ANR-14-CE28-0006 RMT4GRAPH Project.

Deliverable D1:
Report on investigations at $t_0 + 1$ year

1 Project Summary at $t_0 + 1$ year

We first provide a summary of the research activities and contributions as of September 2016 and an update on the technical program.

Important Note: *Since the main funding of the RMT4GRAPH project concerns the hiring of a PhD student who could not be found by October 2014 (date when the project was funded), the project was officially kicked-off on October 1st, 2015 and will be completed on September 30th, 2018.*

1.1 Project Progress

Due to the one-year delay before the official project kick-off, the position of the project as of October 2016 is well ahead of time as compared to the promised output in the original program. Notably, while the original plan was for delivery D1 to provide a mere literature review and position with respect to the project objectives, the present updated D1 delivery shall provide more substantial inputs and results.

In detail, an important progress on the tasks T1.1 and T1.2 of WP1 was made. This progress is the consequence of a major contribution by the PI and his collaborator Florent Benaych-Georges who published together two articles who set the stage for an expectedly fast development of all subsequent works. This fast development was further accelerated by the numerous high potential intern students and visitors who joined the group and contributed to the project. We notably estimate that T1.1 is 1/2 complete and T1.2 is almost complete. Task T1.3 was not investigated so far as our early studies did not reveal the necessity for immediate investigations in this direction. It is proposed to replace T1.3 by the analysis of random matrices with non-linear entries, of more fundamental reach notably in neural network applications than non-Hermitian random matrices. As for WP2, in its original description at submission time, it is almost entirely covered. We therefore updated the content of WP2 to encompass more studies than we originally assumed could be covered in the allotted time. Former T2.1 is almost complete and was expanded to cover new aspects of kernel methods: semi-supervised learning and the technically more challenging support vector machines; T2.2 is also well underway (estimated at 1/3 complete) and T2.3 has known significant developments (however restricted so far to the less challenging linear neural network setting).

1.2 Key Outputs

The major outputs and contributions of the project can be summarized as follows.

WP1. Random Matrix Models for Random Graphs.
This work package proposes the mathematical investigation of new families of random matrices of important use in machine learning applications. In the course of the first (unofficial) year of the project, important advances were made in the mathematical understanding of kernel random matrices, which kick-started many of the investigations of WP2 programmed for a later time. The deeper investigation of structured spiked random matrix models naturally unfolded. In parallel, an (off-project opportunistic) neural network investigation was carried out which also kick-started parts of the project themes and naturally led to the development of a new project branch on random matrices with non-linear or recursively defined entries.

- **Task 1.1. Kernel random matrix models.** By assuming a properly scaled Gaussian-mixture for the input of a kernel random matrix model, we proved that a concentration effect emerges by which the kernel can be linearised by Taylor series expansion. This linearisation produces an asymptotically close approximation of the kernel matrix which, if not completely standard in classical random matrix studies, is amenable to theoretical analysis. This constitutes a major breakthrough which allows for the understanding of many kernel methods in machine learning

when applied to Gaussian mixture inputs.
$\longrightarrow$ *This study unfolded in two major publications [1, 2] in the Electronic Journal of Statistics (EJS) and ESAIM: Probability and Statistics. Many invited talks were solicited on these articles.*

- **Task 1.2. Hermitian models and spikes.** The objective of the task is a deeper study of certain models of spiked random matrices, notably those models exhibiting strong structures such as finitely many classes of vector distributions (as in Gaussian mixtures). The latter models present eigenvectors looking like noisy staircase vectors; the objective is to quantity the staircase levels and the noise variances as a function of the deterministic system parameters. As a follow-up of the preliminaries of the works [1, 2], we obtained such quantitative relations and notably obtained a limiting multivariate Gaussian behavior for the dominant so-many eigenvectors of such models.
$\longrightarrow$ *This study was exploited both in publications to kernel clustering [1] and community detection in graphs [3].*

- **Task 1.3. Random matrices with non-linear or recursive entries.** *(formerly "Task 1.3. Non-Hermitian random matrix models").* This task, initially dedicated to non-Hermitian random matrices, was motivated by several aspects of neural networks (such as stability related to largest isolated eigenvalues) as well as community detection on non-symmetric graphs (with notably the non-backtracking operator method). However, the former aspect is a very specific *qualitative* rather than quantitative aspect of neural networks, disrupting the flow of our quantitative investigation of the performance of neural networks, while the latter non-backtracking operator turned out to be recently supplanted by a more powerful Hermitian operator (the Bethe Hessian matrix). In the meantime, very recent developments within our group working in neural networks put forth the need for a deeper investigation of matrices with recursive dependence between columns or with non-linear entries (however not in the simpler form of kernels of large vectors). Advances were made in the study of random matrices with recursively defined columns, however only so far in the linear setting.
$\longrightarrow$ *The main outcome of this study is the understanding of the performance of the recursively defined echo-state networks [4].*

WP2. Applications to Big Data Processing.
This work package involves the applications of the mathematical toolbox comprising the outcomes of WP1. Our main findings in this respect are threefold: an immediate application to kernel spectral clustering unfolding from the study of kernel random matrices, a parallel application of T1.2 to community detection on realistic random graphs, and an application to the results of T1.3 to the performance analysis of linear echo-state networks.

- **Task 2.1. Applications to machine learning.** The spectral analysis of large dimensional kernel random matrices, based on Gaussian mixture inputs, provided an asymptotic tractable random matrix equivalent of the kernel matrices, from which the existence of isolated (spiked) eigenvalues was studied. This allowed us to exhibit a phase transition phenomenon below which clustering is asymptotically impossible. Beyond the phase transition though, the eigenvectors associated with the spiked eigenvalues carry structural information under the form of a noisy staircase aspect of these eigenvectors. A thorough investigation of the plateaus and joint fluctuations of the eigenvectors about these plateaus allowed for a precise estimation of the probability of correct clustering. More importantly, although the study is confined to idealistic Gaussian inputs, the application to actual image datasets revealed a sharp connection between theory and practice, thereby providing strong indication of the usefulness of our results to actual large dimensional data clustering. A further application to the specific context of subspace clustering (for which a whole new study was made under a different convergence rate) was also performed.
$\longrightarrow$ *These results are reported in [1, 5].*

- **Task 2.2. Signal processing on graphs.** Paralleling the study of kernel methods, but still exploiting the developments on structured spiked random matrix models, we obtained new results on community detection on dense "realistic" graphs. That is, assuming a network with communities and heterogeneous degree distribution for each node, we derive a family of modified spectral clustering algorithms (depending on a regularization parameter to handle the deleterious effects of degree heterogeneity), the performance of which is thoroughly studied as the number of nodes tends to infinity. An on-line method is further derived which selects among the (asymptotically) most appropriate regularization parameter. Comparisons to state-of-the-art methods suggest strong advantages of our novel method and comparable performance in benchmark community detection networks.
$\longrightarrow$ *Early results have been published in [3].*

- **Task 2.3. Neural networks.** *(formerly restricted to "Task 2.3. Echo-state neural networks")* In the scope of an informal collaboration with the computer science department of ENS Paris, a performance study of linear

echo-state networks was performed, which allows to understand some underlying mechanisms of recurrent neural networks. Echo-state networks are convenient to study as opposed to traditional recurrent neural networks as they suppose neuronal connections to be all fixed but for the output layer which is learned by linear regression: while not capable of performing tremendous tasks like back-propagated networks, this setup ensures stability and few hyper-parameters to be tuned. Our initial studies, restricted to (the least interesting) linear networks, provide strong insights and quantitative figures on the adequate tuning of the hyper-parameters and reveal an enhanced notion of memory capacity of the network.
In the course of this study, it turned out that the major difficulty in addressing neural networks performance lies primarily, not in the involved recursive nature of networks such as echo-state networks, but rather in the *non-linearity* of the neuronal activations (which cannot be handled via Taylor series expansions as for kernel random matrices), therefore shifting our investigation from echo-state networks to the much broader scope of neural networks, starting with the simplest *extreme learning machines* to then move on to more involved structures, hence the change of objectives in Task 2.3.
$\longrightarrow$ *These results are reported in [6, 7].*

Publications and Dissemination.

