{N}\|\boldsymbol{N}^{(\ell)}\boldsymbol{a}\|^2 \leqslant (\mu_N^{(\ell)} + 2\sigma_N)\boldsymbol{a}^\top\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{a}$. Without loss of generality, we can assume that $\boldsymbol{a} = \|\boldsymbol{a}\|\boldsymbol{e}_1^{(\prod_{\ell' \neq \ell} n_{\ell'})}$ (replace $\boldsymbol{a}$ and $\boldsymbol{V}$ by $\boldsymbol{O}\boldsymbol{a}$ and $\boldsymbol{O}\boldsymbol{V}$ for a well-chosen $\boldsymbol{O} \in \mathcal{O}_{\prod_{\ell' \neq \ell} n_{\ell'}}(\mathbb{R})$). From Mardia and Khatri (1977), we know that $[\boldsymbol{V}\boldsymbol{V}^\top]_{1,1}$ follows a beta distribution with parameters $\frac{n_\ell}{2}, \frac{\prod_{\ell' \neq \ell} n_{\ell'} - n_\ell}{2}$ so its moments are given by $\mathbb{E}[[\boldsymbol{V}\boldsymbol{V}^\top]_{1,1}^k] = \prod_{r=0}^{k-1}\frac{n_\ell + 2r}{\prod_{\ell' \neq \ell} n_{\ell'} + 2r}$ for all $k \geqslant 1$. This is enough to see that $(\mu_N^{(\ell)} + 2\sigma_N)\mathbb{E}[[\boldsymbol{V}\boldsymbol{V}^\top]_{1,1}] = (\mu_N^{(\ell)} + 2\sigma_N)\frac{n_\ell}{\prod_{\ell' \neq \ell} n_{\ell'}} = \mathcal{O}(1)$ and $(\mu_N^{(\ell)} + 2\sigma_N)^4\mathbb{E}[([\boldsymbol{V}\boldsymbol{V}^\top]_{1,1} - \mathbb{E}[[\boldsymbol{V}\boldsymbol{V}^\top]_{1,1}])^4] = \mathcal{O}(N^{-2})$, whence the almost sure statement $\frac{1}{N}\|\boldsymbol{N}^{(\ell)}\boldsymbol{a}\|^2 = \mathcal{O}(1)$.

bounded (sequences of) vectors $\boldsymbol{a} \in \mathbb{R}^{\prod_{\ell' \neq \ell} n_{\ell'}}$. The only term which remains to evaluate is the block (2,3):

$$\frac{1}{\sigma_N^2 N} \left\| \boldsymbol{L}^{(\ell)} \boldsymbol{N}^{(\ell)\top} \widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi}) \boldsymbol{N}^{(\ell)} \boldsymbol{L}^{(\ell)\top} \right\| \leqslant \frac{1}{\sigma_N} \frac{\left\| \boldsymbol{H}^{(\ell)} \right\|^2}{\sigma_N} \frac{\left\| \boldsymbol{N}^{(\ell)} \left(\bigotimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')} \right) \right\|^2}{N} \left\| \widetilde{\boldsymbol{Q}}_0^{(\ell)}(\tilde{\xi}) \right\| \xrightarrow[N \to +\infty]{\text{a.s.}} 0.$$

Eventually, as the determinant is a continuous function in the entries of the matrix, we have, in the large N limit,

$$\det \begin{bmatrix} \frac{m_{\mathrm{SC}}(\tilde{\xi})}{\sigma_N} \boldsymbol{H}^{(\ell)} \boldsymbol{H}^{(\ell)\top} + \boldsymbol{I}_{r_\ell} & \frac{m_{\mathrm{SC}}(\tilde{\xi})}{\sigma_N} \boldsymbol{H}^{(\ell)} \boldsymbol{H}^{(\ell)\top} & \boldsymbol{0}_{r_\ell \times r_\ell} \\ \boldsymbol{0}_{r_\ell \times r_\ell} & \boldsymbol{I}_{r_\ell} & \boldsymbol{0}_{r_\ell \times r_\ell} \\ m_{\mathrm{SC}}(\tilde{\xi}) \boldsymbol{I}_{r_\ell} & m_{\mathrm{SC}}(\tilde{\xi}) \boldsymbol{I}_{r_\ell} & \boldsymbol{I}_{r_\ell} \end{bmatrix} = 0.$$

Using twice the relation $\det \left[\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{smallmatrix} \right] = \det(\boldsymbol{A}\boldsymbol{D}^{-1} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{D})$ when $\boldsymbol{D}$ is invertible, this simplifies into

$$\det\left(\frac{m_{\mathrm{SC}}(\tilde{\xi})}{\sigma_N} \boldsymbol{H}^{(\ell)} \boldsymbol{H}^{(\ell)\top} + \boldsymbol{I}_{r_\ell} \right) = 0$$

Thus, we seek $\tilde{\xi}_{q_\ell}^{(\ell)} \in \mathbb{R} \setminus [-2,2]$ such that

$$\frac{m_{\mathrm{SC}}(\tilde{\xi}_{q_\ell}^{(\ell)})}{\sigma_N} s_{q_\ell}^2(\boldsymbol{P}^{(\ell)}) + 1 = 0, \qquad q_\ell \in [r_\ell].$$

Injecting the expression $m_{\mathrm{SC}}(\tilde{\xi}_{q_\ell}^{(\ell)}) = -\frac{\sigma_N}{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})}$ into the equation $m_{\mathrm{SC}}^2(\tilde{\xi}_{q_\ell}^{(\ell)}) + \tilde{\xi}_{q_\ell}^{(\ell)} m_{\mathrm{SC}}(\tilde{\xi}_{q_\ell}^{(\ell)}) + 1 = 0$ yields

$$\frac{\sigma_N^2}{s_{q_\ell}^4(\boldsymbol{P}^{(\ell)})} - \tilde{\xi}_{q_\ell}^{(\ell)} \frac{\sigma_N}{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})} + 1 = 0 \iff \tilde{\xi}_{q_\ell}^{(\ell)} = \frac{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})}{\sigma_N} + \frac{\sigma_N}{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})}.$$

As $\tilde{\xi}_{q_\ell}^{(\ell)} > 0$ by definition, it must be strictly greater than 2 (the right edge of the semicircle). This is true only if $\rho_{q_\ell}^{(\ell)} \overset{\text{def}}{=} \frac{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})}{\sigma_N} > 1$.

5.B.3 Eigenvector Alignments

Let $\hat{\boldsymbol{u}}_{i_\ell}^{(\ell)}$, $i_\ell \in [n_\ell]$, denote the i_ℓ-th left singular vector of $\boldsymbol{T}^{(\ell)}$ (sorted in non-increasing order of its corresponding singular value). From the spectral decomposition $\boldsymbol{T}^{(\ell)} \boldsymbol{T}^{(\ell)\top} = \sum_{i_\ell=1}^{n_\ell} s_{i_\ell}^2(\boldsymbol{T}^{(\ell)}) \hat{\boldsymbol{u}}_{i_\ell}^{(\ell)} \hat{\boldsymbol{u}}_{i_\ell}^{(\ell)\top}$, we have,

$$\widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) = \sum_{i_\ell=1}^{n_\ell} \frac{\hat{\boldsymbol{u}}_{q_\ell}^{(\ell)} \hat{\boldsymbol{u}}_{q_\ell}^{(\ell)\top}}{\frac{1}{\sigma_N}\left[s_{i_\ell}^2(\boldsymbol{T}^{(\ell)}) - \mu_N^{(\ell)} \right] - \tilde{z}}.$$

If $\rho_{q_\ell}^{(\ell)} > 1$, $q_\ell \in [r_\ell]$, then $s_{q_\ell}^2(\boldsymbol{T}^{(\ell)})$ is an isolated eigenvalue in the spectrum of $\boldsymbol{T}^{(\ell)} \boldsymbol{T}^{(\ell)\top}$ and $\frac{1}{\sigma_N}\left[s_{q_\ell}^2(\boldsymbol{T}^{(\ell)}) - \mu_N^{(\ell)} \right] \xrightarrow[N \to +\infty]{\text{a.s.}} \tilde{\xi}_{q_\ell}^{(\ell)}$. Hence, for any positively-oriented simple closed complex contour $\gamma_{q_\ell}^{(\ell)}$ circling around $\tilde{\xi}_{q_\ell}^{(\ell)}$, leaving all the other $\tilde{\xi}_{q'_\ell}^{(\ell)}$, $q'_\ell \neq q_\ell$, outside and not crossing $[-2,2]$, Cauchy's integral formula (Proposition 2.15) yields, for N large enough and any $\boldsymbol{a} \in \mathbb{R}^{n_\ell}$,

$$\left\langle \boldsymbol{a}, \hat{\boldsymbol{u}}_{q_\ell}^{(\ell)} \right\rangle^2 = -\frac{1}{2\mathrm{i}\pi} \oint_{\gamma_{q_\ell}^{(\ell)}} \boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) \boldsymbol{a} \, \mathrm{d}\tilde{z} \xrightarrow[N \to +\infty]{\text{a.s.}} -\frac{1}{2\mathrm{i}\pi} \oint_{\gamma_{q_\ell}^{(\ell)}} \boldsymbol{a}^\top \bar{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) \boldsymbol{a} \, \mathrm{d}\tilde{z}$$

by the dominated convergence theorem since, for all $\tilde{z} \in \gamma_{q_\ell}^{(\ell)}$, $\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z})\boldsymbol{a} \to \boldsymbol{a}^\top \bar{\boldsymbol{Q}}^{(\ell)}(\tilde{z})\boldsymbol{a}$ almost surely as $N \to +\infty$ by definition of the deterministic equivalent (Definition 2.17) and $\tilde{z} \in \gamma_{q_\ell}^{(\ell)} \mapsto |\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z})\boldsymbol{a}|$ is almost surely bounded for N large enough because we can choose $\gamma_{q_\ell}^{(\ell)}$ such that $\mathrm{Dist}(\tilde{\xi}_{q_\ell}^{(\ell)}, \gamma_{q_\ell}^{(\ell)}) \geqslant \varepsilon > 0$ and therefore $|\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z})\boldsymbol{a}| \leqslant \|\boldsymbol{a}\|^2 \|\widetilde{\boldsymbol{Q}}^{(\ell)}(\tilde{z})\| \leqslant \|\boldsymbol{a}\|^2/\varepsilon$ almost surely.

Using residue calculus, we can compute,

$$-\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_{q_\ell}^{(\ell)}} \boldsymbol{a}^\top \bar{\boldsymbol{Q}}^{(\ell)}(\tilde{z})\boldsymbol{a}\,\mathrm{d}\tilde{z} = -\lim_{\tilde{z}\to\tilde{\xi}_{q_\ell}^{(\ell)}} \left(\tilde{z}-\tilde{\xi}_{q_\ell}^{(\ell)}\right)\boldsymbol{a}^\top\left(\frac{1}{\sigma_N}\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top} + \frac{1}{m_{\mathrm{SC}}(\tilde{z})}\boldsymbol{I}_{n_\ell}\right)^{-1}\boldsymbol{a}.$$

Note that $\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top} = \boldsymbol{X}^{(\ell)}\boldsymbol{H}^{(\ell)}\boldsymbol{H}^{(\ell)\top}\boldsymbol{X}^{(\ell)\top}$ and there exist an $r_\ell \times r_\ell$ orthogonal matrix $\boldsymbol{O}^{(\ell)} \in \mathcal{O}_{r_\ell}(\mathbb{R})$ such that $\boldsymbol{H}^{(\ell)}\boldsymbol{H}^{(\ell)\top} = \boldsymbol{O}^{(\ell)}\boldsymbol{\Lambda}^{(\ell)}\boldsymbol{O}^{(\ell)\top}$ with $\boldsymbol{\Lambda}^{(\ell)} = \mathrm{Diag}(s_1^2(\boldsymbol{P}^{(\ell)}), \ldots, s_{r_\ell}^2(\boldsymbol{P}^{(\ell)}))$. Hence,

$$\boldsymbol{a}^\top\left(\frac{1}{\sigma_N}\boldsymbol{P}^{(\ell)}\boldsymbol{P}^{(\ell)\top} + \frac{1}{m_{\mathrm{SC}}(\tilde{z})}\boldsymbol{I}_{n_\ell}\right)^{-1}\boldsymbol{a} = \boldsymbol{a}^\top\boldsymbol{X}^{(\ell)}\boldsymbol{O}^{(\ell)}\left(\frac{1}{\sigma_N}\boldsymbol{\Lambda}^{(\ell)} + \frac{1}{m_{\mathrm{SC}}(\tilde{z})}\boldsymbol{I}_{n_\ell}\right)^{-1}\boldsymbol{O}^{(\ell)\top}\boldsymbol{X}^{(\ell)\top}\boldsymbol{a}.$$

Let us therefore compute the following quantity,

$$-\lim_{\tilde{z}\to\tilde{\xi}_{q_\ell}^{(\ell)}} \left(\tilde{z}-\tilde{\xi}_{q_\ell}^{(\ell)}\right)\left(\frac{s_{q'_\ell}^2(\boldsymbol{P}^{(\ell)})}{\sigma_N} + \frac{1}{m_{\mathrm{SC}}(\tilde{z})}\right)^{-1} = \begin{cases} 0 & \text{if } q'_\ell \neq q_\ell \\ \zeta_{q_\ell}^{(\ell)} & \text{if } q'_\ell = q_\ell \end{cases}, \qquad q'_\ell \in [r_\ell],$$

where we have used the fact that $m_{\mathrm{SC}}(\tilde{\xi}_{q'_\ell}^{(\ell)}) = -\frac{\sigma_N}{s_{q'_\ell}^2(\boldsymbol{P}^{(\ell)})}$. In order to handle the case $q'_\ell = q_\ell$, we use L'Hôpital's rule,

$$\begin{aligned}
\zeta_{q_\ell}^{(\ell)} &= -\left(\frac{\mathrm{d}}{\mathrm{d}\tilde{z}}\left[\frac{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})}{\sigma_N} + \frac{1}{m_{\mathrm{SC}}(\tilde{z})}\right]_{\tilde{z}=\tilde{\xi}_{q_\ell}^{(\ell)}}\right)^{-1} \\
&= \frac{m_{\mathrm{SC}}^2(\tilde{\xi}_{q_\ell}^{(\ell)})}{m'_{\mathrm{SC}}(\tilde{\xi}_{q_\ell}^{(\ell)})} \\
&= \frac{\sigma_N^2}{s_{q_\ell}^4(\boldsymbol{P}^{(\ell)})\, m'_{\mathrm{SC}}(\tilde{\xi}_{q_\ell}^{(\ell)})}.
\end{aligned}$$

In order to compute $m'_{\mathrm{SC}}(\tilde{\xi}_{q_\ell}^{(\ell)})$, let us differentiate the relation $m_{\mathrm{SC}}^2(\tilde{z}) + \tilde{z}m_{\mathrm{SC}}(\tilde{z}) + 1 = 0$,

$$\begin{aligned}
2m'_{\mathrm{SC}}(\tilde{z})\, m_{\mathrm{SC}}(\tilde{z}) + m_{\mathrm{SC}}(\tilde{z}) + \tilde{z}m'_{\mathrm{SC}}(\tilde{z}) &= 0, \\
m'_{\mathrm{SC}}(\tilde{z}) &= -\frac{m_{\mathrm{SC}}(\tilde{z})}{2m_{\mathrm{SC}}(\tilde{z}) + \tilde{z}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
m'_{\mathrm{SC}}(\tilde{\xi}_{q_\ell}^{(\ell)}) &= -\frac{-\frac{\sigma_N}{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})}}{-2\frac{\sigma_N}{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})} + \frac{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})}{\sigma_N} + \frac{\sigma_N}{s_{q_\ell}^2(\boldsymbol{P}^{(\ell)})}} \\
&= \frac{1}{\frac{s_{q_\ell}^4(\boldsymbol{P}^{(\ell)})}{\sigma_N^2} - 1}.
\end{aligned}$$

Back to our previous expression of $\zeta_{q_\ell}^{(\ell)}$, we now have,

$$\zeta_{q_\ell}^{(\ell)} = 1 - \frac{\sigma_N^2}{s_{q_\ell}^4(\boldsymbol{P}^{(\ell)})}.$$

Therefore, for all $\boldsymbol{a} \in \mathbb{R}^{n_\ell}$,

$$-\frac{1}{2\mathrm{i}\pi} \oint_{\gamma_{q_\ell}^{(\ell)}} \boldsymbol{a}^\top \bar{\boldsymbol{Q}}^{(\ell)}(\tilde{z}) \boldsymbol{a} \, \mathrm{d}\tilde{z} = \boldsymbol{a}^\top \boldsymbol{X}^{(\ell)} \boldsymbol{O}^{(\ell)} \boldsymbol{Z}_{q_\ell}^{(\ell)} \boldsymbol{O}^{(\ell)\top} \boldsymbol{X}^{(\ell)\top} \boldsymbol{a}$$

where $\boldsymbol{Z}_{q_\ell}^{(\ell)}$ is an $r_\ell \times r_\ell$ matrix with all its entries equal to 0 except $[\boldsymbol{Z}_{q_\ell}^{(\ell)}]_{q_\ell,q_\ell} = \zeta_{q_\ell}^{(\ell)}$. Thus, summing the alignments of $\hat{\boldsymbol{u}}_{q_\ell}^{(\ell)}$ with each column of $\boldsymbol{X}^{(\ell)}$ yields

$$\left\| \boldsymbol{X}^{(\ell)\top} \hat{\boldsymbol{u}}_{q_\ell}^{(\ell)} \right\|^2 \xrightarrow[N\to+\infty]{\text{a.s.}} \zeta_{q_\ell}^{(\ell)} \sum_{q'_\ell=1}^{r_\ell} O_{q'_\ell,q_\ell}^{(\ell)2} = \zeta_{q_\ell}^{(\ell)}.$$

5.C Proof of Lemma 5.10

Our proof of Lemma 5.10 uses the notion of ε-covering. An ε-covering of a *compact* set $\mathscr{K}$ for the norm $\|\cdot\|$ is a *finite* set $\mathscr{C} \subset \mathscr{K}$ such that for all $x \in \mathscr{K}$, there exists $\bar{x} \in \mathscr{C}$ such that $\|x - \bar{x}\| \leqslant \varepsilon$. We also define the covering number $N(\varepsilon, \mathscr{K}, \|\cdot\|)$ as the smallest possible number of elements in $\mathscr{C}$.

Moreover, we recall the definition of the Gamma function $\Gamma(s) = \int_0^{+\infty} t^{s-1} e^{-t} \mathrm{d}t$ and the (upper) incomplete Gamma function $\Gamma(s,x) = \int_x^{+\infty} t^{s-1} e^{-t} \mathrm{d}t$ for $s > 0$ and $x \geqslant 0$.

For our proof, we need to introduce a few preliminary results which are stated and proven below (except Lemma 5.17 for which a reference is given).

Lemma 5.15. *For $\ell \in [d]$ and $\varepsilon > 0$, let $\boldsymbol{\Delta}^{(\ell)} \in \mathbb{R}^{n_\ell \times r_\ell}$ be such that $\|\boldsymbol{\Delta}^{(\ell)}\| \leqslant \varepsilon$ and $\boldsymbol{V}^{(\ell)} \in V_{r_\ell}(\mathbb{R}^{n_\ell})$ be the matrix of its left singular vectors. For all $\boldsymbol{A}^{(\ell')} \in \mathbb{R}^{n_{\ell'} \times r_{\ell'}}$, $\ell' \neq \ell$,*

$$\left\| \mathcal{N}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{\Delta}^{(\ell)}, \ldots, \boldsymbol{A}^{(d)}) \right\|_{\mathrm{F}} \leqslant \varepsilon \left\| \mathcal{N}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{V}^{(\ell)}, \ldots, \boldsymbol{A}^{(d)}) \right\|_{\mathrm{F}}.$$

Proof. Let $\boldsymbol{V}^{(\ell)} \boldsymbol{\Sigma}^{(\ell)} \boldsymbol{W}^{(\ell)\top}$ be the singular value decomposition of $\boldsymbol{\Delta}^{(\ell)}$. We have,

$$\begin{aligned}
\left\| \mathcal{N}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{\Delta}^{(\ell)}, \ldots, \boldsymbol{A}^{(d)}) \right\|_{\mathrm{F}} &= \left\| \boldsymbol{\Delta}^{(\ell)\top} \boldsymbol{N}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{A}^{(\ell')} \right\|_{\mathrm{F}} \\
&= \left\| \boldsymbol{W}^{(\ell)\top} \boldsymbol{\Sigma}^{(\ell)} \boldsymbol{V}^{(\ell)\top} \boldsymbol{N}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{A}^{(\ell')} \right\|_{\mathrm{F}} \leqslant \underbrace{\left\| \boldsymbol{W}^{(\ell)\top} \boldsymbol{\Sigma}^{(\ell)} \right\|}_{=\varepsilon} \underbrace{\left\| \boldsymbol{V}^{(\ell)\top} \boldsymbol{N}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{A}^{(\ell')} \right\|_{\mathrm{F}}}_{= \left\| \mathcal{N}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{V}^{(\ell)}, \ldots, \boldsymbol{A}^{(d)}) \right\|_{\mathrm{F}}}
\end{aligned}$$

using the fact that $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}} \leqslant \|\boldsymbol{A}\| \|\boldsymbol{B}\|_{\mathrm{F}}$. □

Lemma 5.16. *Given $\boldsymbol{A}(\ell) \in V_{r_\ell}(\mathbb{R}^{n_\ell})$, $\ell \in [d]$,*

$$\left\| \mathcal{N}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)}) \right\|_{\mathrm{F}}^2 \sim \chi^2\left(\prod_{\ell \in [d]} r_\ell \right).$$

Proof. Firstly, observe that, for all $(q_1,\dots,q_d) \in \bigtimes_{\ell\in[d]}[r_\ell]$,

$$[\mathcal{N}(\boldsymbol{A}^{(1)},\dots,\boldsymbol{A}^{(d)})]_{q_1,\dots,q_d} = \sum_{i_1,\dots,i_d=1}^{n_1,\dots,n_d} \mathcal{N}_{i_1,\dots,i_d} A^{(1)}_{i_1,q_1}\dots A^{(d)}_{i_d,q_d} \sim \mathcal{N}(0,1)$$

since $\sum_{i_\ell=1}^{n_\ell} A^{(\ell)2}_{i_\ell,q_\ell} = 1$ for all $\ell \in [d]$. Then, we show that all the entries of $\mathcal{N}(\boldsymbol{A}^{(1)},\dots,\boldsymbol{A}^{(d)})$ are independent because their covariance is identity,

$$\begin{aligned}
&\mathbb{E}\Big[[\mathcal{N}(\boldsymbol{A}^{(1)},\dots,\boldsymbol{A}^{(d)})]_{q_1,\dots,q_d}[\mathcal{N}(\boldsymbol{A}^{(1)},\dots,\boldsymbol{A}^{(d)})]_{q'_1,\dots,q'_d}\Big] \\
&= \sum_{i_1,\dots,i_d=1}^{n_1,\dots,n_d}\sum_{i'_1,\dots,i'_d=1}^{n_1,\dots,n_d} \mathbb{E}\Big[\mathcal{N}_{i_1,\dots,i_d}\mathcal{N}_{i'_1,\dots,i'_d}\Big] A^{(1)}_{i_1,q_1}A^{(1)}_{i'_1,q'_1}\dots A^{(d)}_{i_d,q_d}A^{(d)}_{i'_d,q'_d} \\
&= \sum_{i_1,\dots,i_d=1}^{n_1,\dots,n_d} A^{(1)}_{i_1,q_1}A^{(1)}_{i_1,q'_1}\dots A^{(d)}_{i_d,q_d}A^{(d)}_{i_d,q'_d} \\
&= \begin{cases} 1 & \text{if } (q_1,\dots,q_d) = (q'_1,\dots,q'_d) \\ 0 & \text{otherwise} \end{cases}.
\end{aligned}$$

Hence, the result follows from the fact that $\|\mathcal{N}(\boldsymbol{A}^{(1)},\dots,\boldsymbol{A}^{(d)})\|_{\mathrm{F}}^2$ is the sum of $\prod_{\ell\in[d]} r_\ell$ squared independent $\mathcal{N}(0,1)$ variables. □

Lemma 5.17 (Hinrichs et al., 2017, Lemma 4.1)**.** *For $0<\varepsilon<1$, we have the following upper bound on the ε-covering number of the Stiefel manifold $V_r(\mathbb{R}^n)$ for the spectral norm $\|\cdot\|$,*

$$N(\varepsilon, V_r(\mathbb{R}^n), \|\cdot\|) \leqslant \left[\frac{C}{\varepsilon}\right]^{r\left(n-\frac{r+1}{2}\right)}.$$

where $C>0$ is a universal constant.

Lemma 5.18. *$\Gamma(s,x) \leqslant \max(1, e^{s-1})\Gamma(s)e^{-x/2}$ for all $x \geqslant 0$ and $s>0$.*

Proof. Given $s>0$, consider the function $f : x \in [0,+\infty[\mapsto \frac{\Gamma(s,x)}{Ce^{-x/2}}$ with $C>0$. Our goal is to show that $0<f\leqslant 1$ when C is well chosen. f is continuously differentiable on $[0,+\infty[$ and

$$f'(x) = \frac{1}{Ce^{-x/2}}\left(-x^{s-1}e^{-x} + \frac{1}{2}\Gamma(s,x)\right) \geqslant 0 \iff \Gamma(s,x) - 2x^{s-1}e^{-x} \geqslant 0.$$

Consider the function $g : x \in [0,+\infty[\mapsto \Gamma(s,x) - 2x^{s-1}e^{-x}$. g is also continuously differentiable on $[0,+\infty[$ and

$$g'(x) = x^{s-1}e^{-x} - 2(s-1)x^{s-2}e^{-x} \geqslant 0 \iff x \geqslant 2(s-1).$$

We distinguish two cases.

1. If $0<s\leqslant 1$, then g increases monotonically on $[0,+\infty[$. Since $\lim_{x\to+\infty} g(x) = 0$, we necessarily have $g(x)\leqslant 0$ for all $x\in[0,+\infty[$. Hence, $f(x)\leqslant f(0) = \frac{\Gamma(s)}{C}$ and we can choose $C=\Gamma(s)$.

2. If $s>1$, our conclusion stems from the following table.

x	0		$2(s-1)$		$+\infty$
$g'(x)$		$-$	0	$+$	
$g(x)$	$\Gamma(s)$	↘	$g(2(s-1))$	↗	0

Since g is strictly increasing on $[2(s-1),+\infty[$ and $\lim_{x\to+\infty} g(x) = 0$, we necessarily have $g(2(s-1)) < 0$. Hence, since $\Gamma(s) > 0$, the equation $g(x) = 0$ has a unique solution on $[0,+\infty[$ and it lies between 0 and $2(s-1)$. Let $x_0(s) \in [0, 2(s-1)]$ be this unique solution. Then,

$$\sup_{[0,+\infty[} f = f(x_0(s)) = \frac{\Gamma(s, x_0(s))}{Ce^{-x_0(s)/2}}$$

and we can choose $C = \Gamma(s, x_0(s))e^{x_0(s)/2}$. The final result follows from $\Gamma(s, x_0(s)) \leqslant \Gamma(s)$ and $e^{x_0(s)/2} \leqslant e^{s-1}$.

□

We are now ready to prove Lemma 5.10.

Let $\varepsilon > 0$ and $\mathscr{C}_\ell$ be an ε-covering of $V_{r_\ell}(\mathbb{R}^{n_\ell})$ for the spectral norm $\|\cdot\|$, $\ell \in [d]$. Since $\bigtimes_{\ell\in[d]} V_{r_\ell}(\mathbb{R}^{n_\ell})$ is compact, it contains an element $(\boldsymbol{A}_\star^{(1)}, \ldots, \boldsymbol{A}_\star^{(d)})$ such that

$$\sup_{\boldsymbol{A}^{(\ell)}\in V_{r_\ell}(\mathbb{R}^{n_\ell}),\ \ell\in[d]} \left\|\boldsymbol{\mathcal{N}}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)})\right\|_{\mathrm{F}} = \left\|\boldsymbol{\mathcal{N}}(\boldsymbol{A}_\star^{(1)}, \ldots, \boldsymbol{A}_\star^{(d)})\right\|_{\mathrm{F}}.$$

Let $\bar{\boldsymbol{A}}^{(\ell)} \in \mathscr{C}_\ell$ be such that $\boldsymbol{A}_\star^{(\ell)} = \bar{\boldsymbol{A}}^{(\ell)} + \boldsymbol{\Delta}^{(\ell)}$ with $\|\boldsymbol{\Delta}^{(\ell)}\| \leqslant \varepsilon$. Then, using the triangle inequality, Lemma 5.15 and the optimality of $(\boldsymbol{A}_\star^{(1)}, \ldots, \boldsymbol{A}_\star^{(d)})$, we have,

$$\left\|\boldsymbol{\mathcal{N}}(\bar{\boldsymbol{A}}^{(1)} + \boldsymbol{\Delta}^{(1)}, \ldots, \bar{\boldsymbol{A}}^{(d)} + \boldsymbol{\Delta}^{(d)})\right\|_{\mathrm{F}} \leqslant \left\|\boldsymbol{\mathcal{N}}(\bar{\boldsymbol{A}}^{(1)}, \ldots, \bar{\boldsymbol{A}}^{(d)})\right\|_{\mathrm{F}} + S\left\|\boldsymbol{\mathcal{N}}(\boldsymbol{A}_\star^{(1)}, \ldots, \boldsymbol{A}_\star^{(d)})\right\|_{\mathrm{F}}$$

with $S \overset{\text{def}}{=} \sum_{k=1}^{d} \binom{d}{k}\varepsilon^k \leqslant \sum_{k=1}^{d} \varepsilon^k \frac{d^k}{k!} \leqslant e^{\varepsilon d} - 1$. Hence, choosing $\varepsilon = \frac{1}{d}\log\frac{3}{2}$, we get,

$$\left\|\boldsymbol{\mathcal{N}}(\boldsymbol{A}_\star^{(1)}, \ldots, \boldsymbol{A}_\star^{(d)})\right\|_{\mathrm{F}} \leqslant 2\left\|\boldsymbol{\mathcal{N}}(\bar{\boldsymbol{A}}^{(1)}, \ldots, \bar{\boldsymbol{A}}^{(d)})\right\|_{\mathrm{F}}$$

and, from the union bound, for any $t \geqslant 0$,

$$\begin{aligned}\mathbb{P}\left(\left\|\boldsymbol{\mathcal{N}}(\boldsymbol{A}_\star^{(1)}, \ldots, \boldsymbol{A}_\star^{(d)})\right\|_{\mathrm{F}} \geqslant t\right) &\leqslant \mathbb{P}\left(\bigcup_{\boldsymbol{A}^{(\ell)}\in\mathscr{C}_\ell,\ \ell\in[d]} \left\{\left\|\boldsymbol{\mathcal{N}}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)})\right\|_{\mathrm{F}} \geqslant \frac{t}{2}\right\}\right) \\ &\leqslant \sum_{\boldsymbol{A}^{(\ell)}\in\mathscr{C}_\ell,\ \ell\in[d]} \mathbb{P}\left(\left\|\boldsymbol{\mathcal{N}}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)})\right\|_{\mathrm{F}} \geqslant \frac{t}{2}\right).\end{aligned}$$

Thus, combining Lemma 5.16 and 5.17, we have,

$$\mathbb{P}\left(\left\|\boldsymbol{\mathcal{N}}(\boldsymbol{A}_\star^{(1)}, \ldots, \boldsymbol{A}_\star^{(d)})\right\|_{\mathrm{F}} \geqslant t\right) \leqslant \left[\frac{Cd}{\log\frac{3}{2}}\right]^{\sum_{\ell=1}^{d} r_\ell\left(n_\ell - \frac{r_\ell+1}{2}\right)} \mathbb{P}\left(X \geqslant \frac{t^2}{4}\right)$$

where X is a random variable following a $\chi^2(\prod_{\ell\in[d]} r_\ell)$ distribution. Eventually, the probability on the right-hand side can be bounded using Lemma 5.18,

$$\mathbb{P}\left(X \geqslant \frac{t^2}{4}\right) = \frac{\Gamma\left(\frac{1}{2}\prod_{\ell\in[d]} r_\ell, \frac{t^2}{8}\right)}{\Gamma\left(\frac{1}{2}\prod_{\ell\in[d]} r_\ell\right)} \leqslant \max(1, e^{\frac{1}{2}\prod_{\ell\in[d]} r_\ell - 1})e^{-t^2/16}.$$

We get the result stated in Lemma 5.10 with

$$t^2 = 16\left[\left(\sum_{\ell=1}^{d} r_\ell\left(n_\ell - \frac{r_\ell+1}{2}\right)\right)\log\frac{Cd}{\log\frac{3}{2}} + \log\left(\frac{1}{\delta}\max\left(1, e^{\frac{1}{2}\prod_{\ell=1}^{d} r_\ell - 1}\right)\right)\right].$$

5.D Proof of Theorem 5.11

Recall the decomposition $\mathcal{P} = [\![\mathcal{H}; \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(d)}]\!]$. We use the following lemma whenever we state that $\|\mathcal{P}(\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(d)})\|_{\mathrm{F}} = \mathcal{O}(\|\mathcal{P}\|_{\mathrm{F}})$ as $N \to +\infty$.

Lemma 5.19. *For all* $(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)}) \in \bigtimes_{\ell \in [d]} V_{r_\ell}(\mathbb{R}^{n_\ell})$,

$$\left\| \mathcal{P}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)}) \right\|_{\mathrm{F}} \leqslant \|\mathcal{P}\|_{\mathrm{F}} \prod_{\ell \in [d]} \left\| \boldsymbol{X}^{(\ell)\top} \boldsymbol{A}^{(\ell)} \right\|.$$

Proof. The proof relies on the property $\|\boldsymbol{AB}\|_{\mathrm{F}} \leqslant \|\boldsymbol{A}\| \|\boldsymbol{B}\|_{\mathrm{F}}$.

$$\begin{aligned}
\left\| \mathcal{P}(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(d)}) \right\|_{\mathrm{F}} &= \left\| \boldsymbol{A}^{(1)\top} \boldsymbol{X}^{(1)} \boldsymbol{H}^{(1)} \bigboxtimes_{\ell=2}^{d} \boldsymbol{X}^{(\ell)\top} \boldsymbol{A}^{(\ell)} \right\|_{\mathrm{F}} \\
&\leqslant \left\| \boldsymbol{A}^{(1)\top} \boldsymbol{X}^{(1)} \right\| \left\| \boldsymbol{H}^{(1)} \bigboxtimes_{\ell=2}^{d} \boldsymbol{X}^{(\ell)\top} \boldsymbol{A}^{(\ell)} \right\|_{\mathrm{F}} \\
&= \left\| \boldsymbol{A}^{(1)\top} \boldsymbol{X}^{(1)} \right\| \left\| \boldsymbol{A}^{(2)\top} \boldsymbol{X}^{(2)} \boldsymbol{H}^{(2)} \left(\boldsymbol{I}_{r_1} \boxtimes \bigboxtimes_{\ell=3}^{d} \boldsymbol{X}^{(\ell)\top} \boldsymbol{A}^{(\ell)} \right) \right\|_{\mathrm{F}} \\
&\leqslant \left\| \boldsymbol{A}^{(1)\top} \boldsymbol{X}^{(1)} \right\| \left\| \boldsymbol{A}^{(2)\top} \boldsymbol{X}^{(2)} \right\| \left\| \boldsymbol{H}^{(2)} \left(\boldsymbol{I}_{r_1} \boxtimes \bigboxtimes_{\ell=3}^{d} \boldsymbol{X}^{(\ell)\top} \boldsymbol{A}^{(\ell)} \right) \right\|_{\mathrm{F}} \\
&\cdots \\
&\leqslant \left(\prod_{\ell=1}^{d} \left\| \boldsymbol{A}^{(\ell)\top} \boldsymbol{X}^{(\ell)} \right\| \right) \underbrace{\left\| \boldsymbol{H}^{(d)} \bigboxtimes_{\ell=1}^{d} \boldsymbol{I}_{r_\ell} \right\|_{\mathrm{F}}}_{= \|\mathcal{H}\|_{\mathrm{F}} = \|\mathcal{P}\|_{\mathrm{F}}}.
\end{aligned}$$

□

Given $\ell \in [d]$, $\boldsymbol{U}_1^{(\ell)}$ gathers the r_ℓ dominant left singular vectors of $\boldsymbol{T}^{(\ell)} \boxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')}$, i.e., it is solution to

$$\max_{\boldsymbol{U}^{(\ell)} \in V_{r_\ell}(\mathbb{R}^{n_\ell})} \left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{T}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2. \tag{5.19}$$

Consider also a solution $\widetilde{\boldsymbol{U}}_1^{(\ell)}$ to the following related problem

$$\max_{\boldsymbol{U}^{(\ell)} \in V_{r_\ell}(\mathbb{R}^{n_\ell})} \left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2. \tag{5.20}$$

Observe that, using the property $\|\boldsymbol{AB}\|_{\mathrm{F}} \leqslant \|\boldsymbol{A}\| \|\boldsymbol{B}\|_{\mathrm{F}}$, we have,

$$\begin{aligned}
\left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 &= \left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{X}^{(\ell)} \boldsymbol{H}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 \\
&\leqslant \left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{X}^{(\ell)} \right\|^2 \left\| \boldsymbol{H}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 \leqslant \left\| \boldsymbol{H}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2
\end{aligned}$$

and this upper bound is only reached with $\boldsymbol{U}^{(\ell)} = \boldsymbol{X}^{(\ell)} \boldsymbol{O}^{(\ell)}$, for any $r_\ell \times r_\ell$ orthogonal matrix $\boldsymbol{O}^{(\ell)}$. Hence, $\widetilde{\boldsymbol{U}}_1^{(\ell)} = \boldsymbol{X}^{(\ell)} \boldsymbol{O}^{(\ell)}$. The strategy of our proof is to show that, as $N \to +\infty$, Problem (5.19) has the same solutions as Problem (5.20), which are known.

With the decomposition $\mathcal{T} = \mathcal{P} + \frac{1}{\sqrt{N}}\mathcal{N}$, we have,

$$\left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{T}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 = \left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 + \frac{1}{N} \left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{N}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 + \frac{2}{\sqrt{N}} \left\langle \boldsymbol{U}^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')}, \boldsymbol{U}^{(\ell)\top} \boldsymbol{N}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\rangle_{\mathrm{F}}.$$

From Lemma 5.10, $\frac{1}{N} \| \boldsymbol{U}^{(\ell)\top} \boldsymbol{N}^{(\ell)} \boxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \|_{\mathrm{F}}^2 = \mathcal{O}(1)$ almost surely and

$$\frac{2}{\sqrt{N}} \left| \left\langle \boldsymbol{U}^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')}, \boldsymbol{U}^{(\ell)\top} \boldsymbol{N}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\rangle_{\mathrm{F}} \right| \leqslant \frac{2}{\sqrt{N}} \underbrace{\left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}}_{=\mathcal{O}(\|\mathcal{P}\|_{\mathrm{F}})} \underbrace{\left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{N}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}}_{=\mathcal{O}(\sqrt{N})}.$$

Therefore, for all $\boldsymbol{U}^{(\ell)} \in V_{r_\ell}(\mathbb{R}^{n_\ell})$,

$$\left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{T}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 = \left\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 + \mathcal{O}(\|\mathcal{P}\|_{\mathrm{F}}) \qquad \text{almost surely.} \tag{5.21}$$

In particular,

$$\left\| \boldsymbol{U}_1^{(\ell)\top} \boldsymbol{T}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 = \left\| \boldsymbol{U}_1^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 + \mathcal{O}(\|\mathcal{P}\|_{\mathrm{F}}) \qquad \text{almost surely,}$$

$$\left\| \boldsymbol{U}_1^{(\ell)\top} \boldsymbol{T}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 = \left\| \widetilde{\boldsymbol{U}}_1^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 + \mathcal{O}(\|\mathcal{P}\|_{\mathrm{F}}) \qquad \text{almost surely,}$$

where the first equation is simply Equation (5.21) with $\boldsymbol{U}^{(\ell)} = \boldsymbol{U}_1^{(\ell)}$ and the second equation stems from the maximum over $\boldsymbol{U}^{(\ell)} \in V_{r_\ell}(\mathbb{R}^{n_\ell})$ of both sides of Equation (5.21)[17]. Hence,

$$\left\| \boldsymbol{U}_1^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 = \left\| \widetilde{\boldsymbol{U}}_1^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 + \mathcal{O}(\|\mathcal{P}\|_{\mathrm{F}}) \qquad \text{almost surely.} \tag{5.22}$$

Then, consider the singular value decomposition $\boldsymbol{X}^{(\ell)\top} \boldsymbol{U}_1^{(\ell)} = \sum_{q_\ell=1}^{r_\ell} s_{q_\ell}^{(\ell)} \boldsymbol{v}_{q_\ell}^{(\ell)} \boldsymbol{w}_{q_\ell}^{(\ell)\top}$.

$$\begin{aligned}
\left\| \boldsymbol{U}_1^{(\ell)\top} \boldsymbol{P}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 &= \left\| \boldsymbol{U}_1^{(\ell)\top} \boldsymbol{X}^{(\ell)} \boldsymbol{H}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 \\
&= \left\| \sum_{q_\ell=1}^{r_\ell} s_{q_\ell}^{(\ell)} \boldsymbol{w}_{q_\ell}^{(\ell)} \boldsymbol{v}_{q_\ell}^{(\ell)\top} \boldsymbol{H}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 \\
&= \sum_{q_\ell=1}^{r_\ell} s_{q_\ell}^{(\ell)2} \left\| \boldsymbol{v}_{q_\ell}^{(\ell)\top} \boldsymbol{H}^{(\ell)} \bigboxtimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2
\end{aligned}$$

[17] $\max\{\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{T}^{(\ell)} \boxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \|_{\mathrm{F}}^2\} \leqslant \max\{\| \boldsymbol{U}^{(\ell)\top} \boldsymbol{P}^{(\ell)} \boxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \|_{\mathrm{F}}^2\} + \max\{\mathcal{O}(\|\mathcal{P}\|_{\mathrm{F}})\}$ where each max is over $\boldsymbol{U}^{(\ell)} \in V_{r_\ell}(\mathbb{R}^{n_\ell})$. Thus, $\| \boldsymbol{U}_1^{(\ell)\top} \boldsymbol{T}^{(\ell)} \boxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \|_{\mathrm{F}}^2 - \| \widetilde{\boldsymbol{U}}_1^{(\ell)\top} \boldsymbol{P}^{(\ell)} \boxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \|_{\mathrm{F}}^2 = \mathcal{O}(\|\mathcal{P}\|_{\mathrm{F}})$ almost surely.

$$= \sum_{q_\ell=1}^{r_\ell} \left\| \boldsymbol{v}_{q_\ell}^{(\ell)\top} \boldsymbol{H}^{(\ell)} \bigotimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 - \sum_{q_\ell=1}^{r_\ell} \left(1 - s_{q_\ell}^{(\ell)2}\right) \left\| \boldsymbol{v}_{q_\ell}^{(\ell)\top} \boldsymbol{H}^{(\ell)} \bigotimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2$$

$$= \left\| \boldsymbol{H}^{(\ell)} \bigotimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 - \sum_{q_\ell=1}^{r_\ell} \left(1 - s_{q_\ell}^{(\ell)2}\right) \left\| \boldsymbol{v}_{q_\ell}^{(\ell)\top} \boldsymbol{H}^{(\ell)} \bigotimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 .$$

Therefore, because $\| \boldsymbol{H}^{(\ell)} \boxtimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \|_{\mathrm{F}}^2 = \| \widetilde{\boldsymbol{U}}_1^{(\ell)\top} \boldsymbol{P}^{(\ell)} \boxtimes_{\ell' \neq \ell} \boldsymbol{U}_0^{(\ell')} \|_{\mathrm{F}}^2$, Equation (5.22) yields,

$$\sum_{q_\ell=1}^{r_\ell} \left(1 - s_{q_\ell}^{(\ell)2}\right) \left\| \boldsymbol{v}_{q_\ell}^{(\ell)\top} \boldsymbol{H}^{(\ell)} \bigotimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 = \mathcal{O}(\|\mathcal{P}\|_{\mathrm{F}}) \qquad \text{almost surely.}$$

Using the decomposition $\boldsymbol{v}_{q_\ell}^{(\ell)} = \sum_{q'_\ell=1}^{r_\ell} [\boldsymbol{v}_{q_\ell}^{(\ell)}]_{q'_\ell} \boldsymbol{X}^{(\ell)\top} \boldsymbol{x}_{q'_\ell}^{(\ell)}$, we can see that

$$\left\| \boldsymbol{v}_{q_\ell}^{(\ell)\top} \boldsymbol{H}^{(\ell)} \bigotimes_{\ell' \neq \ell} \boldsymbol{X}^{(\ell')\top} \boldsymbol{U}_0^{(\ell')} \right\|_{\mathrm{F}}^2 = \sum_{q'_\ell=1}^{r_\ell} [\boldsymbol{v}_{q_\ell}^{(\ell)}]_{q'_\ell}^2 \left\| \mathcal{P}(\boldsymbol{U}_0^{(1)}, \ldots, \boldsymbol{x}_{q'_\ell}^{(\ell)}, \ldots, \boldsymbol{U}_0^{(d)}) \right\|_{\mathrm{F}}^2 = \Theta(L_N^2).$$

Hence,

$$\sum_{q_\ell=1}^{r_\ell} \left(1 - s_{q_\ell}^{(\ell)2}\right) = \mathcal{O}\left(\frac{\|\mathcal{P}\|_{\mathrm{F}}}{L_N^2}\right) \qquad \text{almost surely,}$$

which is the result stated in Theorem 5.11:

$$\frac{1}{r_\ell} \left\| \boldsymbol{X}^{(\ell)\top} \boldsymbol{U}_1^{(\ell)} \right\|_{\mathrm{F}}^2 = \frac{1}{r_\ell} \sum_{q_\ell=1}^{r_\ell} s_{q_\ell}^{(\ell)2} = 1 + \mathcal{O}\left(\frac{\|\mathcal{P}\|_{\mathrm{F}}}{L_N^2}\right) \qquad \text{almost surely.}$$

Chapter 6

Performance Gaps in Multi-View Clustering under the Nested Matrix-Tensor Model

In the age of artificial intelligence, handling vast amounts of data has become a fundamental aspect of machine learning tasks. Datasets are often high-dimensional and composed of multiple modes, such as various modalities, sensors, sources, types, or domains, naturally leading themselves to be represented as tensors. Tensors offer a richer structure compared to traditional one-dimensional vectors and two-dimensional matrices, making them increasingly relevant in various applications, including statistical learning and data analysis (Landsberg, 2011; Sun et al., 2021).