As of September 2016, the number of publications falling in the scope of the RMT4GRAPH project, submitted articles and on-going works apart, is of **5 conference articles** and **2 journal articles**. These are listed below.

Published Journal Articles.

- R. Couillet, F. Benaych-Georges, "Kernel Spectral Clustering of Large Dimensional Data", Electronic Journal of Statistics, vol. 10, no. 1, pp. 1393-1454, 2016.
- F. Benaych-Georges, R. Couillet, "Spectral Analysis of the Gram Matrix of Mixture Models" (in Press), ESAIM: Probability and Statistics, 2016.

Published Conference Articles.

- R. Couillet, G. Wainrib, H. Sevi, H. Tiomoko Ali, "A Random Matrix Approach to Echo-State Neural Networks", International Conference on Machine Learning (ICML), New York, USA, 2016.
- A. Kammoun, R. Couillet, F. Pascal, M. Slim-Alouini, "Optimal Design of Adaptive Normalized Matched Filter For Large Antenna Arrays", IEEE Statistical Signal Processing Workshop (SSP), Palma de Majorca, Spain, 2016.
- R. Couillet, G. Wainrib, H. Sevi, H. Tiomoko Ali, "Training performance of echo state neural networks", IEEE Statistical Signal Processing Workshop (SSP), Palma de Majorca, Spain, 2016.
- H. Tiomoko Ali, R. Couillet, "Performance analysis of spectral community detection in realistic graph models", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16), Shangai, China, 2016.
- R. Couillet, F. Benaych-Georges, "Understanding Big Data Spectral Clustering", IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP'15), Cancun, Mexico, 2015.

In terms of dissemination, many actions were already undertaken: class lectures in master programs, larger audience workshops, special sessions, invited talks, special issues, etc. The most important of those actions are listed next.

- **Joint ANR-DIONISOS and ANR-RMT4GRAPH Summer School on "Large Random Matrices and High Dimensional Statistical Signal Processing" (Telecom ParisTech, June 7-8, 2016).** In this two-day joint event co-organized by the ANR DIONISOS and the ANR RMT4GRAPH, lectures on advances of random matrix theory in signal processing and machine learning were proposed to a large audience, mostly composed of researchers in France. The summer school was consituted of four 3h-courses as follows: (i) Jamal Najim: Introduction to large random matrix theory, (ii) Philippe Loubaton: Large random matrices for array processing, (iii) Abla Kammoun: Robust estimation in large systems, (iv) Romain Couillet: Random matrices and machine learning.

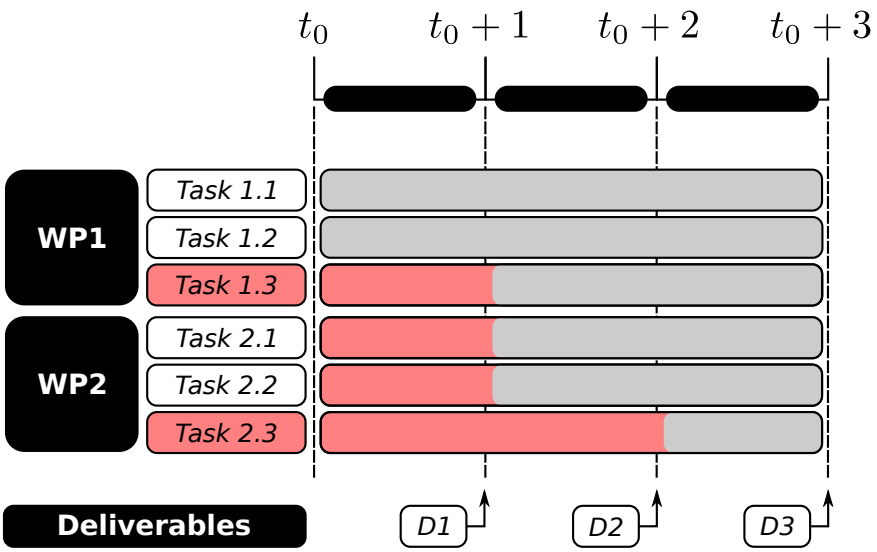
- **Special Session "Random matrices in signal processing and machine learning" at the Statistical Signal Processing Workshop (SSP'16).** A special session at the 2016 Statistical Signal Processing Workshop (SSP'16), Palma de Majorca (Spain), was organized by the PI. The session was composed of 7 poster presentations (ranging from theoretical random matrix theory to applications to machine learning and array processing).
- **Distinguished keynote speaker at EUSIPCO 2016.** The early results of the ANR-RMT4GRAPH project were presented in a large audience at the European Signal Processing Conference (EUSIPCO) in September 2016, as part of the STATOS Thematic Workshop on Machine Learning and BigData. The PI was there invited as a distinguished keynote speaker.
- **Special Issue on Random Matrices in "Revue du Traitement du Signal".** A special issue on Random Matrices and its applications to signal processing and machine learning was edited by the PI. The special issue contains 6 articles ranging from introduction to basic notions of random matrices to advanced applications in robust statistics and machine learning.
- **Invited talks and contributions to local events.** As a follow up of some of the key publications above, several invited talks were given by the PI and co-authors (to ENS Paris twice, to ENS Lyon twice, to the University of Orsay, etc.). A willingness to broadcast the results of RMT4GRAPH was also ensured by proposing talks in various GdR and local meetings.

1.3 Updated Program and Timeline

As previously mentioned, the main changes in the program and timeline are as follows:

- **Task 1.3.** The topic of Task 1.3 was altered from "Non-Hermitian random matrix models" to "Random Matrices with non-linear or recursive entries". While we do not exclude the study of non-Hermitian random matrix models, especially in the context of (the adjacency matrix of) directed graphs, a more pressing need for the study of random matrices with non-linear entries (and possibly recursive) appeared in our early investigations of neural networks. Our initial assumption was that non-linear neural networks would behave similar (from a mathematical viewpoint) to kernel random matrices in allowing for Taylor expansions of the non-linearities; thus Task 1.1 was deemed sufficient for the development of applications within Task 2.3. This turned out not to be a valid guess, which opens a new avenue of research in the direction of random matrices with non-linear entries. Very recent calculus suggest that this direction, although technically complex at the onset, promises strikingly new results of deep importance to machine learning as a whole. The less challenging but equally needed recursive aspect of the new task (mandatory for the study of recurrent neural networks) has already been opened and provided its first results.
- **Task 2.3.** A collaboration with colleagues at ENS Paris accelerated the progress of Task 2.3 for which several publications appeared and a journal article has been submitted. Although restricted to the linear setting so far (due again to the lack of mathematical methods at this point to handle the more interesting non-linear case), these publications constitute a major piece of the originally promised outputs for Task 2.3. As such, we decided on an extension of this task to more general purpose neural networks. To be more exact, since the mathematical key resides in tackling random matrices with non-linear entries, Task 2.3 shall now be organized in a natural progression from the study of elementary *non-linear* neural networks (such as the single-layer extreme learning machine) down to more elaborate settings.

As a consequence of the full exploitation of the one-year delay allowed by the ANR to kick-off the project (which followed from the impossibility to find an appropriate PhD student in time), the PI made significant progress on the technical program prior to the project official kick-off date. This (positively) shifted the project time frame towards earlier results and findings than initially assumed. In particular, Work Package 2, which we assumed could not be provided with any input before conclusive technical results are achieved within Work Package 1, already contains significant contributions. In the figure below are depicted in pale red the modifications to the initial program and timeline.



1.4 Research Group

As argued in the project proposal, the ANR RMT4GRAPH funding serves as a springboard to the building of a research group focusing on the study of *large dimensional machine learning tools and methods*. Along with external fundings, collaborative works and student exchanges, we are happy to say that this objective is already well underway, after only one year within the project duration. As of today, eight PhD and intern students have contributed to the project, with a culminating current five students within the team.

The students, collaborators and contributions (all related to the topic scope of the ANR RMT4GRAPH but not necessarily falling within an ANR support) are listed below.

1.4.1 Collaborators.

The collaborations below all are all informal and do not in particular fall within any funding from the ANR. Nonetheless, the ANR-RMT4GRAPH project is a strong driver of incentives to open up collaborations to research centers sharing common scientific interests. This is in particular the case with Professor Benaych-Georges on kernel random matrices as well as Assistant Professor Gilles Wainrib on (echo-state) neural networks.

- **Florent Benaych-Georges** (professor at Université Paris Descartes), on kernel random matrices.
  - 2 journal articles (EJS, ESAIM Probability and Statistics)
  - 1 conference article (CAMSAP 2015)
- **Gilles Wainrib** (assistant professor at ENS Paris), on neural networks.
  - 2 conference articles (ICML 2016, SSP 2016)
  - 1 journal article submitted (JMLR)
- **Matthew M. McKay** (professor at Hong Kong UST), on sparse PCA and applied robust estimation.
  - 1 journal article (IEEE-TSP)
  - 2 conference article (CAMSAP 2015, SSP 2016)
- **Abla Kammoun** (research scientist at KAUST), on subspace clustering and robust estimation.
  - 1 conference article (Asilomar 2016)
  - 1 journal article under preparation.

1.4.2 Students.

All the students listed below have worked on topics covered by the ANR-RMT4GRAPH project. Hafiz Tiomoko Ali is the only "permanent" PhD student of the group, funded himself by the RMT4GRAPH grant.

- **Hafiz Tiomoko Ali** (PhD student under RMT4GRAPH grant, sep. 2015-2018), on community detection and neural networks.
  - 4 conference articles (ICASSP 2015, Asilomar 2016, ICML 2016, SSP 2016)
  - 1 journal article submitted (JMLR)
  - 1 journal article under preparation (JMVA)
- **Xiaoyi Mai** (intern, under ERC-MORE grant, 2016), on large dimensional semi-supervised learning performance.
  - *Prospective PhD student within the group*
  - 1 conference article under preparation (ICASSP 2017)
- **Zhenyu Liao** (intern, under ERC-MORE grant, 2016), on large dimensional support vector machines performance.
  - *Prospective PhD student within the group*
  - 1 conference article under preparation (ICASSP 2017)
- **Cosme Louart** (intern, under ERC-MORE grant, 2016), on neural networks and random matrices (extreme learning machines).
  - 1 conference article under preparation (ICASSP 2017)
- **Harry Sevi** (intern, under ERC-MORE grant, 2015), on recurrent (echo-state) neural networks.
  - 2 conference articles (ICML 2016, SSP 2016)
  - 1 journal article submitted (JMLR)
- **Liusha Yang** (PhD student, visiting from HKUST, 2015), on financial applications of robust estimation.
  - Visiting student from Hong-Kong UST
  - 2 conference articles (Asilomar 2014, CAMSAP 2015)
  - 1 journal article (IEEE-TSP)
- **Evgeny Kusmenko** (PhD student under ERC-MORE grant, jan. 2015-dec. 2015), on spectral clustering methods.
- **Aymeric Thibault** (intern, under ERC-MORE grant, 2015), on eigenvectors of sample covariance matrices, applied to clustering.

2 Technical Aspects

2.1 Random Matrix Methods (within Tasks 1.1 and 1.2)

2.1.1 Kernel Random Matrices

Motivated by a large family of applications in machine learning tools using *kernel representations*, this section deals with the asymptotic behavior of the *random kernel matrix*

$$K \equiv \{\kappa(x_i, x_j)\}_{i,j=1}^{n}$$

where $\kappa(x, y)$ is some *affinity* function between the vectors $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^p$. Two types of kernel functions are common in the literature: (i) the radial kernel with $\kappa(x, y) = f(\|x - y\|^2)$ and (ii) the inner product kernel $\kappa(x, y) = f(x^\mathsf{T} y)$, for some function f.

Our interest is here on the more popular (and technically more involved) radial kernel. Precisely, with an added normalization for simplicity of exposition, we consider the kernel matrix:

$$K \equiv \left\{ f\left(\frac{1}{p} \|x_i - x_j\|^2 \right) \right\}_{i,j=1}^{n}$$

where f will be taken smooth enough (precisely, we shall request it to be at least three times differentiable).

The objective is to study the asymptotic behavior of K as $n, p \to \infty$ with p/n away from zero and infinity, when the vectors x_i are independent random vectors extracted from a Gaussian mixture of k classes. Precisely, we assume that there exist k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$, and that $x_1, \ldots, x_{n_1} \in \mathcal{C}_1$, up to $x_{n-n_k}, \ldots, x_n \in \mathcal{C}_k$, where

$$x \in \mathcal{C}_a \Leftrightarrow x \sim \mathcal{N}(\mu_a, C_a).$$

For simplicity, we also assume that, for each a, $n_a = |\mathcal{C}_a|$ is such that n_a/n remains away from zero as $n \to \infty$.

Since the ultimate goal is to study the performance of spectral clustering, support vector machines, or semi-supervised learning methods based on kernels, our interest is precisely focused on the case where *the classes are barely distinguishable*. That is, we shall enforce that, as $n, p \to \infty$, the probability to state that x_i belongs to its proper class, based solely on matrix K, should not tend to one (and obviously, should not always be zero). A careful study of the matrix K reveals that the proper conditions for this to be achieved is that the following growth rates be simultaneously ensured.

Assumption 1 (Growth Rate) *As $n \to \infty$, the following conditions hold.*

1. **Data scaling**: *defining $c_0 \triangleq \frac{p}{n}$*

$$0 < \liminf_n c_0 \leq \limsup_n c_0 < \infty$$

2. **Class scaling**: *for each $a \in \{1, \ldots, k\}$, defining $c_a \triangleq \frac{n_a}{n}$,*

$$0 < \liminf_n c_a \leq \limsup_n c_a < \infty.$$

   We shall denote $c \triangleq \{c_a\}_{a=1}^k$.

3. **Mean scaling**: *let $\mu^\circ \triangleq \sum_{i=1}^k \frac{n_i}{n} \mu_i$ and for each $a \in \{1, \ldots, k\}$, $\mu_a^\circ \triangleq \mu_a - \mu^\circ$, then*

$$\limsup_n \max_{1 \leq a \leq k} \|\mu_a^\circ\| < \infty$$

4. **Covariance scaling**: *let $C^\circ \triangleq \sum_{i=1}^k \frac{n_i}{n} C_i$ and for each $a \in \{1, \ldots, k\}$, $C_a^\circ \triangleq C_a - C^\circ$, then*

$$\limsup_n \max_{1 \leq a \leq k} \|C_a\| < \infty$$
$$\limsup_n \max_{1 \leq a \leq k} \frac{1}{\sqrt{n}} \operatorname{tr} C_a^\circ < \infty.$$

For further use, we add the definition

$$\tau \triangleq \frac{2}{p} \operatorname{tr} C^\circ.$$

This quantity is important since it is easily shown that, under Assumption 1,

$$\max_{\substack{1 \leq i,j \leq n \\ i \neq j}} \left\{ \frac{1}{p} \|x_i - x_j\|^2 - \tau \right\} \to 0 \tag{2.1}$$

almost surely. This result is at the same time mathematically extremely interesting as it shall allow for a natural Taylor expansion of all (non-diagonal) entries of K around the limiting τ but is quite intriguing on the onset. Indeed, since all entries of K have the same limit, it seems unlikely that one can retrieve the class-indentity of the x_i's from K.

This is however made possible thanks to a "redundancy" effect within K that shall bring forward *structured dominant eigenvectors containing the class information.*

To proceed with the Taylor expansion of the matrix K, note that it is fundamental not to control the residual matrix terms through the amplitude of their respective entries but through their *operator norms*. As such, while entries are individually Taylor expanded, what matters most in the approximation lies in the spectral norm of the successively obtained matrices. This leads to non obvious (and rather painstaking) calculus. As an example, note that, letting $X = \{X_{ij}\}_{i,j=1}^n$ be a matrix with, say, X_{ij} $\mathcal{N}(0,1)$, it is well-known that $\|X\| = O(\sqrt{n})$, while if instead $X_{ij} = 1$ (therefore of the same order of magnitude as if $\mathcal{N}(0,1)$), then $\|X\| = n$ and the two obtained matrices are not comparable in norm while their entries have comparable amplitudes. In performing this Taylor development, we follow here the tracks of [8] who first proposed the study of kernel random matrices, however only for matrices $[x_1, \ldots, x_n]$ with i.i.d. zero mean columns (in our case, this corresponds to a single-class scenario with $\mu_1 = 0$).