Yet, in the existing literature, there is a notable scarcity of theoretical studies that specifically address the performance gaps between tensor-based methods and traditional (matrix) spectral methods in the context of high-dimensional data analysis. While tensor methods have shown promise in various applications, including multi-view clustering, co-clustering, community detection, and latent variable modeling (Wu et al., 2019; Anandkumar et al., 2014; Papalexakis et al., 2013; Wang et al., 2023), little attention has been devoted to rigorously quantifying the advantages and drawbacks of leveraging the hidden low-rank tensor structure. Filling this gap by conducting a thorough theoretical analysis is crucial for gaining a deeper understanding of the practical implications and potential performance gains associated with tensor-based techniques.

In the specific case of multi-view clustering, Seddik et al. (2023a) recently proposed a spectral tensor method and carried out a precise analysis of its performance in the large-dimensional limit. Their method consists in computing a best rank-one (tensor) approximation of a *nested matrix-tensor model*, which, in particular, generalizes the classical *rank-one spiked tensor model* of Richard and Montanari (2014), and can be described as follows. Assume that we observe m transformations of a $p \times n$ matrix $\boldsymbol{M} = \boldsymbol{\mu}\boldsymbol{y}^\top + \boldsymbol{N}$ representing n points in dimension p split into two clusters centered around $\pm\boldsymbol{\mu}$, with $\boldsymbol{y} \in \{-1,+1\}^n$ and $\boldsymbol{N}$ a Gaussian matrix encoding the "inherent" dispersion of individuals (that is, regardless of measurement errors) around the center of their respective cluster. Mathematically, each view is thus expressed as

$$\boldsymbol{X}_k = f_k(\boldsymbol{\mu}\boldsymbol{y}^\top + \boldsymbol{N}) + \boldsymbol{W}_k, \quad k = 1, \ldots, m, \tag{6.1}$$

where f_k models the transformation applied to $\boldsymbol{M}$ on the k-th view and $\boldsymbol{W}_k$ is an additive observation noise with i.i.d. entries drawn from $\mathcal{N}(0,1)$. The nested-matrix tensor model then arises when we take $f_k(\boldsymbol{M}) = h_k\boldsymbol{M}$, meaning the function f_k simply rescales the matrix $\boldsymbol{M}$ by an unknown coefficient

$h_k \in \mathbb{R}$. With $\boldsymbol{h} = \begin{pmatrix} h_1 & \dots & h_m \end{pmatrix}^\top \in \mathbb{R}^m$ and $\otimes$ denoting the outer product, this gives

$$\mathcal{X} = \left(\boldsymbol{\mu}\boldsymbol{y}^\top + \boldsymbol{N}\right) \otimes \boldsymbol{h} + \mathcal{W} \in \mathbb{R}^{p \times n \times m}. \tag{6.2}$$

By seeking a best rank-one approximation of $\mathcal{X}$ to estimate its latent clustering structure, Seddik et al. (2023a) showed *empirically* that it outperforms an unfolding approach based on applying an SVD to an *unfolding* of $\mathcal{X}$ (as in Chapter 5 and Ben Arous et al., 2023), which is a matrix obtained by rearranging the entries of a tensor (see Section 2.4). However, the tensor-based approach hinges upon solving a problem which is worst-case NP-hard, unlike the unfolding approach. A natural question is thus: what is the exact performance gap that exists between these two approaches, as a function of some measure of difficulty of the problem (typically, a measure of signal-to-noise ratio)?

Here, in order to answer this question, we rigorously study the unfolding method by deploying tools from random matrix theory. Specifically, our main contributions are

- within the framework of the general nested matrix-tensor model, we derive the limiting spectral distribution of the unfoldings of the tensor (Theorems 6.1 and 6.5) and precisely quantify how well the hidden low-rank (tensor) structure can be recovered from them in the high-dimensional regime (Theorems 6.2 and 6.7);
- we perform a similar random matrix analysis of the model when the vector spanning the third mode is known (Theorems 6.9 and 6.10), providing an optimal upper bound on the recovery performance;
- in the context of multi-view clustering, we compare the performance of the tensor and unfolding approaches to the optimal one and specify the gap between them thanks to our theoretical findings (Theorem 6.11), supported by empirical results[1].

Although the above described model arises from a rather particular choice of view transformations f_k, it is amenable to a precise estimation performance analysis, either by means of a tensor spectral estimator as recently done by Seddik et al. (2023a), or via a matrix spectral estimator as we consider in the present chapter. Moreover, from a broader perspective (that is, beyond the multi-view clustering problem considered here), this model can be viewed as a more flexible version of the rank-one spiked model, incorporating a nested structure that allows for versatile data modeling, deviating from a pure rank-one assumption. A common low-rank structure encoding the underlying latent clustering pattern is shared by all slices $\boldsymbol{X}_k$, which represent distinct views of the data. In particular, when the variances of the elements in $\boldsymbol{N}$ approach zero, the rank-one spiked model is retrieved. Hence, we believe that the nested tensor-matrix model (and extensions) can be a useful tool in other contexts in the broader area of statistical learning.

Related work. In the machine learning literature, the notion of "view" is fairly general and models data whose form may differ but all represent the same object seen from different (and complementary) angles (for instance, multiple descriptors of an image, translations of a text or features of a webpage such as its hyperlinks, text and images). Various approaches have been considered to address multi-view clustering problems. For instance, Nie et al. (2016, 2017b,a) consider a graph-based model and construct a similarity matrix by integrating all views with a weighted sum before applying spectral clustering. Other approaches, relying upon a space-learning-based model (Wang et al., 2017; Zhang et al., 2017; Wang et al., 2019; Peng et al., 2019), reconstruct the data in an ideal space where clustering is easy. Zhang et al. (2019) suggest a method which is more suitable for large datasets by mixing

[1] Note that these numerical results are only meant to illustrate our theoretical findings, showing their implications in practice. However, our work does not purport to explain the performance gap between *any* tensor-based and *any* matrix-based multi-view clustering methods.

binary coding and clustering. Tensor methods have also been considered: Wu et al. (2019) propose an essential tensor learning approach for Markov chain-based multi-view spectral clustering. Furthermore, Liu et al. (2021, 2023) design simple yet effective methods for multi-view clustering relying on multiple kernel k-means. Our work differs from these previous contributions in that it is focused on a tensor model having the specific form of Model (6.2). Even though this model corresponds to a rather particular case of the general setting given in Equation (6.1), our results represent a first step towards precisely understanding how tensor methods can contribute to addressing the latter.

Regarding the analysis of performance gaps between tensor- and matrix-based methods, we can mention the recent work by Seddik et al. (2023b), where the authors proposed a data model that consists of a Gaussian mixture assuming a low-rank tensor structure on the population means and further characterized the theoretical performance gap between a simple tensor-based method and a flattening-based method that neglects the low-rank structure. Their study has demonstrated that the tensor approach yields provably better performance compared to treating the data as mere vectors.

Proofs and simulations. All proofs are deferred to the appendix. Python codes to reproduce simulations are available in the following GitHub repository `https://github.com/HugoLebeau/nested_matrix-tensor`.

6.1 Random Matrix Analysis of the Nested Matrix-Tensor Model

Before presenting its practical applications in Section 6.2, we define the nested matrix-tensor model in a general framework. Consider the following statistical model,

$$\mathcal{T} = \beta_T \boldsymbol{M} \otimes \boldsymbol{z} + \frac{1}{\sqrt{n_T}} \mathcal{W} \in \mathbb{R}^{n_1 \times n_2 \times n_3}, \qquad \boldsymbol{M} = \beta_M \boldsymbol{x}\boldsymbol{y}^\top + \frac{1}{\sqrt{n_M}} \boldsymbol{N} \in \mathbb{R}^{n_1 \times n_2}, \tag{6.3}$$

where $n_M = n_1 + n_2$ and $n_T = n_1 + n_2 + n_3$, $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are of unit norm and the entries of $\mathcal{W}$ and $\boldsymbol{N}$ are independent Gaussian random variables[2]: $\mathcal{W}_{i,j,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $N_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. $\boldsymbol{M}$ is a rank-1 signal $\beta_M \boldsymbol{x}\boldsymbol{y}^\top$ corrupted by noise $\boldsymbol{N}$, modeling the data matrix, whereas $\mathcal{T}$ models its multi-view observation $\beta_T \boldsymbol{M} \otimes \boldsymbol{z}$ corrupted by noise $\mathcal{W}$. The positive parameters β_M and β_T control the signal-to-noise ratio (SNR). Our interest is the statistical recovery of $\boldsymbol{x}$, $\boldsymbol{y}$ or $\boldsymbol{z}$ in the regime where $n_1, n_2, n_3 \to +\infty$ with $0 < c_\ell \overset{\text{def}}{=} n_\ell / n_T < 1$ for all $\ell \in [3]$. This models the fact that, in practice, we deal with large tensors whose dimensions have comparable sizes.

Seddik et al. (2023a) have studied the spectral estimator of $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ based on computing the best rank-one approximation of $\mathcal{T}$, that is, by solving

$$(\boldsymbol{x}_\star, \boldsymbol{y}_\star, \boldsymbol{z}_\star) = \underset{(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) \in \mathbb{S}^{n_1-1} \times \mathbb{S}^{n_2-1} \times \mathbb{S}^{n_3-1}}{\arg\max} \langle \mathcal{T}, \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \rangle.$$

Concretely, they used random matrix tools to assess its performance in the recovery of $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, by deploying a recent approach developed by Goulart et al. (2022); Seddik et al. (2022). In this work, we study instead the performance of a spectral approach based on computing the dominant singular vectors of the (matrix) *unfoldings* of $\mathcal{T}$, aiming to precisely quantify the performance gap between these different approaches.

Because $\boldsymbol{M}$ has the structure of a standard spiked matrix model (Benaych-Georges and Nadakuditi, 2011; Couillet and Liao, 2022) with a rank-one perturbation $\beta_M \boldsymbol{x}\boldsymbol{y}^\top$ of a random matrix $\frac{1}{\sqrt{n_m}} \boldsymbol{N}$, we shall assume $\beta_M = \Theta(1)$ since we know that it is in this "non-trivial regime" that the recovery of $\boldsymbol{x}$ or $\boldsymbol{y}$ given $\boldsymbol{M}$ is neither too easy (too high SNR) nor too hard (too small SNR) and a phase-transition

[2] See our discussion on the Gaussian noise assumption in Section 2.2.3.

phenomenon (Baik et al., 2005) between impossible and possible recovery can be observed. However, we shall see that the algorithmic phase transition related to the unfolding approach takes place when $\beta_T = \Theta(n_T^{1/4})$ in lieu of $\beta_T = \Theta(1)$. This is different from the tensor spectral approach for which both β_M and β_T are $\Theta(1)$ in the non-trivial regime, supposing a better performance of the latter method, although, practically, no known algorithm is able to compute it below $\beta_T = \Theta(n_T^{1/4})$ (Richard and Montanari, 2014).

6.1.1 Unfoldings Along the First Two Modes

We start by studying the recovery of $\boldsymbol{y}$ (resp. $\boldsymbol{x}$) from the unfolding $\boldsymbol{T}^{(2)}$ (resp. $\boldsymbol{T}^{(1)}$). In a multi-view clustering perspective — which motivates our work and which will be developed in Section 6.2 — we are especially interested in the recovery of $\boldsymbol{y}$ since it carries the class labels. Therefore, we present our results for $\boldsymbol{T}^{(2)}$ only. As $\boldsymbol{x}$ and $\boldsymbol{y}$ play a symmetric role, it is easy to deduce the results for $\boldsymbol{T}^{(1)}$ from those presented below. The recovery of $\boldsymbol{z}$ from $\boldsymbol{T}^{(3)}$ is dealt with in Section 6.1.2.

Following the model presented in Equation (6.3), the unfolding along the second mode of $\mathcal{T}$ develops as

$$\boldsymbol{T}^{(2)} = \beta_T \beta_M \boldsymbol{y}(\boldsymbol{x} \boxtimes \boldsymbol{z})^\top + \frac{\beta_T}{\sqrt{n_M}} \boldsymbol{N}^\top \left(\boldsymbol{I}_{n_1} \boxtimes \boldsymbol{z}\right)^\top + \frac{1}{\sqrt{n_T}} \boldsymbol{W}^{(2)}. \tag{6.4}$$

Hence, a natural estimator $\hat{\boldsymbol{y}}$ of $\boldsymbol{y}$ is the dominant left singular vector of $\boldsymbol{T}^{(2)}$ or, equivalently, the dominant eigenvector of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$. The latter being symmetric, it is better suited to the tools presented in Chapter 2. Our first step is to characterize the limiting spectral distribution of this random matrix. However, similarly to the model studied in Chapter 5, one must be careful with the fact that the dimensions of $\boldsymbol{T}^{(2)} \in \mathbb{R}^{n_2 \times n_1 n_3}$ do not have sizes of the same order, causing the spectrum of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ to diverge as $n_1, n_2, n_3 \to +\infty$. In fact, its eigenvalues gather in a "bulk" centered around a $\Theta(n_T)$ value and spread on an interval of size $\Theta(\sqrt{n_T})$ — a phenomenon which was first characterized by Ben Arous et al. (2023). For this reason, in Theorem 6.1, we do not specify the LSD of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ per se but of a properly centered-and-scaled version of it, whose spectrum no longer diverges. Moreover, it is expected that the rank-one signal $\beta_T\beta_M\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top$ causes the presence of an isolated eigenvalue in the spectrum of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ with corresponding eigenvector positively correlated with $\boldsymbol{y}$ when it is detectable, i.e., when $\beta_T\beta_M$ is large enough.

Our second step is thus to precisely specify what is meant by "large enough" and characterize the asymptotic position of this spike eigenvalue and the alignment with $\boldsymbol{y}$ of its corresponding eigenvector. It turns out that the signal vanishes if β_T does not scale with n_T. Precisely, $\beta_T^2 n_T / \sqrt{n_1 n_2 n_3}$ must converge to a fixed positive quantity, denoted ρ_T, to reach the "non-trivial regime" — that is, one where the signal and the noise in the model have comparable strengths. However, because the noise in $\boldsymbol{N}$ is also weighted by β_T, this affects the shape of the bulk. Hence, the value of ρ_T influences the limiting spectral distribution of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ and shall appear in its defining equation[3]. Having said all this, we are now ready to introduce the following theorem.

Theorem 6.1 (Limiting Spectral Distribution)**.** *As $n_1, n_2, n_3 \to +\infty$, the centered-and-scaled matrix $\frac{n_T}{\sqrt{n_1 n_2 n_3}} \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - \frac{n_1 n_3}{\sqrt{n_1 n_2 n_3}} \boldsymbol{I}_{n_2}$ has a limiting spectral distribution $\tilde{\nu}$ whose Stieltjes transform $\tilde{m}$ satisfies*

$$\frac{\rho_T c_2}{1-c_3}\tilde{m}^3(\tilde{s}) + \left(1 + \tilde{s}\frac{\rho_T c_2}{1-c_3}\right)\tilde{m}^2(\tilde{s}) + \left(\tilde{s} + \frac{\rho_T(c_2-c_1)}{1-c_3}\right)\tilde{m}(\tilde{s}) + 1 = 0, \qquad \tilde{s} \in \mathbb{C} \setminus \operatorname{Supp}\tilde{\nu}$$

where $\rho_T = \frac{\beta_T^2 n_T}{\sqrt{n_1 n_2 n_3}}$. Moreover, if $\beta_M = 0$ (no signal), no eigenvalue stays outside $\operatorname{Supp}\tilde{\nu}$ almost surely as $n_1, n_2, n_3 \to +\infty$.

[3] This may be at first surprising because ρ_T relates to the strength of the *signal* while the LSD stems from the *noise*. We see that $\boldsymbol{N}$ plays an ambivalent role of both a signal and a noise term.

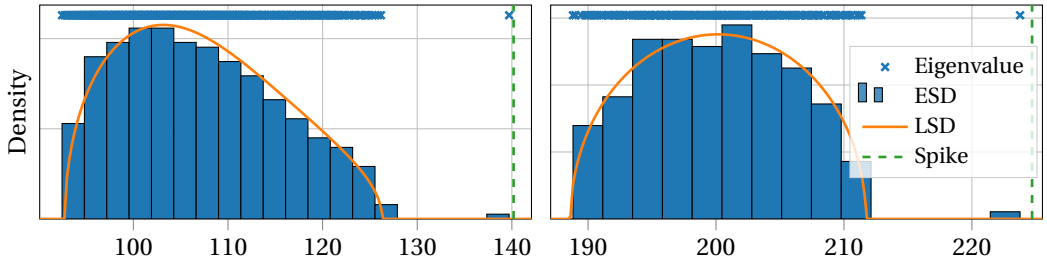


Figure 6.1: **Empirical Spectral Distribution (ESD)** and **Limiting Spectral Distribution (LSD)** of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ (**left**) and $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$ (**right**) with $n_1 = 600$, $n_2 = 400$ and $n_3 = 200$. Both spectra show an **isolated eigenvalue** close to its predicted asymptotic position, represented by the green dashed line. **Left:** $\rho_T = 2$, $\beta_M = 1.5$. The centered-and-scaled LSD $\widetilde{\nu}$ and spike location $\tilde{\xi}$ are defined in Theorems 6.1 and 6.2. **Right:** $\varrho = 4$, $\beta_M = 3$. The LSD is a shifted-and-rescaled semicircle distribution and the normalized spike location is $\varrho + \frac{1}{\varrho}$ as precised in Theorems 6.5 and 6.7.

Proof. See Appendix 6.A. □

As mentioned in Section 2.1, the LSD $\widetilde{\nu}$ is characterized by its Stieltjes transform, uniquely defined as the solution of a polynomial equation[4]. The influence of ρ_T on the LSD of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ is made explicit in this equation and it is interesting to remark that, if $\rho_T = 0$ (i.e., in the absence of signal), this equality reduces to $\widetilde{m}^2(\tilde{s}) + \tilde{s}\widetilde{m}(\tilde{s}) + 1 = 0$, which is a well-known characterization of the Stieltjes transform of the semicircle distribution (Equation (2.1)). Note also that the condition $\rho_T = \Theta(1)$ amounts to saying that $\beta_T = \Theta(n_T^{1/4})$, which coincides with the conjectured "computational threshold" under which no known algorithm is able to detect a signal without prior information (Richard and Montanari, 2014; Ben Arous et al., 2023).

The ESD and LSD of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ with parameters $(\rho_T, \beta_M) = (2, 1.5)$ are represented in the left panel of Figure 6.1. We observe a good agreement between the actual and predicted shape of the bulk. As expected, we see an isolated eigenvalue on the right which only appears for sufficiently high values of ρ_T and β_M. The following theorem specifies this behavior and quantifies the alignment between the signal $\boldsymbol{y}$ and the spike eigenvector $\hat{\boldsymbol{y}} = \boldsymbol{u}_1(\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top})$.

Theorem 6.2 (Spike Behavior)**.**

$$\textit{Let} \qquad \tilde{\xi} = \frac{\rho_T}{\beta_M^2}\left(\frac{c_1}{1-c_3} + \beta_M^2\right)\left(\frac{c_2}{1-c_3} + \beta_M^2\right) + \frac{1}{\rho_T\left(\frac{c_2}{1-c_3} + \beta_M^2\right)}$$

$$\textit{and} \qquad \zeta = 1 - \frac{1}{\beta_M^2\left(\frac{c_2}{1-c_3} + \beta_M^2\right)}\left[\left(\frac{\beta_M^2}{\rho_T\left(\frac{c_2}{1-c_3} + \beta_M^2\right)}\right)^2 + \frac{c_2}{1-c_3}\left(\frac{c_1}{1-c_3} + \beta_M^2\right)\right].$$

If $\zeta > 0$, then the centered-and-scaled matrix $\frac{n_T}{\sqrt{n_1 n_2 n_3}}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - \frac{n_1 n_3}{\sqrt{n_1 n_2 n_3}}\boldsymbol{I}_{n_2}$ has an isolated eigenvalue asymptotically located in $\tilde{\xi}$ almost surely. Furthermore, in this case, the alignment between the corresponding eigenvector $\hat{\boldsymbol{y}} = \boldsymbol{u}_1(\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top})$ and the true signal $\boldsymbol{y}$ converges to ζ almost surely, i.e.,

$$\langle \boldsymbol{y}, \hat{\boldsymbol{y}} \rangle^2 \xrightarrow[n_1,n_2,n_3 \to +\infty]{a.s.} \zeta.$$

[4]Although this is not the only solution to this equation, it is the only one that has the properties of a Stieltjes transform, such as $\Im[\tilde{s}]\Im[\widetilde{m}(\tilde{s})] > 0$ for all $\tilde{s} \in \mathbb{C} \setminus \mathbb{R}$ (see Proposition 2.7).

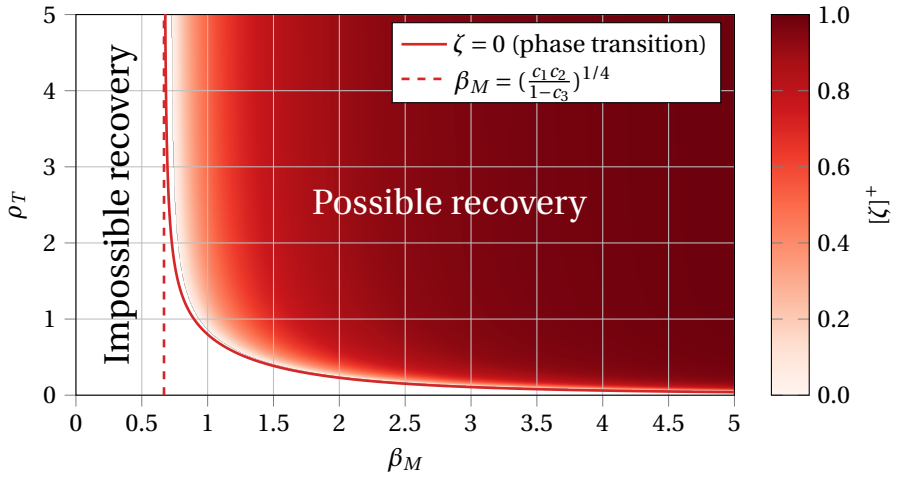


Figure 6.2: **Asymptotic alignment** $[\zeta]^+ = \max(\zeta, 0)$ between the signal $\boldsymbol{y}$ and the dominant eigenvector of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$, as defined in Theorem 6.2, with $c_1 = \frac{1}{2}$, $c_2 = \frac{1}{3}$ and $c_3 = \frac{1}{6}$. The curve $\zeta = 0$ is the position of the **phase transition** between the impossible detectability of the signal (below) and the presence of an isolated eigenvalue in the spectrum of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ with corresponding eigenvector correlated with the signal (above). It has an asymptote $\beta_M = (\frac{c_1 c_2}{1-c_3})^{1/4}$ as $\rho_T \to +\infty$, represented by the red dashed line.

Proof. See Appendix 6.B. □

Naturally, we must assume $\rho_T, \beta_M > 0$ for $\tilde{\xi}$ and ζ to be well defined. The location of the isolated eigenvalue in the spectrum of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ predicted from the expression of $\tilde{\xi}$ is represented as the green dashed line in the left panel of Figure 6.1. In fact, Theorem 6.2 reveals a phase transition phenomenon between impossible and possible recovery of the signal with the estimator $\hat{\boldsymbol{y}} = \boldsymbol{u}_1(\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top})$. This is precisely quantified by the value of $[\zeta]^+ = \max(\zeta, 0)$: the closer it is to 1, the better is the estimation of $\boldsymbol{y}$. The precise dependence of ζ on ρ_T and β_M is hard to interpret directly from its expression. Figure 6.2 displays $[\zeta]^+$ as a function of ρ_T and β_M. The expression of the curve $\zeta = 0$ marking the position of the transition from impossible to possible recovery is given by the following proposition.

Proposition 6.3 (Phase Transition). *If* $\beta_M^4 > \frac{c_1 c_2}{1-c_3}$, *then* $\zeta = 0 \iff \rho_T = \dfrac{\beta_M^2}{\left(\frac{c_2}{1-c_3} + \beta_M^2\right)\sqrt{\beta_M^4 - \frac{c_1 c_2}{1-c_3}}}$.

We see that, if $\beta_M^4 \leqslant \frac{c_1 c_2}{1-c_3}$, it is impossible to find $\rho_T > 0$ such that $\zeta > 0$. This is due to the fact that $\beta_M^4 = \frac{c_1 c_2}{1-c_3}$ corresponds to the position of the phase transition in the estimation of $\boldsymbol{y}$ *from* $\boldsymbol{M}$. If the signal is not detectable from $\boldsymbol{M}$, there is obviously no chance to recover it from $\mathcal{T}$. Moreover, as ρ_T grows, the value of β_M such that $\zeta = 0$ coincides with $(\frac{c_1 c_2}{1-c_3})^{1/4}$ but it goes to $+\infty$ as ρ_T approaches 0. This shows the importance of having $\beta_T = \Theta(n_T^{1/4})$ (in which case it is more convenient to work with the rescaled version ρ_T of β_T): if β_T is an order below ($\rho_T \to 0$) then we are stuck in the "Impossible recovery" zone while if β_T is an order above ($\rho_T \to +\infty$) then estimating from $\mathcal{T}$ is just like estimating from $\boldsymbol{M}$. It is precisely in the regime $\beta_T = \Theta(n_T^{1/4})$ that this phase-transition phenomenon can be observed, thereby justifying its designation as "non-trivial".

Remark 6.4. It should be noted that the aforementioned impossibility of (partially) recovering the sought signal in a given regime refers only to the case where such a recovery is carried out by the

unfolding method. In other words, our discussion concerns algorithmic thresholds pertaining to such method, and not statistical ones.

6.1.2 Unfolding Along the Third Mode

For the sake of completeness, we study in this section the recovery of $\boldsymbol{z}$ from the third unfolding of $\mathcal{T}$. Following the model in Equation (6.3), the expression of $\boldsymbol{T}^{(3)}$ is

$$\boldsymbol{T}^{(3)} = \beta_T \boldsymbol{z}\boldsymbol{m}^\top + \frac{1}{\sqrt{n_T}} \boldsymbol{W}^{(3)} \tag{6.5}$$

where $\boldsymbol{m} = \begin{bmatrix} \boldsymbol{M}_{1,\cdot} & \dots & \boldsymbol{M}_{n_1,\cdot} \end{bmatrix}^\top \in \mathbb{R}^{n_1 n_2}$.

This unfolding has the peculiarity that the rank-one perturbation $\beta_T \boldsymbol{z}\boldsymbol{m}^\top$ mixes signal (the vector $\boldsymbol{z}$) and noise (contained in $\boldsymbol{M}$). Still, as for the previous unfoldings, the dominant left singular vector of $\boldsymbol{T}^{(3)}$ remains a natural estimator of $\boldsymbol{z}$ and we study the asymptotic spectral properties of $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$. Because of the long shape of $\boldsymbol{T}^{(3)}$ (one dimension grows faster than the other), the spectrum of $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$ diverges in the same way as that of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$. Therefore, we must proceed to a similar rescaling. The following theorem states that, after proper rescaling, the distribution of eigenvalues of $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$ approaches the semicircle distribution.

Theorem 6.5 (Limiting Spectral Distribution)**.** *As $n_1, n_2, n_3 \to +\infty$, the empirical spectral distribution of the centered-and-scaled matrix $\frac{n_T}{\sqrt{n_1 n_2 n_3}} \boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top} - \frac{n_1 n_2}{\sqrt{n_1 n_2 n_3}} \boldsymbol{I}_{n_3}$ converges weakly to the semicircle distribution on $[-2,2]$ given by*

$$\mathrm{d}\mu_{\mathrm{SC}}(x) = \frac{1}{2\pi}\sqrt{\left[4 - x^2\right]^+}\mathrm{d}x.$$

Moreover, if $\beta_T = 0$ (no signal), no eigenvalue stays outside $[-2,2]$ almost surely as $n_1, n_2, n_3 \to +\infty$.

Proof. See Appendix 6.C. □

The ESD and LSD of $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$ are plotted in the right panel of Figure 6.1. The result of Theorem 6.5 is not surprising: the "non-trivial" shape of the LSD of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ (Theorem 6.1) is due to the presence of a "signal-noise" $\frac{\beta_T}{\sqrt{n_M}} \boldsymbol{N}^\top \left(\boldsymbol{I}_{n_1} \boxtimes \boldsymbol{z}\right)^\top$ in the expression of $\boldsymbol{T}^{(2)}$ but, when β_T is set to 0, we have observed that the LSD of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ is simply a semicircle. This is coherent with the case that interests us here: in $\boldsymbol{T}^{(3)}$, the "signal-noise" is restrained to the rank-one perturbation and therefore does not impact the LSD of $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$, which is then a semicircle.

Because of the rank-one perturbation $\beta_T \boldsymbol{z}\boldsymbol{m}^\top$, the spectrum of $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$ exhibits an isolated eigenvalue which can be observed in the right panel of Figure 6.1. Our next step is to characterize the behavior of this spike eigenvalue and the correlation with $\boldsymbol{z}$ of its corresponding eigenvector. Before introducing the formal result in Theorem 6.7, let us have a close look at the expression of $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$ to understand, with hand-waving arguments, what should be the non-trivial regime in this case.

$$\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top} = \beta_T^2 \|\boldsymbol{M}\|_{\mathrm{F}}^2 \boldsymbol{z}\boldsymbol{z}^\top + \frac{\beta_T}{\sqrt{n_T}}\left(\boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{W}^{(3)\top} + \boldsymbol{W}^{(3)}\boldsymbol{m}\boldsymbol{z}^\top\right) + \frac{1}{n_T}\boldsymbol{W}^{(3)}\boldsymbol{W}^{(3)\top}$$

Starting from the right, the term $\frac{1}{n_T}\boldsymbol{W}^{(3)}\boldsymbol{W}^{(3)\top}$ is already understood thanks to Theorem 6.5 and yields a semicircle as limiting spectral distribution. The crossed-terms in the middle have zero mean and are expected to vanish. On the left, remains the rank-one term $\boldsymbol{z}\boldsymbol{z}^\top$ weighted by $\beta_T^2\|\boldsymbol{M}\|_{\mathrm{F}}^2$, which is a random quantity because of the noise $\boldsymbol{N}$ in $\boldsymbol{M}$. However, the quantity $\|\boldsymbol{M}\|_{\mathrm{F}}^2$ is expected to rapidly concentrate around its mean $\frac{n_1 n_2}{n_M} + \beta_M^2$. Hence, guessing from the results on the previous unfoldings, we would need the quantity $\beta_T^2 \frac{n_T}{\sqrt{n_1 n_2 n_3}}\left(\frac{n_1 n_2}{n_M} + \beta_M^2\right)$ to converge to a fixed positive value denoted ϱ.

Indeed, this is precisely what is found when this analysis is rigorously carried out (see Appendix 6.C), meaning that $\beta_T = \Theta(n_T^{-1/4})$. In other words, if β_T is a constant ($\beta_T = \Theta(1)$), we are above the non-trivial regime and therefore should expect (asymptotically) exact recovery. This is because the strength of the signal is "boosted" by $\|\boldsymbol{M}\|_{\mathrm{F}}^2 = \Theta(n_T)$.

Remark 6.6. ϱ is defined as the value of $\beta_T^2 \frac{n_T}{\sqrt{n_1 n_2 n_3}}\left(\frac{n_1 n_2}{n_M} + \beta_M^2\right)$ which, since $\beta_M = \Theta(1)$, is the same as that of $\beta_T^2 \frac{n_1 n_2}{n_M} \frac{n_T}{\sqrt{n_1 n_2 n_3}}$. However, we keep the β_M^2 term in the definition of ϱ as it yields better predictions in our simulations.

Theorem 6.7 (Spike Behavior)**.** *If* $\varrho \overset{def}{=} \frac{\beta_T^2 n_T}{\sqrt{n_1 n_2 n_3}}\left(\frac{n_1 n_2}{n_M} + \beta_M^2\right) > 1$*, then the centered-and-scaled matrix* $\frac{n_T}{\sqrt{n_1 n_2 n_3}} \boldsymbol{T}^{(3)} \boldsymbol{T}^{(3)\top} - \frac{n_1 n_2}{\sqrt{n_1 n_2 n_3}} \boldsymbol{I}_{n_3}$ *has an isolated eigenvalue asymptotically located in* $\varrho + \frac{1}{\varrho}$ *almost surely. Furthermore, in this case, the alignment between the corresponding eigenvector* $\hat{\boldsymbol{z}} = \boldsymbol{u}_1(\boldsymbol{T}^{(3)} \boldsymbol{T}^{(3)\top})$ *and the true signal* $\boldsymbol{z}$ *converges to* $1 - \frac{1}{\varrho^2}$ *almost surely, i.e.,*

$$\langle \boldsymbol{z}, \hat{\boldsymbol{z}} \rangle^2 \xrightarrow[n_1,n_2,n_3 \to +\infty]{a.s.} 1 - \frac{1}{\varrho^2}.$$

Proof. See Appendix 6.D. □

Once the quantity ϱ is defined, we recognize in Theorem 6.7 the same results as that of the spiked Wigner model (Section 2.2.1).

Remark 6.8. In practice, we work with *large but finite* tensors. Hence, it makes no sense to say that $\beta_T = \Theta(n_T^{1/4})$ or $\beta_T = \Theta(n_T^{-1/4})$. In fact, the characterization of the "non-trivial" regime is only important here to reveal the relevant quantities, i.e., ρ_T and ϱ, which we will use in practice without worrying on whether β_T is in the right regime or not.

6.1.3 Estimation with Weighted Mean

Before diving into the application of the previous results to multi-view clustering (where we will be interested in the estimation of the class labels contained in $\boldsymbol{y}$), we propose an analysis of a related matrix model corresponding to the optimal estimation of $\boldsymbol{y}$ when $\boldsymbol{z}$ is perfectly known. These results will give us an optimistic upper bound on the performance of the estimation of $\boldsymbol{y}$ from $\mathcal{T}$.

In case $\boldsymbol{z}$ is known, $\boldsymbol{y}$ can be estimated with the following weighted mean of $\mathcal{T}$ along mode 3,

$$\bar{\boldsymbol{T}} = \sum_{k=1}^{n_3} z_k \mathcal{T}_{\cdot,\cdot,k} = \beta_T \beta_M \boldsymbol{x}\boldsymbol{y}^\top + \frac{\varsigma}{\sqrt{n_M}} \boldsymbol{Z} \tag{6.6}$$

where $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\varsigma^2 = \beta_T^2 + \frac{n_M}{n_T}$. It is well known that the dominant right singular vector of $\bar{\boldsymbol{T}}$ is an optimal estimator of $\boldsymbol{y}$ under this model (Onatski et al., 2013; Löffler et al., 2021). Hence, we study the spectrum of $\frac{1}{\varsigma^2} \bar{\boldsymbol{T}}^\top \bar{\boldsymbol{T}}$, which is a sample covariance matrix — a standard model in random matrix theory (Pastur and Shcherbina, 2011; Bai and Silverstein, 2010). Its eigenvalue distribution converges to the Marčenko-Pastur distribution (Marčenko and Pastur, 1967), as expressed in the following theorem.

Theorem 6.9 (Limiting Spectral Distribution)**.** *Let* $E_\pm = \left(\sqrt{\frac{c_1}{1-c_3}} \pm \sqrt{\frac{c_2}{1-c_3}}\right)^2$*. As* $n_1, n_2 \to +\infty$*, the matrix* $\frac{1}{\varsigma^2} \bar{\boldsymbol{T}}^\top \bar{\boldsymbol{T}}$ *has a limiting spectral distribution* $[1 - \frac{c_1}{c_2}]\delta_0 + \eta$ *with* η *explicitly given by*

$$\mathrm{d}\eta(x) = \frac{1}{2\pi \frac{c_2}{1-c_3} x} \sqrt{[x - E_-]^+ [E_+ - x]^+}\, \mathrm{d}x.$$

Similarly to the previous spiked models, the rank-one information $\beta_T\beta_M \boldsymbol{x}\boldsymbol{y}^\top$ induces the presence of an isolated eigenvalue in the spectrum of $\frac{1}{\varsigma^2}\bar{\boldsymbol{T}}^\top\bar{\boldsymbol{T}}$. The following theorem characterize its behavior and that of its corresponding eigenvector.

Theorem 6.10 (Spike Behavior)**.** *If* $\left(\frac{\beta_T\beta_M}{\varsigma}\right)^4 > \frac{c_1c_2}{(1-c_3)^2}$*, then the spectrum of* $\frac{1}{\varsigma^2}\bar{\boldsymbol{T}}^\top\bar{\boldsymbol{T}}$ *exhibits an isolated eigenvalue asymptotically located in* $\frac{\varsigma^2}{\beta_T^2\beta_M^2}\left(\frac{\beta_T^2\beta_M^2}{\varsigma^2}+\frac{c_1}{1-c_3}\right)\left(\frac{\beta_T^2\beta_M^2}{\varsigma^2}+\frac{c_2}{1-c_3}\right)$ *almost surely. Moreover, in this case, the corresponding eigenvector* $\boldsymbol{u}_1(\bar{\boldsymbol{T}}^\top\bar{\boldsymbol{T}})$ *is aligned with the signal* $\boldsymbol{y}$*,*

$$\left\langle \boldsymbol{y}, \boldsymbol{u}_1(\bar{\boldsymbol{T}}^\top\bar{\boldsymbol{T}})\right\rangle^2 \xrightarrow[n_1,n_2,n_3\to+\infty]{a.s.} 1-\frac{\varsigma^2}{\beta_T^2\beta_M^2}\frac{c_2}{1-c_3}\frac{\frac{\beta_T^2\beta_M^2}{\varsigma^2}+\frac{c_1}{1-c_3}}{\frac{\beta_T^2\beta_M^2}{\varsigma^2}+\frac{c_2}{1-c_3}}.$$

These results follow from the "standard" signal-plus-noise model presented in Section 2.2.3.

6.2 Performance Gaps in Multi-View Clustering

We shall now illustrate our results in the context of multi-view clustering. As explained in the introduction, we consider the observation of a tensor $\mathcal{X}\in\mathbb{R}^{p\times n\times m}$ following the nested matrix-tensor model,

$$\mathcal{X} = \left(\boldsymbol{\mu}\bar{\boldsymbol{y}}^\top + \boldsymbol{N}\right)\otimes\boldsymbol{h} + \mathcal{W} \qquad \text{with} \qquad \begin{cases} N_{i,j} & \overset{\text{i.i.d.}}{\sim} & \mathcal{N}\left(0,\frac{1}{p+n}\right) \\ \mathcal{W}_{i,j,k} & \overset{\text{i.i.d.}}{\sim} & \mathcal{N}\left(0,\frac{1}{p+n+m}\right)\end{cases}. \tag{6.7}$$

The two cluster centers are $\pm\boldsymbol{\mu}$ and $\bar{y}_i = \pm\frac{1}{\sqrt{n}}$ depending on the class of the i-th individual. The third vector $\boldsymbol{h}$ encodes the variances along the different views of $\boldsymbol{\mu}\bar{\boldsymbol{y}}^\top + \boldsymbol{N}$. The clustering is performed by estimating the class labels with the dominant left singular vector $\hat{\boldsymbol{y}}$ of $\boldsymbol{X}^{(2)}$. It is thus a direct application of the results of Section 6.1.1, where $(\|\boldsymbol{\mu}\|,\|\boldsymbol{h}\|)$ plays the role of (β_M,β_T). In fact, the behavior of the alignment $\langle\boldsymbol{y},\hat{\boldsymbol{y}}\rangle^2$ given by Theorem 6.2 can be further precised with the following theorem.