With this in mind, let $x_i = \mu_a + \sqrt{p} w_i$, with $w_i \sim \mathcal{N}(0, \frac{1}{p} C_a)$, whenever $x_i \in \mathcal{C}_a$. Then, we start by writing, for $x_i \in \mathcal{C}_a$ and $x_j \in \mathcal{C}_b$,

$$\begin{aligned}\frac{1}{p}\|x_j - x_i\|^2 &= \|w_j - w_i\|^2 + \frac{1}{p}\|\mu_b - \mu_a\|^2 + \frac{2}{\sqrt{p}}(\mu_b - \mu_a)^\mathsf{T}(w_j - w_i) \\ &= \tau + \frac{1}{p}\operatorname{tr} C_a^\circ + \frac{1}{p}\operatorname{tr} C_b^\circ + \psi_j + \psi_i - 2 w_i^\mathsf{T} w_j \\ &\quad + \frac{\|\mu_b - \mu_a\|^2}{p} + \frac{2}{\sqrt{p}}(\mu_b - \mu_a)^\mathsf{T}(w_j - w_i)\end{aligned}$$

in which we introduced the value $\psi_i = \|w_i\|^2 - \frac{1}{p}\operatorname{tr} C_a$ (for $x_i \in \mathcal{C}_a$). The object of this development is to introduce *redundant information* which, at the scale of the matrix K, shall play leading role in the dominant eigenspaces. This information is encapsulated in the deterministic values $\|\mu_b - \mu_a\|$, $\operatorname{tr} C_a^\circ$, etc., which appear in large blocks of size $n_a \times n_b$ in K. As for residues such as ψ_i, they are zero mean random variables independent across i and thus bring no redundancy.

At this point, we then need to evaluate the respective orders of amplitudes of each term in the expansion of $\frac{1}{p}\|x_j - x_i\|^2$ before applying the f function to it. Clearly (as as mentioned earlier), τ is the only term of non-vanishing order, and thus the Taylor expansion can be performed around τ. This means that, τ aside, all terms in the above formula shall be brought to successive powers. The operator norm of the resulting matrices will need to be individually evaluated, resulting finally, after intensive calculus and book-keeping, to the following main result which we only present in the case where $f'(\tau)$ is away from zero (other results are available in the other case).

Theorem 1 (Asymptotic Approximation for K) *In addition to Assumption 1, assume that $f'(\tau)$ remains bounded and bounded away from zero as $n \to \infty$. Then the following holds*

$$K = -2f'(\tau)\left(PW^\mathsf{T}WP + VAV^\mathsf{T}\right) + (f(0) - f(\tau) + \tau f'(\tau))I_n + O_{\|\cdot\|}(n^{-\frac{1}{2}})$$

where V is the $n \times (2k+4)$ matrix defined by

$$\begin{aligned}V &\triangleq \left[\frac{J}{\sqrt{p}}, v_1, \ldots, v_k, \tilde{v}, \psi^\circ, \sqrt{p}(\psi)^2, \sqrt{p}\tilde{\psi}\right] \\ v_a &\triangleq PW^\mathsf{T}\mu_a^\circ \\ \tilde{v} &\triangleq \left\{(WP)_a^\mathsf{T}\mu_a^\circ\right\}_{a=1}^k \\ \tilde{\psi} &\triangleq \operatorname{diag}\left(\left\{t_a \frac{1_{n_a}}{\sqrt{p}}\right\}_{a=1}^k\right)\psi\end{aligned}$$

with $P = I_n - \frac{1}{n}1_n 1_n^\mathsf{T}$, $W = [w_1, \ldots, w_n]$, $(\psi)^2$ the vector with entries ψ_i^2, $Y_a \in \mathbb{R}^{p \times n_a}$ the class-$\mathcal{C}_a$ submatrix of Y and

$A \triangleq A_n + A_{\sqrt{n}} + A_1$, *with* A_n, $A_{\sqrt{n}}$ *and* A_1 *the symmetric matrices*

$$A_n \triangleq -\frac{f(\tau)}{2f'(\tau)} p \begin{bmatrix} 1_k 1_k' & 0_{k\times k+4} \\ * & 0_{k+4\times k+4} \end{bmatrix}$$

$$A_{\sqrt{n}} \triangleq -\frac{1}{2}\sqrt{p} \begin{bmatrix} \{t_a + t_b\}_{a,b=1}^k & 0_{k\times k} & 0_{k\times 1} & 1_k & 0_{k\times 1} & 0_{k\times 1} \\ * & 0_{k\times k} & 0_{k\times 1} & 0_{k\times 1} & 0_{k\times 1} & 0_{k\times 1} \\ * & * & 0 & 0 & 0 & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & * & * & 0 & 0 \\ * & * & * & * & * & 0 \end{bmatrix}$$

$$A_1 \triangleq \begin{bmatrix} A_{1,11} & I_k & -1_k & -\frac{f''(\tau)}{2f'(\tau)} t & -\frac{f''(\tau)}{4f'(\tau)} 1_k & -\frac{f''(\tau)}{2f'(\tau)} 1_k \\ * & 0_{k\times k} & 0_{k\times 1} & 0_{k\times 1} & 0_{k\times 1} & 0_{k\times 1} \\ * & * & 0 & 0 & 0 & 0 \\ * & * & * & -\frac{f''(\tau)}{2f'(\tau)} & 0 & 0 \\ * & * & * & * & 0 & 0 \\ * & * & * & * & * & 0 \end{bmatrix}$$

$$A_{1,11} = \left\{ -\frac{1}{2}\|\mu_b - \mu_a\|^2 - \frac{f''(\tau)}{4f'(\tau)} (t_a + t_b)^2 - \frac{f''(\tau)}{f'(\tau)} \frac{1}{p} \operatorname{tr} C_a C_b \right\}_{a,b=1}^k .$$

The theorem suggests that, aside from a shift by a proportion of the identity matrix, K can be well approximated by a *"sort of" spiked random matrix model*. This model differs from the classically studied spiked models in that:

1. the matrix $PW^\mathsf{T}WP$ is not a straightforward sample covariance matrix model (a model well explored under the spiked hypothesis [9, 10]) as the columns of W are not identically distributed; in fact, $PW^\mathsf{T}WP$ may itself be seen as a spiked model as P introduced an isolated eigenvalue (but this has little consequences in our follow-up investigations);

2. the "information matrix" VAV^T also contains noise terms, usually not a considered setting; more technically challenging, these noise terms are *not independent* of W;

3. the matrix VAV^T contains terms of much higher orders than $PW^\mathsf{T}WP$, which on the onset poses important problems of generalization of classical spiked proof methods [11].

As shall be seen in the subsequent sections, while the spiked study of K is non trivial, alterations of K, as proposed in the machine learning literature (notably the normalized kernel matrix $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ where $D = \operatorname{diag}(K1_n)$), do not suffer all the problems of K itself. In particular, we shall see that the most problematic item 3) above is no longer a difficulty with the matrix $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$.

2.1.2 Structured Spiked Models

Spiked models appear naturally in machine learning methods, where they are often used in so-called *spectral methods*. A spectral method consists in general of the following multi-step approach:

1. Given data $x_1, \ldots, x_n$ (these could be vectors that one wishes to cluster or nodes of a graph from which communities should be extracted), form an affinity matrix $A \in \mathbb{R}^{n\times n}$ with A_{ij} evaluating the proximity between x_i and x_j. In particular, A may be a graph adjacency of (normalized, unnormalized) Laplacian matrix, or the (normalized or unnormalized) affinity kernel of vectors $x_1, \ldots, x_n$.

2. Extract the *so many* dominant eigenvectors $v_1, \ldots, v_\ell$ of A, where by dominant we mean those corresponding to extreme (smallest or largest in general) eigenvalues. The exact count of how many are needed is often directly related to the number of classes one needs the data to be shared into, although, as we shall subsequently see, this may not be the case.