Theorem 6.11 (Performance of Multi-View Spectral Clustering)**.**

$$\textit{Let}\quad (c_p,c_n,c_m) = \frac{(p,n,m)}{p+n+m},\quad \rho = \|\boldsymbol{h}\|^2\frac{p+n+m}{\sqrt{pnm}},$$

$$\textit{and}\quad \zeta = 1-\frac{1}{\|\boldsymbol{\mu}\|^2\left(\frac{c_n}{1-c_m}+\|\boldsymbol{\mu}\|^2\right)}\left[\left(\frac{\|\boldsymbol{\mu}\|^2}{\rho\left(\frac{c_n}{1-c_m}+\|\boldsymbol{\mu}\|^2\right)}\right)^2+\frac{c_n}{1-c_m}\left(\frac{c_p}{1-c_m}+\|\boldsymbol{\mu}\|^2\right)\right].$$

Then, $\sqrt{\frac{n}{1-\zeta}}(\hat{y}_j-\sqrt{\zeta}\bar{y}_j)\xrightarrow[n\to+\infty]{\mathscr{D}}\mathcal{N}(0,1)$ *for all* $j\in[n]$*, i.e.,* $\hat{y}_j$ *approximately follows* $\mathcal{N}\left(\sqrt{\zeta}\bar{y}_j,\frac{1-\zeta}{n}\right)$*. Therefore, the clustering accuracy of the estimator* $\hat{\boldsymbol{y}}$ *converges almost surely to* $\Phi\left(\sqrt{\frac{\zeta}{1-\zeta}}\right)$ *where* $\Phi: x\mapsto\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}\,\mathrm{d}t$ *is the standard Gaussian cumulative distribution function.*

Proof. The proof is similar to that of Theorem 3.4 given in Chapter 3. □

Figure 6.3 compares the performances of the unfolding approach predicted by Theorem 6.11 with that of the "tensor approach" (Seddik et al., 2023a) which performs clustering with a rank-one approximation of $\mathcal{X}$. Moreover, an optimistic upper bound on the best achievable performance, given by

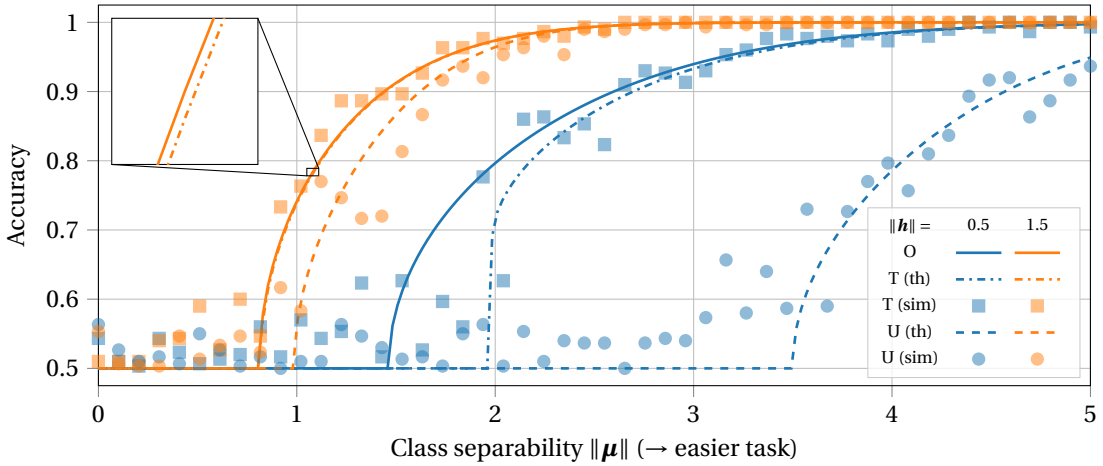


Figure 6.3: Empirical versus theoretical **multi-view clustering performance** with parameters $(p, n, m) = (150, 300, 60)$, varying $\|\boldsymbol{\mu}\|$ and two values of $\|\boldsymbol{h}\|$: 0.5 in blue and 1.5 in orange. The solid curve (O) is an optimistic upper bound given by Theorem 6.10, as it can be reached when the variances along each view are perfectly known. The dash-dotted curve (T) is the performance achieved with a rank-one approximation of $\mathcal{X}$ (Seddik et al., 2023a). The dashed curve (U) is the performance predicted by Theorem 6.11 with the unfolding approach.

the solid curve, can be derived from Theorem 6.10. Empirical accuracies are computed for both approaches and show a good match between theory and simulation results. It appears that the unfolding approach has a later phase transition and a lower performance than the tensor approach. This was expected since they do not have the same non-trivial regime ($\Theta(n_T^{1/4})$ against $\Theta(1)$). As $\|\boldsymbol{h}\|$ increases, the performance gap between both approaches reduces. The performance of the tensor approach rapidly comes very close to the upper bound: the two curves almost coincide for $\|\boldsymbol{h}\| = 1.5$.

These results show the superiority of the tensor approach in terms of accuracy of the multi-view spectral clustering. In particular, by contrast with the unfolding-based estimator, the tensor approach has near-optimal performance, as quantified by Theorem 6.10. Nevertheless, when considering "not too hard" problems (i.e., for which $\|\boldsymbol{\mu}\|$ and $\|\boldsymbol{h}\|$ are not too close to the phase transition threshold), the performances of both methods are close and the unfolding approach may be more interesting given its ease of implementation and lower computational cost.

6.3 Conclusion and Perspectives

In the context of multi-view clustering, we have conducted a thorough random matrix analysis of the unfoldings of a newly-introduced nested matrix-tensor model, which is an extension of the classical rank-one spiked tensor model. Our results shed light on the behavior of the distribution of the singular values of the unfolding in the large-dimensional regime. Moreover, we have characterized the alignments of their dominant eigenvectors with the sought signal, thereby describing the frontier between possible and impossible recovery in the corresponding non-trivial regime. In a previous work, Seddik et al. (2023a) studied the recovery performance of the best rank-one approximation on the same model and showed *empirically* that it performs better than the unfolding approach. Our analysis completes the picture by studying the performance of the latter. A natural avenue to delve deeper into this analysis of multi-view clustering is to extend our results to more general view functions f_k.

From a theoretical standpoint, we have shown the superiority of the tensor approach in terms of signal recovery. In particular, the latter can, in principle, recover the signal at a $\Theta(1)$ signal-to-noise ratio, while the matrix approach needs this ratio to diverge as $n_T^{1/4}$ as shown in Theorem 6.2, consistently with our results from Chapter 5. However, the tensor approach is based on an NP-hard formation and it is conjectured that no polynomial-time algorithm is capable of succeeding at a $\Theta(1)$ signal-to-noise ratio. In practice, for a sufficiently large ratio, one may combine these approaches by initializing a tensor rank-one approximation algorithm (such as HOOI in Algorithm 5) with the estimate given by the unfolding method. Although this procedure allows to compute the best rank-one approximation and therefore reach the optimal performance, it is only efficient above the *computational* threshold (Theorem 5.11). Therefore, the results illustrated in Figure 6.3 are indicative of the existing computational-to-statistical gap. The "tensor approach" consists in computing the best rank-one approximation of the observed tensor as estimator of the sought signal. Although, this is not always feasible because of the practical limitations we have just mentioned, the analysis of this problem provides many insight into the statistical limits of the estimation.

6.A Proof of Theorem 6.1

Denote $\boldsymbol{Q}^{(2)}(s) = \left(\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - s\boldsymbol{I}_{n_2}\right)^{-1}$ the resolvent of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ defined for all $s \in \mathbb{C} \setminus \operatorname{Sp} \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$.

6.A.1 Computations with Stein's Lemma

Before delving into the analysis of $\boldsymbol{Q}^{(2)}$, we will derive a few useful results thanks to Stein's lemma (Lemma 2.18). They are gathered in the following Proposition 6.12.

Proposition 6.12.

$$\mathbb{E}\big[\boldsymbol{W}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big] = \frac{n_1 n_3}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big] - \frac{1}{\sqrt{n_T}}\mathbb{E}\big[(n_2+1)\boldsymbol{Q}^{(2)} + s\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\big], \tag{6.8}$$

$$\begin{aligned}\beta_T\mathbb{E}\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)^\top\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\Big] &= \beta_T^2\mathbb{E}\big[\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big] \\ &\quad - \frac{1}{n_T}\mathbb{E}\big[s\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)2} + \boldsymbol{I}_{n_2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\big] \\ &\quad + \frac{1}{n_T^2}\mathbb{E}\big[(n_1n_3 - n_2 - 2)\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + (n_1n_3 - n_2 - 4)\boldsymbol{Q}^{(2)2}\big] \\ &\quad - \frac{s}{n_T^2}\mathbb{E}[4\boldsymbol{Q}^{(2)3} + 2\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2}],\end{aligned} \tag{6.9}$$

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{N}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big] &= \frac{1}{\sqrt{n_M}}\mathbb{E}\left[(n_1 - n_2 - 1)\boldsymbol{Q}^{(2)} - \left(s - \frac{n_1n_3}{n_T}\right)\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\right] \\ &\quad - \frac{s}{n_T\sqrt{n_M}}\mathbb{E}\big[4\boldsymbol{Q}^{(2)3} + 2\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2} + \boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)}\big] \\ &\quad - \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[(n_2+2)\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + (n_2+4)\boldsymbol{Q}^{(2)2}\big],\end{aligned} \tag{6.10}$$

$$\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top\boldsymbol{N}\boldsymbol{Q}^{(2)}\big] = -\frac{\beta_T}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}^{(2)\top}\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\big]. \tag{6.11}$$

In order to prove these results, we will need the following expressions for the derivatives of $\boldsymbol{Q}^{(2)}$.

Proposition 6.13.

$$\frac{\partial Q^{(2)}_{a,b}}{\partial W^{(2)}_{c,d}} = -\frac{1}{\sqrt{n_T}}\Big(Q^{(2)}_{a,c}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{d,b} + Q^{(2)}_{c,b}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{d,a}\Big) \tag{6.12}$$

$$\frac{\partial Q^{(2)}_{a,b}}{\partial N_{c,d}} = -\frac{\beta_T}{\sqrt{n_M}}\Big(Q^{(2)}_{a,d}\big[\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)\big]_{b,c} + Q^{(2)}_{d,b}\big[\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)\big]_{a,c}\Big) \tag{6.13}$$

Proof. Since $\partial\boldsymbol{Q}^{(2)} = -\boldsymbol{Q}^{(2)}\partial\big(\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\big)\boldsymbol{Q}^{(2)}$,

$$\begin{aligned}\frac{\partial Q^{(2)}_{a,b}}{\partial \mathcal{W}_{i,j,k}} &= -\sum_{e=1}^{n_2}\sum_{f=1}^{n_1}\sum_{g=1}^{n_3}\sum_{h=1}^{n_2} Q^{(2)}_{a,e}\left(\frac{\partial \mathcal{T}_{f,e,g}}{\partial \mathcal{W}_{i,j,k}}\mathcal{T}_{f,h,g} + \mathcal{T}_{f,e,g}\frac{\partial \mathcal{T}_{f,h,g}}{\partial \mathcal{W}_{i,j,k}}\right)Q^{(2)}_{h,b} \\ &= -\frac{1}{\sqrt{n_T}}\left(\sum_{h=1}^{n_2} Q^{(2)}_{a,j}\mathcal{T}_{i,h,k}Q^{(2)}_{h,b} + \sum_{e=1}^{n_2} Q^{(2)}_{a,e}\mathcal{T}_{i,e,k}Q^{(2)}_{j,b}\right) \\ \frac{\partial Q^{(2)}_{a,b}}{\partial \mathcal{W}_{i,j,k}} &= -\frac{1}{\sqrt{n_T}}\Big(Q^{(2)}_{a,j}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{[i,k],b} + Q^{(2)}_{j,b}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{[i,k],a}\Big)\end{aligned}$$

Likewise,

$$\begin{aligned}
\frac{\partial Q^{(2)}_{a,b}}{\partial N_{c,d}} &= -\sum_{e=1}^{n_2}\sum_{f=1}^{n_1}\sum_{g=1}^{n_3}\sum_{h=1}^{n_2} Q^{(2)}_{a,e}\left(\frac{\partial \mathcal{T}_{f,e,g}}{\partial N_{c,d}}\mathcal{T}_{f,h,g} + \mathcal{T}_{f,e,g}\frac{\partial \mathcal{T}_{f,h,g}}{\partial N_{c,d}}\right) Q^{(2)}_{h,b} \\
&= -\frac{\beta_T}{\sqrt{n_M}}\left(\sum_{g=1}^{n_3}\sum_{h=1}^{n_2} Q^{(2)}_{a,d} z_g \mathcal{T}_{c,h,g} Q^{(2)}_{h,b} + \sum_{e=1}^{n_2}\sum_{g=1}^{n_3} Q^{(2)}_{a,e}\mathcal{T}_{c,e,g} z_g Q^{(2)}_{d,b}\right) \\
\frac{\partial Q^{(2)}_{a,b}}{\partial N_{c,d}} &= -\frac{\beta_T}{\sqrt{n_M}}\Big(Q^{(2)}_{a,d}\big[\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\big(\boldsymbol{I}_{n_1}\boxtimes \boldsymbol{z}\big)\big]_{b,c} + Q^{(2)}_{d,b}\big[\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\big(\boldsymbol{I}_{n_1}\boxtimes \boldsymbol{z}\big)\big]_{a,c}\Big).
\end{aligned}$$

□

The identities in Proposition 6.12 are proved using Stein's lemma (Lemma 2.18) and Proposition 6.13.

Proof of Equation (6.8)

$$\begin{aligned}
&\mathbb{E}\big[\boldsymbol{W}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{i,j} = \sum_{k=1}^{n_1 n_3}\sum_{l=1}^{n_2}\mathbb{E}\Big[W^{(2)}_{i,k}T^{(2)}_{l,k}Q^{(2)}_{l,j}\Big] \\
&= \sum_{k=1}^{n_1 n_3}\sum_{l=1}^{n_2}\mathbb{E}\left[\frac{\partial T^{(2)}_{l,k}}{\partial W^{(2)}_{i,k}}Q^{(2)}_{l,j} + T^{(2)}_{l,k}\frac{\partial Q^{(2)}_{l,j}}{\partial W^{(2)}_{i,k}}\right] \\
&= \frac{n_1 n_3}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{1}{\sqrt{n_T}}\sum_{k=1}^{n_1 n_3}\sum_{l=1}^{n_2}\mathbb{E}\Big[T^{(2)}_{l,k}\Big(Q^{(2)}_{l,i}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{k,j} + Q^{(2)}_{i,j}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{k,l}\Big)\Big] \\
&= \frac{n_1 n_3}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{1}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\big(\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big)\big]_{i,j} \\
&= \frac{n_1 n_3}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{1}{\sqrt{n_T}}\mathbb{E}\big[(n_2+1)\boldsymbol{Q}^{(2)} + s\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\big]_{i,j}
\end{aligned}$$

where the last equality comes from $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} = \boldsymbol{I}_{n_2} + s\boldsymbol{Q}^{(2)}$.

Proof of Equation (6.9)

$$\begin{aligned}
&\mathbb{E}\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes \boldsymbol{z}\big)^\top\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\Big]_{i,j} \\
&= \beta_T\mathbb{E}\big[\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big]_{i,j} + \frac{1}{\sqrt{n_T}}\sum_{k=1}^{n_1 n_3}\sum_{l=1}^{n_2}\mathbb{E}\Big[\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes \boldsymbol{z}\big)^\top\Big]_{i,k}W^{(2)}_{l,k}Q^{(2)}_{l,j}\Big] \\
&= \beta_T\mathbb{E}\big[\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big]_{i,j} + \frac{1}{\sqrt{n_T}}\sum_{k=1}^{n_1 n_3}\sum_{l=1}^{n_2}\mathbb{E}\left[\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes \boldsymbol{z}\big)^\top\Big]_{i,k}\frac{\partial Q^{(2)}_{l,j}}{\partial W^{(2)}_{l,k}}\right] \\
&= \beta_T\mathbb{E}\big[\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{1}{n_T}\sum_{k=1}^{n_1 n_3}\sum_{l=1}^{n_2}\mathbb{E}\Big[\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes \boldsymbol{z}\big)^\top\Big]_{i,k}Q^{(2)}_{l,l}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{k,j}\Big] \\
&\qquad - \frac{1}{n_T}\sum_{k=1}^{n_1 n_3}\sum_{l=1}^{n_2}\mathbb{E}\Big[\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes \boldsymbol{z}\big)^\top\Big]_{i,k}Q^{(2)}_{l,j}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{k,l}\Big] \\
&= \beta_T\mathbb{E}\big[\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{1}{n_T}\mathbb{E}\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes \boldsymbol{z}\big)^\top\boldsymbol{T}^{(2)\top}\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\Big]_{i,j}.
\end{aligned}$$

We have $\boldsymbol{T}^{(2)} = \beta_T \boldsymbol{M}^\top (\boldsymbol{I}_{n_1} \boxtimes \boldsymbol{z})^\top + \frac{1}{\sqrt{n_M}} \boldsymbol{W}^{(2)}$ and $\boldsymbol{T}^{(2)} \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} = s\boldsymbol{Q}^{(2)} + \boldsymbol{I}_{n_2}$ thus

$$
\begin{aligned}
&\beta_T \mathbb{E}\Big[\boldsymbol{M}^\top \big(\boldsymbol{I}_{n_1} \boxtimes \boldsymbol{z}\big)^\top \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \Big] \\
&= \beta_T^2 \mathbb{E}\big[\boldsymbol{M}^\top \boldsymbol{M} \boldsymbol{Q}^{(2)} \big] - \frac{1}{n_T} \mathbb{E}\Big[\Big(\boldsymbol{T}^{(2)} - \frac{1}{\sqrt{n_T}} \boldsymbol{W}^{(2)} \Big) \boldsymbol{T}^{(2)\top} \big(\boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2} \big) \Big] \\
&= \beta_T^2 \mathbb{E}\big[\boldsymbol{M}^\top \boldsymbol{M} \boldsymbol{Q}^{(2)} \big] - \frac{1}{n_T} \mathbb{E}\big[\big(s\boldsymbol{Q}^{(2)} + \boldsymbol{I}_{n_2} \big) \big(\boldsymbol{I}_{n_2} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)} \big) \big] + \frac{1}{n_T \sqrt{n_T}} \mathbb{E}\big[\boldsymbol{W}^{(2)} \boldsymbol{T}^{(2)\top} \big(\boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2} \big) \big] \\
&= \beta_T^2 \mathbb{E}\big[\boldsymbol{M}^\top \boldsymbol{M} \boldsymbol{Q}^{(2)} \big] - \frac{1}{n_T} \mathbb{E}\big[s\boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)2} + \boldsymbol{I}_{n_2} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)} \big] \\
&\qquad\qquad + \frac{1}{n_T \sqrt{n_T}} \mathbb{E}\big[\boldsymbol{W}^{(2)} \boldsymbol{T}^{(2)\top} \big(\boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2} \big) \big].
\end{aligned}
$$

We use Stein's lemma (Lemma 2.18) again to handle the last term.

$$
\begin{aligned}
&\mathbb{E}\big[\boldsymbol{W}^{(2)} \boldsymbol{T}^{(2)\top} \big(\boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2} \big) \big]_{i,j} \\
&= \sum_{u=1}^{n_1 n_3} \sum_{v=1}^{n_2} \sum_{w=1}^{n_2} \mathbb{E}\Big[W_{i,u}^{(2)} T_{v,u}^{(2)} \Big(Q_{v,j}^{(2)} Q_{w,w}^{(2)} + Q_{v,w}^{(2)} Q_{w,j}^{(2)} \Big) \Big] \\
&= \sum_{u=1}^{n_1 n_3} \sum_{v=1}^{n_2} \sum_{w=1}^{n_2} \mathbb{E}\left[\frac{\partial T_{v,u}^{(2)}}{\partial W_{i,u}^{(2)}} \Big(Q_{v,j}^{(2)} Q_{w,w}^{(2)} + Q_{v,w}^{(2)} Q_{w,j}^{(2)} \Big) \right] \\
&\qquad + \sum_{u=1}^{n_1 n_3} \sum_{v=1}^{n_2} \sum_{w=1}^{n_2} \mathbb{E}\left[T_{v,u}^{(2)} \left(\frac{\partial Q_{v,j}^{(2)}}{\partial W_{i,u}^{(2)}} Q_{w,w}^{(2)} + Q_{v,j}^{(2)} \frac{\partial Q_{w,w}^{(2)}}{\partial W_{i,u}^{(2)}} + \frac{\partial Q_{v,w}^{(2)}}{\partial W_{i,u}^{(2)}} Q_{w,j}^{(2)} + Q_{v,w}^{(2)} \frac{\partial Q_{w,j}^{(2)}}{\partial W_{i,u}^{(2)}} \right) \right] \\
&= \frac{n_1 n_3}{\sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2} \big]_{i,j} \\
&\qquad - \frac{1}{\sqrt{n_T}} \sum_{u=1}^{n_1 n_3} \sum_{v=1}^{n_2} \sum_{w=1}^{n_2} \mathbb{E}\Big[T_{v,u}^{(2)} \Big(Q_{v,i}^{(2)} \big[\boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{u,j} + Q_{i,j}^{(2)} \big[\boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{u,v} \Big) Q_{w,w}^{(2)} \Big] \\
&\qquad - \frac{1}{\sqrt{n_T}} \sum_{u=1}^{n_1 n_3} \sum_{v=1}^{n_2} \sum_{w=1}^{n_2} \mathbb{E}\Big[T_{v,u}^{(2)} Q_{v,j}^{(2)} \Big(Q_{w,i}^{(2)} \big[\boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{u,w} + Q_{i,w}^{(2)} \big[\boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{u,w} \Big) \Big] \\
&\qquad - \frac{1}{\sqrt{n_T}} \sum_{u=1}^{n_1 n_3} \sum_{v=1}^{n_2} \sum_{w=1}^{n_2} \mathbb{E}\Big[T_{v,u}^{(2)} \Big(Q_{v,i}^{(2)} \big[\boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{u,w} + Q_{i,w}^{(2)} \big[\boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{u,v} \Big) Q_{w,j}^{(2)} \Big] \\
&\qquad - \frac{1}{\sqrt{n_T}} \sum_{u=1}^{n_1 n_3} \sum_{v=1}^{n_2} \sum_{w=1}^{n_2} \mathbb{E}\Big[T_{v,u}^{(2)} Q_{v,w}^{(2)} \Big(Q_{w,i}^{(2)} \big[\boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{u,j} + Q_{i,j}^{(2)} \big[\boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{u,w} \Big) \Big] \\
&= \frac{n_1 n_3}{\sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2} \big]_{i,j} \\
&\qquad\qquad - \frac{1}{\sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(2)} \boldsymbol{T}^{(2)} \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{T}^{(2)} \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} \big]_{i,j} \\
&\qquad - \frac{2}{\sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(2)2} \boldsymbol{T}^{(2)} \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{i,j} - \frac{1}{\sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(2)} \boldsymbol{T}^{(2)} \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)2} + \boldsymbol{Q}^{(2)2} \operatorname{Tr} \boldsymbol{T}^{(2)} \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} \big]_{i,j} \\
&\qquad\qquad - \frac{1}{\sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(2)2} \boldsymbol{T}^{(2)} \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{T}^{(2)} \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)2} \big]_{i,j}.
\end{aligned}
$$

And, again, with $\boldsymbol{T}^{(2)} \boldsymbol{T}^{(2)\top} \boldsymbol{Q}^{(2)} = s\boldsymbol{Q}^{(2)} + \boldsymbol{I}_{n_2}$, we obtain

$$
\mathbb{E}\big[\boldsymbol{W}^{(2)} \boldsymbol{T}^{(2)\top} \big(\boldsymbol{Q}^{(2)} \operatorname{Tr} \boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2} \big) \big]
$$

$$
\begin{aligned}
&= \frac{n_1 n_3}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big] \\
&\quad - \frac{1}{\sqrt{n_T}}\mathbb{E}\big[s\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)} + n_2\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\big] \\
&\quad - \frac{2}{\sqrt{n_T}}\mathbb{E}\big[s\boldsymbol{Q}^{(2)3} + \boldsymbol{Q}^{(2)2}\big] - \frac{1}{\sqrt{n_T}}\mathbb{E}\big[s\boldsymbol{Q}^{(2)3} + \boldsymbol{Q}^{(2)2} + s\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + n_2\boldsymbol{Q}^{(2)2}\big] \\
&\quad - \frac{1}{\sqrt{n_T}}\mathbb{E}\big[s\boldsymbol{Q}^{(2)3} + \boldsymbol{Q}^{(2)2} + s\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\big] \\
&= \frac{1}{\sqrt{n_T}}\mathbb{E}\big[(n_1 n_3 - n_2 - 2)\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + (n_1 n_3 - n_2 - 4)\boldsymbol{Q}^{(2)2}\big] \\
&\quad - \frac{s}{\sqrt{n_T}}\mathbb{E}[4\boldsymbol{Q}^{(2)3} + 2\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2}].
\end{aligned}
$$

Eventually, this gives,

$$
\begin{aligned}
\beta_T\mathbb{E}\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)^\top\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\Big] &= \beta_T^2\mathbb{E}\big[\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big] - \frac{1}{n_T}\mathbb{E}\big[s\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)2} + \boldsymbol{I}_{n_2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\big] \\
&\quad + \frac{1}{n_T^2}\mathbb{E}\big[(n_1 n_3 - n_2 - 2)\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + (n_1 n_3 - n_2 - 4)\boldsymbol{Q}^{(2)2}\big] \\
&\quad - \frac{s}{n_T^2}\mathbb{E}[4\boldsymbol{Q}^{(2)3} + 2\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2}].
\end{aligned}
$$

Proof of Equation (6.10)

$$
\begin{aligned}
&\mathbb{E}\big[\boldsymbol{N}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big]_{i,j} = \sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\mathbb{E}\Big[N_{k,i}M_{k,l}Q^{(2)}_{l,j}\Big] \\
&= \frac{n_1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} + \sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\mathbb{E}\left[M_{k,l}\frac{\partial Q^{(2)}_{l,j}}{\partial N_{k,i}}\right] \\
&= \frac{n_1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} \\
&\quad - \frac{\beta_T}{\sqrt{n_M}}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\mathbb{E}\Big[M_{k,l}\Big(Q^{(2)}_{l,i}\big[\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)\big]_{j,k} + Q^{(2)}_{i,j}\big[\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)\big]_{l,k}\Big)\Big] \\
&= \frac{n_1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{\beta_T}{\sqrt{n_M}}\mathbb{E}\Big[\boldsymbol{Q}^{(2)}\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)^\top\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\Big(\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)^\top\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\Big)\Big]_{i,j}
\end{aligned}
$$

and, since $\beta_T\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)^\top = \boldsymbol{T}^{(2)} - \frac{1}{\sqrt{n_T}}\boldsymbol{W}^{(2)}$, with the relation $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} = s\boldsymbol{Q}^{(2)} + \boldsymbol{I}_{n_2}$ we have,

$$
\begin{aligned}
\mathbb{E}\big[\boldsymbol{N}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big]_{i,j} &= \frac{n_1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{1}{\sqrt{n_M}}\mathbb{E}\big[s\boldsymbol{Q}^{(2)2} + \boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + n_2\boldsymbol{Q}^{(2)}\big]_{i,j} \\
&\quad + \frac{1}{\sqrt{n_M n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\boldsymbol{W}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\big(\boldsymbol{W}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big)\big]_{i,j}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n_1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{1}{\sqrt{n_M}}\mathbb{E}\big[(n_2+1)\boldsymbol{Q}^{(2)} + s\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\big]_{i,j} \\
&\quad + \frac{1}{\sqrt{n_M n_T}}\sum_{u=1}^{n_2}\sum_{v=1}^{n_1 n_3}\sum_{w=1}^{n_2}\mathbb{E}\left[\frac{\partial Q^{(2)}_{i,u}}{\partial W^{(2)}_{u,v}}T^{(2)}_{w,v}Q^{(2)}_{w,j} + Q^{(2)}_{i,u}\frac{\partial T^{(2)}_{w,v}}{\partial W^{(2)}_{u,v}}Q^{(2)}_{w,j} + Q^{(2)}_{i,u}T^{(2)}_{w,v}\frac{\partial Q^{(2)}_{w,j}}{\partial W^{(2)}_{u,v}}\right] \\
&\quad + \frac{1}{\sqrt{n_M n_T}}\sum_{u=1}^{n_2}\sum_{v=1}^{n_1 n_3}\sum_{w=1}^{n_2}\mathbb{E}\left[\frac{\partial Q^{(2)}_{i,j}}{\partial W^{(2)}_{u,v}}T^{(2)}_{w,v}Q^{(2)}_{w,u} + Q^{(2)}_{i,j}\frac{\partial T^{(2)}_{w,v}}{\partial W^{(2)}_{u,v}}Q^{(2)}_{w,u} + Q^{(2)}_{i,j}T^{(2)}_{w,v}\frac{\partial Q^{(2)}_{w,u}}{\partial W^{(2)}_{u,v}}\right]
\end{aligned}
$$

where we have used Stein's lemma (Lemma 2.18) again. Hence,

$$
\begin{aligned}
&\mathbb{E}\big[\boldsymbol{N}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big]_{i,j} \\
&= \frac{n_1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{1}{\sqrt{n_M}}\mathbb{E}\big[(n_2+1)\boldsymbol{Q}^{(2)} + s\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\big]_{i,j} + \frac{n_1 n_3}{n_T\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)2} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\big]_{i,j} \\
&\qquad - \frac{1}{n_T\sqrt{n_M}}\sum_{u=1}^{n_2}\sum_{v=1}^{n_1 n_3}\sum_{w=1}^{n_2}\mathbb{E}\Big[\Big(Q^{(2)}_{i,u}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{v,u} + Q^{(2)}_{u,u}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{v,i}\Big)T^{(2)}_{w,v}Q^{(2)}_{w,j}\Big] \\
&\qquad - \frac{1}{n_T\sqrt{n_M}}\sum_{u=1}^{n_2}\sum_{v=1}^{n_1 n_3}\sum_{w=1}^{n_2}\mathbb{E}\Big[Q^{(2)}_{i,u}T^{(2)}_{w,v}\Big(Q^{(2)}_{w,u}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{v,j} + Q^{(2)}_{u,j}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{v,w}\Big)\Big] \\
&\qquad - \frac{1}{n_T\sqrt{n_M}}\sum_{u=1}^{n_2}\sum_{v=1}^{n_1 n_3}\sum_{w=1}^{n_2}\mathbb{E}\Big[\Big(Q^{(2)}_{i,u}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{v,j} + Q^{(2)}_{u,j}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{v,i}\Big)T^{(2)}_{w,v}Q^{(2)}_{w,u}\Big] \\
&\qquad\quad - \frac{1}{n_T\sqrt{n_M}}\sum_{u=1}^{n_2}\sum_{v=1}^{n_1 n_3}\sum_{w=1}^{n_2}\mathbb{E}\Big[Q^{(2)}_{i,j}T^{(2)}_{w,v}\Big(Q^{(2)}_{w,u}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{v,u} + Q^{(2)}_{u,u}\big[\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big]_{v,w}\Big)\Big]
\end{aligned}
$$

which bunches up into

$$
\begin{aligned}
&\mathbb{E}\big[\boldsymbol{N}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big]_{i,j} \\
&\qquad = \frac{n_1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big]_{i,j} - \frac{1}{\sqrt{n_M}}\mathbb{E}\left[(n_2+1)\boldsymbol{Q}^{(2)} + \left(s - \frac{n_1 n_3}{n_T}\right)\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\right]_{i,j} \\
&\qquad\quad - \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)2}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\big]_{i,j} \\
&\qquad\quad - \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)2}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\operatorname{Tr}\big(\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big)\big]_{i,j} \\
&\qquad\quad - \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)2}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)2}\big]_{i,j} \\
&\qquad\quad - \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\operatorname{Tr}\big(\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big) + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\operatorname{Tr}\big(\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big)\big]_{i,j}
\end{aligned}
$$

and, using again the relation $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} = s\boldsymbol{Q}^{(2)} + \boldsymbol{I}_{n_2}$, we get,

$$
\begin{aligned}
&\mathbb{E}\big[\boldsymbol{N}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big] \\
&= \frac{n_1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big] - \frac{1}{\sqrt{n_M}}\mathbb{E}\left[(n_2+1)\boldsymbol{Q}^{(2)} + \left(s - \frac{n_1 n_3}{n_T}\right)\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\right] \\
&\quad - \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[s\boldsymbol{Q}^{(2)3} + \boldsymbol{Q}^{(2)2} + s\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)3} + \boldsymbol{Q}^{(2)2} + s\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + n_2\boldsymbol{Q}^{(2)2}\big] \\
&\qquad\qquad - \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[s\boldsymbol{Q}^{(2)3} + \boldsymbol{Q}^{(2)2} + s\boldsymbol{Q}^{(2)3} + \boldsymbol{Q}^{(2)2}\big] \\
&\qquad\qquad - \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[s\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)} + n_2\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\big]
\end{aligned}
$$

$$= \frac{n_1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big] - \frac{1}{\sqrt{n_M}}\mathbb{E}\bigg[(n_2+1)\boldsymbol{Q}^{(2)} + \Big(s - \frac{n_1 n_3}{n_T}\Big)\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\bigg]$$
$$- \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[s\big(4\boldsymbol{Q}^{(2)3} + 2\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2} + \boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)}\big)\big]$$
$$- \frac{1}{n_T\sqrt{n_M}}\mathbb{E}\big[(n_2+4)\boldsymbol{Q}^{(2)2} + (n_2+2)\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\big].$$

Proof of Equation (6.11)

$$\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top\boldsymbol{N}\boldsymbol{Q}^{(2)}\big]_{i,j} = \sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\mathbb{E}\Big[\big[\boldsymbol{y}\boldsymbol{x}^\top\big]_{i,k}N_{k,l}Q^{(2)}_{l,j}\Big]$$
$$= \sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\mathbb{E}\left[\big[\boldsymbol{y}\boldsymbol{x}^\top\big]_{i,k}\frac{\partial Q^{(2)}_{l,j}}{\partial N_{k,l}}\right]$$
$$= -\frac{\beta_T}{\sqrt{n_M}}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\mathbb{E}\Big[\big[\boldsymbol{y}\boldsymbol{x}^\top\big]_{i,k}\Big(Q^{(2)}_{l,l}\big[\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)\big]_{j,k} + Q^{(2)}_{l,j}\big[\boldsymbol{Q}^{(2)}\boldsymbol{T}^{(2)}\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)\big]_{l,k}\Big)\Big]$$
$$= -\frac{\beta_T}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}^{(2)\top}\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\big]_{i,j}.$$

6.A.2 Asymptotic Behavior of the Resolvent

Since $\boldsymbol{Q}^{(2)-1}\boldsymbol{Q}^{(2)} = \boldsymbol{I}_{n_2}$, we have $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)} - s\boldsymbol{Q}^{(2)} = \boldsymbol{I}_{n_2}$. Hence, using Equation (6.4) and taking the expectation,

$$\beta_T\mathbb{E}\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)^\top\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\Big] + \frac{1}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{W}^{(2)}\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\big] - s\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big] = \boldsymbol{I}_{n_2}$$

and, injecting Equation (6.8), this yields

$$\beta_T\mathbb{E}\Big[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)^\top\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}\Big]$$
$$+ \frac{n_1 n_3}{n_T}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big] - \frac{1}{n_T}\mathbb{E}\big[(n_2+1)\boldsymbol{Q}^{(2)} + s\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\big] - s\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big] = \boldsymbol{I}_{n_2}.$$

Then, we use Equation (6.9) to develop $\beta_T\mathbb{E}[\boldsymbol{M}^\top\big(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\big)^\top\boldsymbol{T}^{(2)\top}\boldsymbol{Q}^{(2)}]$. After rearranging the terms and using the decomposition $\boldsymbol{M} = \beta_M\boldsymbol{x}\boldsymbol{y}^\top + \frac{1}{\sqrt{n_M}}\boldsymbol{N}$, we find

$$\frac{s}{n_T}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\big] + \Big(s + \frac{n_2 - n_1 n_3}{n_T}\Big)\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big] + \boldsymbol{I}_{n_2}$$
$$= -\frac{1}{n_T}\mathbb{E}\big[\boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)2}\big] + \beta_T^2\beta_M\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big] + \frac{\beta_T^2}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{N}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big]$$
$$- \frac{1}{n_T}\mathbb{E}\big[s\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)2} + \boldsymbol{I}_{n_2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\big]$$
$$+ \frac{1}{n_T^2}\mathbb{E}\big[(n_1 n_3 - n_2 - 2)\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + (n_1 n_3 - n_2 - 4)\boldsymbol{Q}^{(2)2}\big]$$
$$- \frac{s}{n_T^2}\mathbb{E}[4\boldsymbol{Q}^{(2)3} + 2\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2}].$$

We can now use Equation (6.10) to develop $\mathbb{E}[\boldsymbol{N}^\top \boldsymbol{M}\boldsymbol{Q}^{(2)}]$.

$$\begin{aligned}
\frac{s}{n_T}\mathbb{E}\big[\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\big] + \left(s+\frac{n_2-n_1n_3}{n_T}\right)\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big] + \boldsymbol{I}_{n_2} \\
= -\frac{1}{n_T}\mathbb{E}\big[\boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)2}\big] + \beta_T^2\beta_M\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big] \\
+ \frac{\beta_T^2}{n_M}\mathbb{E}\left[(n_1-n_2-1)\boldsymbol{Q}^{(2)} - \left(s-\frac{n_1n_3}{n_T}\right)\big(\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)2}\big)\right] \\
- s\frac{\beta_T^2}{n_Tn_M}\mathbb{E}\big[4\boldsymbol{Q}^{(2)3} + 2\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2} + \boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)}\big] \\
- \frac{\beta_T^2}{n_Tn_M}\mathbb{E}\big[(n_2+2)\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + (n_2+4)\boldsymbol{Q}^{(2)2}\big] \\
- \frac{1}{n_T}\mathbb{E}\big[s\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + s\boldsymbol{Q}^{(2)2} + \boldsymbol{I}_{n_2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\big] \\
+ \frac{1}{n_T^2}\mathbb{E}\big[(n_1n_3-n_2-2)\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)} + (n_1n_3-n_2-4)\boldsymbol{Q}^{(2)2}\big] \\
- \frac{s}{n_T^2}\mathbb{E}[4\boldsymbol{Q}^{(2)3} + 2\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2}].
\end{aligned}$$

After a few rearrangements this becomes

$$\begin{aligned}
\left(s-\frac{n_1n_3-n_2-2}{n_T}-\frac{\beta_T^2}{n_M}(n_1-n_2-1)\right)\mathbb{E}\big[\boldsymbol{Q}^{(2)}\big] + \left(\frac{1}{n_T}\mathbb{E}\big[\operatorname{Tr}\boldsymbol{Q}^{(2)}\big]+1\right)\boldsymbol{I}_{n_2} \\
+ \left(s\left(\frac{\beta_T^2}{n_M}+\frac{2}{n_T}\right) - \frac{n_1n_3-n_2-2}{n_T}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\right)\mathbb{E}\big[\boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)}\big] \\
+ \left(s\left(\frac{\beta_T^2}{n_M}+\frac{2}{n_T}\right) - \frac{n_1n_3-n_2-4}{n_T}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\right)\mathbb{E}\big[\boldsymbol{Q}^{(2)2}\big] \\
+ \frac{s}{n_T}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\mathbb{E}\big[4\boldsymbol{Q}^{(2)3} + 2\boldsymbol{Q}^{(2)2}\operatorname{Tr}\boldsymbol{Q}^{(2)} + \boldsymbol{Q}^{(2)}\operatorname{Tr}\boldsymbol{Q}^{(2)2} + \boldsymbol{Q}^{(2)}\operatorname{Tr}^2\boldsymbol{Q}^{(2)}\big] = \beta_T^2\beta_M\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{Q}^{(2)}\big].
\end{aligned}$$

Here, we must be careful that, due to the $n_2 \times n_1n_3$ rectangular shape of $\boldsymbol{T}^{(2)}$, the spectrum of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ diverges in the limit $n_1, n_2, n_3 \to +\infty$. In order to bypass this obstacle, we shall perform a change of variable $(s, \boldsymbol{Q}^{(2)}) \curvearrowright (\tilde{s}, \widetilde{\boldsymbol{Q}}^{(2)})$. Let $\tilde{s} = \frac{n_T s - n_1n_3}{\sqrt{n_1n_2n_3}}$ and $\widetilde{\boldsymbol{Q}}^{(2)}(\tilde{s}) = \left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}} - \tilde{s}\boldsymbol{I}_{n_2}\right)^{-1}$. With this rescaling, note that we have $\boldsymbol{Q}^{(2)}(s) = \frac{n_T}{\sqrt{n_1n_2n_3}}\widetilde{\boldsymbol{Q}}^{(2)}(\tilde{s})$, and the previous equation becomes

$$\begin{aligned}
\left(\tilde{s}+\frac{n_2+2}{\sqrt{n_1n_2n_3}} - \rho_T\frac{n_1-n_2-1}{n_M}\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)}\big] + \left(\frac{1}{\sqrt{n_1n_2n_3}}\mathbb{E}\big[\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}\big]+1\right)\boldsymbol{I}_{n_2} \\
+ \left(\tilde{s}\left(\frac{\rho_T}{n_M}+\frac{2}{\sqrt{n_1n_2n_3}}\right) + \frac{1}{n_2} + \frac{n_2+2}{\sqrt{n_1n_2n_3}}\left(\frac{\rho_T}{n_M}+\frac{1}{\sqrt{n_1n_2n_3}}\right)\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}\big] \\
+ \left(\tilde{s}\left(\frac{\rho_T}{n_M}+\frac{2}{\sqrt{n_1n_2n_3}}\right) + \frac{1}{n_2} + \frac{n_2+4}{\sqrt{n_1n_2n_3}}\left(\frac{\rho_T}{n_M}+\frac{1}{\sqrt{n_1n_2n_3}}\right)\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)2}\big] \\
+ \left(\frac{\tilde{s}}{\sqrt{n_1n_2n_3}}+\frac{1}{n_2}\right)\left(\frac{\rho_T}{n_M}+\frac{1}{\sqrt{n_1n_2n_3}}\right)\mathbb{E}\big[4\widetilde{\boldsymbol{Q}}^{(2)3} + 2\widetilde{\boldsymbol{Q}}^{(2)2}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)} + \widetilde{\boldsymbol{Q}}^{(2)}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)2} + \widetilde{\boldsymbol{Q}}^{(2)}\operatorname{Tr}^2\widetilde{\boldsymbol{Q}}^{(2)}\big] \\
= \rho_T\beta_M\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top\boldsymbol{M}\widetilde{\boldsymbol{Q}}^{(2)}\big]
\end{aligned}$$

where we have introduced the $\mathcal{O}(1)$ quantity $\rho_T = \frac{\beta_T^2 n_T}{\sqrt{n_1 n_2 n_3}}$.

Remark 6.14. We see, here, that ρ_T must be bounded to avoid any diverging coefficient in the previous equation. In fact, we must chose $\beta_T = \Theta(n_T^{1/4})$ (implying $\rho_T = \Theta(1)$) in order to place ourselves in the non-trivial regime where both the noise and the signal are $\Theta(1)$. This is different from the estimation of $\boldsymbol{y}$ with a rank-one approximation of $\mathcal{T}$, where β_T could be kept $\Theta(1)$ (Seddik et al., 2023a) (but at the cost of an NP-hard computation (Hillar and Lim, 2013)).

Let us reformulate the previous equation by keeping only the non-vanishing terms on the left-hand side of the equal sign.

$$\begin{aligned}
\rho_T \frac{n_2}{n_M}\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(2)}\left(\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}}{n_2}\right)^2\right] + \left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(2)}\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}}{n_2}\right] + \left(\tilde{s}+\rho_T\frac{n_2-n_1}{n_M}\right)\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)}] + \boldsymbol{I}_{n_2} \\
= -\left(\frac{n_2+2}{\sqrt{n_1 n_2 n_3}} + \frac{\rho_T}{n_M}\right)\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)}] - \frac{n_2}{\sqrt{n_1 n_2 n_3}}\mathbb{E}\left[\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}}{n_2}\right]\boldsymbol{I}_{n_2} \\
-\left(\tilde{s}\frac{2n_2}{\sqrt{n_1 n_2 n_3}} + \frac{n_2+2}{\sqrt{n_1 n_2 n_3}}\left(\rho_T\frac{n_2}{n_M} + \frac{n_2}{\sqrt{n_1 n_2 n_3}}\right)\right)\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(2)}\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}}{n_2}\right] \\
-\frac{n_2}{\sqrt{n_1 n_2 n_3}}\left(1+\tilde{s}\left(\rho_T\frac{n_2}{n_M} + \frac{n_2}{\sqrt{n_1 n_2 n_3}}\right)\right)\mathbb{E}\left[\widetilde{\boldsymbol{Q}}^{(2)}\left(\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}}{n_2}\right)^2\right] \\
-\left(\tilde{s}\left(\frac{\rho_T}{n_M} + \frac{2}{\sqrt{n_1 n_2 n_3}}\right) + \frac{1}{n_2} + \frac{n_2+4}{\sqrt{n_1 n_2 n_3}}\left(\frac{\rho_T}{n_M} + \frac{1}{\sqrt{n_1 n_2 n_3}}\right)\right)\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)2}] \\
-4\left(\frac{\tilde{s}}{\sqrt{n_1 n_2 n_3}} + \frac{1}{n_2}\right)\left(\frac{\rho_T}{n_M} + \frac{1}{\sqrt{n_1 n_2 n_3}}\right)\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)3}] \\
-\left(\tilde{s}\frac{n_2}{\sqrt{n_1 n_2 n_3}} + 1\right)\left(\frac{\rho_T}{n_M} + \frac{1}{\sqrt{n_1 n_2 n_3}}\right)\mathbb{E}\left[2\widetilde{\boldsymbol{Q}}^{(2)2}\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}}{n_2} + \widetilde{\boldsymbol{Q}}^{(2)}\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)2}}{n_2}\right] \\
+\rho_T\beta_M\mathbb{E}[\boldsymbol{y}\boldsymbol{x}^\top\boldsymbol{M}\widetilde{\boldsymbol{Q}}^{(2)}]. \qquad (6.14)
\end{aligned}$$

6.A.3 Concentration of Bilinear Forms and Traces

We have the following concentration results

$$\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}^{(2)} - \mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)}])\boldsymbol{b} \xrightarrow[n_1,n_2,n_3\to+\infty]{\text{a.s.}} 0 \qquad \text{and} \qquad \frac{1}{n_2}\operatorname{Tr}\boldsymbol{A}(\widetilde{\boldsymbol{Q}}^{(2)} - \mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)}]) \xrightarrow[n_1,n_2,n_3\to+\infty]{\text{a.s.}} 0 \qquad (6.15)$$

for all bounded (sequences of) vectors $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^{n_2}$ and matrices $\boldsymbol{A}\in\mathbb{R}^{n_2\times n_2}$.

The proof relies on the use of the Poincaré-Nash inequality (Lemma 2.19) and Lemma 2.20. Because of the "double-noise" structure of our model, the computations can be quite cumbersome but present no conceptual difficulty compared to those performed similarly in Section 2.3.1 so we just give the main steps.

Recall that we assume $\beta_T = \Theta(n_T^{-1/4})$ in the non-trivial regime.

For bilinear forms, we find

$$\operatorname{Var}(\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b}) \leqslant \frac{4n_T^2}{n_1 n_2 n_3}\left(\frac{\beta_T^2}{n_M} + \frac{1}{n_T}\right)\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\left[\left\|\boldsymbol{T}^{(2)}\right\|^2\left\|\widetilde{\boldsymbol{Q}}^{(2)}\right\|^4\right] = \mathcal{O}_{\tilde{s}}(n_T^{-1/2}), \qquad (6.16)$$

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}^{(2)} - \mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)}])\boldsymbol{b}\right|^4\right] \\
&\leqslant \frac{16n_T^2}{n_1 n_2 n_3}\left(\frac{\beta_T^2}{n_M} + \frac{1}{n_T}\right)\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\left[\left|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}^{(2)} - \mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)}])\boldsymbol{b}\right|^2\left\|\boldsymbol{T}^{(2)}\right\|^2\left\|\widetilde{\boldsymbol{Q}}^{(2)}\right\|^4\right] + \operatorname{Var}(\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b})^2 \\
&= \mathcal{O}_{\tilde{s}}(n_T^{-1}),
\end{aligned}$$

$$\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}^{(2)}-\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)}])\boldsymbol{b}\big|^6\Big]$$
$$\leqslant \frac{36n_T^2}{n_1n_2n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}^{(2)}-\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)}])\boldsymbol{b}\big|^4\big\|\boldsymbol{T}^{(2)}\big\|^2\big\|\widetilde{\boldsymbol{Q}}^{(2)}\big\|^4\Big]+\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}^{(2)}-\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(2)}])\boldsymbol{b}\big|^4\Big]^2$$
$$=\mathcal{O}_{\tilde{s}}(n_T^{-3/2}).$$

For traces, we find

$$\mathrm{Var}\left(\frac{1}{n_2}\operatorname{Tr}\boldsymbol{A}\widetilde{\boldsymbol{Q}}^{(2)}\right)\leqslant\frac{4}{n_2}\frac{n_T^2}{n_1n_2n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\|\boldsymbol{A}\|^2\mathbb{E}\Big[\big\|\boldsymbol{T}^{(2)}\big\|^2\big\|\widetilde{\boldsymbol{Q}}^{(2)}\big\|^4\Big]=\mathcal{O}_{\tilde{s}}(n_T^{-3/2}). \tag{6.17}$$

We also note here two results which are not needed for the concentration (6.15) but will be useful below and are derived with the Poincaré-Nash inequality (Lemma 2.19) as well.

$$\mathrm{Var}\left(\frac{1}{n_2}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)2}\right)\leqslant\frac{16}{n_2}\frac{n_T^2}{n_1n_2n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\mathbb{E}\Big[\big\|\boldsymbol{T}^{(2)}\big\|^2\big\|\widetilde{\boldsymbol{Q}}^{(2)}\big\|^6\Big]=\mathcal{O}_{\tilde{s}}(n_T^{-3/2}), \tag{6.18}$$

$$\mathrm{Var}\left(\left[\frac{1}{n_2}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}\right]^2\right)\leqslant\frac{16}{n_2}\frac{n_T^2}{n_1n_2n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\mathbb{E}\left[\left|\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}}{n_2}\right|^2\big\|\boldsymbol{T}^{(2)}\big\|^2\big\|\widetilde{\boldsymbol{Q}}^{(2)}\big\|^4\right]=\mathcal{O}_{\tilde{s}}(n_T^{-3/2}). \tag{6.19}$$

6.A.4 Expansion of the Mean Empirical Stieltjes Transform

Let $\widetilde{\boldsymbol{Q}}_0^{(2)}$ denote the resolvent of the model with $\beta_M=0$ (no signal). Using the resolvent identity (Proposition 2.21), it is easy to see that $\frac{1}{n_2}\operatorname{Tr}(\widetilde{\boldsymbol{Q}}_0^{(2)}-\widetilde{\boldsymbol{Q}}^{(2)})=\mathcal{O}_{\tilde{s}}(n_T^{-1})$, which shows that the rank-one signal does not change the limiting spectral distribution of $\frac{n_T}{\sqrt{n_1n_2n_3}}\Big[\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-\frac{n_1n_3}{n_T}\boldsymbol{I}_{n_2}\Big]$ and we can assume $\beta_M=0$ from now on.

Define $\widetilde{m}_{n_2}(\tilde{s})\overset{\text{def}}{=}\frac{1}{n_2}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}(\tilde{s})$. We seek an explicit expansion of $\mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})]$ up to a term dominated by n_T^{-1}. Let us apply $\frac{1}{n_2}\operatorname{Tr}$ to Equation (6.14) and keep explicitly only the terms which are not dominated by n_T^{-1}.

$$\begin{aligned}\rho_T\frac{n_2}{n_M}\mathbb{E}\big[\widetilde{m}_{n_2}^3(\tilde{s})\big]+\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}^2(\tilde{s})\big]+\left(\tilde{s}+\rho_T\frac{n_2-n_1}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]+1\\ =-\left(\frac{2n_2}{\sqrt{n_1n_2n_3}}+\frac{\rho_T}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]-\frac{n_2}{\sqrt{n_1n_2n_3}}\left(2\tilde{s}+\frac{n_2}{\sqrt{n_1n_2n_3}}+\rho_T\frac{n_2}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}^2(\tilde{s})\big]\\ -\frac{n_2}{\sqrt{n_1n_2n_3}}\left(1+\tilde{s}\left(\rho_T\frac{n_2}{n_M}+\frac{n_2}{\sqrt{n_1n_2n_3}}\right)\right)\mathbb{E}\big[\widetilde{m}_{n_2}^3(\tilde{s})\big]\\ -\left(\tilde{s}\frac{\rho_T}{n_M}+\frac{1}{n_2}\right)\mathbb{E}\big[\widetilde{m}'_{n_2}(\tilde{s})\big]-3\frac{\rho_T}{n_M}\mathbb{E}\big[\widetilde{m}'_{n_2}(\tilde{s})\widetilde{m}_{n_2}(\tilde{s})\big]+\mathcal{O}_{\tilde{s}}(n_T^{-3/2})\end{aligned} \tag{6.20}$$

where the derivative of $\widetilde{m}_{n_2}$ appeared because $\frac{1}{n_2}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)2}=\widetilde{m}'_{n_2}(\tilde{s})$. From Equations (6.17), (6.18) and (6.19) together with the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}\big[\widetilde{m}'_{n_2}(\tilde{s})\widetilde{m}_{n_2}(\tilde{s})\big]&=\mathbb{E}\big[\widetilde{m}'_{n_2}(\tilde{s})\big]\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]+\mathrm{Cov}(\widetilde{m}'_{n_2}(\tilde{s}),\widetilde{m}_{n_2}(\tilde{s}))=\mathbb{E}[\widetilde{m}'_{n_2}(\tilde{s})]\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]+\mathcal{O}_{\tilde{s}}(n_T^{-3/2}),\\ \mathbb{E}\big[\widetilde{m}_{n_2}^2(\tilde{s})\big]&=\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^2+\mathrm{Var}(\widetilde{m}_{n_2}(\tilde{s}))=\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^2+\mathcal{O}_{\tilde{s}}(n_T^{-3/2}),\\ \mathbb{E}\big[\widetilde{m}_{n_2}^3(\tilde{s})\big]&=\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^3+\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]\mathrm{Var}(\widetilde{m}_{n_2}(\tilde{s}))+\mathrm{Cov}(\widetilde{m}_{n_2}^2(\tilde{s}),\widetilde{m}_{n_2}(\tilde{s}))=\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^3+\mathcal{O}_{\tilde{s}}(n_T^{-3/2}).\end{aligned}$$

Hence, Equation (6.20) becomes

$$\begin{aligned}
&\rho_T \frac{n_2}{n_M}\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^3 + \left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^2 + \left(\tilde{s}+\rho_T\frac{n_2-n_1}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big] + 1 \\
&\quad = -\left(\frac{2n_2}{\sqrt{n_1n_2n_3}}+\frac{\rho_T}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big] - \frac{n_2}{\sqrt{n_1n_2n_3}}\left(2\tilde{s}+\frac{n_2}{\sqrt{n_1n_2n_3}}+\rho_T\frac{n_2}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^2 \\
&\qquad - \frac{n_2}{\sqrt{n_1n_2n_3}}\left(1+\tilde{s}\left(\rho_T\frac{n_2}{n_M}+\frac{n_2}{\sqrt{n_1n_2n_3}}\right)\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^3 \\
&\qquad - \left(\tilde{s}\frac{\rho_T}{n_M}+\frac{1}{n_2}\right)\mathbb{E}\big[\widetilde{m}'_{n_2}(\tilde{s})\big] - 3\frac{\rho_T}{n_M}\mathbb{E}\big[\widetilde{m}'_{n_2}(\tilde{s})\big]\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big] + \mathscr{O}_{\tilde{s}}(n_T^{-3/2}).
\end{aligned}$$

Consider the equation $\rho_T\frac{n_2}{n_M}\widetilde{m}^3(\tilde{s}) + (1+\tilde{s}\rho_T\frac{n_2}{n_M})\widetilde{m}^2(\tilde{s}) + (\tilde{s}+\rho_T\frac{n_2-n_1}{n_M})\widetilde{m}(\tilde{s}) + 1 = 0$. By expliciting the roots of this polynomial in $\widetilde{m}(\tilde{s})$, one can show that, for all $\tilde{s}\in\mathbb{C}\setminus\mathbb{R}$ it has a unique solution satisfying the properties of a Stieltjes transform (in particular, $\Im[\tilde{s}]\Im[\widetilde{m}(\tilde{s})] > 0$). Moreover, we define in this way the Stieltjes transform of a compactly supported probability distribution on $\mathbb{R}$, which we denote $\widetilde{\nu}$. Let us now subtract the relation $\rho_T\frac{n_2}{n_M}\widetilde{m}^3(\tilde{s}) + (1+\tilde{s}\rho_T\frac{n_2}{n_M})\widetilde{m}^2(\tilde{s}) + (\tilde{s}+\rho_T\frac{n_2-n_1}{n_M})\widetilde{m}(\tilde{s}) + 1 = 0$ to our previous equation in $\mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})]$. With the relations $a^3-b^3 = (a-b)(a^2+ab+b^2)$ and $a^2-b^2 = (a-b)(a+b)$, we find

$$\begin{aligned}
&\big(\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big] - \widetilde{m}(\tilde{s})\big)\left[\rho_T\frac{n_2}{n_M}\Big(\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^2 + \mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]\widetilde{m}(\tilde{s}) + \widetilde{m}^2(\tilde{s})\Big)\right. \\
&\qquad \left. + \left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\Big(\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big] + \widetilde{m}(\tilde{s})\Big) + \tilde{s} + \rho_T\frac{n_2-n_1}{n_M}\right] \\
&\quad = -\left(\frac{2n_2}{\sqrt{n_1n_2n_3}}+\frac{\rho_T}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big] - \frac{n_2}{\sqrt{n_1n_2n_3}}\left(2\tilde{s}+\frac{n_2}{\sqrt{n_1n_2n_3}}+\rho_T\frac{n_2}{n_M}\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^2 \\
&\qquad - \frac{n_2}{\sqrt{n_1n_2n_3}}\left(1+\tilde{s}\left(\rho_T\frac{n_2}{n_M}+\frac{n_2}{\sqrt{n_1n_2n_3}}\right)\right)\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^3 \\
&\qquad - \left(\tilde{s}\frac{\rho_T}{n_M}+\frac{1}{n_2}\right)\mathbb{E}\big[\widetilde{m}'_{n_2}(\tilde{s})\big] - 3\frac{\rho_T}{n_M}\mathbb{E}\big[\widetilde{m}'_{n_2}(\tilde{s})\big]\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big] + \mathscr{O}_{\tilde{s}}(n_T^{-3/2}). \qquad (6.21)
\end{aligned}$$

Let us define $g_{n_2}(\tilde{s}) \overset{\text{def}}{=} \frac{-1}{\rho_T\frac{n_2}{n_M}\left(\mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})]^2+\mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})]\widetilde{m}(\tilde{s})+\widetilde{m}^2(\tilde{s})\right)+\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\left(\mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})]+\widetilde{m}(\tilde{s})\right)+\tilde{s}+\rho_T\frac{n_2-n_1}{n_M}}$ as well as $\Delta_{n_2}(\tilde{s}) \overset{\text{def}}{=} \mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})] - \widetilde{m}(\tilde{s})$. Then, simply replacing $\mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})]$ by $\widetilde{m}(\tilde{s}) + \Delta_{n_2}(\tilde{s})$ in the expression of $g_{n_2}(\tilde{s})$, we obtain

$$\begin{aligned}
g_{n_2}(\tilde{s}) &= -\left[3\rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + 2\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s}) + \tilde{s} + \rho_T\frac{n_2-n_1}{n_M}\right. \\
&\qquad \left. + \rho_T\frac{n_2}{n_M}\left(3\widetilde{m}(\tilde{s})\Delta^{(2)}_{n_2}(\tilde{s}) + \Delta^{(2)2}_{n_2}(\tilde{s})\right) + \left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\Delta^{(2)}_{n_2}(\tilde{s})\right]^{-1} \\
&= g(\tilde{s}) \times \frac{1}{1 - g(\tilde{s})\Delta^{(2)}_{n_2}(\tilde{s})\left(\rho_T\frac{n_2}{n_M}\left(3\widetilde{m}(\tilde{s}) + \Delta^{(2)}_{n_2}(\tilde{s}) + \tilde{s}\right) + 1\right)} \qquad (6.22)
\end{aligned}$$

where $g(\tilde{s}) \overset{\text{def}}{=} \frac{-1}{3\rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + 2\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s}) + \tilde{s} + \rho_T\frac{n_2-n_1}{n_M}}$ and, differentiating the relation verified by $\widetilde{m}$ with respect to the complex variable $\tilde{s}$, we can also remark that $g(\tilde{s}) = \frac{\widetilde{m}'(\tilde{s})}{\widetilde{m}(\tilde{s}) + \rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s})}$. Here, we want to prove that $g(\tilde{s}) = \mathscr{O}_{\tilde{s}}(1)$, that is, there exist two polynomials P, Q with positive coefficients such that $g(\tilde{s}) \leqslant \frac{P(|\tilde{s}|)}{|\Im\tilde{s}|Q(|\Im\tilde{s}|)}$. In fact, in Lemma 6.15 below, we show a slightly stronger result which will also be useful later.

Lemma 6.15. *There exists a constant $C_0 > 0$ such that, $|g(\tilde{s})| \leqslant \frac{\max(2|\tilde{s}|, C_0)}{\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^2}$ for all $\tilde{s}\in\mathbb{C}\setminus\mathrm{Supp}\,\widetilde{\nu}$.*

Proof. $g(\tilde{s}) = \frac{\tilde{m}'(\tilde{s})}{\tilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\tilde{m}^2(\tilde{s})}$ and we know that $|\tilde{m}'(\tilde{s})| \leqslant \mathrm{Dist}(\tilde{s},\mathrm{Supp}\,\tilde{\nu})^{-2}$ (from the properties of Stieltjes transforms, see Proposition 2.7) so the only difficulty is in bounding $|\tilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\tilde{m}^2(\tilde{s})|$ from below. Since $\tilde{m}(\tilde{s}) = \int_{\mathbb{R}}\frac{\mathrm{d}\tilde{\nu}(t)}{t-\tilde{s}}$ where $\tilde{\nu}$ is a compactly supported probability distribution on $\mathbb{R}$, we have $|\tilde{s}(\tilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\tilde{m}^2(\tilde{s}))| \to 1$ as $|\tilde{s}| \to +\infty$. Hence, there exists a constant $M>0$ such that $|\tilde{s}| \geqslant M \implies \frac{1}{2|\tilde{s}|} \leqslant |\tilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\tilde{m}^2(\tilde{s})| \leqslant \frac{3}{2|\tilde{s}|}$. Conversely, $\tilde{s}\mapsto\tilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\tilde{m}^2(\tilde{s})$ never equals 0 on $\{z\in\mathbb{C}\setminus\mathrm{Supp}\,\tilde{\nu}\mid |z|\leqslant M\}$ and has a non-zero limit as $\tilde{s}$ approaches $\mathrm{Supp}\,\tilde{\nu}$ with $|\tilde{s}|\leqslant M$ (these statements are not straightforward but can be verified trough a study of the function $x\mapsto\tilde{m}(x)$ on $\mathbb{R}\setminus\mathrm{Supp}\,\tilde{\nu}$, which requires to derive an explicit expression of $\tilde{m}$ from the cubic equation it satisfies) therefore there exists a constant $C_0>0$ such that $|\tilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\tilde{m}^2(\tilde{s})| \geqslant \frac{1}{C_0}$ for all $\tilde{s}\in\mathbb{C}\setminus\mathrm{Supp}\,\tilde{\nu}$ such that $|\tilde{s}|\leqslant M$. In the end, we have $|g(\tilde{s})| \leqslant \frac{2|\tilde{s}|}{\mathrm{Dist}(\tilde{s},\mathrm{Supp}\,\tilde{\nu})^2}$ if $|\tilde{s}|\geqslant M$ and $|g(\tilde{s})|\leqslant\frac{C_0}{\mathrm{Dist}(\tilde{s},\mathrm{Supp}\,\tilde{\nu})^2}$ otherwise. □

From Equations (6.21) and (6.22), we have

$$\Delta_{n_2}(\tilde{s})\left(1-g(\tilde{s})\Delta^{(2)}_{n_2}(\tilde{s})\left(\rho_T\frac{n_2}{n_M}\left(3\tilde{m}(\tilde{s})+\Delta^{(2)}_{n_2}(\tilde{s})+\tilde{s}\right)+1\right)\right) = g(\tilde{s})\times\mathscr{O}_{\tilde{s}}(n_T^{-1/2})$$

thus, since $|\tilde{m}(\tilde{s})|\leqslant|\Im\tilde{s}|^{-1}$ and $g(\tilde{s})=\mathscr{O}_{\tilde{s}}(1)$ by Lemma 6.15, this implies that $\Delta_{n_2}(\tilde{s})=\mathscr{O}_{\tilde{s}}(n_T^{-1/2})$. We can then use this relation in Equation (6.22) and find that $g_{n_2}(\tilde{s}) = g(\tilde{s})+\mathscr{O}_{\tilde{s}}(n_T^{-1/2})$.

Remark 6.16. $\Delta_{n_2}(\tilde{s})=\mathscr{O}_{\tilde{s}}(n_T^{-1/2})$ shows that $\mathbb{E}[\tilde{m}_{n_2}(\tilde{s})]$ converges to $\tilde{m}(\tilde{s})$ for all $\tilde{s}\in\mathbb{C}\setminus\mathbb{R}$ and therefore that $\tilde{\nu}$ is the limiting spectral distribution of $\frac{n_T}{\sqrt{n_1n_2n_3}}\left[\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-\frac{n_1n_3}{n_T}\boldsymbol{I}_{n_2}\right]$.

In Equation (6.21) there are also $\mathbb{E}[\tilde{m}'_{n_2}(\tilde{s})]$ appearing in the last line. To handle these expectations, consider the relation obtained after applying $\frac{1}{n_2}\mathrm{Tr}$ to Equation (6.14) and differentiate it with respect to the complex variable $\tilde{s}$ (differentiation under $\mathbb{E}$ is possible because the integrand can be upper bounded on every compact subset of $\mathbb{C}\setminus\mathbb{R}$), this yields

$$\begin{aligned}\left(3\rho_T\frac{n_2}{n_M}\mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big]^2+2\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big]+\tilde{s}+\rho_T\frac{n_2-n_1}{n_M}\right)\mathbb{E}\big[\tilde{m}'_{n_2}(\tilde{s})\big]&\\ = -\left(\mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big]+\rho_T\frac{n_2}{n_M}\mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big]^2\right)+\mathscr{O}_{\tilde{s}}(n_T^{-1/2})&\end{aligned}$$

where we have also used the relations

$$\mathbb{E}\big[\tilde{m}^2_{n_2}(\tilde{s})\big] = \mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big]^2+\mathscr{O}_{\tilde{s}}(n_T^{-3/2}),\qquad \mathbb{E}\big[\tilde{m}'_{n_2}(\tilde{s})\tilde{m}_{n_2}(\tilde{s})\big] = \mathbb{E}\big[\tilde{m}'_{n_2}(\tilde{s})\big]\mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big]+\mathscr{O}_{\tilde{s}}(n_T^{-3/2})$$
$$\text{and}\qquad \mathbb{E}\big[\tilde{m}'_{n_2}(\tilde{s})\tilde{m}^2_{n_2}(\tilde{s})\big] = \mathbb{E}\big[\tilde{m}'_{n_2}(\tilde{s})\big]\mathbb{E}\big[\tilde{m}^2_{n_2}(\tilde{s})\big]+\mathscr{O}_{\tilde{s}}(n_T^{-3/2}).$$

Hence, since $\mathbb{E}[\tilde{m}_{n_2}(\tilde{s})] = \tilde{m}(\tilde{s})+\mathscr{O}(n_T^{-1/2})$, we find that

$$\mathbb{E}\big[\tilde{m}'_{n_2}(\tilde{s})\big] = g_{n_2}(\tilde{s})\left(\tilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\tilde{m}^2(\tilde{s})\right)+\mathscr{O}_{\tilde{s}}(n_T^{-1/2})$$

and Equation (6.21) becomes

$$\begin{aligned}\mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big] = \tilde{m}(\tilde{s})+&\left(\frac{2n_2}{\sqrt{n_1n_2n_3}}+\frac{\rho_T}{n_M}\right)g_{n_2}(\tilde{s})\mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big]\\ &+\frac{n_2}{\sqrt{n_1n_2n_3}}\left(2\tilde{s}+\rho_T\frac{n_2}{n_M}+\frac{n_2}{\sqrt{n_1n_2n_3}}\right)g_{n_2}(\tilde{s})\mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big]^2\\ &+\frac{n_2}{\sqrt{n_1n_2n_3}}\left(1+\tilde{s}\left(\rho_T\frac{n_2}{n_M}+\frac{n_2}{\sqrt{n_1n_2n_3}}\right)\right)g_{n_2}(\tilde{s})\mathbb{E}\big[\tilde{m}_{n_2}(\tilde{s})\big]^3\end{aligned}$$

$$+\left(\widetilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s})\right)\left(\frac{\rho_T}{n_M}\left(3\mathbb{E}\left[\widetilde{m}_{n_2}(\tilde{s})\right]+\tilde{s}\right)+\frac{1}{n_2}\right)g_{n_2}^2(\tilde{s})+\mathcal{O}_{\tilde{s}}(n_T^{-3/2}) \quad (6.23)$$

or, more simply,

$$\mathbb{E}\left[\widetilde{m}_{n_2}(\tilde{s})\right]=\widetilde{m}(\tilde{s})+\frac{2n_2}{\sqrt{n_1n_2n_3}}g_{n_2}(\tilde{s})\mathbb{E}\left[\widetilde{m}_{n_2}(\tilde{s})\right]+\frac{n_2}{\sqrt{n_1n_2n_3}}\left(2\tilde{s}+\rho_T\frac{n_2}{n_M}\right)g_{n_2}(\tilde{s})\mathbb{E}\left[\widetilde{m}_{n_2}(\tilde{s})\right]^2$$
$$+\frac{n_2}{\sqrt{n_1n_2n_3}}\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)g_{n_2}(\tilde{s})\mathbb{E}\left[\widetilde{m}_{n_2}(\tilde{s})\right]^3+\mathcal{O}_{\tilde{s}}(n_T^{-1}). \quad (6.24)$$

Let us now replace every $\mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})]$ in the right-hand side of Equation (6.24) by $\widetilde{m}(\tilde{s})+\mathcal{O}_{\tilde{s}}(n_T^{-1/2})$.

$$\mathbb{E}\left[\widetilde{m}_{n_2}(\tilde{s})\right]=\widetilde{m}(\tilde{s})\left[1+\frac{n_2}{\sqrt{n_1n_2n_3}}g_{n_2}(\tilde{s})\left(2+\left(2\tilde{s}+\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s})+\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}^2(\tilde{s})\right)\right]+\mathcal{O}_{\tilde{s}}(n_T^{-1}). \quad (6.25)$$

We can now inject Equation (6.25) in Equation (6.23). After a few simplifications, this yields

$$\mathbb{E}\left[\widetilde{m}_{n_2}(\tilde{s})\right]=\widetilde{m}(\tilde{s})+\frac{n_2}{\sqrt{n_1n_2n_3}}g_{n_2}(\tilde{s})\widetilde{m}(\tilde{s})q(\tilde{s})$$
$$+\frac{n_2}{n_1n_3}g_{n_2}^2(\tilde{s})\widetilde{m}(\tilde{s})q(\tilde{s})\left(2+2\left(2\tilde{s}+\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s})+3\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}^2(\tilde{s})\right)$$
$$+g_{n_2}(\tilde{s})\widetilde{m}(\tilde{s})\left(\frac{\rho_T}{n_M}+\frac{n_2}{n_1n_3}\left(\widetilde{m}(\tilde{s})+\tilde{s}\widetilde{m}^2(\tilde{s})\right)\right)$$
$$+\frac{1}{n_2}g_{n_2}^2(\tilde{s})\left(\widetilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s})\right)\left(\rho_T\frac{n_2}{n_M}(3\widetilde{m}(\tilde{s})+\tilde{s})+1\right)+\mathcal{O}_{\tilde{s}}(n_T^{-3/2}). \quad (6.26)$$

where $q(\tilde{s})\overset{\text{def}}{=}2+\left(2\tilde{s}+\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s})+\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}^2(\tilde{s})$.

We just need to find an expansion of $g_{n_2}(\tilde{s})$ up to a $\mathcal{O}_{\tilde{s}}(n_T^{-1})$ term to conclude this section. From the fact that $\Delta_{n_2}(\tilde{s})=\mathcal{O}_{\tilde{s}}(n_T^{-1/2})$, Equation (6.22) becomes

$$g_{n_2}(\tilde{s})=g(\tilde{s})\left(1+g(\tilde{s})\Delta_{n_2}^{(2)}(\tilde{s})\left(\rho_T\frac{n_2}{n_M}\left(3\widetilde{m}(\tilde{s})+\Delta_{n_2}^{(2)}(\tilde{s})+\tilde{s}\right)+1\right)+\mathcal{O}_{\tilde{s}}(n_T^{-1})\right)$$
$$=g(\tilde{s})\left(1+g(\tilde{s})\Delta_{n_2}^{(2)}(\tilde{s})\left(\rho_T\frac{n_2}{n_M}(3\widetilde{m}(\tilde{s})+\tilde{s})+1\right)\right)+\mathcal{O}_{\tilde{s}}(n_T^{-1})$$

Let us use Equation (6.25) (with $g_{n_2}(\tilde{s})$ replaced by $g(\tilde{s})+\mathcal{O}_{\tilde{s}}(n_T^{-1/2})$) to develop the $\Delta_{n_2}^{(2)}(\tilde{s})$ in the last equality.

$$g_{n_2}(\tilde{s})=g(\tilde{s})\left(1+\frac{n_2}{\sqrt{n_1n_2n_3}}g^2(\tilde{s})\widetilde{m}(\tilde{s})q(\tilde{s})\left(\rho_T\frac{n_2}{n_M}(3\widetilde{m}(\tilde{s})+\tilde{s})+1\right)\right)+\mathcal{O}_{\tilde{s}}(n_T^{-1}).$$

We can now inject this relation in Equation (6.26).

$$\mathbb{E}\left[\widetilde{m}_{n_2}(\tilde{s})\right]=\widetilde{m}(\tilde{s})+h(\tilde{s})+\mathcal{O}_{\tilde{s}}(n_T^{-3/2}) \quad (6.27)$$

with

$$h(\tilde{s})\overset{\text{def}}{=}\frac{n_2}{\sqrt{n_1n_2n_3}}g(\tilde{s})\widetilde{m}(\tilde{s})q(\tilde{s})+\frac{n_2}{n_1n_3}g^3(\tilde{s})\widetilde{m}^2(\tilde{s})q^2(\tilde{s})\left(\rho_T\frac{n_2}{n_M}(3\widetilde{m}(\tilde{s})+\tilde{s})+1\right)$$
$$+\frac{n_2}{n_1n_3}g^2(\tilde{s})\widetilde{m}(\tilde{s})q(\tilde{s})\left(2+2\left(2\tilde{s}+\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s})+3\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}^2(\tilde{s})\right)$$
$$+g(\tilde{s})\widetilde{m}(\tilde{s})\left(\frac{\rho_T}{n_M}+\frac{n_2}{n_1n_3}\left(\widetilde{m}(\tilde{s})+\tilde{s}\widetilde{m}^2(\tilde{s})\right)\right)$$
$$+\frac{1}{n_2}g^2(\tilde{s})\left(\widetilde{m}(\tilde{s})+\rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s})\right)\left(\rho_T\frac{n_2}{n_M}(3\widetilde{m}(\tilde{s})+\tilde{s})+1\right).$$

6.A.5 Confinement of the Spectrum

With the method presented in Section 2.3.2, we prove that for all $\varepsilon > 0$, there exists an integer n_0 such that $\mathrm{Dist}(\frac{n_T\lambda - n_1 n_3}{\sqrt{n_1 n_2 n_3}}, \mathrm{Supp}\,\widetilde{\nu}) \leqslant \varepsilon$ for all $\lambda \in \mathrm{Sp}\,\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ as soon as $n_T \geqslant n_0$.

Let $\varepsilon > 0$ and $\mathscr{S}_\varepsilon = \{x \in \mathbb{R} \mid \mathrm{Dist}(x, \mathrm{Supp}\,\widetilde{\nu}) < \varepsilon\}$. Let $\varphi : \mathbb{R} \to [0,1]$ be an infinitely differentiable function which equals 1 on $\mathrm{Supp}\,\widetilde{\nu}$ and 0 on $\mathbb{R} \setminus \mathscr{S}_\varepsilon$. We also define $\psi = 1 - \varphi$. In order to show that $\mathrm{Tr}\,\psi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right) \to 0$ almost surely as $n_1, n_2, n_3 \to +\infty$, we prove the convergence of the expectation and the $\mathscr{O}(n_T^{-3/2})$ decay of the fourth moment.

Firstly, we show that $\mathbb{E}\left[\mathrm{Tr}\,\psi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right] \to 0$ as $n_1, n_2, n_3 \to +\infty$. Just as presented in Section 2.3.2, using the Helffer-Sjöstrand formula (Proposition 2.12) and the development of $\mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})]$ (Equation (6.27)), we find that

$$\mathbb{E}\left[\frac{1}{n_2}\mathrm{Tr}\,\varphi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right] = 1 + \frac{2}{\pi}\Re\int_{\mathbb{C}^+} \frac{\partial \Phi_q[\varphi]}{\partial \bar{\tilde{s}}}(\tilde{s}) h(\tilde{s})\,\mathrm{d}\tilde{s} + \mathscr{O}(n_T^{-3/2}).$$

We can rely on Lemma 2.14 to evaluate the integral: h is analytic on $\mathbb{C} \setminus \mathrm{Supp}\,\widetilde{\nu}$, $h(\tilde{s}) \to 0$ as $|\tilde{s}| \to +\infty$ and $h(\bar{\tilde{s}}) = \overline{h(\tilde{s})}$. Hence, we just need to find an upper bound to $|h(\tilde{s})|$ as in the third assumption of Lemma 2.14. To this end, let us remark that, because $\mathrm{Supp}\,\widetilde{\nu}$ is compact, $\frac{|\tilde{s}|}{\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})} \to 1$ as $|\tilde{s}| \to +\infty$ and thus there exists a constant $M > 0$ such that $|\tilde{s}| \geqslant M \implies \frac{1}{2}\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu}) \leqslant |\tilde{s}| \leqslant \frac{3}{2}\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})$. Then, for any integer $\kappa \geqslant 1$,

- if $|\tilde{s}| \leqslant M$, $\frac{|\tilde{s}|}{\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^\kappa} \leqslant \frac{M}{\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^\kappa} \leqslant M \max(\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^{-\kappa}, 1)$,
- if $|\tilde{s}| \geqslant M$, $\frac{|\tilde{s}|}{\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^\kappa} \leqslant \frac{3}{2}\frac{1}{\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^{\kappa-1}} \leqslant \frac{3}{2}\max(\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^{-\kappa}, 1)$.

Thus, more simply, $\frac{|\tilde{s}|}{\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^\kappa} \leqslant C_s \max(\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^{-\kappa}, 1)$ with $C_s = \max(M, \frac{3}{2})$. Let us start by finding an upper bound to $|q(\tilde{s})|$.

$$\left|q(\tilde{s})\right| \leqslant 2 + \frac{2|\tilde{s}| + \rho_T\frac{n_2}{n_M}}{\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})} + \frac{1 + |\tilde{s}|\rho_T\frac{n_2}{n_M}}{\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^2} \leqslant C_q \max(\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^{-2}, 1)$$

with $C_q = 2 + 2C_s + \rho_T\frac{n_2}{n_M} + 1 + C_s\rho_T\frac{n_2}{n_M}$. Then, we can upper bound $|g(\tilde{s})|$ using Lemma 6.15.

$$\left|g(\tilde{s})\right| \leqslant C_g \max(\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^{-2}, 1)$$

with $C_g = \max(2C_s, C_0)$. We are now ready to upper bound $|h(\tilde{s})|$.

$$|h(\tilde{s})| \leqslant C_h \max(\mathrm{Dist}(\tilde{s}, \mathrm{Supp}\,\widetilde{\nu})^{-13}, 1)$$

with

$$\begin{aligned} C_h = \frac{n_2}{\sqrt{n_1 n_2 n_3}} C_g C_q + \frac{n_2}{n_1 n_3} C_g^3 C_q^2 \left(\rho_T\frac{n_2}{n_M}(3 + C_s) + 1\right) \\ + \frac{n_2}{n_1 n_3} C_g^2 C_q \left(2 + 2\left(2C_s + \rho_T\frac{n_2}{n_M}\right) + 3\left(1 + C_s\rho_T\frac{n_2}{n_M}\right)\right) \\ + C_g\left(\frac{\rho_T}{n_M} + \frac{n_2}{n_1 n_3}(1 + C_s)\right) + \frac{C_g^2}{n_2}\left(1 + \rho_T\frac{n_2}{n_M}\right)\left(\rho_T\frac{n_2}{n_M}(3 + C_s) + 1\right). \end{aligned}$$

Hence, since $\varphi(x)=1$ for all $x\in\operatorname{Supp}\widetilde{\nu}$ and $\lim_{y\to+\infty}-\mathrm{i}yh(\mathrm{i}y)=0$, Lemma 2.14 gives us

$$\frac{2}{\pi}\Re\int_{\mathbb{C}^+}\frac{\partial\Phi_q[\varphi]}{\partial\tilde{s}}(\tilde{s})h(\tilde{s})\,\mathrm{d}\tilde{s}=\lim_{y\downarrow 0}\frac{1}{\pi}\int_{\operatorname{Supp}\widetilde{\nu}}\Im[h(x+\mathrm{i}y)]\,\mathrm{d}x=0$$

and therefore $\mathbb{E}\left[\frac{1}{n_2}\operatorname{Tr}\varphi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right]=1+\mathcal{O}(n_T^{-3/2})$, which is equivalent to the desired result: $\mathbb{E}\left[\operatorname{Tr}\psi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right]=\mathcal{O}(n_T^{-1/2})$.

Secondly, we show that the convergence is almost sure thanks to Lemma 2.20. Indeed, with the Poincaré-Nash inequality (Lemma 2.19), we have

$$\operatorname{Var}\left(\operatorname{Tr}\psi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right)\leqslant A+B$$

with $A=\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbb{E}\left[\left|\frac{\partial\operatorname{Tr}\varphi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)}{\partial N_{i,j}}\right|^2\right]$ and $B=\sum_{u=1}^{n_2}\sum_{v=1}^{n_1n_3}\mathbb{E}\left[\left|\frac{\partial\operatorname{Tr}\varphi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)}{\partial W^{(2)}_{u,v}}\right|^2\right]$.

Skipping a few cumbersome computations, we find

$$A=4\frac{n_T^2}{n_1n_2n_3}\frac{\beta_T^2}{n_M}\mathbb{E}\left[\left\|\left(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\right)^\top\boldsymbol{T}^{(2)\top}\varphi'\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right\|_{\mathrm{F}}^2\right],$$
$$B=4\frac{n_T^2}{n_1n_2n_3}\frac{1}{n_T}\mathbb{E}\left[\left\|\boldsymbol{T}^{(2)\top}\varphi'\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right\|_{\mathrm{F}}^2\right].$$

Hence, using the inequality $\|\boldsymbol{AB}\|_{\mathrm{F}}\leqslant\|\boldsymbol{A}\|\|\boldsymbol{B}\|_{\mathrm{F}}$ and the fact that $\|\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}\|=1$ in A, we obtain

$$\operatorname{Var}\left(\operatorname{Tr}\psi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right)\leqslant 4\frac{n_T^2}{n_1n_2n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\mathbb{E}\left[\left\|\boldsymbol{T}^{(2)\top}\varphi'\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right\|_{\mathrm{F}}^2\right].$$

Moreover, we have

$$\frac{n_T}{\sqrt{n_1n_2n_3}}\left\|\boldsymbol{T}^{(2)\top}\varphi'\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right\|_{\mathrm{F}}^2$$
$$=\operatorname{Tr}u\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)+\frac{n_1n_3}{\sqrt{n_1n_2n_3}}\operatorname{Tr}\varphi'^2\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)$$

with $u:x\mapsto x\varphi'^2(x)$. Thus,

$$\operatorname{Var}\left(\operatorname{Tr}\psi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right)\leqslant 4\frac{n_T}{\sqrt{n_1n_2n_3}}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\mathbb{E}\left[\operatorname{Tr}u\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right]$$
$$+4\frac{n_T}{n_2}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\mathbb{E}\left[\operatorname{Tr}\varphi'^2\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right].$$

Since u and φ'^2 are compactly supported infinitely differentiable functions which equal 0 on $\operatorname{Supp}\widetilde{\nu}$, with the Helffer-Sjöstrand formula (Proposition 2.12), we find that

$$\operatorname{Var}\left(\operatorname{Tr}\psi\left(\frac{n_T\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}-n_1n_3\boldsymbol{I}_{n_2}}{\sqrt{n_1n_2n_3}}\right)\right)=4\frac{n_T}{\sqrt{n_1n_2n_3}}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\times\mathcal{O}(n_T^{-1/2})+4\frac{n_T}{n_2}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\times\mathcal{O}(n_T^{-1/2})$$

$$= \mathcal{O}(n_T^{-1}).$$

Unfortunately, this is not enough to apply Lemma 2.20, so we must evaluate the moment of order $\kappa = 4$.

$$\begin{aligned}
&\mathbb{E}\left[\left|\operatorname{Tr}\psi\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right) - \mathbb{E}\left[\operatorname{Tr}\psi\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right]\right|^4\right] \\
&\qquad = \operatorname{Var}\left(\left(\operatorname{Tr}\psi\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right) - \mathbb{E}\left[\operatorname{Tr}\psi\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right]\right)^2\right) \\
&\qquad\qquad + \left|\mathbb{E}\left[\left(\operatorname{Tr}\psi\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right) - \mathbb{E}\left[\operatorname{Tr}\psi\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right]\right)^2\right]\right|^2.
\end{aligned}$$

From our previous result, the rightmost term is $\mathcal{O}(n_T^{-2})$. To handle the first term of the right-hand side, we proceed similarly as above, with the Poincaré-Nash inequality (Lemma 2.19), and we find that

$$\begin{aligned}
&\operatorname{Var}\left(\left(\operatorname{Tr}\psi\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right) - \mathbb{E}\left[\operatorname{Tr}\psi\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right]\right)^2\right) \\
&\leqslant 16\frac{n_T^2}{n_1 n_2 n_3}\left(\frac{\beta_T^2}{n_M} + \frac{1}{n_T}\right)\mathbb{E}\left[\left|\left(\operatorname{Tr}\varphi - \mathbb{E}\operatorname{Tr}\varphi\right)\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right|^2 \left\|\boldsymbol{T}^{(2)\top}\varphi'\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right\|_{\mathrm{F}}^2\right] \\
&= 16\frac{n_T}{\sqrt{n_1 n_2 n_3}}\left(\frac{\beta_T^2}{n_M} + \frac{1}{n_T}\right)\mathbb{E}\left[\left|\left(\operatorname{Tr}\varphi - \mathbb{E}\operatorname{Tr}\varphi\right)\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right|^2 \operatorname{Tr} u\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right] \\
&\qquad + 16\frac{n_T}{n_2}\left(\frac{\beta_T^2}{n_M} + \frac{1}{n_T}\right)\mathbb{E}\left[\left|\left(\operatorname{Tr}\varphi - \mathbb{E}\operatorname{Tr}\varphi\right)\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right|^2 \operatorname{Tr}\varphi'^2\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right)\right] \\
&= \mathcal{O}(n_T^{-3/2}).
\end{aligned}$$

Finally, we can conclude with Lemma 2.20:

$$\operatorname{Tr}\psi\left(\frac{n_T \boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}}\right) \xrightarrow[n_1,n_2,n_3\to+\infty]{} 0 \qquad \text{almost surely.}$$

6.B Proof of Theorem 6.2

6.B.1 Convergence of Bilinear Forms

Recall that $\widetilde{\boldsymbol{Q}}_0^{(2)}$ denotes the resolvent of the model without signal ($\beta_M = 0$). That is,

$$\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{s}) = \left(\frac{n_T \boldsymbol{T}_0^{(2)}\boldsymbol{T}_0^{(2)\top} - n_1 n_3 \boldsymbol{I}_{n_2}}{\sqrt{n_1 n_2 n_3}} - \tilde{s}\boldsymbol{I}_{n_2}\right)^{-1} \qquad \text{with} \qquad \boldsymbol{T}_0^{(2)} \overset{\text{def}}{=} \frac{\beta_T}{\sqrt{n_M}}\boldsymbol{N}^\top\left(\boldsymbol{I}_{n_1} \boxtimes \boldsymbol{z}\right)^\top + \frac{1}{\sqrt{n_T}}\boldsymbol{W}^{(2)}.$$

We want to show the following result: for all bounded (sequences of) vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{n_2}$,

$$\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{s})\boldsymbol{b} - \widetilde{m}(\tilde{s})\langle \boldsymbol{a}, \boldsymbol{b}\rangle \xrightarrow[n_1,n_2,n_3\to+\infty]{} 0 \qquad \text{almost surely.}$$

Since we already know that $\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{s}) - \mathbb{E}[\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{s})])\boldsymbol{b} \to$ almost surely (see Section 6.A.3), we only need to show the convergence of the expectation. Let us multiply Equation (6.14) when $\beta_M = 0$ by $\boldsymbol{a}^\top$ on the left and $\boldsymbol{b}$ on the right.

$$\rho_T \frac{n_2}{n_M}\mathbb{E}\big[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b}\widetilde{m}_{n_2}^2(\tilde{s})\big] + \left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\mathbb{E}\big[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b}\widetilde{m}_{n_2}(\tilde{s})\big] + \left(\tilde{s}+\rho_T\frac{n_2-n_1}{n_M}\right)\mathbb{E}\big[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b}\big] + \langle \boldsymbol{a},\boldsymbol{b}\rangle = \mathcal{O}_{\tilde{s}}(n_T^{-1/2})$$

From Equations (6.16), (6.17) and (6.19) together with the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b}\widetilde{m}_{n_2}^2(\tilde{s})\big] &= \mathbb{E}\big[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b}\big]\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big]^2 + \mathcal{O}_{\tilde{s}}(n_T^{-1})\\ \text{and}\qquad \mathbb{E}\big[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b}\widetilde{m}_{n_2}(\tilde{s})\big] &= \mathbb{E}\big[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b}\big]\mathbb{E}\big[\widetilde{m}_{n_2}(\tilde{s})\big] + \mathcal{O}_{\tilde{s}}(n_T^{-1}).\end{aligned}$$

Moreover $\mathbb{E}[\widetilde{m}_{n_2}(\tilde{s})] = \widetilde{m}(\tilde{s}) + \mathcal{O}_{\tilde{s}}(n_T^{-1/2})$ so we obtain

$$\underbrace{\left(\rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + \left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s}) + \left(\tilde{s}+\rho_T\frac{n_2-n_1}{n_M}\right)\right)}_{=-\frac{1}{\widetilde{m}(\tilde{s})}}\mathbb{E}\big[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}^{(2)}\boldsymbol{b}\big] + \langle \boldsymbol{a},\boldsymbol{b}\rangle = \mathcal{O}_{\tilde{s}}(n_T^{-1/2}),$$

which gives the desired result.