3. Construct the matrix $V = [v_1, \ldots, v_\ell] \in \mathbb{R}^{n\times \ell}$ and, up to a possible additional row normalization, proceed to a popular low-dimensional clustering of the *rows* of A, seen as n vectors in $\mathbb{R}^\ell$, in k classes. Popular clustering methods are k-means, or EM (expectation maximization, which assumes the n vectors in $\mathbb{R}^\ell$ are issued from a multivariate Gaussian mixture).

4. The result of the clustering provides classes for the data $x_1, \dots, x_n$.

Our interest is to understand why such procedures work out and what are their limits. More precisely, under the various settings of investigation (community detection over graph or kernel spectral clustering), the interest is set over a refined analysis of the *content of the dominant eigenvectors*. Classical spiked model (say $Y = X + B$ with B the low rank perturbation and X a noise matrix) analyses are usually restricted to showing (i) the existence of finitely many isolated eigenvalues in Y outside a main *bulk* of close-by eigenvalues, and (ii) that the eigenvectors associated with these eigenvalues are *somehow aligned* to the deterministic eigenvectors of B (but orthogonal when the eigenvalues are not isolated). This raises a notion of phase transition by which, beyond a certain threshold, some eigenvalues tend to isolate and the eigenvectors associated with these eigenvalues are all the more correlated to B that the eigenvalues are far from the others.

In the applicative context of clustering, this shall carry an important piece of information: below some data-dependent threshold, clustering will be asymptotically impossible using spectral methods as no relevant information is carried within the dominant eigenvectors. However, the fact the mere information that dominant eigenvectors "correlate to some extent" to the eigenvectors of B will not provide us with an accurate enough measure of performance of spectral clustering. To this end, we need extra investigation. Precisely, in the context of a k-class clustering, the dominant eigenvectors take the form of *noisy step functions*, each step being mapped to one of the k classes. Our additional investigation then consists in:

1. evaluating the *average level of each plateau* of the eigenvector step functions
2. studying the fluctuations around each plateau average: in the simplest cases, each eigenvector entry within a given class fluctuates like a Gaussian random variable with a given (class-dependent) variance; in more advanced settings (such as in heterogeneous graphs), individual entries fluctuate with their own variance.

Let us place ourself in a generic spiked model context $Y = X + B \in \mathbb{R}^{n \times n}$, where B is a finite-rank ℓ matrix and X a rank $O(n)$-rank matrix, possibly dependent of B. Both matrices are supposed to be of operator norm $O(1)$. The relevant information is further supposed to be carried within the eigenvectors of B and, as $n \to \infty$, we assume that the eigenvalues of X converge to (one or several) connected components.

Assuming that the dominant eigenvectors of Y will be shaped as step functions with steps of sizes $n_1, \dots, n_k$ ($\sum_{a=1}^k n_a = n$), we may then write an individual eigenvector u_i^Y (say, the i-th dominant eigenvector of Y) as

$$u_i^Y = \sum_{a=1}^k \alpha_a^i \frac{j_a}{\sqrt{n_a}} + \sigma_a^i \omega_a^i$$

where $j_a \in \mathbb{R}^n$ is the indicator vector for the indices of plateau a (or class $\mathcal{C}_a$), $\omega_a^i \in \mathbb{R}^n$ is a random vector, orthogonal to j_a, of unit norm, supported on the indices of $\mathcal{C}_a$, where its entries are identically distributed. The scalars $\alpha_a^i \in \mathbb{R}$ and $\sigma_a^i \geq 0$ are the coefficients of alignment to j_a and the standard deviation of the fluctuations around $\alpha_a^i \frac{j_a}{\sqrt{n_a}}$, respectively.

Assuming when needed unit multiplicity for the eigenvalue associated with u_i^Y, our objective is now twofold:

1. *Class-wise Eigenvector Means.* We first wish to retrieve the values of the α_a^i's. For this, note that

$$\alpha_a^i = (u_i^Y)^\mathsf{T} \frac{j_a}{\sqrt{n_a}}. \tag{2.2}$$

   We shall evaluate these quantities by obtaining an estimator for the $k \times k$ matrix $\frac{1}{p} J^\mathsf{T} \hat{u}_i^\mathsf{T} \hat{u}_i^\mathsf{T} J$ with $J = [j_1, \dots, j_k] \in \mathbb{R}^{n \times k}$. The diagonal entries of the latter will allow us to retrieve $|\alpha_a^i|$ and the off-diagonal entries will be used to decide on the signs of $\alpha_1^i, \dots, \alpha_k^i$ (up to a convention in the sign of u_i^Y).

   As per classical spiked model analysis, each isolated eigenvalue-eigenvector pair (λ_i^Y, u_i^Y) is mapped to a corresponding eigenvalue-eigenvector pair (λ_i^B, u_i^B) of matrix B (for simplicity, let us take λ_i^B of unit multiplicity). Then the evaluation of α_a^i can be performed based on the Cauchy integral relation

$$\frac{1}{n} j_a^\mathsf{T} u_i^Y (u_i^Y)^\mathsf{T} j_a = -\frac{1}{2\pi\imath} \oint_{\Gamma_i} \frac{1}{n} j_a^\mathsf{T} (Y - zI_n)^{-1} j_a dz$$

for Γ_i a complex contour circling around λ_i^Y. A formal analysis of quadratic forms of the type $d^\mathsf{T}(Y-zI_n)^{-1}d$, exploiting in particular Woodbury's identity $(Y-zI_n)^{-1} = Q - QU_B(I_\ell + V_B^\mathsf{T} Q U_B)^{-1} V_B^\mathsf{T} Q$ with $Q = (X-zI_n)^{-1}$ and for some $U_B \in \mathbb{R}^{n\times \ell}$, $V_B \in \mathbb{R}^{n\times \ell}$ such that $U_B V_B^\mathsf{T} = B$, allows one to relate $\frac{1}{n} j_a^\mathsf{T} u_i^Y (u_i^Y)^\mathsf{T} j_a$ to deterministic functions of the resolvent Q of the matrix X. The method of *deterministic equivalents* [12, 13] provides asymptotic deterministic approximations for objects of the type $d_1^\mathsf{T} Q d_2$, which finally provides a deterministic complex integral representation for $\frac{1}{n} j_a^\mathsf{T} u_i^Y (u_i^Y)^\mathsf{T} j_a$. The evaluation of this resolvent can be made explicit using a residue calculus approach, which completes the method.

2. *Class-wise Eigenvector Inner and Cross Fluctuations.* Our second objective is to evaluate the quantities

$$\sigma_a^{i,j} \triangleq \left(u_i^Y - \alpha_a^i \frac{j_a}{\sqrt{n_a}}\right)^\mathsf{T} \mathcal{D}(j_a) \left(u_j^Y - \alpha_a^j \frac{j_a}{\sqrt{n_a}}\right) = (u_i^Y)^\mathsf{T} \mathcal{D}(j_a) u_j^Y - \alpha_a^i \alpha_a^j$$

between the fluctuations of two eigenvectors indexed by i and j on the subblock indexing $\mathcal{C}_a$, where $\mathcal{D}(x) \equiv \mathrm{diag}(x)$. In particular, letting $i = j$, $\sigma_a^{i,i} = (\sigma_a^i)^2$ from the previous definition (2.2). For this, it is sufficient to exploit the previous estimates and to evaluate the quantities $(u_i^Y)^\mathsf{T} \mathcal{D}(j_a)(u^Y)_j$. But, to this end, for lack of a better approach, we shall resort to estimating the more involved object

$$\frac{1}{n} J^\mathsf{T} u_i^Y (u_i^Y)^\mathsf{T} \mathcal{D}(j_a) u_j^Y (u_j^Y)^\mathsf{T} J$$

from which $(u_i^Y)^\mathsf{T} \mathcal{D}(j_a) u_j^Y$ can be extracted by division of any entry m, l by $\alpha_m^i \alpha_l^i$.