6.B.2 Isolated Eigenvalue

We seek the asymptotic position $\tilde{\xi}$ of an eigenvalue of $\frac{n_T}{\sqrt{n_1n_2n_3}}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - \frac{n_1n_3}{\sqrt{n_1n_2n_3}}\boldsymbol{I}_{n_2}$ which is *not* in the limiting spectrum of $\frac{n_T}{\sqrt{n_1n_2n_3}}\boldsymbol{T}_0^{(2)}\boldsymbol{T}_0^{(2)\top} - \frac{n_1n_3}{\sqrt{n_1n_2n_3}}\boldsymbol{I}_{n_2}$. Because $\tilde{\xi}\in \mathrm{Sp}\Big(\frac{n_T}{\sqrt{n_1n_2n_3}}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - \frac{n_1n_3}{\sqrt{n_1n_2n_3}}\boldsymbol{I}_{n_2}\Big)$, it must verify

$$\det\left(\frac{n_T}{\sqrt{n_1n_2n_3}}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - \frac{n_1n_3}{\sqrt{n_1n_2n_3}}\boldsymbol{I}_{n_2} - \tilde{\xi}\boldsymbol{I}_{n_2}\right) = 0.$$

With the decomposition $\boldsymbol{T}^{(2)} = \beta_T\beta_M\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top + \boldsymbol{T}_0^{(2)}$, this becomes

$$\begin{aligned}\det\Bigg(\frac{n_T}{\sqrt{n_1n_2n_3}}\Big(\beta_T^2\beta_M^2\boldsymbol{y}\boldsymbol{y}^\top + \beta_T\beta_M\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}_0^{(2)\top} + \beta_T\beta_M\boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\boldsymbol{y}^\top\Big)\\ + \frac{n_T}{\sqrt{n_1n_2n_3}}\boldsymbol{T}_0^{(2)}\boldsymbol{T}_0^{(2)\top} - \frac{n_1n_3}{\sqrt{n_1n_2n_3}}\boldsymbol{I}_{n_2} - \tilde{\xi}\boldsymbol{I}_{n_2}\Bigg) = 0\end{aligned}$$

and, for n_T large enough, $\det\Big(\frac{n_T}{\sqrt{n_1n_2n_3}}\boldsymbol{T}_0^{(2)}\boldsymbol{T}_0^{(2)\top} - \frac{n_1n_3}{\sqrt{n_1n_2n_3}}\boldsymbol{I}_{n_2} - \tilde{\xi}\boldsymbol{I}_{n_2}\Big) \neq 0$ by the confinement of the spectrum of $\frac{n_T}{\sqrt{n_1n_2n_3}}\boldsymbol{T}_0^{(2)}\boldsymbol{T}_0^{(2)\top} - \frac{n_1n_3}{\sqrt{n_1n_2n_3}}\boldsymbol{I}_{n_2}$ proven in Section 6.A.5, therefore

$$\det\left(\frac{n_T}{\sqrt{n_1n_2n_3}}\Big(\beta_T^2\beta_M^2\boldsymbol{y}\boldsymbol{y}^\top + \beta_T\beta_M\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}_0^{(2)\top} + \beta_T\beta_M\boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\boldsymbol{y}^\top\Big)\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi}) + \boldsymbol{I}_{n_2}\right) = 0.$$

Notice that the sum $\beta_T^2\beta_M^2\boldsymbol{y}\boldsymbol{y}^\top + \beta_T\beta_M\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}_0^{(2)\top} + \beta_T\beta_M\boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\boldsymbol{y}^\top$ can be written as the matrix product $\begin{bmatrix}\beta_T\beta_M\boldsymbol{y} & \beta_T\beta_M\boldsymbol{y} & \boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\end{bmatrix}\begin{bmatrix}\beta_T\beta_M\boldsymbol{y}^\top\\ (\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}_0^{(2)\top}\\ \beta_T\beta_M\boldsymbol{y}^\top\end{bmatrix}$. Hence, with Sylvester's identity (Proposition 2.22), our $n_2\times n_2$ determinant becomes a 3×3 determinant:

$$\det\left(\frac{n_T}{\sqrt{n_1n_2n_3}}\begin{bmatrix}\beta_T^2\beta_M^2\boldsymbol{y}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{y} & \beta_T^2\beta_M^2\boldsymbol{y}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{y} & \beta_T\beta_M\boldsymbol{y}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\\ \beta_T\beta_M(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}_0^{(2)\top}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{y} & \beta_T\beta_M(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}_0^{(2)\top}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{y} & (\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}_0^{(2)\top}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\\ \beta_T^2\beta_M^2\boldsymbol{y}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{y} & \beta_T^2\beta_M^2\boldsymbol{y}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{y} & \beta_T\beta_M\boldsymbol{y}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\end{bmatrix} + \boldsymbol{I}_3\right) = 0. \quad (6.28)$$

We have $\frac{n_T}{\sqrt{n_1n_2n_3}}\beta_T^2\beta_M^2\boldsymbol{y}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{y} \to \rho_T\beta_M^2\widetilde{m}(\tilde{\xi})$ almost surely so this is the limit of the entries $(1,1)$, $(1,2)$, $(3,1)$ and $(3,2)$. Then, the entries $(1,3)$, $(2,1)$, $(2,2)$ and $(3,3)$ vanish almost surely since

$$\frac{n_T}{\sqrt{n_1 n_2 n_3}}\beta_T\beta_M\left|\boldsymbol{y}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\right| \\ \leqslant \frac{\rho_T\beta_M}{\sqrt{n_M}}\left|\boldsymbol{y}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right| + \frac{n_T}{\sqrt{n_1 n_2 n_3}}\frac{\beta_T\beta_M}{\sqrt{n_T}}\left|\boldsymbol{y}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{W}^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\right|$$

where $\frac{1}{\sqrt{n_T}}\|\boldsymbol{W}^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\| = \mathcal{O}(1)$ almost surely[5] so the second term in the right-hand side is $\mathcal{O}(n_T^{-1/4})$ almost surely and the first term is handled with the following lemma.

Lemma 6.17. *For all bounded (sequences of) vectors $\boldsymbol{a}\in\mathbb{R}^{n_2}$, $\boldsymbol{b}\in\mathbb{R}^{n_1}$, $\frac{1}{\sqrt{n_M}}|\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}| \to 0$ almost surely as $n_1, n_2, n_3 \to +\infty$.*

Proof. The proof is performed with the standard approach relying on Stein's lemma (Lemma 2.18) and the Poincaré-Nash inequality (Lemma 2.19).

With the Poincaré-Nash inequality we find that

$$\mathrm{Var}\left(\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\right) \\ \leqslant \frac{\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2}{n_M}\mathbb{E}\left[2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^2 + 4\frac{n_T^2}{n_1 n_2 n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\|\boldsymbol{N}\|^2\left\|\boldsymbol{T}_0^{(2)}\right\|^2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^4\right] = \mathcal{O}(n_T^{-1/2}),$$

$$\mathbb{E}\left[\left|\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b} - \mathbb{E}\left[\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\right]\right|^4\right] \\ \leqslant \frac{\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2}{n_M^2}\mathbb{E}\left[\left|\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b} - \mathbb{E}\left[\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\right]\right|^2\right. \\ \left.\times\left(8\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^2 + 16\frac{n_T^2}{n_1 n_2 n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\|\boldsymbol{N}\|^2\left\|\boldsymbol{T}_0^{(2)}\right\|^2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^4\right)\right] \\ + \mathrm{Var}\left(\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\right)^2 = \mathcal{O}(n_T^{-1}),$$

$$\mathbb{E}\left[\left|\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b} - \mathbb{E}\left[\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\right]\right|^6\right] \\ \leqslant \frac{\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2}{n_M^3}\mathbb{E}\left[\left|\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b} - \mathbb{E}\left[\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\right]\right|^4\right. \\ \left.\times\left(18\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^2 + 36\frac{n_T^2}{n_1 n_2 n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\|\boldsymbol{N}\|^2\left\|\boldsymbol{T}_0^{(2)}\right\|^2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^4\right)\right] \\ + \mathbb{E}\left[\left|\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b} - \mathbb{E}\left[\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\right]\right|^4\right]^2 = \mathcal{O}(n_T^{-2}).$$

By Lemma 2.20, this shows that $\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b} - \mathbb{E}[\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}] \to 0$ almost surely. Thus, we just need to show that $\mathbb{E}[\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}] \to 0$.

With Stein's lemma, we find that

$$\frac{1}{\sqrt{n_M}}\mathbb{E}\left[\boldsymbol{a}^\top \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\right] = -\frac{\beta_T}{n_M}\frac{n_T}{\sqrt{n_1 n_2 n_3}}\boldsymbol{a}^\top\mathbb{E}\left[\left(\widetilde{\boldsymbol{Q}}_0^{(2)2}(\tilde{\xi}) + \widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\,\mathrm{Tr}\,\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right)\boldsymbol{T}_0^{(2)}\right](\boldsymbol{b}\boxtimes\boldsymbol{z})$$

and, with the decomposition $\boldsymbol{T}_0^{(2)} = \frac{\beta_T}{\sqrt{n_M}}\boldsymbol{N}^\top(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z}) + \frac{1}{\sqrt{n_T}}\boldsymbol{W}^{(2)}$, it becomes

[5] See footnote 16 in Chapter 5.

$$\frac{1}{\sqrt{n_M}}\mathbb{E}\Big[\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\Big]=-\frac{\rho_T}{n_M\sqrt{n_M}}\mathbb{E}\Big[\boldsymbol{a}^\top\Big(\widetilde{\boldsymbol{Q}}_0^{(2)2}(\tilde{\xi})+\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\Big)\boldsymbol{N}^\top\boldsymbol{b}\Big]$$
$$-\frac{\beta_T}{n_M\sqrt{n_T}}\frac{n_T}{\sqrt{n_1n_2n_3}}\mathbb{E}\Big[\boldsymbol{a}^\top\Big(\widetilde{\boldsymbol{Q}}_0^{(2)2}(\tilde{\xi})+\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\Big)\boldsymbol{W}_0^{(2)}(\boldsymbol{b}\boxtimes\boldsymbol{z})\Big].$$

Almost surely, $\|\boldsymbol{N}^\top\boldsymbol{b}\|,\|\boldsymbol{W}^{(2)}(\boldsymbol{b}\boxtimes\boldsymbol{z})\|=\mathcal{O}(n_T^{1/2})$ therefore we just have

$$\frac{1}{\sqrt{n_M}}\mathbb{E}\Big[\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\Big]=-\frac{\rho_T}{n_M\sqrt{n_M}}\mathbb{E}\Big[\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\Big]+\mathcal{O}(n_T^{-1/4})$$

or, equivalently,

$$\frac{1}{\sqrt{n_M}}\mathbb{E}\left[\left(1+\rho_T\frac{n_2}{n_M}\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})}{n_2}\right)\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}\right]=\mathcal{O}(n_T^{-1/4}).$$

Since $\frac{1}{n_2}\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\to\tilde{m}(\tilde{\xi})$ almost surely, this shows that $\mathbb{E}[\frac{1}{\sqrt{n_M}}\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{b}]=\mathcal{O}(n_T^{-1/4})$. □

We just need to find the limiting behavior of the entry $(2,3)$. To this end, notice that

$$\frac{n_T}{\sqrt{n_1n_2n_3}}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{T}_0^{(2)\top}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})$$
$$=\frac{\rho_T}{n_M}\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}+\frac{1}{\sqrt{n_1n_2n_3}}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{W}^{(2)}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{W}^{(2)\top}(\boldsymbol{x}\boxtimes\boldsymbol{z})$$
$$+\frac{n_T}{\sqrt{n_1n_2n_3}}\frac{\beta_T}{\sqrt{n_Tn_M}}\Big(\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{W}^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})+(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top\boldsymbol{W}^{(2)\top}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\Big).$$

Since $\|\boldsymbol{N}^\top\boldsymbol{x}\|,\|\boldsymbol{W}^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\|=\mathcal{O}(n_T^{1/2})$ almost surely, we can see that the only non-vanishing term is $\frac{\rho_T}{n_M}\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}$. Moreover, we have the following lemma.

Lemma 6.18. $\frac{\rho_T}{n_M}\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\to\frac{\rho_T\frac{n_2}{n_M}\tilde{m}(\tilde{\xi})}{1+\rho_T\frac{n_2}{n_M}\tilde{m}(\tilde{\xi})}$ *almost surely as* $n_1,n_2,n_3\to+\infty$.

Proof. Firstly, let us show that $\frac{\rho_T}{n_M}\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}-\frac{\rho_T}{n_M}\boldsymbol{x}^\top\mathbb{E}[\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top]\boldsymbol{x}\to0$ almost surely. With the Poincaré-Nash inequality (Lemma 2.19), we find that

$$\operatorname{Var}\left(\frac{\rho_T}{n_M}\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right)$$
$$\leqslant8\frac{\rho_T^2}{n_M^2}\mathbb{E}\left[\|\boldsymbol{N}\|^2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^2+\frac{n_T^2}{n_1n_2n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\|\boldsymbol{N}\|^4\left\|\boldsymbol{T}_0^{(2)}\right\|^2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^4\right]=\mathcal{O}(n_T^{-1/2}),$$
$$\mathbb{E}\left[\left|\frac{\rho_T}{n_M}\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}-\mathbb{E}\left[\frac{\rho_T}{n_M}\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right]\right|^4\right]$$
$$\leqslant32\frac{\rho_T^4}{n_M^4}\mathbb{E}\left[\left|\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}-\mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right]\right|^2\frac{n_T^2}{n_1n_2n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\|\boldsymbol{N}\|^4\left\|\boldsymbol{T}_0^{(2)}\right\|^2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^4\right]$$
$$+32\frac{\rho_T^4}{n_M^4}\mathbb{E}\left[\left|\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}-\mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right]\right|^2\|\boldsymbol{N}\|^2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^2\right]$$
$$+\operatorname{Var}\left(\frac{\rho_T}{n_M}\boldsymbol{x}^\top\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right)^2=\mathcal{O}(n_T^{-1}),$$

$$
\begin{aligned}
&\mathbb{E}\left[\left|\frac{\rho_T}{n_M}\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}-\mathbb{E}\left[\frac{\rho_T}{n_M}\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right]\right|^6\right]\\
&\leqslant 72\frac{\rho_T^6}{n_M^6}\mathbb{E}\left[\left|\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}-\mathbb{E}\left[\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right]\right|^4\frac{n_T^2}{n_1n_2n_3}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\|\boldsymbol{N}\|^4\left\|\boldsymbol{T}_0^{(2)}\right\|^2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^4\right]\\
&\qquad+72\frac{\rho_T^6}{n_M^6}\mathbb{E}\left[\left|\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}-\mathbb{E}\left[\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right]\right|^4\|\boldsymbol{N}\|^2\left\|\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right\|^2\right]\\
&\qquad\qquad+\mathbb{E}\left[\left|\frac{\rho_T}{n_M}\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}-\mathbb{E}\left[\frac{\rho_T}{n_M}\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right]\right|^4\right]^2=\mathcal{O}(n_T^{-2}).
\end{aligned}
$$

Hence $\frac{\rho_T}{n_M}\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}-\frac{\rho_T}{n_M}\boldsymbol{x}^\top\mathbb{E}[\boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top]\boldsymbol{x}\to 0$ almost surely by Lemma 2.20.

Secondly, we just need to show that $\mathbb{E}[\frac{\rho_T}{n_M}\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}]\to\frac{\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi})}{1+\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi})}$ in order to conclude. With Stein's lemma (Lemma 2.18), we find that

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\rho_T}{n_M}\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right]\\
&\quad=\frac{\rho_T}{n_M}\mathbb{E}\left[\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})-\frac{n_T}{\sqrt{n_1n_2n_3}}\frac{\beta_T}{\sqrt{n_M}}\boldsymbol{x}^\top\boldsymbol{N}\left(\widetilde{\boldsymbol{Q}}_0^{(2)2}(\tilde{\xi})+\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right)\boldsymbol{T}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\right]\\
&\quad=\frac{\rho_T}{n_M}\mathbb{E}\left[\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})-\frac{\rho_T}{n_M}\boldsymbol{x}^\top\boldsymbol{N}\left(\widetilde{\boldsymbol{Q}}_0^{(2)2}(\tilde{\xi})+\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right)\boldsymbol{N}^\top\boldsymbol{x}\right.\\
&\qquad\qquad\left.-\frac{n_T}{\sqrt{n_1n_2n_3}}\frac{\beta_T}{\sqrt{n_Mn_T}}\boldsymbol{x}^\top\boldsymbol{N}\left(\widetilde{\boldsymbol{Q}}_0^{(2)2}(\tilde{\xi})+\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\right)\boldsymbol{W}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\right]
\end{aligned}
$$

where the second line stems from the decomposition $\boldsymbol{T}_0^{(2)}=\frac{\beta_T}{\sqrt{n_M}}\boldsymbol{N}^\top(\boldsymbol{I}_{n_1}\boxtimes\boldsymbol{z})^\top+\frac{1}{\sqrt{n_T}}\boldsymbol{W}^{(2)}$. Then, recalling that $\widetilde{m}_{n_2}=\frac{1}{n_2}\operatorname{Tr}\widetilde{\boldsymbol{Q}}_0^{(2)}$, we have

$$
\mathbb{E}\left[\frac{\rho_T}{n_M}\left(1+\frac{n_2}{n_M}\widetilde{m}_{n_2}(\tilde{\xi})\right)\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}\right]=\rho_T\frac{n_2}{n_M}\mathbb{E}\left[\widetilde{m}_{n_2}(\tilde{\xi})\right]+\mathcal{O}(n_T^{-1/4})
$$

since $\|\boldsymbol{W}_0^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\|=\mathcal{O}(n_T^{1/2})$ almost surely. Then, using the Cauchy-Schwarz inequality, we have $|\operatorname{Cov}(\widetilde{m}_{n_2}(\tilde{\xi}),\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x})|\leqslant\sqrt{\operatorname{Var}(\widetilde{m}_{n_2}(\tilde{\xi})\operatorname{Var}(\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}))}=\mathcal{O}(n_T^{-5/4})$ whence the relation $\mathbb{E}[\widetilde{m}_{n_2}(\tilde{\xi})\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}]=\mathbb{E}[\widetilde{m}_{n_2}(\tilde{\xi})]\mathbb{E}[\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}]+\mathcal{O}(n_T^{-5/4})$, which we use with $\mathbb{E}[\widetilde{m}_{n_2}(\tilde{\xi})]=\widetilde{m}(\tilde{\xi})+\mathcal{O}(n_T^{-1/2})$ (see Section 6.A.4) to obtain $\frac{\rho_T}{n_M}\mathbb{E}[\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}_0^{(2)}(\tilde{\xi})\boldsymbol{N}^\top\boldsymbol{x}]=\frac{\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi})}{1+\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi})}+\mathcal{O}(n_T^{-1/4})$. □

Eventually, as the determinant is a continuous function in the entries of the matrix, Equation (6.28) becomes asymptotically almost surely

$$
\det\left(\begin{bmatrix}\rho_T\beta_M^2\widetilde{m}(\tilde{\xi}) & \rho_T\beta_M^2\widetilde{m}(\tilde{\xi}) & 0\\ 0 & 0 & \frac{\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi})}{1+\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi})}\\ \rho_T\beta_M^2\widetilde{m}(\tilde{\xi}) & \rho_T\beta_M^2\widetilde{m}(\tilde{\xi}) & 0\end{bmatrix}+\boldsymbol{I}_3\right)=0.
$$

Computing this determinant gives $1+\rho_T\left(\frac{n_2}{n_M}+\beta_M^2\right)\widetilde{m}(\tilde{\xi})=0$. Then, injecting $\widetilde{m}(\tilde{\xi})=\frac{-1}{\rho_T\left(\frac{n_2}{n_M}+\beta_M^2\right)}$ in the equation $\rho_T\frac{n_2}{n_M}\widetilde{m}^3(\tilde{s})+(1+\tilde{s}\rho_T\frac{n_2}{n_M})\widetilde{m}^2(\tilde{s})+(\tilde{s}+\rho_T\frac{n_2-n_1}{n_M})\widetilde{m}(\tilde{s})+1=0$ defining $\widetilde{m}$, we obtain

$$
\tilde{\xi}=\frac{\rho_T}{\beta_M^2}\left(\frac{n_1}{n_M}+\beta_M^2\right)\left(\frac{n_2}{n_M}+\beta_M^2\right)+\frac{1}{\rho_T\left(\frac{n_2}{n_M}+\beta_M^2\right)}.
$$

6.B.3 Deterministic Equivalent of the Resolvent

In order to determine the alignment of the dominant eigenvector of $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top}$ with $\boldsymbol{y}$ in the next section, let us exhibit a deterministic equivalent of the resolvent $\widetilde{\boldsymbol{Q}}^{(2)}$.

Applying the change of variable $(s, \boldsymbol{Q}^{(2)}) \curvearrowright (\tilde{s}, \widetilde{\boldsymbol{Q}}^{(2)})$ to Equation (6.11) yields

$$\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}^{(2)}\big] = -\frac{\beta_T}{\sqrt{n_M}}\frac{n_T}{\sqrt{n_1 n_2 n_3}}\mathbb{E}\big[\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top \boldsymbol{T}^{(2)\top}\big(\widetilde{\boldsymbol{Q}}^{(2)}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)} + \widetilde{\boldsymbol{Q}}^{(2)2}\big)\big].$$

Hence, we have

$$\begin{aligned}
\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top \boldsymbol{M}\widetilde{\boldsymbol{Q}}^{(2)}\big] &= \beta_M \boldsymbol{y}\boldsymbol{y}^\top \mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)}\big] + \frac{1}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top \boldsymbol{N}\widetilde{\boldsymbol{Q}}^{(2)}\big] \\
&= \beta_M \boldsymbol{y}\boldsymbol{y}^\top \mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)}\big] - \frac{\beta_T}{n_M}\frac{n_T}{\sqrt{n_1 n_2 n_3}}\mathbb{E}\big[\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top \boldsymbol{T}^{(2)\top}\big(\widetilde{\boldsymbol{Q}}^{(2)}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)} + \widetilde{\boldsymbol{Q}}^{(2)2}\big)\big] \\
&= \beta_M \boldsymbol{y}\boldsymbol{y}^\top \mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)}\big] - \frac{\beta_T}{n_M}\frac{n_T}{\sqrt{n_1 n_2 n_3}}\mathbb{E}\left[\left(\beta_T \boldsymbol{y}\boldsymbol{x}^\top \boldsymbol{M} + \frac{1}{\sqrt{n_T}}\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top \boldsymbol{W}^{(2)\top}\right)\big(\widetilde{\boldsymbol{Q}}^{(2)}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)} + \widetilde{\boldsymbol{Q}}^{(2)2}\big)\right] \\
&= \beta_M \boldsymbol{y}\boldsymbol{y}^\top \mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)}\big] - \rho_T \frac{n_2}{n_M}\mathbb{E}\left[\frac{\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}}{n_2}\boldsymbol{y}\boldsymbol{x}^\top \boldsymbol{M}\widetilde{\boldsymbol{Q}}^{(2)}\right] \\
&\quad - \mathbb{E}\left[\frac{\rho_T}{n_M}\boldsymbol{y}\boldsymbol{x}^\top \boldsymbol{M}\widetilde{\boldsymbol{Q}}^{(2)2} + \frac{\beta_T}{n_M}\frac{n_T}{\sqrt{n_1 n_2 n_3}}\frac{1}{\sqrt{n_T}}\boldsymbol{y}(\boldsymbol{x}\boxtimes\boldsymbol{z})^\top \boldsymbol{W}^{(2)\top}\big(\widetilde{\boldsymbol{Q}}^{(2)}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)} + \widetilde{\boldsymbol{Q}}^{(2)2}\big)\right]
\end{aligned}$$

Since $\|\boldsymbol{M}\| = \mathcal{O}(1)$ and $\frac{1}{\sqrt{n_T}}\|\boldsymbol{W}^{(2)}(\boldsymbol{x}\boxtimes\boldsymbol{z})\| = \mathcal{O}(1)$ almost surely, the norm of the last expectation vanishes as $n_1, n_2, n_3 \to +\infty$. Moreover $\frac{1}{n_2}\operatorname{Tr}\widetilde{\boldsymbol{Q}}^{(2)}(\tilde{s}) \to \widetilde{m}(\tilde{s})$ almost surely, therefore

$$\left\|\left(1 + \rho_T \frac{n_2}{n_M}\widetilde{m}(\tilde{s})\right)\mathbb{E}\big[\boldsymbol{y}\boldsymbol{x}^\top \boldsymbol{M}\widetilde{\boldsymbol{Q}}^{(2)}\big] - \beta_M \boldsymbol{y}\boldsymbol{y}^\top \mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)}\big]\right\| \to 0.$$

Thus, from Equation (6.14), we have

$$\left\|\left(\rho_T \frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + \left(1 + \tilde{s}\rho_T \frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s}) + \left(\tilde{s} + \rho_T \frac{n_2 - n_1}{n_M}\right)\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)}\big] + \boldsymbol{I}_{n_2} - \frac{\rho_T \beta_M^2}{1 + \rho_T \frac{n_2}{n_M}\widetilde{m}(\tilde{s})}\boldsymbol{y}\boldsymbol{y}^\top \mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(2)}\big]\right\| \to 0$$

and, since $\rho_T \frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + \left(1 + \tilde{s}\rho_T \frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s}) + \left(\tilde{s} + \rho_T \frac{n_2 - n_1}{n_M}\right) = \frac{-1}{\widetilde{m}(\tilde{s})}$, we can define the following deterministic equivalent (Definition 2.17) of $\widetilde{\boldsymbol{Q}}^{(2)}(\tilde{s})$:

$$\bar{\boldsymbol{Q}}(\tilde{s}) \overset{\text{def}}{=} \widetilde{m}(\tilde{s})\left(\frac{\rho_T \beta_M^2 \widetilde{m}(\tilde{s})}{1 + \rho_T \frac{n_2}{n_M}\widetilde{m}(\tilde{s})} + \boldsymbol{I}_{n_2}\right)^{-1}. \tag{6.29}$$

6.B.4 Eigenvector Alignment

The eigendecomposition of $\widetilde{\boldsymbol{Q}}^{(2)}(\tilde{s})$ is given by $\sum_{i=1}^{n}\frac{1}{\tilde{\lambda}_i - \tilde{s}}\boldsymbol{u}_i\boldsymbol{u}_i^\top$ where $(\tilde{\lambda}_i, \boldsymbol{u}_i)_{1\leqslant i\leqslant n}$ are the eigenvalue-eigenvector pairs of $\frac{n_T}{\sqrt{n_1 n_2 n_3}}\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)\top} - \frac{n_1 n_3}{\sqrt{n_1 n_2 n_3}}\boldsymbol{I}_{n_2}$. Hence, thanks to Cauchy's integral formula, we have

$$\langle \boldsymbol{y}, \hat{\boldsymbol{y}}\rangle^2 = -\frac{1}{2i\pi}\oint_{\widetilde{\gamma}} \boldsymbol{y}^\top \widetilde{\boldsymbol{Q}}^{(2)}(\tilde{s})\boldsymbol{y}\,\mathrm{d}\tilde{s}$$

where $\widetilde{\gamma}$ is a positively-oriented simple closed complex contour circling around the isolated eigenvalue only. The asymptotic value ζ of $\langle \boldsymbol{y}, \hat{\boldsymbol{y}}\rangle^2$ can then be computed with the deterministic equivalent defined in Equation (6.29),

$$\zeta = -\frac{1}{2i\pi}\oint_{\widetilde{\gamma}} \boldsymbol{y}^\top \bar{\boldsymbol{Q}}^{(2)}(\tilde{s})\boldsymbol{y}\,\mathrm{d}\tilde{s}.$$

Using residue calculus (Proposition 2.16), we shall compute,

$$\zeta = -\lim_{\tilde{s}\to\tilde{\xi}} \frac{\left(\tilde{s}-\tilde{\xi}\right)\widetilde{m}(\tilde{s})}{\frac{\rho_T\beta_M^2\widetilde{m}(\tilde{s})}{1+\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{s})}+1}.$$

The limit can be expressed using L'Hôpital's rule,

$$\begin{aligned}
\zeta = \frac{-1}{\frac{\mathrm{d}}{\mathrm{d}\tilde{s}}\left[\frac{\rho_T\beta_M^2}{1+\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{s})}+\frac{1}{\widetilde{m}(\tilde{s})}\right]_{\tilde{s}=\tilde{\xi}}} &= \left[\left(\frac{1}{\widetilde{m}^2(\tilde{\xi})}+\frac{\rho_T^2\beta_M^2\frac{n_2}{n_M}}{\left(1+\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi})\right)^2}\right)\widetilde{m}'(\tilde{\xi})\right]^{-1} \\
&= \left[\left(1+\frac{1}{\beta_M^2}\frac{n_2}{n_M}\right)\frac{\widetilde{m}'(\tilde{\xi})}{\widetilde{m}^2(\tilde{\xi})}\right]^{-1} \qquad \text{since } 1+\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi}) = -\rho_T\beta_M^2\widetilde{m}(\tilde{\xi}) \\
&= \left[\frac{\rho_T^2}{\beta_M^2}\left(\frac{n_2}{n_M}+\beta_M^2\right)^3\widetilde{m}'(\tilde{\xi})\right]^{-1} \qquad \text{since } \widetilde{m}(\tilde{\xi}) = \frac{-1}{\rho_T\left(\frac{n_2}{n_M}+\beta_M^2\right)}.
\end{aligned}$$

In order to find an expression for $\widetilde{m}'(\tilde{\xi})$, let us differentiate $\rho_T\frac{n_2}{n_M}\widetilde{m}^3(\tilde{s}) + (1+\tilde{s}\rho_T\frac{n_2}{n_M})\widetilde{m}^2(\tilde{s}) + (\tilde{s}+\rho_T\frac{n_2-n_1}{n_M})\widetilde{m}(\tilde{s}) + 1 = 0$,

$$\left(3\rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + 2\left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s}) + \tilde{s} + \rho_T\frac{n_2-n_1}{n_M}\right)\widetilde{m}'(\tilde{s}) + \rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + \widetilde{m}(\tilde{s}) = 0.$$

Since $\left(\rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + \left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s}) + \tilde{s} + \rho_T\frac{n_2-n_1}{n_M}\right)\widetilde{m}(\tilde{s}) + 1 = 0$, it simplifies into

$$\begin{aligned}
&\left(2\rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + \left(1+\tilde{s}\rho_T\frac{n_2}{n_M}\right)\widetilde{m}(\tilde{s}) - \frac{1}{\widetilde{m}(\tilde{s})}\right)\widetilde{m}'(\tilde{s}) + \rho_T\frac{n_2}{n_M}\widetilde{m}^2(\tilde{s}) + \widetilde{m}(\tilde{s}) = 0 \\
\text{and} \qquad &\left(2\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{s}) + 1 + \tilde{s}\rho_T\frac{n_2}{n_M} - \frac{1}{\widetilde{m}^2(\tilde{s})}\right)\widetilde{m}'(\tilde{s}) + \rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{s}) + 1 = 0.
\end{aligned}$$

Using successively $\widetilde{m}(\tilde{\xi}) = \frac{-1}{\rho_T\left(\frac{n_2}{n_M}+\beta_M^2\right)}$ and $\tilde{\xi} = \frac{\rho_T}{\beta_M^2}\left(\frac{n_1}{n_M}+\beta_M^2\right)\left(\frac{n_2}{n_M}+\beta_M^2\right) + \frac{1}{\rho_T\left(\frac{n_2}{n_M}+\beta_M^2\right)}$, we obtain

$$\begin{aligned}
\widetilde{m}'(\tilde{\xi}) &= \frac{-1-\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi})}{2\rho_T\frac{n_2}{n_M}\widetilde{m}(\tilde{\xi}) + 1 + \tilde{\xi}\rho_T\frac{n_2}{n_M} - \frac{1}{\widetilde{m}^2(\tilde{\xi})}} \\
&= \frac{-1+\frac{\rho_T\frac{n_2}{n_M}}{\rho_T\left(\frac{n_2}{n_M}+\beta_M^2\right)}}{-\frac{2\rho_T\frac{n_2}{n_M}}{\rho_T\left(\frac{n_2}{n_M}+\beta_M^2\right)} + 1 + \tilde{\xi}\rho_T\frac{n_2}{n_M} - \rho_T^2\left(\frac{n_2}{n_M}+\beta_M^2\right)^2} \\
&= \frac{-\rho_T\beta_M^2}{-\rho_T\frac{n_2}{n_M} + \rho_T\beta_M^2 + \tilde{\xi}\rho_T^2\frac{n_2}{n_M}\left(\frac{n_2}{n_M}+\beta_M^2\right) - \rho_T^3\left(\frac{n_2}{n_M}+\beta_M^2\right)^3} \\
\widetilde{m}'(\tilde{\xi}) &= \frac{-\rho_T\beta_M^2}{\rho_T\beta_M^2 + \frac{\rho_T^3}{\beta_M^2}\frac{n_2}{n_M}\left(\frac{n_1}{n_M}+\beta_M^2\right)\left(\frac{n_2}{n_M}+\beta_M^2\right)^2 - \rho_T^3\left(\frac{n_2}{n_M}+\beta_M^2\right)^3}.
\end{aligned}$$

Hence,

$$\zeta = \frac{-\rho_T\beta_M^2 - \frac{\rho_T^3}{\beta_M^2}\frac{n_2}{n_M}\left(\frac{n_1}{n_M}+\beta_M^2\right)\left(\frac{n_2}{n_M}+\beta_M^2\right)^2 + \rho_T^3\left(\frac{n_2}{n_M}+\beta_M^2\right)^3}{\rho_T^3\left(\frac{n_2}{n_M}+\beta_M^2\right)^3}$$

$$= 1 - \frac{\rho_T\beta_M^2}{\rho_T^3\left(\frac{n_2}{n_M}+\beta_M^2\right)^3} - \frac{1}{\beta_M^2}\frac{n_2}{n_M}\frac{\frac{n_1}{n_M}+\beta_M^2}{\frac{n_2}{n_M}+\beta_M^2}$$

$$\zeta = 1 - \frac{1}{\beta_M^2\left(\frac{n_2}{n_M}+\beta_M^2\right)}\left[\left(\frac{\beta_M^2}{\rho_T\left(\frac{n_2}{n_M}+\beta_M^2\right)}\right)^2 + \frac{n_2}{n_M}\left(\frac{n_1}{n_M}+\beta_M^2\right)\right].$$

6.C Proof of Theorem 6.5

For all $s \in \mathbb{C} \setminus \operatorname{Sp} \boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$, let $\boldsymbol{Q}^{(3)}(s) = \left(\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top} - s\boldsymbol{I}_{n_3}\right)^{-1}$.

6.C.1 Preliminary Results

Let us derive a few useful results for the upcoming analysis.

Proposition 6.19.

$$\mathbb{E}\left[\boldsymbol{W}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\right] = \frac{n_1 n_2}{\sqrt{n_T}}\mathbb{E}\left[\boldsymbol{Q}^{(3)}\right] - \frac{1}{\sqrt{n_T}}\mathbb{E}\left[(n_3+1)\boldsymbol{Q}^{(3)} + s\left(\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)}\right)\right], \tag{6.30}$$

$$\mathbb{E}\left[\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\right] = \beta_T\mathbb{E}\left[\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\right] - \frac{1}{n_T}\mathbb{E}\left[\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{T}^{(3)\top}\left(\boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)} + \boldsymbol{Q}^{(3)2}\right)\right], \tag{6.31}$$

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{Q}^{(3)}\right] = {}& \left(\frac{n_1 n_2}{n_M} + \beta_M^2 - \frac{4}{n_M}\right)\mathbb{E}\left[\boldsymbol{Q}^{(3)}\right] - \frac{4}{n_M}\left(s + \frac{n_3 - n_1 n_2}{n_T} + \frac{2}{n_T}\right)\mathbb{E}\left[\boldsymbol{Q}^{(3)2}\right] \\ & - \frac{4n_3}{n_M n_T}\mathbb{E}\left[\boldsymbol{Q}^{(3)}\frac{\operatorname{Tr}\boldsymbol{Q}^{(3)}}{n_3}\right] - \frac{8}{n_M n_T}s\mathbb{E}\left[\boldsymbol{Q}^{(3)3}\right] - \frac{8n_3}{n_M n_T}s\mathbb{E}\left[\boldsymbol{Q}^{(3)2}\frac{\operatorname{Tr}\boldsymbol{Q}^{(3)}}{n_3}\right] \\ & - \frac{2\beta_T^2}{n_M^2}\mathbb{E}\left[(n_3 - n_1 n_2 + 2)\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)} + \left[\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\boldsymbol{z}\right]\boldsymbol{Q}^{(3)}\right] \\ & - \frac{2\beta_T^2 n_3}{n_M^2}s\mathbb{E}\left[\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\frac{\operatorname{Tr}\boldsymbol{Q}^{(3)}}{n_3}\right] \\ & - \frac{2\beta_T^2}{n_M^2}s\mathbb{E}\left[\left[\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\boldsymbol{z}\right]\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)2}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\right].\end{aligned} \tag{6.32}$$

In order to prove theses results, we will need the following proposition.

Proposition 6.20.

$$\frac{\partial Q_{a,b}^{(3)}}{\partial W_{c,d}^{(3)}} = -\frac{1}{\sqrt{n_T}}\left(Q_{a,c}^{(3)}\left[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\right]_{d,b} + Q_{c,b}^{(3)}\left[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\right]_{d,a}\right) \tag{6.33}$$

$$\frac{\partial Q_{a,b}^{(3)}}{\partial N_{c,d}} = -\frac{\beta_T}{\sqrt{n_M}}\left(\left[\boldsymbol{Q}^{(3)}\boldsymbol{z}\right]_a\left[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\right]_{[c,d],b} + \left[\boldsymbol{Q}^{(3)}\boldsymbol{z}\right]_b\left[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\right]_{[c,d],a}\right) \tag{6.34}$$

Proof. Since $\partial \boldsymbol{Q}^{(3)} = -\boldsymbol{Q}^{(3)} \partial\big(\boldsymbol{T}^{(3)} \boldsymbol{T}^{(3)\top}\big) \boldsymbol{Q}^{(3)}$,

$$\begin{aligned}
\frac{\partial Q^{(3)}_{a,b}}{\partial W^{(3)}_{c,d}} &= -\left[\boldsymbol{Q}^{(3)} \frac{\partial \boldsymbol{T}^{(3)} \boldsymbol{T}^{(3)\top}}{\partial W^{(3)}_{c,d}} \boldsymbol{Q}^{(3)} \right]_{a,b} \\
&= -\sum_{e,f,g=1}^{n_3, n_1 n_2, n_3} Q^{(3)}_{a,e} \left(\frac{\partial T^{(3)}_{e,f}}{\partial W^{(3)}_{c,d}} T^{(3)}_{g,f} + T^{(3)}_{e,f} \frac{\partial T^{(3)}_{g,f}}{\partial W^{(3)}_{c,d}} \right) Q^{(3)}_{g,b} \\
&= -\frac{1}{\sqrt{n_T}} \sum_{e,f,g=1}^{n_3, n_1 n_2, n_3} Q^{(3)}_{a,e} \Big(\delta_{e,c} \delta_{f,d} T^{(3)}_{g,f} + T^{(3)}_{e,f} \delta_{g,c} \delta_{f,d} \Big) Q^{(3)}_{g,b} \\
\frac{\partial Q^{(3)}_{a,b}}{\partial W^{(3)}_{c,d}} &= -\frac{1}{\sqrt{n_T}} \Big(Q^{(3)}_{a,c} \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{d,b} + Q^{(3)}_{c,b} \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{d,a} \Big).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{\partial Q^{(3)}_{a,b}}{\partial N_{c,d}} &= -\sum_{e,f,g,h=1}^{n_3, n_1, n_2, n_3} Q^{(3)}_{a,e} \left(\frac{\partial \mathcal{T}_{f,g,e}}{\partial N_{c,d}} \mathcal{T}_{f,g,h} + \mathcal{T}_{f,g,e} \frac{\partial \mathcal{T}_{f,g,h}}{\partial N_{c,d}} \right) Q^{(3)}_{h,b} \\
&= -\frac{\beta_T}{\sqrt{n_M}} \sum_{e,h=1}^{n_3} Q^{(3)}_{a,e} \big(z_e \mathcal{T}_{c,d,h} + \mathcal{T}_{c,d,e} z_h \big) Q^{(3)}_{h,b} \\
\frac{\partial Q^{(3)}_{a,b}}{\partial N_{c,d}} &= -\frac{\beta_T}{\sqrt{n_M}} \Big(\big[\boldsymbol{Q}^{(3)} \boldsymbol{z} \big]_a \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{[c,d],b} + \big[\boldsymbol{Q}^{(3)} \boldsymbol{z} \big]_b \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{[c,d],a} \Big).
\end{aligned}$$

□

Proof of Equation (6.30) With Stein's lemma (Lemma 2.18) and Proposition 6.20, we have

$$\begin{aligned}
\mathbb{E}\big[\boldsymbol{W}^{(3)} \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{i,j} &= \sum_{k,l=1}^{n_1 n_2, n_3} \mathbb{E}\Big[W^{(3)}_{i,k} T^{(3)}_{l,k} Q^{(3)}_{l,j} \Big] \\
&= \sum_{k,l=1}^{n_1 n_2, n_3} \mathbb{E}\left[\frac{\partial T^{(3)}_{l,k}}{\partial W^{(3)}_{i,k}} Q^{(3)}_{l,j} + T^{(3)}_{l,k} \frac{\partial Q^{(3)}_{l,j}}{\partial W^{(3)}_{i,k}} \right] \\
&= \sum_{k,l=1}^{n_1 n_2, n_3} \mathbb{E}\left[\frac{\delta_{l,i} \delta_{k,k}}{\sqrt{n_T}} Q^{(3)}_{l,j} \right] \\
&\quad - \frac{1}{\sqrt{n_T}} \sum_{k,l=1}^{n_1 n_2, n_3} \mathbb{E}\Big[T^{(3)}_{l,k} \Big(Q^{(3)}_{l,i} \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{k,j} + Q^{(3)}_{i,j} \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{k,l} \Big) \Big] \\
&= \frac{n_1 n_2}{\sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(3)} \big]_{i,j} - \frac{1}{\sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(3)} \boldsymbol{T}^{(3)} \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} + \boldsymbol{Q}^{(3)} \operatorname{Tr} \boldsymbol{T}^{(3)} \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{i,j} \\
\mathbb{E}\big[\boldsymbol{W}^{(3)} \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{i,j} &= \frac{n_1 n_2}{\sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(3)} \big]_{i,j} - \frac{1}{\sqrt{n_T}} \mathbb{E}\big[(n_3 + 1) \boldsymbol{Q}^{(3)} + s\big(\boldsymbol{Q}^{(3)} \operatorname{Tr} \boldsymbol{Q}^{(3)} + \boldsymbol{Q}^{(3)2} \big) \big]_{i,j}
\end{aligned}$$

where the last equality comes from $\boldsymbol{T}^{(3)} \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} = \boldsymbol{I}_{n_3} + s \boldsymbol{Q}^{(3)}$.

Proof of Equation (6.31) This is again Stein's lemma (Lemma 2.18) and Proposition 6.20.

$$\mathbb{E}\big[\boldsymbol{z} \boldsymbol{m}^\top \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big]_{i,j} = \beta_T \mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{F}}^2 \boldsymbol{z} \boldsymbol{z}^\top \boldsymbol{Q}^{(3)} \big]_{i,j} + \frac{1}{\sqrt{n_T}} \sum_{k,l=1}^{n_1 n_2, n_3} \mathbb{E}\Big[z_i m_k W^{(3)}_{l,k} Q^{(3)}_{l,j} \Big]$$

$$
\begin{aligned}
&= \beta_T \mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{F}}^2 \boldsymbol{z}\boldsymbol{z}^\top \boldsymbol{Q}^{(3)}\big]_{i,j} + \frac{1}{\sqrt{n_T}} \sum_{k,l=1}^{n_1 n_2, n_3} \mathbb{E}\left[z_i m_k \frac{\partial Q^{(3)}_{l,j}}{\partial W^{(3)}_{l,k}} \right] \\
&= \beta_T \mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{F}}^2 \boldsymbol{z}\boldsymbol{z}^\top \boldsymbol{Q}^{(3)}\big]_{i,j} \\
&\qquad - \frac{1}{n_T} \sum_{k,l=1}^{n_1 n_2, n_3} \mathbb{E}\Big[z_i m_k \Big(Q^{(3)}_{l,l} \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{k,j} + Q^{(3)}_{l,j} \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{k,l} \Big) \Big] \\
\mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{i,j} &= \beta_T \mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{F}}^2 \boldsymbol{z}\boldsymbol{z}^\top \boldsymbol{Q}^{(3)}\big]_{i,j} - \frac{1}{n_T} \mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{T}^{(3)\top} \big(\boldsymbol{Q}^{(3)} \operatorname{Tr} \boldsymbol{Q}^{(3)} + \boldsymbol{Q}^{(3)2}\big)\big]_{i,j}.
\end{aligned}
$$

Proof of Equation (6.32)

$$
\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{F}}^2 \boldsymbol{Q}^{(3)}\big] &= \sum_{u,v=1}^{n_1,n_2} \mathbb{E}\left[\left(\beta_M x_u y_v + \frac{1}{\sqrt{n_M}} N_{u,v} \right)^2 \boldsymbol{Q}^{(3)} \right] \\
&= \beta_M^2 \mathbb{E}\big[\boldsymbol{Q}^{(3)}\big] + \frac{1}{\sqrt{n_M}} \sum_{u,v=1}^{n_1,n_2} \mathbb{E}\left[\left(\frac{N_{u,v}}{\sqrt{n_M}} + 2\beta_M x_u y_v \right) N_{u,v} \boldsymbol{Q}^{(3)} \right] \\
\mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{F}}^2 \boldsymbol{Q}^{(3)}\big] &= \left(\beta_M^2 + \frac{n_1 n_2}{n_M} \right) \mathbb{E}\big[\boldsymbol{Q}^{(3)}\big] + \frac{1}{\sqrt{n_M}} \sum_{u,v=1}^{n_1,n_2} \mathbb{E}\left[\left(\frac{N_{u,v}}{\sqrt{n_M}} + 2\beta_M x_u y_v \right) \frac{\partial \boldsymbol{Q}^{(3)}}{\partial N_{u,v}} \right]
\end{aligned}
$$

where we have used Stein's lemma (Lemma 2.18) to derive the last equality. Hence, using Proposition 6.20, it becomes

$$
\begin{aligned}
&\mathbb{E}\left[\left(\|\boldsymbol{M}\|_{\mathrm{F}}^2 - \left(\beta_M^2 + \frac{n_1 n_2}{n_M} \right) \right) \boldsymbol{Q}^{(3)} \right]_{i,j} \\
&= -\frac{\beta_T}{n_M} \sum_{u,v=1}^{n_1,n_2} \mathbb{E}\left[\left(\frac{N_{u,v}}{\sqrt{n_M}} + 2\beta_M x_u y_v \right) \left(\big[\boldsymbol{Q}^{(3)} \boldsymbol{z}\big]_i \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{[u,v],j} + \big[\boldsymbol{Q}^{(3)} \boldsymbol{z}\big]_j \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{[u,v],i} \right) \right] \\
&= -\frac{\beta_T}{n_M} \sum_{u,v=1}^{n_1,n_2} \mathbb{E}\left[\left(2m_{[u,v]} - \frac{N_{u,v}}{\sqrt{n_M}} \right) \left(\big[\boldsymbol{Q}^{(3)} \boldsymbol{z}\big]_i \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{[u,v],j} + \big[\boldsymbol{Q}^{(3)} \boldsymbol{z}\big]_j \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{[u,v],i} \right) \right] \\
&= -2\frac{\beta_T}{n_M} \mathbb{E}\Big[\boldsymbol{Q}^{(3)} \boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} + \big(\boldsymbol{Q}^{(3)} \boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} \big)^\top \Big]_{i,j} \\
&\qquad + \frac{\beta_T}{n_M \sqrt{n_M}} \sum_{u,v=1}^{n_1,n_2} \mathbb{E}\Big[N_{u,v} \Big(\big[\boldsymbol{Q}^{(3)} \boldsymbol{z}\big]_i \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{[u,v],j} + \big[\boldsymbol{Q}^{(3)} \boldsymbol{z}\big]_j \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{[u,v],i} \Big) \Big]
\end{aligned}
$$

and, since $\beta_T \boldsymbol{z}\boldsymbol{m}^\top = \boldsymbol{T}^{(3)} - \frac{1}{\sqrt{n_T}} \boldsymbol{W}^{(3)}$ and $\boldsymbol{T}^{(3)} \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} = \boldsymbol{I}_{n_3} + s\boldsymbol{Q}^{(3)}$, we have,

$$
\beta_T \boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)} = \boldsymbol{I}_{n_3} + s\boldsymbol{Q}^{(3)} - \frac{1}{\sqrt{n_T}} \boldsymbol{W}^{(3)} \boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}.
$$

Thus, our expression turns into

$$
\begin{aligned}
&\mathbb{E}\left[\left(\|\boldsymbol{M}\|_{\mathrm{F}}^2 - \left(\beta_M^2 + \frac{n_1 n_2}{n_M} \right) \right) \boldsymbol{Q}^{(3)} \right]_{i,j} \\
&= -\frac{4}{n_M} \mathbb{E}\big[\boldsymbol{Q}^{(3)} + s\boldsymbol{Q}^{(3)2}\big]_{i,j} + \frac{2}{n_M \sqrt{n_T}} \mathbb{E}\big[\boldsymbol{Q}^{(3)} \big(\boldsymbol{W}^{(3)} \boldsymbol{T}^{(3)\top} + \boldsymbol{T}^{(3)} \boldsymbol{W}^{(3)\top} \big) \boldsymbol{Q}^{(3)}\big]_{i,j} \\
&\qquad + \frac{\beta_T}{n_M \sqrt{n_M}} \sum_{u,v=1}^{n_1,n_2} \mathbb{E}\Big[N_{u,v} \Big(\big[\boldsymbol{Q}^{(3)} \boldsymbol{z}\big]_i \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{[u,v],j} + \big[\boldsymbol{Q}^{(3)} \boldsymbol{z}\big]_j \big[\boldsymbol{T}^{(3)\top} \boldsymbol{Q}^{(3)}\big]_{[u,v],i} \Big) \Big] \\
&= -\frac{4}{n_M} \mathbb{E}\big[\boldsymbol{Q}^{(3)} + s\boldsymbol{Q}^{(3)2}\big]_{i,j} + \frac{2}{n_M \sqrt{n_T}} \big[\boldsymbol{A}_1 + \boldsymbol{A}_1^\top\big]_{i,j} + \frac{\beta_T}{n_M \sqrt{n_M}} \big[\boldsymbol{A}_2 + \boldsymbol{A}_2^\top\big]_{i,j}
\end{aligned}
$$

with

$$\boldsymbol{A}_1 = \mathbb{E}\big[\boldsymbol{Q}^{(3)}\boldsymbol{W}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big], \qquad [\boldsymbol{A}_2]_{i,j} = \sum_{u,v=1}^{n_1,n_2} \mathbb{E}\Big[N_{u,v}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]_i \big[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{[u,v],j}\Big].$$

Let us develop $\boldsymbol{A}_1$ with Stein's lemma (Lemma 2.18) on $\boldsymbol{W}^{(3)}$.

$$\begin{aligned}
[\boldsymbol{A}_1]_{i,j} &= \sum_{a,b,c=1}^{n_3,n_1n_2,n_3} \mathbb{E}\left[\frac{\partial Q^{(3)}_{i,a}}{\partial W^{(3)}_{a,b}} T^{(3)}_{c,b} Q^{(3)}_{c,j} + Q^{(3)}_{i,a}\frac{\partial T^{(3)}_{c,b}}{\partial W^{(3)}_{a,b}} Q^{(3)}_{c,j} + Q^{(3)}_{i,a} T^{(3)}_{c,b}\frac{\partial Q^{(3)}_{c,j}}{\partial W^{(3)}_{a,b}}\right] \\
&= \frac{n_1n_2}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(3)2}\big]_{i,j} \\
&\quad - \frac{1}{\sqrt{n_T}} \sum_{a,b,c=1}^{n_3,n_1n_2,n_3} \mathbb{E}\Big[\Big(Q^{(3)}_{i,a}\big[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{b,a} + Q^{(3)}_{a,a}\big[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{b,i}\Big) T^{(3)}_{c,b} Q^{(3)}_{c,j}\Big] \\
&\quad - \frac{1}{\sqrt{n_T}} \sum_{a,b,c=1}^{n_3,n_1n_2,n_3} \mathbb{E}\Big[Q^{(3)}_{i,a} T^{(3)}_{c,b}\Big(Q^{(3)}_{c,a}\big[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{b,j} + Q^{(3)}_{a,j}\big[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{b,c}\Big)\Big] \\
&= \frac{n_1n_2}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(3)2}\big]_{i,j} - \frac{1}{\sqrt{n_T}}\mathbb{E}\big[\big(2\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)}\big)\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)} + \boldsymbol{Q}^{(3)2}\operatorname{Tr}\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{i,j} \\
[\boldsymbol{A}_1]_{i,j} &= \frac{n_1n_2}{\sqrt{n_T}}\mathbb{E}\big[\boldsymbol{Q}^{(3)2}\big]_{i,j} - \frac{1}{\sqrt{n_T}}\mathbb{E}\big[(n_3+2)\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)} + 2s\boldsymbol{Q}^{(3)}\big(\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)}\big)\big]_{i,j}
\end{aligned}$$

since $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)} = \boldsymbol{I}_{n_3} + s\boldsymbol{Q}^{(3)}$. Next, we develop $\boldsymbol{A}_2$ with Stein's lemma (Lemma 2.18) on $\boldsymbol{N}$.

$$\begin{aligned}
[\boldsymbol{A}_2]_{i,j} &= \sum_{u,v=1}^{n_1,n_2}\sum_{a,b=1}^{n_3,n_3} \mathbb{E}\left[\frac{\partial Q^{(3)}_{i,a}}{\partial N_{u,v}} z_a T^{(3)}_{b,[u,v]} Q^{(3)}_{b,j} + Q^{(3)}_{i,a} z_a \frac{\partial T^{(3)}_{b,[u,v]}}{\partial N_{u,v}} Q^{(3)}_{b,j} + Q^{(3)}_{i,a} z_a T^{(3)}_{b,[u,v]}\frac{\partial Q^{(3)}_{b,j}}{\partial N_{u,v}}\right] \\
&= \frac{\beta_T n_1n_2}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\big]_{i,j} \\
&\quad - \frac{\beta_T}{\sqrt{n_M}}\sum_{u,v=1}^{n_1,n_2}\sum_{a,b=1}^{n_3,n_3}\mathbb{E}\Big[\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]_i\big[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{[u,v],a} z_a T^{(3)}_{b,[u,v]} Q^{(3)}_{b,j}\Big] \\
&\quad - \frac{\beta_T}{\sqrt{n_M}}\sum_{u,v=1}^{n_1,n_2}\sum_{a,b=1}^{n_3,n_3}\mathbb{E}\Big[\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]_a\big[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{[u,v],i} z_a T^{(3)}_{b,[u,v]} Q^{(3)}_{b,j}\Big] \\
&\quad - \frac{\beta_T}{\sqrt{n_M}}\sum_{u,v=1}^{n_1,n_2}\sum_{a,b=1}^{n_3,n_3}\mathbb{E}\Big[Q^{(3)}_{i,a} z_a T^{(3)}_{b,[u,v]}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]_b\big[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{[u,v],j}\Big] \\
&\quad - \frac{\beta_T}{\sqrt{n_M}}\sum_{u,v=1}^{n_1,n_2}\sum_{a,b=1}^{n_3,n_3}\mathbb{E}\Big[Q^{(3)}_{i,a} z_a T^{(3)}_{b,[u,v]}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]_j\big[\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{[u,v],b}\Big] \\
&= \frac{\beta_T n_1n_2}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\big]_{i,j} - \frac{\beta_T}{\sqrt{n_M}}\mathbb{E}\big[\big(\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)} + \big[\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]\boldsymbol{Q}^{(3)}\big)\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big]_{i,j} \\
&\qquad - \frac{\beta_T}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\big(\boldsymbol{Q}^{(3)}\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)} + \boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big)\big]_{i,j} \\
[\boldsymbol{A}_2]_{i,j} &= \frac{\beta_T n_1n_2}{\sqrt{n_M}}\mathbb{E}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\big]_{i,j} - \frac{\beta_T}{\sqrt{n_M}}\mathbb{E}\big[(n_3+2)\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)} + \big[\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]\boldsymbol{Q}^{(3)}\big]_{i,j} \\
&\qquad - \frac{\beta_T}{\sqrt{n_M}}s\mathbb{E}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\big(\boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)} + 2\boldsymbol{Q}^{(3)2}\big) + \big[\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]\boldsymbol{Q}^{(3)2}\big]_{i,j}
\end{aligned}$$

since, again, $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)} = \boldsymbol{I}_{n_3} + s\boldsymbol{Q}^{(3)}$. Eventually, we have

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{Q}^{(3)}\big] = \left(\beta_M^2 + \frac{n_1n_2}{n_M}\right)\mathbb{E}\big[\boldsymbol{Q}^{(3)}\big] - \frac{4}{n_M}\mathbb{E}\big[\boldsymbol{Q}^{(3)} + s\boldsymbol{Q}^{(3)2}\big] \\
+ \frac{4n_1n_2}{n_Mn_T}\mathbb{E}\big[\boldsymbol{Q}^{(3)2}\big] - \frac{4}{n_Mn_T}\mathbb{E}\big[(n_3+2)\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)} + 2s\boldsymbol{Q}^{(3)}\big(\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)}\big)\big] \\
+ \frac{2\beta_T^2n_1n_2}{n_M^2}\mathbb{E}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\big] - \frac{2\beta_T^2}{n_M^2}\mathbb{E}\big[(n_3+2)\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)} + \big[\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]\boldsymbol{Q}^{(3)}\big] \\
- \frac{2\beta_T^2}{n_M^2}s\mathbb{E}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)} + \big[\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]\boldsymbol{Q}^{(3)2}\big] \\
- \frac{2\beta_T^2}{n_M^2}s\mathbb{E}\big[\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)2}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\big]
\end{aligned}$$

which gives Equation (6.32) after rearranging the terms.

6.C.2 Mean Behavior of the Resolvent

Since $\boldsymbol{Q}^{(3)-1}\boldsymbol{Q}^{(3)} = \boldsymbol{I}_{n_3}$,

$$\beta_T\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)} + \frac{1}{\sqrt{n_T}}\boldsymbol{W}^{(3)}\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)} - s\boldsymbol{Q}^{(3)} = \boldsymbol{I}_{n_3}$$

and, from Equation (6.30), we have

$$\beta_T\mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big] + \frac{n_1n_2}{n_T}\mathbb{E}\big[\boldsymbol{Q}^{(3)}\big] - \frac{1}{n_T}\mathbb{E}\big[(n_3+1)\boldsymbol{Q}^{(3)} + s\big(\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)}\operatorname{Tr}\boldsymbol{Q}^{(3)}\big)\big] - s\mathbb{E}\big[\boldsymbol{Q}^{(3)}\big] = \boldsymbol{I}_{n_3}.$$

Let us rearrange the terms

$$s\frac{n_3}{n_T}\mathbb{E}\left[\frac{\operatorname{Tr}\boldsymbol{Q}^{(3)}}{n_3}\boldsymbol{Q}^{(3)}\right] + \left(s + \frac{n_3 - n_1n_2}{n_T}\right)\mathbb{E}\big[\boldsymbol{Q}^{(3)}\big] + \boldsymbol{I}_{n_3} = \beta_T\mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{T}^{(3)\top}\boldsymbol{Q}^{(3)}\big] - \frac{1}{n_T}\mathbb{E}\big[\boldsymbol{Q}^{(3)} + s\boldsymbol{Q}^{(3)2}\big]$$

in order to see that we need to following rescaling to counteract the divergence of the spectrum of $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$,

$$\tilde{s} = \frac{n_Ts - n_1n_2}{\sqrt{n_1n_2n_3}}, \qquad \widetilde{\boldsymbol{Q}}^{(3)}(\tilde{s}) = \left(\frac{n_T\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top} - n_1n_2\boldsymbol{I}_{n_3}}{\sqrt{n_1n_2n_3}} - \tilde{s}\boldsymbol{I}_{n_3}\right)^{-1}.$$

Hence, our equation becomes

$$\begin{aligned}
\mathbb{E}\big[\widetilde{m}_{n_3}(\tilde{s})\widetilde{\boldsymbol{Q}}^{(3)}\big] + \tilde{s}\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)}\big] + \boldsymbol{I}_{n_3} - \frac{\beta_Tn_T}{\sqrt{n_1n_2n_3}}\mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{T}^{(3)\top}\widetilde{\boldsymbol{Q}}^{(3)}\big] \\
= -\frac{n_3+1}{\sqrt{n_1n_2n_3}}\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)}\big] - \left(\frac{1}{\sqrt{n_1n_2n_3}}\tilde{s} + \frac{1}{n_3}\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)2}\big] - \frac{n_3}{\sqrt{n_1n_2n_3}}\tilde{s}\mathbb{E}\big[\widetilde{m}_{n_3}(\tilde{s})\widetilde{\boldsymbol{Q}}^{(3)}\big] \qquad (6.35)
\end{aligned}$$

where we have introduced $\widetilde{m}_{n_3}(\tilde{s}) \overset{\text{def}}{=} \frac{1}{n_3}\operatorname{Tr}\boldsymbol{Q}^{(3)}(\tilde{s})$. From Equation (6.31), we have

$$\mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{T}^{(3)\top}\widetilde{\boldsymbol{Q}}^{(3)}\big] + \frac{n_3}{\sqrt{n_1n_2n_3}}\mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{T}^{(3)\top}\big(\widetilde{m}_{n_3}(\tilde{s})\widetilde{\boldsymbol{Q}}^{(3)} + \widetilde{\boldsymbol{Q}}^{(3)2}\big)\big] = \beta_T\mathbb{E}\big[\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{z}\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}^{(3)}\big]$$

and, with Equation (6.32),

$$
\begin{aligned}
&\mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{T}^{(3)\top}\widetilde{\boldsymbol{Q}}^{(3)}\big] + \frac{n_3}{\sqrt{n_1n_2n_3}}\mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{T}^{(3)\top}\big(\widetilde{m}_{n_3}(\tilde{s})\widetilde{\boldsymbol{Q}}^{(3)} + \widetilde{\boldsymbol{Q}}^{(3)2}\big)\big] \\
&= \beta_T \boldsymbol{z}\boldsymbol{z}^\top\Bigg[\left(\frac{n_1n_2}{n_M} + \beta_M^2 - \frac{4}{n_M}\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)}\big] + \frac{4}{n_M}\left(\tilde{s} + \frac{n_3+2}{\sqrt{n_1n_2n_3}}\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)2}\big] - \frac{4}{n_M}\frac{n_3}{\sqrt{n_1n_2n_3}}\mathbb{E}\big[\widetilde{m}_{n_3}(\tilde{s})\widetilde{\boldsymbol{Q}}^{(3)}\big] \\
&\quad - \frac{8}{n_M}\left(\frac{1}{\sqrt{n_1n_2n_3}}\tilde{s} + \frac{1}{n_3}\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)3}\big] - \frac{8}{n_M}\left(\frac{n_3}{\sqrt{n_1n_2n_3}}\tilde{s} + 1\right)\mathbb{E}\big[\widetilde{m}_{n_3}(\tilde{s})\widetilde{\boldsymbol{Q}}^{(3)2}\big] \\
&\quad - \frac{2\beta_T^2}{n_M^2}\frac{n_T}{\sqrt{n_1n_2n_3}}\mathbb{E}\big[(n_3 - n_1n_2 + 2)\widetilde{\boldsymbol{Q}}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}^{(3)} + \big[\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}^{(3)}\boldsymbol{z}\big]\widetilde{\boldsymbol{Q}}^{(3)}\big] \\
&\quad - \frac{2\beta_T^2 n_3}{n_M^2}\frac{n_T}{\sqrt{n_1n_2n_3}}\left(\tilde{s} + \frac{n_1n_2}{\sqrt{n_1n_2n_3}}\right)\mathbb{E}\big[\widetilde{m}_{n_3}(\tilde{s})\boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\big] \\
&\quad - \frac{2\beta_T^2}{n_M^2}\frac{n_T}{\sqrt{n_1n_2n_3}}\left(\tilde{s} + \frac{n_1n_2}{\sqrt{n_1n_2n_3}}\right)\mathbb{E}\big[\big[\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\boldsymbol{z}\big]\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)2} + \boldsymbol{Q}^{(3)2}\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{Q}^{(3)}\big]\Bigg].
\end{aligned}
$$

Denoting $\mathscr{O}_{\tilde{s}}^{\|\cdot\|}(v_n)$ a quantity whose spectral norm $\|\cdot\|$ is $\mathscr{O}_{\tilde{s}}(v_n)$, the previous equation can be summarized (after multiplication by $\frac{\beta_T n_T}{\sqrt{n_1n_2n_3}}$) as

$$
\begin{aligned}
&\frac{\beta_T n_T}{\sqrt{n_1n_2n_3}}\mathbb{E}\big[\boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{T}^{(3)\top}\widetilde{\boldsymbol{Q}}^{(3)}\big] + \mathscr{O}_{\tilde{s}}^{\|\cdot\|}(\beta_T) \\
&\qquad = \frac{\beta_T^2 n_T}{\sqrt{n_1n_2n_3}}\boldsymbol{z}\boldsymbol{z}^\top\left[\left(\frac{n_1n_2}{n_M} + \beta_M^2\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)}\big] + \mathscr{O}_{\tilde{s}}^{\|\cdot\|}(n_T^{-1} + \beta_T^2 n_T^{-1})\right] \qquad (6.36)
\end{aligned}
$$

since $\|\boldsymbol{m}\|, \|\boldsymbol{T}^{(3)}\| = \mathscr{O}(\sqrt{n_T})$ almost surely[6] and $\|\widetilde{\boldsymbol{Q}}^{(3)}\| \leqslant |\Im\tilde{s}|^{-1}$. In the non-trivial regime, the quantity $\frac{\beta_T n_T}{\sqrt{n_1n_2n_3}}\mathbb{E}[\boldsymbol{z}\boldsymbol{m}^\top \boldsymbol{T}^{(3)\top}\widetilde{\boldsymbol{Q}}^{(3)}]$ appearing in Equation (6.35) must neither vanish nor diverge ; this is possible if, and only if, $\frac{\beta_T^2 n_T}{\sqrt{n_1n_2n_3}}\left(\frac{n_1n_2}{n_M} + \beta_M^2\right) = \Theta(1)$, i.e., $\beta_T = \Theta(n_T^{-1/4})$, which is what we assume from now on and we define the $\Theta(1)$ quantity $\varrho \overset{\text{def}}{=} \frac{\beta_T^2 n_T}{\sqrt{n_1n_2n_3}}\left(\frac{n_1n_2}{n_M} + \beta_M^2\right)$.

Remark 6.21. In fact, with the assumption $\beta_M = \Theta(1)$, we have $\frac{\beta_T^2 n_T}{\sqrt{n_1n_2n_3}}\left(\frac{n_1n_2}{n_M} + \beta_M^2\right) = \beta_T^2\frac{n_T}{n_M}\sqrt{\frac{n_1n_2}{n_3}} + \Theta(n_T^{-1})$ so we could have kept only the dominant term. However, for finite horizon considerations (as it is the case in practice), adding the $\Theta(n_T^{-1})$ term leads to slightly more precise predictions. Indeed, with the dominant term only, we consider a "worst-case scenario" $\beta_M = 0$.

6.C.3 Concentration of Bilinear Forms and Traces

Before going further, we must, as usual, show the following almost sure convergences,

$$
\boldsymbol{a}^\top\big(\widetilde{\boldsymbol{Q}}^{(3)} - \mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)}\big]\big)\boldsymbol{b} \xrightarrow[n_1,n_2,n_3\to+\infty]{\text{a.s.}} 0 \qquad \text{and} \qquad \frac{1}{n_3}\operatorname{Tr}\boldsymbol{A}\big(\widetilde{\boldsymbol{Q}}^{(3)} - \mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)}\big]\big) \xrightarrow[n_1,n_2,n_3\to+\infty]{\text{a.s.}} 0
$$

for all bounded (sequences of) vectors $\boldsymbol{a},\boldsymbol{b} \in \mathbb{R}^{n_3}$ and matrices $\boldsymbol{A} \in \mathbb{R}^{n_3\times n_3}$.

With the Poincaré-Nash inequality (Lemma 2.19), we find

$$
\operatorname{Var}\big(\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}^{(3)}\boldsymbol{b}\big) \leqslant 4\left(\frac{\beta_T^2}{n_M} + \frac{1}{n_T}\right)\frac{n_T^2}{n_1n_2n_3}\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\Big[\big\|\boldsymbol{T}^{(3)}\big\|^2\big\|\widetilde{\boldsymbol{Q}}^{(3)}\big\|^4\Big] = \mathscr{O}_{\tilde{s}}(n_T^{-1}),
$$

[6] $\|\boldsymbol{m}\|^2 = \|\boldsymbol{M}\|_{\mathrm{F}}^2 \leqslant 2\beta_M^2 + \frac{2}{n_M}\|\boldsymbol{N}\|_{\mathrm{F}}^2 \leqslant 2\beta_M^2 + \frac{2}{n_M}\min(n_1,n_2)\|\boldsymbol{N}\|^2$ and we already know that $\frac{1}{n_M}\|\boldsymbol{N}\|^2 = \mathscr{O}(1)$ almost surely.

$$\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}^{(3)}-\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(3)}])\boldsymbol{b}\big|^4\Big]$$
$$\leqslant 16\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\frac{n_T^2}{n_1n_2n_3}\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2\mathbb{E}\Big[\big|\boldsymbol{a}^\top(\widetilde{\boldsymbol{Q}}^{(3)}-\mathbb{E}[\widetilde{\boldsymbol{Q}}^{(3)}])\boldsymbol{b}\big|^2\big\|\boldsymbol{T}^{(3)}\big\|^2\big\|\widetilde{\boldsymbol{Q}}^{(3)}\big\|^4\Big]+\mathrm{Var}\big(\boldsymbol{a}^\top\widetilde{\boldsymbol{Q}}^{(3)}\boldsymbol{b}\big)^2$$
$$=\mathscr{O}_{\bar{s}}(n_T^{-2}),$$

$$\text{and}\qquad \mathrm{Var}\left(\frac{1}{n_3}\mathrm{Tr}\,\boldsymbol{A}\widetilde{\boldsymbol{Q}}^{(3)}\right)\leqslant\frac{4}{n_3^2}\left(\frac{\beta_T^2}{n_M}+\frac{1}{n_T}\right)\frac{n_T^2}{n_1n_2n_3}\|\boldsymbol{A}\|^2\mathbb{E}\Big[\big\|\boldsymbol{T}^{(3)}\big\|^2\big\|\widetilde{\boldsymbol{Q}}^{(3)}\big\|^4\Big]=\mathscr{O}_{\bar{s}}(n_T^{-3}).$$

Hence, by Lemma 2.20, we have the desired convergences.

6.C.4 Limiting Spectral Distribution and Confinement of the Spectrum

Denote $\widetilde{\boldsymbol{Q}}_0^{(3)}$ the resolvent of $\frac{n_T}{\sqrt{n_1n_2n_3}}\Big[\frac{1}{n_T}\boldsymbol{W}^{(3)}\boldsymbol{W}^{(3)\top}-\frac{n_1n_2}{n_T}\boldsymbol{I}_{n_3}\Big]$, that is, the model with $\beta_T=0$ (no signal). With the resolvent identity (Proposition 2.21), we have

$$\frac{1}{n_3}\mathrm{Tr}\Big(\widetilde{\boldsymbol{Q}}_0^{(3)}-\widetilde{\boldsymbol{Q}}^{(3)}\Big)=\frac{1}{n_3}\frac{n_T}{\sqrt{n_1n_2n_3}}\mathrm{Tr}\left(\widetilde{\boldsymbol{Q}}_0^{(3)}\left[\beta_T^2\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{z}\boldsymbol{z}^\top+\frac{\beta_T}{\sqrt{n_T}}\big(\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{W}^{(3)\top}+\boldsymbol{W}^{(3)}\boldsymbol{m}\boldsymbol{z}^\top\big)\right]\widetilde{\boldsymbol{Q}}^{(3)}\right)$$
$$=\mathscr{O}_{\bar{s}}(n_T^{-1}).$$

Therefore, the limiting spectral distribution of $\frac{n_T}{\sqrt{n_1n_2n_3}}\Big[\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}-\frac{n_1n_2}{n_T}\boldsymbol{I}_{n_3}\Big]$ is the same as that of $\frac{n_T}{\sqrt{n_1n_2n_3}}\Big[\frac{1}{n_T}\boldsymbol{W}^{(3)}\boldsymbol{W}^{(3)\top}-\frac{n_1n_2}{n_T}\boldsymbol{I}_{n_3}\Big]$. This is good news because we can then invoke the results of Chapter 5 (in particular Corollary 5.2 and Remark 5.4) to state

1. the limiting spectral distribution of $\frac{n_T}{\sqrt{n_1n_2n_3}}\Big[\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}-\frac{n_1n_2}{n_T}\boldsymbol{I}_{n_3}\Big]$ is the standard semicircle law μ_{SC},
2. for all $\varepsilon>0$, $\max_{\lambda\in\mathrm{Sp}\,\frac{1}{n_T}\boldsymbol{W}^{(3)}\boldsymbol{W}^{(3)\top}}\mathrm{Dist}\Big(\frac{\sqrt{n_1n_2n_3}\lambda-n_1n_2}{n_T},\mathrm{Supp}\,\mu_{\mathrm{SC}}\Big)<\varepsilon$ for n_1,n_2,n_3 large enough.

6.D Proof of Theorem 6.7

6.D.1 Isolated Eigenvalue

We seek the asymptotic position $\tilde{\xi}$ of an eigenvalue of $\frac{n_T}{\sqrt{n_1n_2n_3}}\Big[\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}-\frac{n_1n_2}{n_T}\boldsymbol{I}_{n_3}\Big]$ which is not in the limiting spectrum of $\frac{n_T}{\sqrt{n_1n_2n_3}}\Big[\frac{1}{n_T}\boldsymbol{W}^{(3)}\boldsymbol{W}^{(3)\top}-\frac{n_1n_2}{n_T}\boldsymbol{I}_{n_3}\Big]$. It must verify

$$\det\left(\frac{n_T}{\sqrt{n_1n_2n_3}}\left[\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}-\frac{n_1n_2}{n_T}\boldsymbol{I}_{n_3}\right]-\tilde{\xi}\boldsymbol{I}_{n_3}\right)=0.$$

Expanding $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)\top}$ with the decomposition $\boldsymbol{T}^{(3)}=\beta_T\boldsymbol{z}\boldsymbol{m}^\top+\frac{1}{\sqrt{n_T}}\boldsymbol{W}^{(3)}$, we can factorize this determinant by $\det\Big(\frac{n_T}{\sqrt{n_1n_2n_3}}\Big[\frac{1}{n_T}\boldsymbol{W}^{(3)}\boldsymbol{W}^{(3)\top}-\frac{n_1n_2}{n_T}\boldsymbol{I}_{n_3}\Big]-\tilde{\xi}\boldsymbol{I}_{n_3}\Big)$, which is not zero for n_1,n_2,n_3 large enough by the confinement of the spectrum. Therefore,

$$\det\left(\boldsymbol{I}_{n_3}+\frac{n_T}{\sqrt{n_1n_2n_3}}\left[\beta_T^2\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{z}\boldsymbol{z}^\top+\frac{\beta_T}{\sqrt{n_T}}\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{W}^{(3)\top}+\frac{\beta_T}{\sqrt{n_T}}\boldsymbol{W}^{(3)}\boldsymbol{m}\boldsymbol{z}^\top\right]\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\right)=0$$

where we recall that $\widetilde{\boldsymbol{Q}}_0^{(3)}$ denotes the resolvent of $\frac{n_T}{\sqrt{n_1n_2n_3}}\Big[\frac{1}{n_T}\boldsymbol{W}^{(3)}\boldsymbol{W}^{(3)\top}-\frac{n_1n_2}{n_T}\boldsymbol{I}_{n_3}\Big]$. Notice that the sum $\beta_T^2\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{z}\boldsymbol{z}^\top+\frac{\beta_T}{\sqrt{n_T}}\boldsymbol{z}\boldsymbol{m}^\top\boldsymbol{W}^{(3)\top}+\frac{\beta_T}{\sqrt{n_T}}\boldsymbol{W}^{(3)}\boldsymbol{m}\boldsymbol{z}^\top$ can be written as the following matrix product,

$\left[\beta_T\|\boldsymbol{M}\|_{\mathrm{F}}\boldsymbol{z}\ \beta_T\|\boldsymbol{M}\|_{\mathrm{F}}\boldsymbol{z}\ \frac{1}{\|\boldsymbol{M}\|_{\mathrm{F}}\sqrt{n_T}}\boldsymbol{W}^{(3)}\boldsymbol{m}\right]\begin{bmatrix}\beta_T\|\boldsymbol{M}\|_{\mathrm{F}}\boldsymbol{z}^\top\\ \frac{1}{\|\boldsymbol{M}\|_{\mathrm{F}}\sqrt{n_T}}\boldsymbol{m}^\top\boldsymbol{W}^{(3)\top}\\ \beta_T\|\boldsymbol{M}\|_{\mathrm{F}}\boldsymbol{z}^\top\end{bmatrix}$. Hence, with Sylvester's identity (Proposition 2.22), we have

$$\det\left(\boldsymbol{I}_3+\frac{n_T}{\sqrt{n_1n_2n_3}}\begin{bmatrix}\beta_T^2\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{z} & \beta_T^2\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{z} & \frac{\beta_T}{\sqrt{n_T}}\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{W}^{(3)}\boldsymbol{m}\\ \frac{\beta_T}{\sqrt{n_T}}\boldsymbol{m}^\top\boldsymbol{W}^{(3)\top}\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{z} & \frac{\beta_T}{\sqrt{n_T}}\boldsymbol{m}^\top\boldsymbol{W}^{(3)\top}\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{z} & \frac{1}{n_T\|\boldsymbol{M}\|_{\mathrm{F}}^2}\boldsymbol{m}^\top\boldsymbol{W}^{(3)\top}\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{W}^{(3)}\boldsymbol{m}\\ \beta_T^2\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{z} & \beta_T^2\|\boldsymbol{M}\|_{\mathrm{F}}^2\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{z} & \frac{\beta_T}{\sqrt{n_T}}\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{W}^{(3)}\boldsymbol{m}\end{bmatrix}\right)=0.$$

From the convergence of bilinear forms proven in Section 5.B.1, $\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}_0^{(3)}(\tilde{\xi})\boldsymbol{z}\to m_{\mathrm{SC}}(\tilde{\xi})$ almost surely. Moreover, Lemma 6.22 below justifies that entries (1,1), (1,2), (3,1) and (3,2) converge almost surely to $\varrho m_{\mathrm{SC}}(\tilde{\xi})$ (recall that $\beta_T=\Theta(n_T^{-1/4})$ in the non-trivial regime).

Lemma 6.22. *$\frac{\beta_T^2 n_T}{\sqrt{n_1n_2n_3}}\|\boldsymbol{M}\|_{\mathrm{F}}^2\to\varrho$ almost surely as $n_1,n_2,n_3\to+\infty$.*

Proof. We have $\|\boldsymbol{M}\|_{\mathrm{F}}^2=\beta_M^2+\sum_{i=1}^{n_1}\sum_{j=2}^{n_2}\left(2\frac{\beta_M}{\sqrt{n_M}}x_iy_jN_{i,j}+\frac{1}{n_M}N_{i,j}^2\right)$ where $N_{i,j}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$ therefore

$$\mathbb{E}\left[\frac{\beta_T^2n_T}{\sqrt{n_1n_2n_3}}\|\boldsymbol{M}\|_{\mathrm{F}}^2\right]=\frac{\beta_T^2n_T}{\sqrt{n_1n_2n_3}}\left(\beta_M^2+\frac{n_1n_2}{n_M}\right)=\varrho,$$
$$\mathrm{Var}\left(\frac{\beta_T^2n_T}{\sqrt{n_1n_2n_3}}\|\boldsymbol{M}\|_{\mathrm{F}}^2\right)=\left(\frac{\beta_T^2n_T}{\sqrt{n_1n_2n_3}}\right)^2\left[4\frac{\beta_M^2}{n_M}+2\frac{n_1n_2}{n_M^2}\right]=\mathcal{O}(n_T^{-2}).$$

Thus, Lemma 2.20 gives the stated result. □

Then, the next lemma shows that entries (1,3), (2,1), (2,2) and (3,3) are $\mathcal{O}(n_T^{-1/4})$ while entry (2,3) is $\mathcal{O}(n_T^{-1/2})$ almost surely, i.e., they all vanish asymptotically.