The actual computation of $\frac{1}{n} J^\mathsf{T} u_i^Y (u_i^Y)^\mathsf{T} \mathcal{D}(j_a) u_j^Y (u_j^Y)^\mathsf{T} J$ can be obtained via a double complex integral using twice the Cauchy theorem

$$\frac{1}{n} J^\mathsf{T} u_i^Y (u_i^Y)^\mathsf{T} \mathcal{D}_a u_j^Y (u_j^Y)^\mathsf{T} J = -\frac{1}{4\pi} \oint_{\Gamma_i} \oint_{\Gamma_j} \frac{1}{n} J^\mathsf{T} (Y - zI_n)^{-1} \mathcal{D}_a (Y - \tilde{z} I_n)^{-1} J \, dz d\tilde{z}$$

which is treated similarly as above using twice the Woodbury identity and requiring now to obtain deterministic equivalents for quantities of the type $d^\mathsf{T}(X - zI_n)^{-1} D (X - \tilde{z} I_n)^{-1} d$ for deterministic vector and matrix d, D.

This general approach is exploited in both sections below concerning kernel spectral clustering and community detection on graphs.

2.2 Kernel Spectral Clustering (within Task 2.1)

In this section, exploiting the mathematical results on both the large dimensional approximation of kernel matrices and the dominant eigenvectors of structured random matrix models, we discuss the study of so-called *kernel spectral clustering* methods. Most particularly, our focus is on the popular Ng–Weiss–Jordan algorithm [14] which consists in performing spectral clustering on the matrix

$$L = nD^{-\frac{1}{2}} K D^{-\frac{1}{2}}$$

where $D = \mathrm{diag}(K1_n)$ and n is used here as a (practically irrelevant) normalization factor. Many different matrix structures have been disputed in the machine learning literature (see [15, 16] for a review). Our own motivation for the study of L above is driven by the observation that, in the limit $n, p \to \infty$, aside from the main dominant eigenvalue of L equal to n and of unit multiplicity, all other eigenvalues are of order $O(1)$. Besides, the eigenvector associated with the eigenvalue n is exactly known to be $D^{\frac{1}{2}} 1_n$. Therefore, the spectral study of L reduces to the study of $D^{\frac{1}{2}} 1_n$ on the one hand, and of the (more interesting) matrix

$$L' \equiv L - n \frac{D^{\frac{1}{2}} 1_n 1_n^\mathsf{T} D^{\frac{1}{2}}}{1_n^\mathsf{T} D 1_n} \tag{2.3}$$

on the other.

We consider the k-class mixture Gaussian setting of Section 2.1.1. Using the results in Section 2.1.1 and the approach consisting in controlling operator norms of the various matrices in play, we easily find that D is a diagonal matrix dominated by $nf(\tau)$ in the first order, which is then easily developed in Taylor series, and similarly for D^α for any α. This further allows to evaluate the matrix $D^{\frac{1}{2}} 1_n 1_n^\mathsf{T} D^{\frac{1}{2}}$, the scalar $1_n^\mathsf{T} D 1_n$ and its inverse, to finally retrieve (after further painstaking calculus) the approximation for L' as follows.

Theorem 2 (Random Equivalent for L') *Let Assumption 1 hold and L' be defined as in (2.3). Then, as $n \to \infty$,*

$$\left\| L' - \hat{L}' \right\| \to 0$$

almost surely, where $\hat{L}'$ is given by

$$\hat{L}' \triangleq -2\frac{f'(\tau)}{f(\tau)}\left(PW^\mathsf{T}WP + UBU^\mathsf{T}\right) + 2\frac{f'(\tau)}{f(\tau)}F(\tau)I_n$$

with $F(\tau) = \frac{f(0)-f(\tau)+\tau f'(\tau)}{2f'(\tau)}$ and

$$U \triangleq \left[\frac{1}{\sqrt{p}}J, PW^\mathsf{T}M, \psi\right]$$

$$B \triangleq \begin{bmatrix} B_{11} & I_k - 1_k c^\mathsf{T} & \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t \\ I_k - c1_k^\mathsf{T} & 0_{k\times k} & 0_{k\times 1} \\ \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t^\mathsf{T} & 0_{1\times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix}$$

$$B_{11} = M^\mathsf{T}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^\mathsf{T} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}F(\tau)1_k 1_k^\mathsf{T}$$

where we defined the deterministic matrices and vectors

$$M \triangleq [\mu_1^\circ, \ldots, \mu_k^\circ] \in \mathbb{R}^{p\times k}$$

$$t \triangleq \left\{\frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ\right\}_{a=1}^k \in \mathbb{R}^k$$

$$T \triangleq \left\{\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ\right\}_{a,b=1}^k \in \mathbb{R}^{k\times k}$$

and the case $f'(\tau) = 0$ is obtained through extension by continuity ($f'(\tau)B$ being well defined as $f'(\tau) \to 0$).

Of interest here is the fact that both $PW^\mathsf{T}WP$ and UBU^T are matrices of operator norm $O(1)$, with $U \in \mathbb{R}^{n\times(2k+1)}$ of finite rank (but not in general of rank k). The following important comments can be made:

- the eigenvectors of UBU^T are related to the matrix $J = [j_1, \ldots, j_k]$ of the class-wise canonical vectors;
- the coefficients in B, which serve as "weights" on the vectors, depend on the following properties of the classes: their means M, their covariance traces t and their cross-covariance traces T;
- more importantly, the major drivers for the success or failure of spectral clustering are the values of the successive derivative of f evaluated at τ;
- for $f''(\tau) = 0$, T is discarded and therefore has no impact on the clustering; similarly, for $f'(\tau) = 0$, M is discarded and has no impact; finally, for $5f'(\tau) = 4f''(\tau)$, t is discarded and then plays no role in clustering.

As pointed out previously, the study of the eigenvectors of L then boils down to (i) the study of $D^{\frac{1}{2}}1_n$ and (ii) the study of the eigenvectors in the spiked random matrix model $PW^\mathsf{T}WP + UBU^\mathsf{T}$. These results take on involved forms and do not provide further insights in themselves on the performance of spectral clustering. These are documented in full in [1] and the companion article [2].

Of particular interest though is the application of our analysis to realistic spectral clustering. Indeed, while our study focuses on a Gaussian mixture assumption for the vectors $x_1, \ldots, x_n$, practical applications of spectral clustering concern the classification of images, time series, etc., which have no reason to be close to Gaussian vectors. In the second part of our study, we apply the theoretical results to the popular MNIST database clustering. That is, considering here images of zeros, ones and twos as the vectors x_i (as depicted in Figure 1), we apply spectral clustering à la Ng–Weiss–Jordan and compare the empirical results to the theoretical results that one would obtain if the x_i's were truly Gaussian vectors with means and covariances computed empirically from the $60,000$ images of the training database.

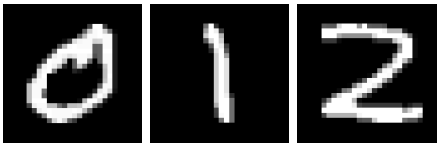


Figure 1: Samples from the MNIST database.

The results are provided in Figure 2 which depicts the four dominant eigenvectors of a snapshot of $n = 192$ evenly divided images versus the theoretical findings if the (vectorized) images were truly Gaussian vectors. These vectors (call them $u_1, \ldots, u_4$) are then collected into a matrix $U = [u_1, \ldots, u_4]$ the rows of which are used to perform clustering using k-means or EM. Since a four-dimensional representation is not practical, we depict in Figure 3 the 2-dimensional representations of $[u_1, u_2]$ and $[u_2, u_3]$. In both figures are provided in blue lines the means and one- (and two-) standard deviations obtained theoretically.

An astounding closeness is observed between theory and practice while, we recall, the dataset under investigation is far from a family of Gaussian vectors. This strongly suggests that our study, although initially motivated by a deeper understanding of the inner mechanism of spectral clustering, appropriately models more exotic datasets. As a matter of fact, it is quite intuitive from our analysis to understand the following phenomenon: real datasets that are difficult to cluster are inherently separable in the first place through their differences in empirical means, and then, if not enough, through their differences in second-order statistics. When $f''(\tau) = 0$, only the first order statistics play a significant role, which explains the shortcomings of classical approaches based on principal component analysis. However, when refining the function f so to push forward more differences visible only in second orders, more advanced clustering can be realized, hence the need for more elaborate kernels, such as Gaussian kernels $f(t) = \exp(-t^2/\sigma^2)$.