Lemma 6.23. *$\|\boldsymbol{W}^{(3)}\boldsymbol{m}\|=\mathcal{O}(n_T)$ almost surely as $n_1,n_2,n_3\to+\infty$.*

Proof. $\|\boldsymbol{W}^{(3)}\boldsymbol{m}\|^2=\boldsymbol{m}^\top\boldsymbol{W}^{(3)\top}\boldsymbol{W}^{(3)}\boldsymbol{m}=\boldsymbol{m}^\top\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{m}$ where $\boldsymbol{D}$ is an $n_3\times n_3$ diagonal matrix of (non-zero) eigenvalues of $\boldsymbol{W}^{(3)\top}\boldsymbol{W}^{(3)}$ (we assume n_T sufficiently large so that $n_1n_2>n_3$) and $\boldsymbol{V}$ is uniformly distributed on the Stiefel manifold $V_{n_3}(\mathbb{R}^{n_1n_2})$ (Chikuse, 2003, Theorem 2.2.1). Hence, $\frac{\|\boldsymbol{W}^{(3)}\boldsymbol{m}\|^2}{n_T\|\boldsymbol{m}\|^2}\leqslant\frac{n_1n_2+2\sqrt{n_1n_2n_3}}{n_T}\frac{\boldsymbol{m}^\top}{\|\boldsymbol{m}\|}\boldsymbol{V}\boldsymbol{V}^\top\frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}$. Although $\frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}$ is random, it is independent of $\boldsymbol{W}^{(3)}$ and therefore $\boldsymbol{V}$. Thus, given $\frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}$, we can choose $\boldsymbol{O}\in\mathcal{O}_{n_1n_2}(\mathbb{R})$ such that $\boldsymbol{O}\frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}=\boldsymbol{e}_1^{(n_1n_2)}$. The orthogonal matrix $\boldsymbol{O}$ is random but independent of $\boldsymbol{V}$ as well so $\frac{\boldsymbol{m}^\top}{\|\boldsymbol{m}\|}\boldsymbol{V}\boldsymbol{V}^\top\frac{\boldsymbol{m}}{\|\boldsymbol{m}\|}=\boldsymbol{e}_1^{(n_1n_2)\top}\boldsymbol{O}\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{O}^\top\boldsymbol{e}_1^{(n_1n_2)}$ is identically distributed to $\boldsymbol{e}_1^{(n_1n_2)\top}\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{e}_1^{(n_1n_2)}$ (Chikuse, 2003, Theorem 2.2.1). According to Mardia and Khatri (1977), $[\boldsymbol{V}\boldsymbol{V}^\top]_{1,1}$ follows a beta distribution with parameters $\frac{n_3}{2},\frac{n_1n_2-n_3}{2}$ so its moments are given by $\mathbb{E}[[\boldsymbol{V}\boldsymbol{V}^\top]_{1,1}^k]=\prod_{r=0}^{k-1}\frac{n_3+2r}{n_1n_2+2r}$ for all $k\geqslant1$. This is enough to see that $\frac{n_1n_2+2\sqrt{n_1n_2n_3}}{n_T}\mathbb{E}[[\boldsymbol{V}\boldsymbol{V}^\top]_{1,1}]=\frac{n_1n_2+2\sqrt{n_1n_2n_3}}{n_T}\frac{n_3}{n_1n_2}=\mathcal{O}(1)$ and $\left(\frac{n_1n_2+2\sqrt{n_1n_2n_3}}{n_T}\right)^4\mathbb{E}[|[\boldsymbol{V}\boldsymbol{V}^\top]_{1,1}-\mathbb{E}[[\boldsymbol{V}\boldsymbol{V}^\top]_{1,1}]|^4]=\mathcal{O}(n_T^{-2})$. Therefore, $\frac{\|\boldsymbol{W}^{(3)}\boldsymbol{m}\|^2}{n_T\|\boldsymbol{m}\|^2}$ is upper bounded by a quantity which is $\mathcal{O}(1)$ almost surely. Since $\|\boldsymbol{m}\|^2=\|\boldsymbol{M}\|_{\mathrm{F}}^2=\mathcal{O}(n_T)$ almost surely, this shows that $\frac{1}{n_T^2}\|\boldsymbol{W}^{(3)}\boldsymbol{m}\|^2=\mathcal{O}(1)$ almost surely. □

Eventually, since the determinant is a continuous function in the entries of the matrix, we have asymptotically

$$\det\begin{bmatrix}1+\varrho m_{\mathrm{SC}}(\tilde{\xi}) & \varrho m_{\mathrm{SC}}(\tilde{\xi}) & 0\\ 0 & 1 & 0\\ \varrho m_{\mathrm{SC}}(\tilde{\xi}) & \varrho m_{\mathrm{SC}}(\tilde{\xi}) & 1\end{bmatrix}=0$$

which simply gives $1+\varrho m_{\mathrm{SC}}(\tilde{\xi})=0$ or $m_{\mathrm{SC}}(\tilde{\xi})=\frac{-1}{\varrho}$. Hence, from the relation $m_{\mathrm{SC}}^2(\tilde{\xi})+\tilde{\xi}m_{\mathrm{SC}}(\tilde{\xi})+1=0$, we find

$$\tilde{\xi}=\varrho+\frac{1}{\varrho}.$$

6.D.2 Deterministic Equivalent

In order to evaluate the alignment $\langle \boldsymbol{z},\hat{\boldsymbol{z}}\rangle$ below, we compute a deterministic equivalent (Definition 2.17) of the resolvent $\widetilde{\boldsymbol{Q}}^{(3)}$. From Equations (6.35) and (6.36), we have

$$\left\|\mathbb{E}\big[\widetilde{m}_{n_3}(\tilde{s})\widetilde{\boldsymbol{Q}}^{(3)}\big]+\tilde{s}\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)}\big]+\boldsymbol{I}_{n_3}-\varrho\boldsymbol{z}\boldsymbol{z}^\top\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)}\big]\right\|\xrightarrow[n_1,n_2,n_3\to+\infty]{}0.$$

Moreover, $\widetilde{m}_{n_3}(\tilde{s})\to m_{\mathrm{SC}}(\tilde{s})$ almost surely and $m_{\mathrm{SC}}(\tilde{s})+\tilde{s}=\frac{-1}{\widetilde{m}(\tilde{s})}$ therefore

$$\left\|\boldsymbol{I}_{n_3}-\left(\varrho\boldsymbol{z}\boldsymbol{z}^\top+\frac{1}{m_{\mathrm{SC}}(\tilde{s})}\boldsymbol{I}_{n_3}\right)\mathbb{E}\big[\widetilde{\boldsymbol{Q}}^{(3)}\big]\right\|\xrightarrow[n_1,n_2,n_3\to+\infty]{}0$$

and we can define the following deterministic equivalent

$$\bar{\boldsymbol{Q}}^{(3)}(\tilde{s})=\left(\varrho\boldsymbol{z}\boldsymbol{z}^\top+\frac{1}{m_{\mathrm{SC}}(\tilde{s})}\boldsymbol{I}_{n_3}\right)^{-1}.$$

6.D.3 Eigenvector Alignment

Following Cauchy's integral formula,

$$\langle \boldsymbol{z},\hat{\boldsymbol{z}}\rangle^2=-\frac{1}{2\mathrm{i}\pi}\oint_{\tilde{\gamma}}\boldsymbol{z}^\top\widetilde{\boldsymbol{Q}}^{(3)}(\tilde{s})\boldsymbol{z}\,\mathrm{d}\tilde{s}$$

where $\tilde{\gamma}$ is a positively-oriented simple closed complex contour circling around the isolated eigenvalue only. Hence, using the previously found deterministic equivalent $\bar{\boldsymbol{Q}}^{(3)}$, we can compute the asymptotic value ζ of $\langle \boldsymbol{z},\hat{\boldsymbol{z}}\rangle^2$,

$$\zeta=-\frac{1}{2\mathrm{i}\pi}\oint_{\tilde{\gamma}}\boldsymbol{z}^\top\bar{\boldsymbol{Q}}^{(3)}(\tilde{s})\boldsymbol{z}\,\mathrm{d}\tilde{s}.$$

This reduces to residue calculus (Proposition 2.16),

$$\zeta=-\lim_{\tilde{s}\to\tilde{\xi}}\big(\tilde{s}-\tilde{\xi}\big)\left[\varrho+\frac{1}{m_{\mathrm{SC}}(\tilde{s})}\right]^{-1}=\frac{-1}{\frac{\mathrm{d}}{\mathrm{d}\tilde{s}}\left[\varrho+\frac{1}{m_{\mathrm{SC}}(\tilde{s})}\right]_{\tilde{s}=\tilde{\xi}}}=\frac{m_{\mathrm{SC}}^2(\tilde{\xi})}{m'_{\mathrm{SC}}(\tilde{\xi})}.$$

Recall that $m_{\mathrm{SC}}(\tilde{\xi})=\frac{-1}{\varrho}$. Differentiating $m_{\mathrm{SC}}^2(\tilde{s})+\tilde{s}m_{\mathrm{SC}}(\tilde{s})+1=0$, we have,

$$\begin{aligned}
2m'_{\mathrm{SC}}(\tilde{\xi})m_{\mathrm{SC}}(\tilde{\xi})+m_{\mathrm{SC}}(\tilde{\xi})+\tilde{\xi}m'_{\mathrm{SC}}(\tilde{\xi})=0 &\iff m'_{\mathrm{SC}}(\tilde{\xi})=-\frac{m_{\mathrm{SC}}(\tilde{\xi})}{2m_{\mathrm{SC}}+\tilde{\xi}}\\
&\iff m'_{\mathrm{SC}}(\tilde{\xi})=\frac{1}{\varrho\tilde{\xi}-2}\\
&\iff m'_{\mathrm{SC}}(\tilde{\xi})=\frac{1}{\varrho^2-1}.
\end{aligned}$$

Hence, we conclude,

$$\zeta=1-\frac{1}{\varrho^2}.$$

Chapter 7

Conclusion and Perspectives

STATISTICAL learning has experienced tremendous development over the past decades and is now more commonly referred to as *machine* learning to emphasize the central role of computers in this field of study. Relying on the computing and storage power granted by the spectacular technological advances of the early 21st century, enormous efforts are put into digging up a new invisible and intangible gold: *information*. Yet, with great power comes great responsibility. Besides the human and environmental impact linked with the production of new technologies, the massive development of artificial intelligence transforms the structure of our societies worldwide and comes with a multitude of ethical and technical questions. The research community is at the forefront in proposing solutions to overcome this turmoil.

This thesis is concerned with the technical aspect of the picture, as it provides new insights into a resource-efficient processing of large-dimensional data. In order to avoid using unnecessary amounts of resources in learning tasks, a thorough understanding of the statistical mechanisms at stake is required. Overall, we promote a paradigm shift towards a prudent use of computing resources based on solid theoretical results yielding the optimal setting for a given desired level of performance.

Initially motivated by the growing dimension of the data (p) relatively to the number of samples (n), such that p is now of the same order as n, the use of random matrix theory proves to be useful not only for large-dimensional but also multidimensional data. However, the increased level of complexity captured by the processing of large and multidimensional arrays (tensors) instead of matrices comes with a cost: a computational-to-statistical gap. Henceforth, the existence of an informative solution to a statistical estimation problem no longer implies that such a solution can actually be computed, not because of a lack of efficient methods suggesting that further research is needed, but because of a computational barrier due to the complexity of the problem. Thus, we distinguish three *phases*: impossible, hard and easy. If the statistical barrier coincides with the computational one, then the hard phase does not exist and an informative solution can be computed with a polynomial-time algorithm as soon as it exists. Otherwise, in the hard phase, computing the solution is not practically doable as it would require an exponential amount of resources.

Each matrix or tensor model considered in this thesis falls under the denomination of a *signal-plus-noise* model: the data is assumed to be the addition of a signal — which we wish to reconstruct — and a noise. A natural sparsity assumption is that the information sought decomposes into a few simple algebraic terms. Hence, in the large-dimensional regime, while the energy of the noise is diluted in every direction, the energy of the signal *spikes* in a finite number of directions and can therefore be reconstructed via a spectral decomposition of the observation. This simple yet powerful idea, coupled with the well-established tools of random matrix theory, allows us to characterize the achievable reconstruction performance depending on the strength of the signal.

Resource-Efficient Learning

Among the methods limiting the use of memory resources in machine learning, *online* learning consists in extracting information from a continuous data stream. It is particularly useful when trying to learn from a large dataset under low-memory constraints: instead of loading every data point, they can be viewed one by one and the learning task is performed on the thus generated data stream. In Chapter 4, we have rigorously shown that, with an astute usage of the available memory, an online version of kernel spectral clustering outperforms the usual batch clustering. It relies on a banded version of the kernel matrix $\boldsymbol{K}_L = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X} \odot \boldsymbol{T}$ which can be computed in an online fashion. Our random matrix results describe under which setting clustering is possible (i.e., for which values of signal-to-noise ratio and available memory, see Figure 4.6) by exhibiting the phase transition between impossible and possible recovery. Moreover, they provide theoretical performance guarantees which allow to decide the best trade-off between performance and memory usage depending on the application.

Nevertheless, this problem still presents many challenges. Although the two-class setting is now well understood, solutions for the multi-class setting remain too heuristic and still lack a precise characterization because of the intricate behavior of the eigenvectors used for the estimation. In order to counteract this phenomenon, a clever choice of weights applied to the entries of the banded kernel matrix could make the unmixing easier to perform. Moreover, optimality results on standard spectral clustering (Onatski et al., 2013; Löffler et al., 2021) suggest that the proposed approach is optimal (in the sense that it achieves non-trivial estimation as soon as theoretically possible). Proving this optimality (e.g., with similar tools as Nguyen and Couillet, 2023) would further strengthen our results.

In the broader context of resource-efficient learning, another pertinent approach, which is not limited to clustering, is the computation on the fly of a summary of the dataset on which the learning task in then performed. This method, known as *sketching* (Gribonval et al., 2021) is useful for learning on data streams but is also relevant in contexts of federated learning and privacy-preserving learning. It relies on the simple idea that, instead of the data matrix $\boldsymbol{X} \in \mathbb{R}^{p\times n}$, the learning task can be performed on a "sketch" of the data $\boldsymbol{z} = \frac{1}{n}\sum_{i=1}^n \phi(\boldsymbol{x}_i) \in \mathbb{C}^m$ with a well-chosen transformation $\phi : \mathbb{R}^p \to \mathbb{C}^m$. In particular, a common choice is $\phi(\boldsymbol{x}) = f(\boldsymbol{W}\boldsymbol{x})$ where $\boldsymbol{W} \in \mathbb{R}^{m\times p}$ is a matrix of random frequencies (e.g., with i.i.d. $\mathcal{N}(0,\sigma^2)$ entries) and f is applied pointwise. Then, the sketch can be written as $\boldsymbol{z} = \frac{1}{n} f(\boldsymbol{W}\boldsymbol{X})\mathbf{1}_n$, which shows that the information in the $m \times n$ matrix $f(\boldsymbol{W}\boldsymbol{X})$ is projected on the direction $\mathbf{1}_n$. Thus, the random matrix analysis of $f(\boldsymbol{W}\boldsymbol{X})$ can disclose the performance of this method and suggest the best trade-off between the size of the sketch (m) and the desired performance. A similar model is studied in a supervised learning framework (single-layer neural networks) by Louart et al. (2018) who combine random matrix theory and concentration of measure tools (Talagrand, 1995; Ledoux, 2001; Louart and Couillet, 2021). Moreover, if $\boldsymbol{X} = \boldsymbol{P} + \boldsymbol{N}$ is a spiked model, then $f(\boldsymbol{W}\boldsymbol{X}) = f(\boldsymbol{W}\boldsymbol{P} + \boldsymbol{W}\boldsymbol{N})$ is a *non-linear* spiked model and its random matrix analysis can be performed via a signal-plus-noise decomposition (Moniri and Hassani, 2024).

In essence, resource-efficient learning is about using wisely the bare minimum of resources to achieve a task. Be it with spectral clustering, sketching or any other method, random matrix theory remains an essential tool to disclose the relevant parameters and provide performance guarantees, thereby allowing an informed use of the available resources.

Algorithmic Tensor Estimation

Because of their multidimensional structure, tensors are particularly memory-demanding, but they can also capture richer information. Efficiently extracting the latter from a signal-plus-noise tensor model is therefore a crucial task. In practice, the low-rank tensor approximation problem is hindered by a computational barrier: there exists a *hard phase* delimited from the *easy phase* by a computational threshold diverging as $N^{\frac{d-2}{4}}$ (in terms of signal-to-noise ratio). Hence, it is necessary to study the re-

construction performance of standard low-rank tensor approximation algorithms in order to identify where they stand relatively to this threshold. In particular, the celebrated AMP algorithm achieving outstanding performances on a number of statistical learning tasks (see Feng et al., 2022 and references therein) does not succeed at low-rank tensor approximation unless the signal-to-noise ratio is greater than a $\Theta(N^{\frac{d-1}{2}})$ threshold (Lesieur et al., 2017).

By studying the properties of "long" random matrices — with a number of columns growing polynomially (N^{d-1}) in the number of rows (N) — stemming from the unfoldings of a tensor following the general spiked tensor model $\mathcal{T} = \mathcal{P} + \frac{1}{\sqrt{N}}\mathcal{N}$, we have characterized in Chapter 5 the reconstruction performance of the truncated multilinear singular value decomposition (MLSVD) as a low-multilinear-rank approximation technique. This study reveals the existence of a phase transition in each principal direction of each mode of the signal. That is, a singular value $s_{q_\ell}(\boldsymbol{P}^{(\ell)})$ of the unfolding of $\mathcal{P}$ along the ℓ-th mode must be greater than a quantity $\sqrt{\sigma_N} = \Theta(N^{\frac{d-2}{4}})$ for the corresponding singular vector to be detected in the MLSVD of $\mathcal{T}$. However, it is known that the truncated MLSVD does *not* yield the best low-multilinear-rank approximation for tensors of order $d \geqslant 3$ (the Eckart-Young-Mirsky theorem does not generalize to tensors of order greater than 2) but it can serve as a good initialization to the higher-order orthogonal iteration (HOOI) which computes a solution to the best low-multilinear-rank approximation problem. Our Theorem 5.11 shows that, as soon as the initialization of this algorithm recovers a sufficient amount of energy from the sought signal (see the formal statement in Theorem 5.11), then its convergence is guaranteed *in a single iteration* as $N \to +\infty$. With a truncated MLSVD initialization, this condition is satisfied above a threshold behaving like $\sqrt{\sigma_N}$. In other words, we have shown two major points.

1. The truncated MLSVD — a very cheap but suboptimal low-multilinear-rank approximation technique — detects the principal directions of a signal as soon as the corresponding singular values (their strength) is above a $\Theta(N^{\frac{d-2}{4}})$ threshold.

2. The convergence of HOOI depends solely on the quality of its initialization.

Hence, without prior information, the truncated MLSVD is the best possible initialization and allows the convergence of HOOI near the computational threshold. Our analysis therefore provides insight into the computational barrier: the exact characterization of the reconstruction performance of the truncated MLSVD describes the minimal amount of signal needed to initialize the HOOI in the right bassin of attraction.

This "unfolding approach" also proves useful in Chapter 6 where it is applied to an order-3 tensor following a nested matrix-tensor model representing a multi-view clustering setting. The model can be seen as an extension of the rank-one spiked tensor model with a matrix noise and a tensor noise. This allows to perform a theoretical analysis of the performance of multi-view clustering depending on a matrix signal-to-noise ratio β_M representing the "inner difficulty" of the problem and a tensor signal-to-noise ratio β_T representing the difficulty due to the fact that we can only see the problem through multiple views. Our results lead to a phase diagram (Figure 6.2) showing under which settings the reconstruction of the rank-one signal from the multiple views is possible. In particular, because of the computational barrier, β_T must scale as $n_T^{1/4}$ (n_T controls the size of the tensor), which is reminiscent of the $\Theta(N^{\frac{d-2}{4}})$ computational threshold (with $N = n_T$ and $d = 3$). Comparing our results with a previous work of Seddik et al. (2023a) on the *statistical* limits to this problem, we are able to precisely quantify and display the computational-to-statistical gap. Notably, this gap is all the larger given that the views are less informative.

Further work is needed to comprehensively characterize the reconstruction performance of multi-view clustering. Our results provide new theoretical guarantees, but are not yet fully satisfying for the practical user, because the nature of the transformations that generate each view is usually complex

and therefore cannot be fully captured by our yet too simplistic model. A interesting direction for future research is thus to study the non-linear spiked tensor model induced by Equation (6.1) with more general view functions f_k. Such an analysis would provide valuable insight to any scientist faced with a multi-view clustering problem. Moreover, it could come with enlightening results on the influence of the non-linearities on the computational barrier.

A myriad of questions concerning the computational limits to tensor estimation still remain open. A noteworthy problem is the computation of the best (CP-)rank-R approximation. Although we argue that this is problem is encompassed by our analysis in Chapter 5 because of the equivalence between low CP-rank and low multilinear rank, the way it is practically solved raises interesting questions. Indeed, computing the best rank-R approximation of a very large tensor is costly and an indirect and less-memory-demanding method is usually preferred (Bro and Andersson, 1998). Given a large tensor $\mathcal{T}$ it is cheaper to first compute a "compressed version" with a truncated MLSVD (of sufficiently high multilinear rank) $[\![\mathcal{G}; \boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(d)}]\!] \approx \mathcal{T}$ and then compute a best rank-R approximation of the core tensor $\mathcal{G} \approx \sum_{r=1}^{R} \lambda_r \bigotimes_{\ell=1}^{d} \boldsymbol{x}_r^{(\ell)}$ which is of smaller dimension and from which a rank-R approximation of $\mathcal{T}$ can be deduced: $\mathcal{T} \approx \sum_{r=1}^{R} \lambda_r \bigotimes_{\ell=1}^{d} [\boldsymbol{U}^{(\ell)} \boldsymbol{x}_r^{(\ell)}]$. We have characterized the first step of this method (computing the truncated MLSVD) in Chapter 5. Therefore, it is worth studying the decomposition of the core tensor in order to complete a global analysis of this approach. An instructive result would help choosing a "good" multilinear-rank for the first step — it should not be too small, otherwise we loose information, but not too large either, otherwise the computation remains costly.

Despite their highly resource-intensive nature, tensors allow a richer data representation. Hence, a rigorous analysis of tensor estimation algorithms is necessary to promote relevant and efficient approaches within the limits imposed by the computational barrier.

Statistical Limits to Tensor Estimation

Although the statistical threshold cannot be reached in practice because of the computational-to-statistical gap, studying the statistical limits to a tensor estimation problem remains a relevant objective. Indeed, this allows to characterize the best achievable performances, even in the easy phase which, in practice, is *not* infinitely far from the statistical threshold as N is finite. To illustrate this point, consider the low-multilinear-rank approximation problem. We have shown in Chapter 5 that, when properly initialized, HOOI converges to an optimum of the likelihood in the easy phase. As $N \to +\infty$, this easy phase is infinitely far from the statistical threshold, hence the perfect alignment of the resulting solution stated in Theorem 5.11. But, in practice, N is finite and the computed solution is not perfectly aligned with the signal as illustrated in Figure 5.2. Therefore, characterizing these alignments at finite (but large) N requires to study the statistical limits to the estimation near the statistical threshold as $N \to +\infty$, though paradoxical it may sound.

However, studying the maximum likelihood estimator is delicate and this is more commonly performed with tools from statistical physics not considered in this thesis (Lesieur et al., 2017; Kadmon and Ganguli, 2018; Ben Arous et al., 2019; Jagannath et al., 2020). Still, recent works of Goulart et al. (2022) and Seddik et al. (2022) have shown how random matrix theory can be a powerful tool to study this problem as well. A solution to $\min_{\mathcal{X} \in \mathcal{M}} \|\mathcal{T} - \mathcal{X}\|_{\mathrm{F}}^2$ where $\mathcal{M}$ is a (closed) set of low-rank tensors can be associated to a random matrix model $\boldsymbol{\Phi}$ (see Section 1.2.3) whose spectral properties allow us to infer features of the solution — in particular, its alignment with a sought signal planted in $\mathcal{T}$.

A promising line of investigation is to apply the approach developed in Seddik et al. (2022) to models of interest in statistical learning. Processing large amounts of multimodal data is a common challenge of the big data era. Tensors as multi-way arrays have inherently a high volume, scaling exponentially with the number of modes, which can rapidly reach a memory limit. Often, the information sought in such gigantic raw data is a latent and low-dimensional structure. Hence, some techniques have been developed to compute decompositions (e.g., CPD or MLSVD, see Section 2.4.2) from *in-*

complete tensors (Vervliet et al., 2014). At first, they were used to tackle the issue of missing values, which is common in chemometrics data (Tomasi and Bro, 2005; Acar et al., 2011). Nowadays, they can be used to compute tensor decompositions on intentionally punctured data. That is, in order to reduce the volume of a tensor, only a portion of it is actually stored and inference is performed on its incomplete version.

A simple yet effective procedure to reduce the volume of a huge tensor $\mathcal{T}$ is therefore to store a portion $\varepsilon \in [0,1]$ of its entries chosen at random. Concretely, this is modeled by the entrywise multiplication $\mathcal{T} \odot \mathcal{B}$ where $\mathcal{B}$ has i.i.d. $\mathscr{B}(\varepsilon)$ entries (i.e., 1 with probability ε and 0 otherwise). Under the standard rank-one spiked tensor model $\mathcal{T} = \beta \bigotimes_{\ell=1}^{d} \boldsymbol{x}^{(\ell)} + \frac{1}{\sqrt{N}}\mathcal{N}$, we can infer the reconstruction performance of a rank-one approximation of $\mathcal{T} \odot \mathcal{B}$ from the results of Seddik et al. (2022) as per the universality result for random tensors of Gurau (2014). This shows that the impact of this puncturing policy on the reconstruction performance is a scaling by $1/\sqrt{\varepsilon}$ of the signal-to-noise ratio (SNR). That is, the statistical threshold is multiplied by $1/\sqrt{\varepsilon}$ and the alignments of the estimate with the signal observed at an SNR β without puncturing are found at an SNR $\beta/\sqrt{\varepsilon}$ with puncturing. Random puncturing can therefore rapidly become limiting if we are faced with drastic memory constraints. Nevertheless, demonstrating these results directly proves delicate because of mixing terms with the mask $\mathcal{B}$ whose mean behavior is intuitively well understood but practically tricky to untangle. As of now, our toolbox is manifestly not rich enough to tackle these seemingly simple questions on random tensors.

In case of strong memory limitations, more clever data-dependent puncturing approaches must be considered. In particular, similar procedures as those studied by Liao et al. (2021) on matrices could be applied to tensor data as well: *sparsification* keeps only the entries whose absolute value is above a fixed threshold while *quantization* replace every entry by their closest predefined "quantum" (hence only a small number of values, the quanta, are stored) and, in some applications, even *binarization* (keeping only the sign of the entries) may be relevant. As such data-dependent puncturing is expected to achieve a better performance versus cost trade-off, studying these procedures trough the associated random matrix model $\boldsymbol{\Phi}$ would provide valuable results to reduce the memory footprint of tensors but also further support this technique to study random tensor models.

Moreover, this approach is currently limited to the rank-one case and further work is needed to study the statistical limits to the best low-multilinear-rank approximation of the more general spiked tensor model $\mathcal{T} = \mathcal{P} + \frac{1}{\sqrt{N}}\mathcal{N}$ presented in Chapter 5. In particular, this is needed to access the theoretical curves behind the performance of HOOI presented in Figure 5.2. Heading in this direction requires the definition of a generalized random matrix model $\boldsymbol{\Phi}$ on which we could deploy the powerful random matrix machinery.

A long way is still to be run in order to comprehensively characterize the statistical limits to tensor estimation. Despite the complexity of the problem, many questions remain unanswered, thereby providing a flourishing source of exciting challenges. These theoretical considerations pave the way towards new tools for the study of large random tensors, allowing a more clever and prudent approach to intricate estimation problems.

Bibliography

Karim M. Abadir and Jan R. Magnus. *Matrix Algebra*. Econometric Exercises. Cambridge University Press, Cambridge, 2005. ISBN 978-0-521-53746-9. doi: 10.1017/CBO9780511810800. URL https://www.cambridge.org/core/books/matrix-algebra/BCE8FD2D62006D4061F88E02615B5622.

Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, April 2009. ISBN 978-1-4008-3024-4. doi: 10.1515/9781400830244. URL https://www.degruyter.com/document/doi/10.1515/9781400830244/html. Publication Title: Optimization Algorithms on Matrix Manifolds.

Evrim Acar, Canan Aykut-Bingol, Haluk Bingol, Rasmus Bro, and Bülent Yener. Multiway analysis of epilepsy tensors. *Bioinformatics (Oxford, England)*, 23(13):i10–18, July 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm210.

Evrim Acar, Daniel M. Dunlavy, Tamara G. Kolda, and Morten Mørup. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56, March 2011. ISSN 0169-7439. doi: 10.1016/j.chemolab.2010.08.004. URL https://www.sciencedirect.com/science/article/pii/S0169743910001437.

Marcel R. Ackermann, Marcus Märtens, Christoph Raupach, Kamil Swierkot, Christiane Lammersen, and Christian Sohler. StreamKM++: A clustering algorithm for data streams. *ACM Journal of Experimental Algorithmics*, 17:2.4:2.1–2.4:2.30, 2012. ISSN 1084-6654. doi: 10.1145/2133803.2184450. URL https://doi.org/10.1145/2133803.2184450.

Charu C. Aggarwal, Philip S. Yu, Jiawei Han, and Jianyong Wang. A Framework for Clustering Evolving Data Streams. In Johann-Christoph Freytag, Peter Lockemann, Serge Abiteboul, Michael Carey, Patricia Selinger, and Andreas Heuer, editors, *Proceedings 2003 VLDB Conference*, pages 81–92. Morgan Kaufmann, San Francisco, January 2003. ISBN 978-0-12-722442-8. doi: 10.1016/B978-012722442-8/50016-1. URL https://www.sciencedirect.com/science/article/pii/B9780127224428500161.

Charu C. Aggarwal, Jiawei Han, Jianyong Wang, and Philip S. Yu. A Framework for Projected Clustering of High Dimensional Data Streams. In Mario A. Nascimento, M. Tamer Özsu, Donald Kossmann, Renée J. Miller, José A. Blakeley, and Berni Schiefer, editors, *Proceedings 2004 VLDB Conference*, pages 852–863. Morgan Kaufmann, St Louis, January 2004. ISBN 978-0-12-088469-8. doi: 10.1016/B978-012088469-8.50075-9. URL https://www.sciencedirect.com/science/article/pii/B9780120884698500759.

Naoum Ilitch Akhiezer and Israel Markovich Glazman. *Theory of Linear Operators in Hilbert Space*. Dover Books on Mathematics. Dover Publications, December 1993. ISBN 978-0-486-67748-4.

Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham Kakade. A Tensor Spectral Approach to Learning Mixed Membership Community Models. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 867–881. PMLR, June 2013. URL https://proceedings.mlr.press/v30/Anandkumar13.html. ISSN: 1938-7228.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor Decompositions for Learning Latent Variable Models. *Journal of Machine Learning Research*, 15(80): 2773–2832, 2014. ISSN 1533-7928. URL http://jmlr.org/papers/v15/anandkumar14b.html.

Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2009. ISBN 978-0-521-19452-5. doi: 10.1017/CBO9780511801334. URL https://www.cambridge.org/core/books/an-introduction-to-random-matrices/8992DA8EB0386651E8DA8214A1FC7241.

Luc Angelelli, Anderson Andrei Da Silva, Yiannis Georgiou, Michael Mercier, Gregory Mounié, and Denis Trystram. Towards a Multi-objective Scheduling Policy for Serverless-based Edge-Cloud Continuum. In *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 485–497, Bangalore, India, May 2023. IEEE. ISBN 979-8-3503-0119-9. doi: 10.1109/CCGrid57682.2023.00052. URL https://ieeexplore.ieee.org/document/10171469/.

Zhidong Bai and Jack W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1): 316–345, January 1998. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1022855421. URL https://projecteuclid.org/journals/annals-of-probability/volume-26/issue-1/No-eigenvalues-outside-the-support-of-the-limiting-spectral-distribution/10.1214/aop/1022855421.full. Publisher: Institute of Mathematical Statistics.

Zhidong Bai and Jack W. Silverstein. Exact Separation of Eigenvalues of Large Dimensional Sample Covariance Matrices. *The Annals of Probability*, 27(3):1536–1555, July 1999. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1022677458. URL https://projecteuclid.org/journals/annals-of-probability/volume-27/issue-3/Exact-Separation-of-Eigenvalues-of-Large-Dimensional-Sample-Covariance-Matrices/10.1214/aop/1022677458.full. Publisher: Institute of Mathematical Statistics.

Zhidong Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Zhidong Bai and Yanqing Yin. Convergence to the Semicircle Law. *The Annals of Probability*, 16(2): 863–875, April 1988. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176991792. URL https://projecteuclid.org/journals/annals-of-probability/volume-16/issue-2/Convergence-to-the-Semicircle-Law/10.1214/aop/1176991792.full. Publisher: Institute of Mathematical Statistics.

Zhidong Bai, Jack W. Silverstein, and Yanqing Yin. A note on the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 26(2):166–168, August 1988. ISSN 0047-259X. doi: 10.1016/0047-259X(88)90078-4. URL https://www.sciencedirect.com/science/article/pii/0047259X88900784.

Jinho Baik and Jack W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006. ISSN 0047-259X. doi: 10.1016/j.jmva.2005.08.003. URL https://www.sciencedirect.com/science/article/pii/S0047259X0500134X.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, September 2005. ISSN 0091-1798, 2168-894X. doi: 10.1214/009117905000000233. URL https://projecteuclid.org/journals/annals-of-probability/volume-33/issue-5/Phase-transition-of-the-largest-eigenvalue-for-nonnull-complex-sample/10.1214/009117905000000233.full. Publisher: Institute of Mathematical Statistics.

Cornelis P. Baldé, Ruediger Kuehr, Tales Yamamoto, Rosie McDonald, Elena D'Angelo, Shahana Althaf, Garam Bel, Otmar Deubzer, Elena Fernandez-Cubillo, Vanessa Forti, Vanessa Gray, Sunil Herat, Shunichi Honda, Giulia Iattoni, Deepali S. Khetriwal, Vittoria Luda di Cortemiglia, Yuliya Lobuntsova, Nnorom Innocent, Noémie Pralat, and Michelle Wagner. The Global E-waste Monitor 2024. Technical report, United Nations Institute for Training and Research, 2024. URL https://ewastemonitor.info/the-global-e-waste-monitor-2024/.

Afonso S. Bandeira, Amelia Perry, and Alexander S. Wein. Notes on computational-to-statistical gaps: predictions using statistical physics. *Portugaliae Mathematica*, 75(2):159–186, December 2018. ISSN 0032-5155. doi: 10.4171/pm/2014. URL https://ems.press/journals/pm/articles/15908.

Marwa Banna, Florence Merlevède, and Magda Peligrad. On the limiting spectral distribution for a large class of symmetric random matrices with correlated entries. *Stochastic Processes and their Applications*, 125(7):2700–2726, July 2015. ISSN 0304-4149. doi: 10.1016/j.spa.2015.01.010. URL https://www.sciencedirect.com/science/article/pii/S0304414915000290.

Gérard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The Landscape of the Spiked Tensor Model. *Communications on Pure and Applied Mathematics*, 72(11):2282–2330, 2019. ISSN 1097-0312. doi: 10.1002/cpa.21861. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21861.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor PCA. *The Annals of Probability*, 48(4):2052–2087, July 2020. ISSN 0091-1798, 2168-894X. doi: 10.1214/19-AOP1415. URL https://projecteuclid.org/journals/annals-of-probability/volume-48/issue-4/Algorithmic-thresholds-for-tensor-PCA/10.1214/19-AOP1415.full. Publisher: Institute of Mathematical Statistics.

Gérard Ben Arous, Daniel Zhengyu Huang, and Jiaoyang Huang. Long random matrices and tensor unfolding. *The Annals of Applied Probability*, 33(6B):5753–5780, December 2023. ISSN 1050-5164, 2168-8737. doi: 10.1214/23-AAP1958. URL https://projecteuclid.org/journals/annals-of-applied-probability/volume-33/issue-6B/Long-random-matrices-and-tensor-unfolding/10.1214/23-AAP1958.full. Publisher: Institute of Mathematical Statistics.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011. ISSN 0001-8708. doi: 10.1016/j.aim.2011.02.007. URL https://www.sciencedirect.com/science/article/pii/S0001870811000570.

Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, October 2012. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.04.019. URL https://www.sciencedirect.com/science/article/pii/S0047259X12001108.

Xuan Bi, Xiwei Tang, Yubai Yuan, Yanqing Zhang, and Annie Qu. Tensors in Statistics. *Annual Review of Statistics and Its Application*, 8(Volume 8, 2021):345–368, March 2021. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-042720-020816. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-042720-020816. Publisher: Annual Reviews.

Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2012.

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973. ISSN 0025-5718, 1088-6842. doi: 10.1090/S0025-5718-1973-0348991-3. URL https://www.ams.org/mcom/1973-27-123/S0025-5718-1973-0348991-3/.

Chris Brew and Sabine Schulte im Walde. Spectral clustering for german verbs. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 117–124, 2002. URL https://aclanthology.org/W02-1016.pdf.

Rasmus Bro and Claus A. Andersson. Improving the speed of multiway algorithms: Part II: Compression. *Chemometrics and Intelligent Laboratory Systems*, 42(1):105–113, August 1998. ISSN 0169-7439. doi: 10.1016/S0169-7439(98)00011-2. URL https://www.sciencedirect.com/science/article/pii/S0169743998000112.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *The Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral. The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations. *The Annals of Probability*, 37(1), January 2009. ISSN 0091-1798. doi: 10.1214/08-AOP394. URL http://arxiv.org/abs/0706.0136. arXiv:0706.0136 [math].

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Advanced Series. Cengage Learning, 2001.

François Chapon, Romain Couillet, Walid Hachem, and Xavier Mestre. The Outliers Among the Singular Values of Large Rectangular Random Matrices with Additive Fixed Rank Deformation. *Markov Processes And Related Fields*, 20(2):183–228, 2014. URL https://math-mprf.org/journal/articles/id1331/.

Antoine Chatalic, Remi Gribonval, and Nicolas Keriven. Large-Scale High-Dimensional Clustering with Fast Sketching. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4714–4718, Calgary, AB, April 2018. IEEE. ISBN 978-1-5386-4658-8. doi: 10.1109/ICASSP.2018.8461328. URL https://ieeexplore.ieee.org/document/8461328/.

Hongbin Chen, Jean-Christophe Mourrat, and Jiaming Xia. Statistical inference of finite-rank tensors. *Annales Henri Lebesgue*, 5:1161–1189, 2022. ISSN 2644-9463. doi: 10.5802/ahl.146. URL https://ahl.centre-mersenne.org/articles/10.5802/ahl.146/.

Louis H. Y. Chen. An inequality for the multivariate normal distribution. *Journal of Multivariate Analysis*, 12(2):306–315, June 1982. ISSN 0047-259X. doi: 10.1016/0047-259X(82)90022-7. URL https://www.sciencedirect.com/science/article/pii/0047259X82900227.

Wei-Kuo Chen, Madeline Handschy, and Gilad Lerman. Phase transition in random tensors with multiple independent spikes. *The Annals of Applied Probability*, 31(4):1868–1913, August 2021. ISSN 1050-5164, 2168-8737. doi: 10.1214/20-AAP1636. URL https://projecteuclid.org/journals/annals-of-applied-probability/volume-31/issue-4/Phase-transition-in-random-tensors-with-multiple-independent-spikes/10.1214/20-AAP1636.full. Publisher: Institute of Mathematical Statistics.

Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 02(04):1350010, October 2013. ISSN 2010-3263. doi: 10.1142/S2010326313500010X. URL https://www.worldscientific.com/doi/abs/10.1142/S2010326313500010X. Publisher: World Scientific Publishing Co.

Antoine Chevreuil and Philippe Loubaton. On The Non-Detectability of Spiked Large Random Tensors. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pages 443–447, June 2018. doi: 10.1109/SSP.2018.8450752. URL https://ieeexplore.ieee.org/document/8450752.

Yasuko Chikuse. *Statistics on Special Manifolds*, volume 174 of *Lecture Notes in Statistics*. Springer, New York, NY, 2003. ISBN 978-0-387-00160-9 978-0-387-21540-2. doi: 10.1007/978-0-387-21540-2. URL http://link.springer.com/10.1007/978-0-387-21540-2.

Andrzej Cichocki, Danilo P. Mandic, Anh Huy Phan, Cesar F. Caiafa, Guoxu Zhou, Qibin Zhao, and Lieven De Lathauwer. Tensor Decompositions for Signal Processing Applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2):145–163, March 2015. ISSN 1558-0792. doi: 10.1109/MSP.2013.2297439. URL https://ieeexplore.ieee.org/document/7038247. Conference Name: IEEE Signal Processing Magazine.

Vincent Cohen-Addad, Benjamin Guedj, Varun Kanade, and Guy Rom. Online k-means Clustering. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1126–1134. PMLR, March 2021. URL https://proceedings.mlr.press/v130/cohen-addad21a.html. ISSN: 2640-3498.

Pierre Comon. Tensors versus Matrices, usefulness and unexpected properties. In IEEE, editor, *IEEE Workshop on Statistical Signal Processing*, pages 780–788, Cardiff, United Kingdom, September 2009. IEEE. URL https://hal.archives-ouvertes.fr/hal-00417258.

Pierre Comon. Tensors: a Brief Introduction. *IEEE Signal Processing Magazine*, 31(3):44–53, May 2014. doi: 10.1109/MSP.2014.2298533. URL https://hal.archives-ouvertes.fr/hal-00923279. Publisher: Institute of Electrical and Electronics Engineers.

Pierre Comon, Xavier Luciani, and André L. F. de Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics*, 23(7-8):393–405, 2009. ISSN 1099-128X. doi: 10.1002/cem.1236. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.1236. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.1236.

Allie Constantine and Ismail Wolff. The Dark Side of Technology: Coltan Mining in the DRC and its Human Rights and Environmental Impacts, November 2023. URL https://globalforestcoalition.org/the-dark-side-of-technology-coltan-mining-in-the-drc-and-its-human-rights-and-environmental-impacts/.

Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454, 2016. doi: 10.1214/16-EJS1144. URL https://hal.archives-ouvertes.fr/hal-01215343. Publisher: Shaker Heights, OH : Institute of Mathematical Statistics.

Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, Cambridge, 2022. ISBN 978-1-009-12323-5. doi: 10.1017/9781009128490. URL https://www.cambridge.org/core/books/random-matrix-methods-for-machine-learning/6B681EB69E58B5F888EDB689C160C682.

Romain Couillet, Florent Chatelain, and Nicolas Le Bihan. Two-way kernel matrix puncturing: towards resource-efficient PCA and spectral clustering. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2156–2165. PMLR, July 2021. URL https://proceedings.mlr.press/v139/couillet21a.html. ISSN: 2640-3498.

Romain Couillet, Denis Trystram, and Thierry Ménissier. The Submerged Part of the AI-ceberg. *IEEE Signal Processing Magazine*, 39(5):10–17, September 2022. ISSN 1558-0792. doi: 10.1109/MSP.2022.3182938. URL https://ieeexplore.ieee.org/document/9869565. Conference Name: IEEE Signal Processing Magazine.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the Best Rank-1 and Rank-(R1, R2, ..., RN) Approximation of Higher-Order Tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, January 2000a. ISSN 0895-4798. doi: 10.1137/S0895479898346995. URL https://epubs.siam.org/doi/10.1137/S0895479898346995. Publisher: Society for Industrial and Applied Mathematics.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A Multilinear Singular Value Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, January 2000b. ISSN 0895-4798. doi: 10.1137/S0895479896305696. URL https://epubs.siam.org/doi/10.1137/S0895479896305696. Publisher: Society for Industrial and Applied Mathematics.

Vin de Silva and Lek-Heng Lim. Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, January 2008. ISSN 0895-4798. doi: 10.1137/06066518X. URL https://epubs.siam.org/doi/10.1137/06066518X. Publisher: Society for Industrial and Applied Mathematics.

Charanpal Dhanjal, Romaric Gaudel, and Stéphan Clémençon. Efficient eigen-updating for spectral graph clustering. *Neurocomputing*, 131:440–452, May 2014. ISSN 0925-2312. doi: 10.1016/j.neucom.2013.11.015. URL https://www.sciencedirect.com/science/article/pii/S0925231214000125.

Chris H. Q. Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst D. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 107–114, November 2001. doi: 10.1109/ICDM.2001.989507. URL https://ieeexplore.ieee.org/document/989507.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936. ISSN 1860-0980. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

Samuel F. Edwards and Raymund C. Jones. The eigenvalue spectrum of a large symmetric random matrix. *Journal of Physics A: Mathematical and General*, 9(10):1595, October 1976. ISSN 0305-4470. doi: 10.1088/0305-4470/9/10/011. URL https://dx.doi.org/10.1088/0305-4470/9/10/011.

Noureddine El Karoui. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, 19(6):2362–2405, December 2009. ISSN 1050-5164, 2168-8737. doi: 10.1214/08-AAP548. URL https://projecteuclid.org/journals/annals-of-applied-probability/volume-19/issue-6

/Concentration-of-measure-and-spectra-of-random-matrices--Applications/10.1214/08-AAP548.full. Publisher: Institute of Mathematical Statistics.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, February 2010. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS648. URL https://projecteuclid.org/journals/annals-of-statistics/volume-38/issue-1/The-spectrum-of-kernel-random-matrices/10.1214/08-AOS648.full. Publisher: Institute of Mathematical Statistics.