Further comments and analyses are made in the complete version of the article [1].

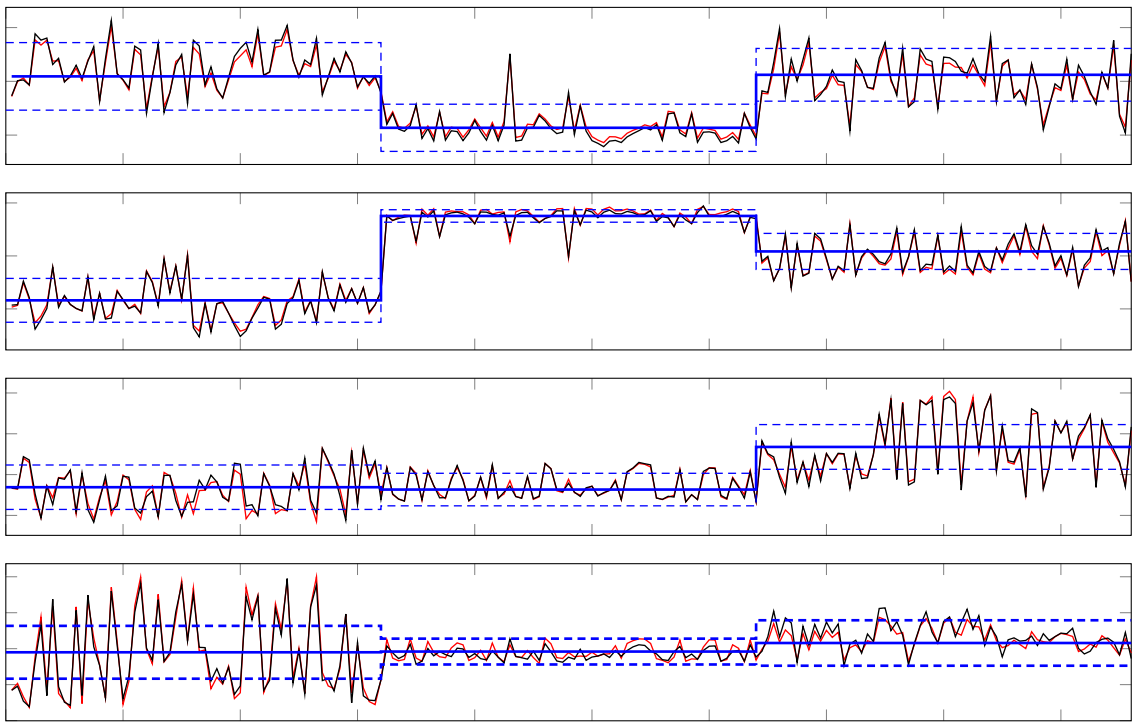


Figure 2: Leading four eigenvectors of L (red) versus $\hat{L}$ (black) and theoretical class-wise means and standard deviations (blue); MNIST data.

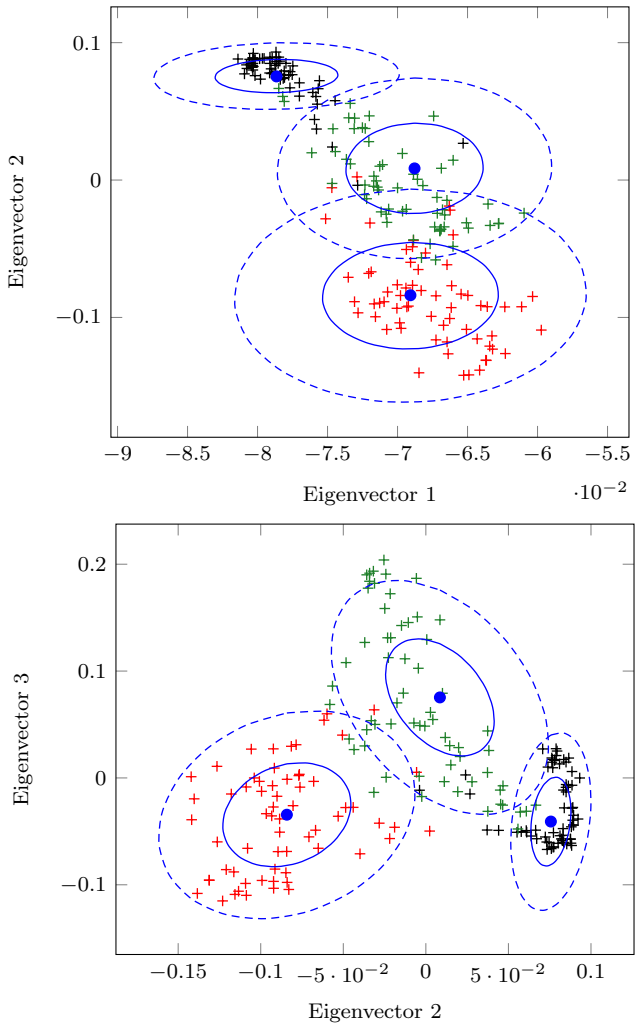


Figure 3: Two dimensional representation of the eigenvectors one and two (top) and two and three (bottom) of L, for the MNIST dataset. In blue, theoretical means and one- and two-standard deviations of fluctuations.

2.3 Community Detection on Graphs (within Task 2.2)

The investigation under this task is the study of community detection algorithms, a subject of heavy interest today in network mining [17]. Community detection algorithms on graphs are meant to find hidden clusters (based on closed similarities between the nodes) from an observation of the nodes links on this graph. Spectral algorithms, whose steps have been described in Section 2.1.2, are one of the popular methods to discover those hidden clusters. Our objective is to understand the performances and limits of those spectral methods, usually based on different normalizations of the adjacency (or modularity) matrix, on more realistic structured graph models. Most of the works in community detection consider the basic model for community structured graphs, the Stochastic Block Model (SBM). This model defines a matrix of edges probabilities $\mathbf{B}$ of size $K \times K$ (K being the number of communities) where B_{ab} represents the probability that node i belonging to community a can get connected to node j belonging to community b. The main limitation of this model though, is that it is more suited for homogeneous graphs where all nodes have the same average degree in each community. A degree-corrected version of the SBM, the Degree-Corrected SBM (DC-SBM), was proposed to take into account degree heterogeneity inside communities. Denoting $\mathcal{G}$ a K-class graph of n vertices with communities $\mathcal{C}_1, \dots, \mathcal{C}_K$ and letting q_i, $1 \leq i \leq n$, be some intrinsic weights which affect the probability for node i to connect to any other network node, the DC-SBM assumes an adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, with A_{ij} independent Bernoulli random variables with parameter $P_{ij} = q_i q_j C_{ab}$, for $i \in \mathcal{C}_a$ and $j \in \mathcal{C}_b$, where C_{ab} is a class-wise correction factor. Our analysis is based on this more realistic scenario.

Real world networks are in general sparse in the sense that the degrees of the nodes are insensitive to the addition of new nodes into the graph or equivalently the degree of each node scales in $\mathcal{O}(1)$ when the number of nodes n grows large. When the degrees scale instead like $\mathcal{O}(\log n)$ or $\mathcal{O}(n)$, the network is said to be dense. The standard spectral algorithms based on the network matrix (adjacency, modularity, Laplacian) of strongly sparse graphs are generally suboptimal in the sense that they fail to detect the communities down to the transition where the detection is theoretically feasible [18]. New operators (non-backtracking [18], Bethe Hessian [19]) based on statistical physics have recently been proposed and are shown to perform well down to the aforementioned sparse regime. However, those former methods were developed by assuming a Stochastic Block Model and we show through some simulations that they completely fail to detect communities in some pure heterogeneous graphs as well as the other classical spectral algorithms. To illustrate the aforementioned limitations of spectral methods under the DCSBM model, the top two graphs of Figure 4 provide 2D representations of dominant eigenvector 1 versus eigenvector 2 for the standard modularity matrix and the sparsity-improved BH matrix, when half the nodes connect with low probability q_1 and half the nodes with high probability q_2.