Martin Ester, H. Kriegel, J. Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, 1996.

Hadi Fanaee-T and João Gama. EigenEvent: An Algorithm for Event Detection from Complex Data Streams in Syndromic Surveillance. *Intelligent Data Analysis*, 19(3):597–616, June 2015. ISSN 1088467X, 15714128. doi: 10.3233/IDA-150734. URL http://arxiv.org/abs/1406.3496. arXiv:1406.3496 [cs, stat].

Michael J. Feldman. Spiked singular values and vectors under extreme aspect ratios. *Journal of Multivariate Analysis*, 196:105187, July 2023. ISSN 0047-259X. doi: 10.1016/j.jmva.2023.105187. URL https://www.sciencedirect.com/science/article/pii/S0047259X23000337.

Michael J. Feldman and David L. Donoho. Sharp Recovery Thresholds of Tensor PCA Spectral Algorithms. *Advances in Neural Information Processing Systems*, 36:56628–56640, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b14d76c7266be21b338527cd25deac45-Abstract-Conference.html.

Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A Unifying Tutorial on Approximate Message Passing. *Foundations and Trends® in Machine Learning*, 15(4):335–536, May 2022. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000092. URL https://www.nowpublishers.com/article/Details/MAL-092. Publisher: Now Publishers, Inc.

Bernd Fritzke. A Growing Neural Gas Network Learns Topologies. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1995. URL https://papers.nips.cc/paper/1994/hash/d56b9fc4b0f1be8871f5e1c40c0067e7-Abstract.html.

Evgeny Frolov and Ivan Oseledets. Tensor methods and recommender systems. *WIREs Data Mining and Knowledge Discovery*, 7(3):e1201, 2017. ISSN 1942-4795. doi: 10.1002/widm.1201. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1201. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1201.

Delphine Féral and Sandrine Péché. The Largest Eigenvalue of Rank One Deformation of Large Wigner Matrices. *Communications in Mathematical Physics*, 272(1):185–228, May 2007. ISSN 1432-0916. doi: 10.1007/s00220-007-0209-3. URL https://doi.org/10.1007/s00220-007-0209-3.

Zoltán Füredi and János Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1 (3):233–241, September 1981. ISSN 1439-6912. doi: 10.1007/BF02579329. URL https://doi.org/10.1007/BF02579329.

David Gamarnik, Cristopher Moore, and Lenka Zdeborová. Disordered systems insights on computational hardness. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114015, November 2022. ISSN 1742-5468. doi: 10.1088/1742-5468/ac9cc8. URL https://dx.doi.org/10.1088/1742-5468/ac9cc8. Publisher: IOP Publishing and SISSA.

Stuart Geman. A Limit Theorem for the Norm of Random Matrices. *The Annals of Probability*, 8(2): 252–261, April 1980. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176994775. URL https://projecteuclid.org/journals/annals-of-probability/volume-8/issue-2/A-Limit-Theorem-for-the-Norm-of-Random-Matrices/10.1214/aop/1176994775.full. Publisher: Institute of Mathematical Statistics.

Jeffrey S. Geronimo and Theodore P. Hill. Necessary and sufficient condition that the limit of Stieltjes transforms is a Stieltjes transform. *Journal of Approximation Theory*, 121(1):54–60, March 2003. ISSN 0021-9045. doi: 10.1016/S0021-9045(02)00042-4. URL https://www.sciencedirect.com/science/article/pii/S0021904502000424.

Diane Gershon. Dealing with the data deluge. *Nature*, 416(6883):889–891, April 2002. ISSN 1476-4687. doi: 10.1038/416889a. URL https://www.nature.com/articles/416889a. Publisher: Nature Publishing Group.

Mohammed Ghesmoune, Mustapha Lebbah, and Hanene Azzag. State-of-the-art on clustering data streams. *Big Data Analytics*, 1(1):13, December 2016. ISSN 2058-6345. doi: 10.1186/s41044-016-0011-3. URL https://doi.org/10.1186/s41044-016-0011-3.

José Henrique de M. Goulart, Romain Couillet, and Pierre Comon. A Random Matrix Perspective on Random Tensors. *Journal of Machine Learning Research*, 23(264):1–36, 2022. ISSN 1533-7928. URL http://jmlr.org/papers/v23/21-1038.html.

Lars Grasedyck, Daniel Kressner, and Christine Tobler. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78, 2013. ISSN 1522-2608. doi: 10.1002/gamm.201310004. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gamm.201310004. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gamm.201310004.

Robert M. Gray. Toeplitz and Circulant Matrices: A Review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, January 2006. ISSN 1567-2190, 1567-2328. doi: 10.1561/0100000006. URL https://www.nowpublishers.com/article/Details/CIT-006. Publisher: Now Publishers, Inc.

Rémi Gribonval, Antoine Chatalic, Nicolas Keriven, Vincent Schellekens, Laurent Jacques, and Philip Schniter. Sketching Data Sets for Large-Scale Learning: Keeping only what you need. *IEEE Signal Processing Magazine*, 38(5):12–36, September 2021. doi: 10.1109/MSP.2021.3092574. URL https://hal.inria.fr/hal-03350599. Publisher: Institute of Electrical and Electronics Engineers.

Thomas Guhr, Axel Mueller-Groeling, and Hans A. Weidenmueller. Random Matrix Theories in Quantum Physics: Common Concepts. *Physics Reports*, 299(4-6):189–425, June 1998. ISSN 03701573. doi: 10.1016/S0370-1573(97)00088-4. URL http://arxiv.org/abs/cond-mat/9707301. arXiv:cond-mat/9707301.

Razvan Gurau. Universality for random tensors. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(4):1474–1525, November 2014. ISSN 0246-0203. doi: 10.1214/13-AIHP567. URL https://projecteuclid.org/journals/annales-de-linstitut-henri-poincare-probabilites-et-statistiques/volume-50/issue-4/Universality-for-random-tensors/10.1214/13-AIHP567.full. Publisher: Institut Henri Poincaré.

Louis Guttman. Enlargement Methods for Computing the Inverse Matrix. *The Annals of Mathematical Statistics*, 17(3):336–343, September 1946. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177730946. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-17/issue-3/Enlargement-Methods-for-Computing-the-Inverse-Matrix/10.1214/aoms/1177730946.full. Publisher: Institute of Mathematical Statistics.

Wolfgang Hackbusch. *Tensor Spaces and Numerical Tensor Calculus*. Springer Series in Computational Mathematics. Springer, 2012. URL https://link.springer.com/book/10.1007/978-3-642-28027-6.

Shigefumi Hata and Hiroya Nakao. Localization of Laplacian eigenvectors on random networks. *Scientific Reports*, 7(1):1121, April 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-01010-0. URL https://www.nature.com/articles/s41598-017-01010-0. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Complex networks;Nonlinear phenomena Subject_term_id: complex-networks;nonlinear-phenomena.

Christopher J. Hillar and Lek-Heng Lim. Most Tensor Problems Are NP-Hard. *Journal of the ACM*, 60 (6):45:1–45:39, November 2013. ISSN 0004-5411. doi: 10.1145/2512329. URL https://doi.org/10.1145/2512329.

Aicke Hinrichs, Joscha Prochno, and Jan Vybíral. Entropy numbers of embeddings of Schatten classes. *Journal of Functional Analysis*, 273(10):3241–3261, November 2017. ISSN 0022-1236. doi: 10.1016/j.jfa.2017.08.008. URL https://www.sciencedirect.com/science/article/pii/S0022123617303221.

Frank L. Hitchcock. The Expression of a Tensor or a Polyadic as a Sum of Products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927. ISSN 1467-9590. doi: 10.1002/sapm192761164. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sapm192761164. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sapm192761164.

Samuel B. Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Proceedings of the 28th Conference on Learning Theory*, pages 956–1006. PMLR, June 2015. URL https://proceedings.mlr.press/v40/Hopkins15.html. ISSN: 1938-7228.

Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, STOC '16, pages 178–191, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4132-5. doi: 10.1145/2897518.2897529. URL https://doi.org/10.1145/2897518.2897529.

Samuel B. Hopkins, Pravesh K. Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The Power of Sum-of-Squares for Detecting Hidden Structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE Computer Society, October 2017. ISBN 978-1-5386-3464-6. doi: 10.1109/FOCS.2017.72. URL https://www.computer.org/csdl/proceedings-article/focs/2017/3464a720/12OmNxFsmCD. ISSN: 0272-5428.

Jiaoyang Huang, Daniel Z. Huang, Qing Yang, and Guang Cheng. Power Iteration for Tensor PCA. *Journal of Machine Learning Research*, 23(128):1–47, 2022. ISSN 1533-7928. URL http://jmlr.org/papers/v23/21-1290.html.

Borbála Hunyadi, Patrick Dupont, Wim Van Paesschen, and Sabine Van Huffel. Tensor decompositions and data fusion in epileptic electroencephalography and functional magnetic resonance imaging data. *WIREs Data Mining and Knowledge Discovery*, 7(1):e1197, 2017. ISSN 1942-4795. doi: 10.1002/widm.1197. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1197. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1197.

Aukosh Jagannath, Patrick Lopatto, and Léo Miolane. Statistical thresholds for tensor PCA. *The Annals of Applied Probability*, 30(4):1910–1933, August 2020. ISSN 1050-5164, 2168-8737. doi: 10.1214/19-AAP1547. URL https://projecteuclid.org/journals/annals-of-applied-probability/volume-30/issue-4/Statistical-thresholds-for-tensor-PCA/10.1214/19-AAP1547.full. Publisher: Institute of Mathematical Statistics.

Mathilde Jay, Vladimir Ostapenco, Laurent Lefevre, Denis Trystram, Anne-Cécile Orgerie, and Benjamin Fichel. An experimental comparison of software-based power meters: focus on CPU and GPU. In *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 106–118, May 2023. doi: 10.1109/CCGrid57682.2023.00020. URL https://ieeexplore.ieee.org/document/10171575.

Edwin Thompson Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, April 2001. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1009210544. URL https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-2/On-the-distribution-of-the-largest-eigenvalue-in-principal-components/10.1214/aos/1009210544.full. Publisher: Institute of Mathematical Statistics.

Arun Kadavankandy and Romain Couillet. Asymptotic Gaussian Fluctuations of Spectral Clustering Eigenvectors. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 694–698, December 2019. doi: 10.1109/CAMSAP45676.2019.9022474.

Jonathan Kadmon and Surya Ganguli. Statistical mechanics of low-rank tensor decomposition. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/b3848d61bbbc6207c6668a8a9e2730ed-Abstract.html.

Charilaos I. Kanatsoulis, Xiao Fu, Nicholas D. Sidiropoulos, and Wing-Kin Ma. Hyperspectral Super-Resolution: A Coupled Tensor Factorization Approach. *IEEE Transactions on Signal Processing*, 66(24):6503–6517, December 2018. ISSN 1941-0476. doi: 10.1109/TSP.2018.2876362. URL https://ieeexplore.ieee.org/abstract/document/8494792. Conference Name: IEEE Transactions on Signal Processing.

Arie Kapteyn, Heinz Neudecker, and Tom Wansbeek. An approach ton-mode components analysis. *Psychometrika*, 51(2):269–275, June 1986. ISSN 1860-0980. doi: 10.1007/BF02293984. URL https://doi.org/10.1007/BF02293984.

Alexandros Karatzoglou, Xavier Amatriain, Linas Baltrunas, and Nuria Oliver. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 79–86, New York, NY, USA, September 2010. Association for Computing Machinery. ISBN 978-1-60558-906-0. doi: 10.1145/1864708.1864727. URL https://doi.org/10.1145/1864708.1864727.

Jonathan P. Keating. Random matrices and the Riemann zeta-function. *Highlights of Mathematical Physics (London)*, pages 153–163, 2002. URL https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a44ad1b8200947f72514485e3656028d3681b90c. Publisher: Citeseer.

Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. Sketching for large-scale learning of mixture models. *Information and Inference: A Journal of the IMA*, 7(3):447–508, September 2018. ISSN 2049-8764. doi: 10.1093/imaiai/iax015. URL https://doi.org/10.1093/imaiai/iax015.

Chiheon Kim, Afonso S. Bandeira, and Michel X. Goemans. Community detection in hypergraphs, spiked tensor models, and Sum-of-Squares. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 124–128, July 2017. doi: 10.1109/SAMPTA.2017.8024470. URL https://ieeexplore.ieee.org/document/8024470.

Tamara G. Kolda. A Counterexample to the Possibility of an Extension of the Eckart–Young Low-Rank Approximation Theorem for the Orthogonal Rank Tensor Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 24(3):762–767, January 2003. ISSN 0895-4798. doi: 10.1137/S0895479801394465. URL https://epubs.siam.org/doi/10.1137/S0895479801394465. Publisher: Society for Industrial and Applied Mathematics.

Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, August 2009. doi: 10.1137/07070111X. URL https://epubs.siam.org/doi/10.1137/07070111X. Publisher: Society for Industrial and Applied Mathematics.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Pieter M. Kroonenberg and Jan de Leeuw. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45(1):69–97, March 1980. ISSN 1860-0980. doi: 10.1007/BF02293599. URL https://doi.org/10.1007/BF02293599.

Joseph M. Landsberg. *Tensors: Geometry and Applications*, volume 128 of *Graduate Studies in Mathematics*. American Mathematical Society, December 2011. ISBN 978-0-8218-6907-9 978-0-8218-8481-2 978-0-8218-8483-6 978-1-4704-0923-4. doi: 10.1090/gsm/128. URL http://www.ams.org/gsm/128. ISSN: 1065-7339.

Jean-François Le Gall. *Measure Theory, Probability, and Stochastic Processes*, volume 295 of *Graduate Texts in Mathematics*. Springer International Publishing, Cham, 2022. ISBN 978-3-031-14204-8 978-3-031-14205-5. doi: 10.1007/978-3-031-14205-5. URL https://link.springer.com/10.1007/978-3-031-14205-5.

Hugo Lebeau, Romain Couillet, and Florent Chatelain. Une analyse par matrices aléatoires de l'apprentissage en ligne : traiter des grandes données avec des ressources mémoire limitées. In *28° Colloque sur le traitement du signal et des images*, pages p. 713–716, Nancy, 2022a. GRETSI - Groupe de Recherche en Traitement du Signal et des Images. Issue: 001-0178.

Hugo Lebeau, Romain Couillet, and Florent Chatelain. A Random Matrix Analysis of Data Stream Clustering: Coping With Limited Memory Resources. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12253–12281. PMLR, 2022b. URL https://proceedings.mlr.press/v162/lebeau22a.html. ISSN: 2640-3498.

Hugo Lebeau, Romain Couillet, and Florent Chatelain. HOSVD Tronquée : Analyse d'une Approximation Tensorielle Rapide. In *29° Colloque sur le traitement du signal et des images*, pages p. 17–20, Grenoble, 2023. GRETSI - Groupe de Recherche en Traitement du Signal et des Images. Issue: 2023-1060.

Hugo Lebeau, Florent Chatelain, and Romain Couillet. Asymptotic Gaussian Fluctuations of Eigenvectors in Spectral Clustering. *IEEE Signal Processing Letters*, 31:1920–1924, 2024a. ISSN 1558-2361. doi: 10.1109/LSP.2024.3422886. URL https://ieeexplore.ieee.org/document/10584249. Conference Name: IEEE Signal Processing Letters.

Hugo Lebeau, Mohamed El Amine Seddik, and José Henrique de Morais Goulart. Performance Gaps in Multi-view Clustering under the Nested Matrix-Tensor Model. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=ILqA090eq2.

Hugo Lebeau, Florent Chatelain, and Romain Couillet. A Random Matrix Approach to Low-Multilinear-Rank Tensor Approximation. *Journal of Machine Learning Research*, 26(7):1–64, 2025. ISSN 1533-7928. URL http://jmlr.org/papers/v26/24-0193.html.

Michel Ledoux. *The Concentration of Measure Phenomenon*. Number 89 in Mathematical Surveys and Monographs. American Mathematical Society, 2001.

Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1297–1301. PMLR, June 2017. URL https://proceedings.mlr.press/v65/lelarge17a.html. ISSN: 2640-3498.

Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 511–515, June 2017. doi: 10.1109/ISIT.2017.8006580. URL http://arxiv.org/abs/1701.08010. arXiv:1701.08010 [cond-mat, stat].

Nan Li and Baoxin Li. Tensor completion for on-board compression of hyperspectral images. In *2010 IEEE International Conference on Image Processing*, pages 517–520, September 2010. doi: 10.1109/ICIP.2010.5651225. ISSN: 2381-8549.

Run-Ze Li. The characteristic functions of spherical matrix distributions. *Statistics & Probability Letters*, 17(4):273–279, July 1993. ISSN 0167-7152. doi: 10.1016/0167-7152(93)90202-T. URL https://www.sciencedirect.com/science/article/pii/016771529390202T.

Zhenyu Liao, Romain Couillet, and Michael W. Mahoney. Sparse Quantized Spectral Clustering. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pBqLS-7KYAF.

Edo Liberty. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '13, pages 581–588, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2487623. URL https://doi.org/10.1145/2487575.2487623.

Edo Liberty, Ram Sriharsha, and Maxim Sviridenko. An Algorithm for Online K-Means Clustering. In *2016 Proceedings of the Meeting on Algorithm Engineering and Experiments (ALENEX)*, Proceedings, pages 81–89. Society for Industrial and Applied Mathematics, December 2015. doi: 10.1137/1.9781611974317.7. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974317.7.

Tianqi Liu, Ming Yuan, and Hongyu Zhao. Characterizing Spatiotemporal Transcriptome of the Human Brain Via Low-Rank Tensor Decomposition. *Statistics in Biosciences*, 14(3):485–513, December 2022. ISSN 1867-1772. doi: 10.1007/s12561-021-09331-5. URL https://doi.org/10.1007/s12561-021-09331-5.

Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and Effective Regularized Incomplete Multi-View Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2634–2646, August 2021. ISSN 1939-3539. doi: 10.1109/TPAMI.2020.2974828. URL https://ieeexplore.ieee.org/document/9001210. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Xinwang Liu, En Zhu, Jiyuan Liu, Timothy Hospedales, Yang Wang, and Meng Wang. SimpleMKKM: Simple Multiple Kernel K-Means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5174–5186, April 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2022.3198638. URL https://ieeexplore.ieee.org/document/9857664. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Cosme Louart and Romain Couillet. Concentration of Measure and Large Random Matrices with an application to Sample Covariance Matrices, January 2021. URL http://arxiv.org/abs/1805.08295. arXiv:1805.08295 [math].

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, April 2018. ISSN 1050-5164, 2168-8737. doi: 10.1214/17-AAP1328. URL https://projecteuclid.org/journals/annals-of-applied-probability/volume-28/issue-2/A-random-matrix-approach-to-neural-networks/10.1214/17-AAP1328.full. Publisher: Institute of Mathematical Statistics.

Philippe Loubaton. On the Almost Sure Location of the Singular Values of Certain Gaussian Block-Hankel Large Random Matrices. *Journal of Theoretical Probability*, 29(4):1339–1443, December 2016. ISSN 1572-9230. doi: 10.1007/s10959-015-0614-z. URL https://doi.org/10.1007/s10959-015-0614-z.

Federica Lucivero. Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives. *Science and Engineering Ethics*, 26(2):1009–1030, April 2020. ISSN 1471-5546. doi: 10.1007/s11948-019-00171-7. URL https://doi.org/10.1007/s11948-019-00171-7.

Anna Lytova and Leonid Pastur. Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *The Annals of Probability*, 37(5):1778–1840, 2009. Publisher: Institute of Mathematical Statistics.

Matthias Löffler, Anderson Y. Zhang, and Harrison H. Zhou. Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, October 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS2044. URL https://projecteuclid.org/journals/annals-of-statistics/volume-49/issue-5/Optimality-of-spectral-clustering-in-the-Gaussian-mixture-model/10.1214/20-AOS2044.full. Publisher: Institute of Mathematical Statistics.

Xiaoyi Mai and Romain Couillet. A Random Matrix Analysis and Improvement of Semi-Supervised Learning for Large Dimensional Data. *Journal of Machine Learning Research*, 19(79):1–27, 2018. ISSN 1533-7928. URL http://jmlr.org/papers/v19/17-421.html.

Kanti V. Mardia and Chinubhai G. Khatri. Uniform distribution on a Stiefel manifold. *Journal of Multivariate Analysis*, 7(3):468–473, September 1977. ISSN 0047-259X. doi: 10.1016/0047-259X(77)90087-2. URL https://www.sciencedirect.com/science/article/pii/0047259X77900872.

Vladimir A. Marčenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 72(114)(4):457–483, 1967. Publisher: IOP Publishing.

Donella H. Meadows, Dennis L. Meadows, Jørgen Randers, and William W. Behrens. *The Limits to Growth*. Potomac Associates - Universe Books, 1972. ISBN 0-87663-165-0.

Florence Merlevède, Costel Peligrad, and Magda Peligrad. On the universality of spectral limit for random matrices with martingale differences entries. *Random Matrices: Theory and Applications*, 04(01):1550003, January 2015. ISSN 2010-3263. doi: 10.1142/S2010326315500033. URL https:

//www.worldscientific.com/doi/abs/10.1142/S2010326315500033. Publisher: World Scientific Publishing Co.

Leonid Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, January 1960. ISSN 0033-5606. doi: 10.1093/qmath/11.1.50. URL https://doi.org/10.1093/qmath/11.1.50.

Behrad Moniri and Hamed Hassani. Signal-Plus-Noise Decomposition of Nonlinear Spiked Random Matrix Models, May 2024. URL http://arxiv.org/abs/2405.18274. arXiv:2405.18274 [cs, eess, math, stat].

Hugh L. Montgomery. The pair correlation of zeros of the zeta function. In *Proc. Symp. Pure Math*, volume 24, pages 181–193, 1973. URL https://www-personal.umich.edu/~hlm/paircor1.pdf.

Chaitanya Muralidhara, Andrew M. Gross, Robin R. Gutell, and Orly Alter. Tensor Decomposition Reveals Concurrent Evolutionary Convergences and Divergences and Correlations with Structural Motifs in Ribosomal RNA. *PLOS ONE*, 6(4):e18768, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0018768. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018768. Publisher: Public Library of Science.

Andrew Ng, Michael Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002. URL https://papers.nips.cc/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html.

Minh-Toan Nguyen and Romain Couillet. Asymptotic Bayes risk of semi-supervised multitask learning on Gaussian mixture. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pages 5063–5078. PMLR, April 2023. URL https://proceedings.mlr.press/v206/nguyen23c.html. ISSN: 2640-3498.

Lê Nguyen Hoang and El Mahdi El Mhamdi. *Le fabuleux chantier*. EDP Sciences, 2019. ISBN 978-2-7598-2361-1.

Lê Nguyen Hoang and Jean-Lou Fourquet. *La Dictature des Algorithmes*. Tallandier, 2024. ISBN 979-10-210-6058-6.

Feiping Nie, Jing Li, and Xuelong Li. Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1881–1887, New York, New York, USA, 2016. AAAI Press. ISBN 978-1-57735-770-4.

Feiping Nie, Guohao Cai, and Xuelong Li. Multi-View Clustering and Semi-Supervised Classification with Adaptive Neighbours. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), February 2017a. ISSN 2374-3468. doi: 10.1609/aaai.v31i1.10909. URL https://ojs.aaai.org/index.php/AAAI/article/view/10909. Number: 1.

Feiping Nie, Jing Li, and Xuelong Li. Self-weighted multiview clustering with multiple graphs. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2564–2570, Melbourne, Australia, 2017b. AAAI Press. ISBN 978-0-9992411-0-3.

Huazhong Ning, Wei Xu, Yun Chi, Yihong Gong, and Thomas S. Huang. Incremental spectral clustering by efficiently updating the eigen-system. *Pattern Recognition*, 43(1):113–127, January 2010. ISSN 0031-3203. doi: 10.1016/j.patcog.2009.06.001. URL https://www.sciencedirect.com/science/article/pii/S0031320309002209.

Larsson Omberg, Gene H. Golub, and Orly Alter. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences*, 104(47):18371–18376, November 2007. doi: 10.1073/pnas.0709146104. URL https://www.pnas.org/doi/full/10.1073/pnas.0709146104. Publisher: Proceedings of the National Academy of Sciences.

Larsson Omberg, Joel R. Meyerson, Kayta Kobayashi, Lucy S. Drury, John F. X. Diffley, and Orly Alter. Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. *Molecular Systems Biology*, 5(1):312, January 2009. ISSN 1744-4292. doi: 10.1038/msb.2009.70. URL https://www.embopress.org/doi/full/10.1038/msb.2009.70. Publisher: John Wiley & Sons, Ltd.

Alexei Onatski, Marcelo J. Moreira, and Marc Hallin. Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics*, 41(3):1204–1231, June 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1100. URL https://projecteuclid.org/journals/annals-of-statistics/volume-41/issue-3/Asymptotic-power-of-sphericity-tests-for-high-dimensional-data/10.1214/13-AOS1100.full. Publisher: Institute of Mathematical Statistics.

OpenAI. GPT-4 Technical Report, March 2024. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Evangelos E. Papalexakis, Nicholas D. Sidiropoulos, and Rasmus Bro. From K-Means to Higher-Way Co-Clustering: Multilinear Decomposition With Sparse Latent Factors. *IEEE Transactions on Signal Processing*, 61(2):493–506, January 2013. ISSN 1941-0476. doi: 10.1109/TSP.2012.2225052. URL https://ieeexplore.ieee.org/document/6331561. Conference Name: IEEE Transactions on Signal Processing.

Leonid Andreevich Pastur and Mariya Shcherbina. *Eigenvalue Distribution of Large Random Matrices*. Number 171 in Mathematical Surveys and Monographs. American Mathematical Society, 2011.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon Emissions and Large Neural Network Training, April 2021. URL http://arxiv.org/abs/2104.10350. arXiv:2104.10350 [cs].

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *undefined*, 2007. URL https://www.semanticscholar.org/paper/ASYMPTOTICS-OF-SAMPLE-EIGENSTRUCTURE-FOR-A-LARGE-Paul/bc1eb681a5561f4f38153d274824d95fbc76e90b.

Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. COMIC: Multi-view Clustering Without Parameter Selection. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5092–5101. PMLR, May 2019. URL https://proceedings.mlr.press/v97/peng19a.html. ISSN: 2640-3498.

Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira. Statistical limits of spiked tensor models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(1):230–264, February 2020. ISSN 0246-0203. doi: 10.1214/19-AIHP960. URL https://projecteuclid.org/journals/annales-de-linstitut-henri-poincare-probabilites-et-statistiques/volume-56/issue-1/Statistical-limits-of-spiked-tensor-models/10.1214/19-AIHP960.full. Publisher: Institut Henri Poincaré.

Johanna Pohl, Lorenz M. Hilty, and Matthias Finkbeiner. How LCA contributes to the environmental assessment of higher order effects of ICT application: A review of different approaches. *Journal of*

Cleaner Production, 219:698–712, May 2019. ISSN 0959-6526. doi: 10.1016/j.jclepro.2019.02.018. URL https://www.sciencedirect.com/science/article/pii/S095965261930397X.

Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory*. Cambridge University Press, 2020.

Sandrine Péché. The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probability Theory and Related Fields*, 134(1):127–173, January 2006. ISSN 1432-2064. doi: 10.1007/s00440-005-0466-z. URL https://doi.org/10.1007/s00440-005-0466-z.

Stephan Rabanser, Oleksandr Shchur, and Stephan Günnemann. Introduction to Tensor Decompositions and their Applications in Machine Learning. *arXiv:1711.10781 [cs, stat]*, November 2017. URL http://arxiv.org/abs/1711.10781. arXiv: 1711.10781.

Neil C. Rabinowitz, Robbe L. Goris, Marlene Cohen, and Eero P. Simoncelli. Attention stabilizes the shared gain of V4 populations. *eLife*, 4:e08998, November 2015. ISSN 2050-084X. doi: 10.7554/eLife.08998. URL https://doi.org/10.7554/eLife.08998. Publisher: eLife Sciences Publications, Ltd.

Steffen Rendle and Lars Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 81–90, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 978-1-60558-889-6. doi: 10.1145/1718487.1718498. URL https://doi.org/10.1145/1718487.1718498.

Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/hash/b5488aeff42889188d03c9895255cecc-Abstract.html.

Valentina Ros, Gerard Ben Arous, Giulio Biroli, and Chiara Cammarota. Complex Energy Landscapes in Spiked-Tensor and Simple Glassy Models: Ruggedness, Arrangements of Local Minima, and Phase Transitions. *Physical Review X*, 9(1):011003, January 2019. doi: 10.1103/PhysRevX.9.011003. URL https://link.aps.org/doi/10.1103/PhysRevX.9.011003. Publisher: American Physical Society.

Walter Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, second edition edition, 1991. ISBN 0-07-100944-2.

Berkant Savas and Lars Eldén. Handwritten digit classification using higher order singular value decomposition. *Pattern Recognition*, 40(3):993–1003, March 2007. ISSN 0031-3203. doi: 10.1016/j.patcog.2006.08.004. URL https://www.sciencedirect.com/science/article/pii/S0031320306003542.

Joel L. Schiff. *Normal Families*. Universitext. Springer, New York, NY, 1993. ISBN 978-0-387-97967-0 978-1-4612-0907-2. doi: 10.1007/978-1-4612-0907-2. URL http://link.springer.com/10.1007/978-1-4612-0907-2.

Isaac Jacob Schoenberg. Metric Spaces and Completely Monotone Functions. *Annals of Mathematics*, 39(4):811–841, 1938. ISSN 0003-486X. doi: 10.2307/1968466. URL https://www.jstor.org/stable/1968466. Publisher: Annals of Mathematics.

Hanne Schultz. Non-commutative polynomials of independent Gaussian random matrices. The real and symplectic cases. *Probability Theory and Related Fields*, 131(2):261–309, February 2005. ISSN 1432-2064. doi: 10.1007/s00440-004-0366-7. URL https://doi.org/10.1007/s00440-004-0366-7.

Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random Matrix Theory Proves that Deep Learning Representations of GAN-data Behave as Gaussian Mixtures. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8573–8582. PMLR, November 2020. URL https://proceedings.mlr.press/v119/seddik20a.html. ISSN: 2640-3498.

Mohamed El Amine Seddik, Maxime Guillaud, and Romain Couillet. When Random Tensors meet Random Matrices, November 2022. URL http://arxiv.org/abs/2112.12348. arXiv:2112.12348 [math, stat].

Mohamed El Amine Seddik, Mastane Achab, Henrique Goulart, and Merouane Debbah. A Nested Matrix-Tensor Model for Noisy Multi-view Clustering, May 2023a. URL http://arxiv.org/abs/2305.19992. arXiv:2305.19992 [cs, stat].

Mohamed El Amine Seddik, Malik Tiomoko, Alexis Decurninge, Maxim Panov, and Maxime Guillaud. Learning from Low Rank Tensor Data: A Random Tensor Theory Perspective. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 1858–1867. PMLR, July 2023b. URL https://proceedings.mlr.press/v216/seddik23a.html. ISSN: 2640-3498.

Jeffrey S. Seely, Matthew T. Kaufman, Stephen I. Ryu, Krishna V. Shenoy, John P. Cunningham, and Mark M. Churchland. Tensor Analysis Reveals Distinct Population Structure that Parallels the Different Computational Roles of Areas M1 and V1. *PLOS Computational Biology*, 12(11):e1005164, November 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005164. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005164. Publisher: Public Library of Science.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000. ISSN 1939-3539. doi: 10.1109/34.868688. URL https://ieeexplore.ieee.org/document/868688. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Nicholas D. Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, and Christos Faloutsos. Tensor Decomposition for Signal Processing and Machine Learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, July 2017. ISSN 1941-0476. doi: 10.1109/TSP.2017.2690524. Conference Name: IEEE Transactions on Signal Processing.

Jack W. Silverstein and Zhidong Bai. On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995. ISSN 0047-259X. doi: 10.1006/jmva.1995.1051. URL https://www.sciencedirect.com/science/article/pii/S0047259X85710512.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *The Third International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1409.1556. arXiv: 1409.1556.

Antonius G. M. Steerneman and Frederieke van Perlo-ten Kleij. Spherical distributions: Schoenberg (1938) revisited. *Expositiones Mathematicae*, 23(3):281–287, September 2005. ISSN 0723-0869. doi: 10.1016/j.exmath.2005.01.002. URL https://www.sciencedirect.com/science/article/pii/S0723086905000034.

Charles M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135–1151, November 1981. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345632. URL https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-6

/Estimation-of-the-Mean-of-a-Multivariate-Normal-Distribution/10.1214/aos/1176345632.full. Publisher: Institute of Mathematical Statistics.

Gilbert Wright Stewart and Ji-guang Sun. *Matrix perturbation theory*. Computer science and scientific computing. Academic Press, Boston, 1990. ISBN 978-0-12-670230-9. URL http://catdir.loc.gov/catdir/toc/els031/90033378.html. OCLC: 21227976.

Will Wei Sun, Botao Hao, and Lexin Li. Tensors in Modern Statistical Learning. In *Wiley StatsRef: Statistics Reference Online*, pages 1–25. John Wiley & Sons, Ltd, 2021. ISBN 978-1-118-44511-2. doi: 10.1002/9781118445112.stat08319. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08319. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat08319.

Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 81(1):73–205, December 1995. ISSN 1618-1913. doi: 10.1007/BF02699376. URL https://doi.org/10.1007/BF02699376.

Terence Tao. *Topics in Random Matrix Theory*. Number 132 in Graduate Studies in Mathematics. American Mathematical Society, 2012.

Dimitris K. Tasoulis, Gordon Ross, and Niall M. Adams. Visualising the Cluster Structure of Data Streams. In Michael R. Berthold, John Shawe-Taylor, and Nada Lavrač, editors, *Advances in Intelligent Data Analysis VII*, Lecture Notes in Computer Science, pages 81–92, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-74825-0. doi: 10.1007/978-3-540-74825-0_8.

Joan Thiesen, Torben S. Christensen, Thomas G. Kristensen, Rikke D. Andersen, Brit Brunoe, Trine K. Gregersen, Mikkel Thrane, and Bo P. Weidema. Rebound effects of price differences. *The International Journal of Life Cycle Assessment*, 13(2):104–114, March 2008. ISSN 1614-7502. doi: 10.1065/lca2006.12.297. URL https://doi.org/10.1065/lca2006.12.297.

Edward Charles Titchmarsh. *The Theory Of Functions*. Oxford University Press, 1939. URL http://archive.org/details/in.ernet.dli.2015.2588.

Giorgio Tomasi and Rasmus Bro. PARAFAC and missing values. *Chemometrics and Intelligent Laboratory Systems*, 75(2):163–180, February 2005. ISSN 0169-7439. doi: 10.1016/j.chemolab.2004.07.003. URL https://www.sciencedirect.com/science/article/pii/S0169743904001741.

Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors, July 2014. URL http://arxiv.org/abs/1407.1870. arXiv:1407.1870 [math, stat].

William F. Trench. Some Spectral Properties of Hermitian Toeplitz Matrices. *SIAM Journal on Matrix Analysis and Applications*, 15(3):938–942, July 1994. ISSN 0895-4798. doi: 10.1137/S0895479892239007. URL https://epubs.siam.org/doi/10.1137/S0895479892239007. Publisher: Society for Industrial and Applied Mathematics.

Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, September 1966. ISSN 1860-0980. doi: 10.1007/BF02289464. URL https://doi.org/10.1007/BF02289464.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998. ISBN 978-0-521-78450-4. doi: 10.1017/CBO9780511802256. URL https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D.

M. Alex O. Vasilescu. Human motion signatures: analysis, synthesis, recognition. In *2002 International Conference on Pattern Recognition*, volume 3, pages 456–460 vol.3, August 2002. doi: 10.1109/ICPR.2002.1047975. ISSN: 1051-4651.

M. Alex O. Vasilescu and Demetri Terzopoulos. Multilinear subspace analysis of image ensembles. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–93, June 2003. doi: 10.1109/CVPR.2003.1211457. ISSN: 1063-6919.

Nico Vervliet, Otto Debals, Laurent Sorber, and Lieven De Lathauwer. Breaking the Curse of Dimensionality Using Decompositions of Incomplete Tensors: Tensor-based scientific computing in big data analysis. *IEEE Signal Processing Magazine*, 31(5):71–79, September 2014. ISSN 1558-0792. doi: 10.1109/MSP.2014.2329429.

Nico Vervliet, Otto Debals, Laurent Sorber, Marc Van Barel, and Lieven De Lathauwer. Tensorlab 3.0, March 2016. URL https://www.tensorlab.net/.

Nico Vervliet, Otto Debals, and Lieven De Lathauwer. Exploiting Efficient Representations in Large-Scale Tensor Decompositions. *SIAM Journal on Scientific Computing*, 41(2):A789–A815, January 2019. ISSN 1064-8275. doi: 10.1137/17M1152371. URL https://epubs.siam.org/doi/abs/10.1137/17M1152371. Publisher: Society for Industrial and Applied Mathematics.

Dan Voiculescu. Limit laws for Random matrices and free products. *Inventiones mathematicae*, 104 (1):201–220, December 1991. ISSN 1432-1297. doi: 10.1007/BF01245072. URL https://doi.org/10.1007/BF01245072.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007. ISSN 1573-1375. doi: 10.1007/s11222-007-9033-z. URL https://doi.org/10.1007/s11222-007-9033-z.

Jinxin Wang, Yuen-Man Pun, Xiaolu Wang, Peng Wang, and Anthony Man-Cho So. Projected Tensor Power Method for Hypergraph Community Recovery. In *Proceedings of the 40th International Conference on Machine Learning*, pages 36285–36307. PMLR, July 2023. URL https://proceedings.mlr.press/v202/wang23af.html. ISSN: 2640-3498.

Xiaobo Wang, Xiaojie Guo, Zhen Lei, Changqing Zhang, and Stan Z. Li. Exclusivity-Consistency Regularized Multi-View Subspace Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 923–931, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Wang_Exclusivity-Consistency_Regularized_Multi-View_CVPR_2017_paper.html.

Xiaobo Wang, Zhen Lei, Xiaojie Guo, Changqing Zhang, Hailin Shi, and Stan Z. Li. Multi-view subspace clustering with intactness-aware similarity. *Pattern Recognition*, 88:50–63, April 2019. ISSN 0031-3203. doi: 10.1016/j.patcog.2018.09.009. URL https://www.sciencedirect.com/science/article/pii/S0031320318303285.

Joachim Weidmann. *Linear Operators in Hilbert Spaces*, volume 68 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1980. ISBN 978-1-4612-6029-5 978-1-4612-6027-1. doi: 10.1007/978-1-4612-6027-1. URL http://link.springer.com/10.1007/978-1-4612-6027-1.

Alexander S. Wein, Ahmed El Alaoui, and Cristopher Moore. The Kikuchi Hierarchy and Tensor PCA. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1446–1468. IEEE Computer Society, November 2019. ISBN 978-1-7281-4952-3. doi: 10.1109/FOCS.2019.000-2. URL https://www.computer.org/csdl/proceedings-article/focs/2019/495200b446/1grNBDzixFe.

Beth Whitehead, Deborah Andrews, Amip Shah, and Graeme Maidment. Assessing the environmental impact of data centres part 1: Background, energy use and metrics. *Building and Environment*, 82: 151–159, December 2014. ISSN 0360-1323. doi: 10.1016/j.buildenv.2014.08.021. URL https://www.sciencedirect.com/science/article/pii/S036013231400273X.

Eugene P. Wigner. Characteristic Vectors of Bordered Matrices With Infinite Dimensions. *Annals of Mathematics*, 62(3):548–564, 1955. ISSN 0003-486X. doi: 10.2307/1970079. URL https://www.jstor.org/stable/1970079. Publisher: Annals of Mathematics.

Eugene P. Wigner. On the Distribution of the Roots of Certain Symmetric Matrices. *Annals of Mathematics*, 67(2):325–327, 1958. ISSN 0003-486X. doi: 10.2307/1970008. URL https://www.jstor.org/stable/1970008. Publisher: Annals of Mathematics.

Alex H. Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I. Ryu, Krishna V. Shenoy, Mark Schnitzer, Tamara G. Kolda, and Surya Ganguli. Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron*, 98(6):1099–1115.e8, June 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.05.015. URL https://www.sciencedirect.com/science/article/pii/S0896627318303878.

Eric Williams. Environmental effects of information and communications technologies. *Nature*, 479 (7373):354–358, November 2011. ISSN 1476-4687. doi: 10.1038/nature10682. URL https://www.nature.com/articles/nature10682. Publisher: Nature Publishing Group.

John Wishart. The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population. *Biometrika*, 20A(1/2):32–52, 1928. ISSN 0006-3444. doi: 10.2307/2331939. URL https://www.jstor.org/stable/2331939. Publisher: [Oxford University Press, Biometrika Trust].

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. Sustainable AI: Environmental Implications, Challenges and Opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, April 2022. URL https://proceedings.mlsys.org/paper_files/paper/2022/hash/462211f67c7d858f663355eff93b745e-Abstract.html.

Jianlong Wu, Zhouchen Lin, and Hongbin Zha. Essential Tensor Learning for Multi-View Spectral Clustering. *IEEE Transactions on Image Processing*, 28(12):5910–5922, December 2019. ISSN 1941-0042. doi: 10.1109/TIP.2019.2916740. Conference Name: IEEE Transactions on Image Processing.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv:1708.07747 [cs, stat]*, September 2017. URL http://arxiv.org/abs/1708.07747. arXiv: 1708.07747.

Yangyang Xu. On the convergence of higher-order orthogonal iteration. *Linear and Multilinear Algebra*, 66(11):2247–2265, November 2018. ISSN 0308-1087. doi: 10.1080/03081087.2017.1391743. URL https://doi.org/10.1080/03081087.2017.1391743. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03081087.2017.1391743.

Shinjae Yoo, Hao Huang, and Shiva Prasad Kasiviswanathan. Streaming spectral clustering. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 637–648, May 2016. doi: 10.1109/ICDE.2016.7498277.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, September 2016. ISSN 0001-8732, 1460-6976. doi: 10.1080/00018732.2016.1211393. URL http://arxiv.org/abs/1511.02476. arXiv:1511.02476 [cond-mat, stat].

Anru Zhang and Dong Xia. Tensor SVD: Statistical and Computational Limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, November 2018. ISSN 1557-9654. doi: 10.1109/TIT.2018.2841377. URL https://ieeexplore.ieee.org/document/8368145. Conference Name: IEEE Transactions on Information Theory.

Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent Multi-View Subspace Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4279–4287, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhang_Latent_Multi-View_Subspace_CVPR_2017_paper.html.

Liangpei Zhang, Lefei Zhang, Dacheng Tao, and Xin Huang. Tensor Discriminative Locality Alignment for Hyperspectral Image Spectral–Spatial Feature Extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):242–256, January 2013. ISSN 1558-0644. doi: 10.1109/TGRS.2012.2197860. URL https://ieeexplore.ieee.org/abstract/document/6213108. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, SIGMOD '96, pages 103–114, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 978-0-89791-794-0. doi: 10.1145/233269.233324. URL https://doi.org/10.1145/233269.233324.

Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary Multi-View Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1774–1782, July 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2847335. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor Regression with Applications in Neuroimaging Data Analysis. *Journal of the American Statistical Association*, 108(502):540–552, June 2013. ISSN 0162-1459. doi: 10.1080/01621459.2013.776499. URL https://doi.org/10.1080/01621459.2013.776499. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2013.776499.

Alaettin Zubaroğlu and Volkan Atalay. Data stream clustering: a review. *Artificial Intelligence Review*, 54(2):1201–1236, February 2021. ISSN 1573-7462. doi: 10.1007/s10462-020-09874-x. URL https://doi.org/10.1007/s10462-020-09874-x.