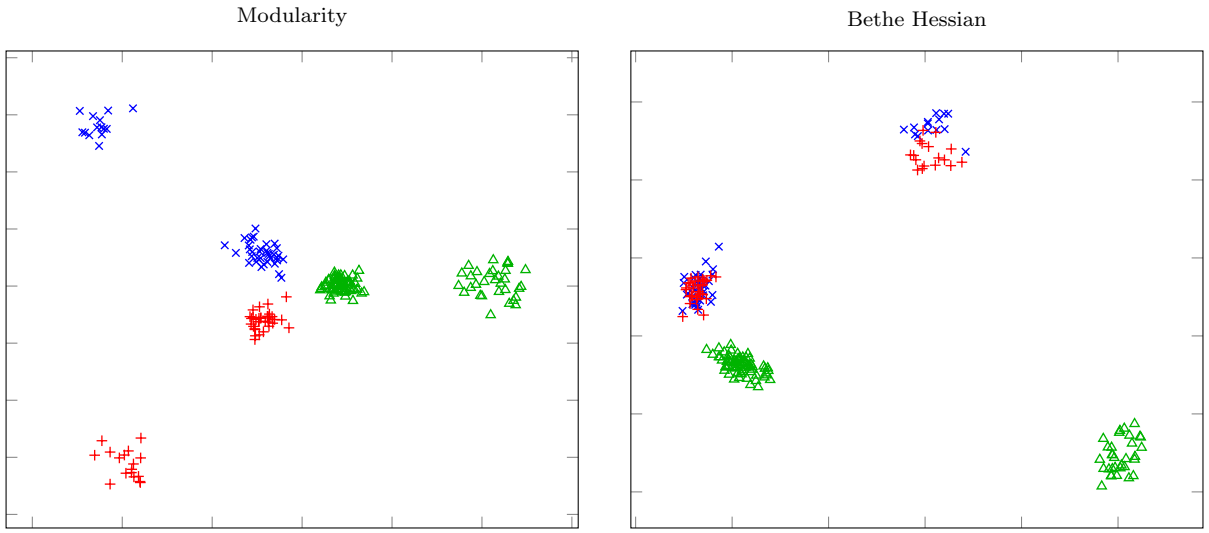


Figure 4: Two dominant eigenvectors (x-y axes) for $n = 2000$, $K = 3$ classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ of sizes $|\mathcal{C}_1| = |\mathcal{C}_2| = \frac{n}{4}$, $|\mathcal{C}_3| = \frac{n}{2}$, intrinsic probabilities taking two values $q_1 = 0.1$, $q_2 = 0.5$, matrix of weights $\mathbf{C} = \mathbf{1}_3\mathbf{1}_3^\mathsf{T} + \frac{100}{\sqrt{n}}\mathbf{I}_3$. Colors correspond to ground truth classes.

For both methods, this erroneously induces the detection of extra communities and even a confusion of genuine communities in the BH approach. We have come to understand that those extra communities are induced by some biases

created by the heterogeneity of the intrinsic probabilities q_i's; intuitively, nodes which have the same intrinsic connection probability tend to create their own sub-clusters inside each community and this creates somehow additional sub communities inside the genuine communities. To correct this, we have first proposed to normalize the adjacency/modularity matrix by $\mathbf{D}_{\hat{q}}^{-1}$ where $\mathbf{D}_{\hat{q}}$ is a diagonal matrix with estimates $\hat{q}_i$'s of the intrinsic weights on the diagonal. In order to achieve non trivial asymptotic error rates, our analysis is based on a non trivial regime where the class-wise correction factors C_{ab} differ by $\mathcal{O}(\frac{1}{\sqrt{n}})$. This trivial regime is ensured by the following growth rate conditions.

Assumption 2 *As $n \to \infty$, K remains fixed and, for all $i \in \{1, \ldots, n\}$:*

1. *$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}$ for $a, b \in \{1, \ldots, K\}$, where $M_{ab} = \mathcal{O}(1)$; we shall denote $\mathbf{M} = \{M_{ab}\}_{a,b=1}^{K}$.*
2. *$q_i \in (0, 1)$, $i \in \{1, \ldots, n\}$, are i.i.d. random variables with probability measure μ having compact support in $(0, 1)$. We shall denote $m_\mu = \int t\mu(dt)$.*
3. *$\frac{n_i}{n} \to c_i > 0$ and we will denote $\mathbf{c} = \{c_k\}_{k=1}^{K}$.*

Under Assumption 2, it is easily shown that

$$\max_{1 \le i \le n} \left| \hat{q}_i - \frac{d_i}{\sqrt{\mathbf{d}^\mathsf{T} \mathbf{1}_n}} \right| \to 0$$

so that the degree of node i is, up to a constant, uniformly consistent estimator of the intrinsic probability q_i.

Our study is then based on the matrix:

$$\mathbf{L} = 2m \frac{1}{\sqrt{n}} \mathbf{D}^{-1} \left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{2m} \right] \mathbf{D}^{-1} \tag{2.4}$$

where $\mathbf{D}$ is a diagonal matrix with the node degrees d_i's on the diagonal and m is the total number of edges.

The matrix $\mathbf{L}$ has non independent entries since $\mathbf{D}$ (and $\mathbf{d}$) depend on $\mathbf{A}$ and it does not follow a standard random matrix model. Our strategy is to approximate $\mathbf{L}_\alpha$ by a more tractable random matrix which asymptotically preserves eigenvalue distribution and isolated eigenvectors. The entries A_{ij} being random variables of mean $q_i q_j (1 + \frac{M_{g_i g_j}}{\sqrt{n}})$ and variance $q_i q_j (1 - q_i q_j) + \mathcal{O}(n^{-\frac{1}{2}})$, we may write A_{ij} as the sum of its mean and a random variable X_{ij} having zero mean and the same variance as A_{ij} . From there, we next provide a Taylor expansion of $\mathbf{d}\mathbf{d}^\mathsf{T}$, $(\mathbf{d}^\mathsf{T}\mathbf{1})^{-1}$, $(\mathbf{d}^\mathsf{T}\mathbf{1})^\alpha$ and $\mathbf{D}^{-\alpha}$ around their dominant terms, where $\mathbf{d} = \mathbf{A}\mathbf{1}_n$ and $\mathbf{D} = \mathcal{D}(\mathbf{d})$. By gathering all those expansions consistently following the structure of Equation (2.4) and by only keeping non-vanishing operator norm terms, we obtain the corresponding approximate of $\mathbf{L}_\alpha$ as follows:

Theorem 3 *Let Assumption 2 hold and let $\mathbf{L}$ be given by (2.4). Then, as $n \to \infty$, $\|\mathbf{L} - \tilde{\mathbf{L}}\| \to 0$ in operator norm, almost surely, where*

$$\begin{aligned}
\tilde{\mathbf{L}} &= \frac{1}{m_\mu^2} \left[\frac{1}{\sqrt{n}} \mathbf{D}_q^{-1} \mathbf{X} \mathbf{D}_q^{-1} + \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\mathsf{T} \right], \\
\mathbf{U} &= \left[\tfrac{\mathbf{J}}{\sqrt{n}} \quad \tfrac{\mathbf{D}_q^{-1} \mathbf{X} \mathbf{1}_n}{\mathbf{q}^\mathsf{T} \mathbf{1}_n} \right], \\
\mathbf{\Lambda} &= \begin{bmatrix} \left(\mathbf{I}_K - \mathbf{1}_K \mathbf{c}^T\right) \mathbf{M} \left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right) & -\mathbf{1}_K \\ -\mathbf{1}_K^T & 0 \end{bmatrix}.
\end{aligned}$$

with $\mathbf{D}_\mathbf{q} = \mathcal{D}(\mathbf{q})$ and $\mathbf{X} = \{X_{ij}\}_{i,j=1}^{n}$ has independent (up to symmetry) entries of zero mean and variances $\sigma_{ij}^2 = q_i q_j (1 - q_i q_j) + \mathcal{O}(n^{-\frac{1}{2}})$.

As for the kernel spectral clustering (Section 2), the matrix $\tilde{\mathbf{L}}$ follows an additive spike random matrix model $\tilde{\mathbf{L}} = \frac{1}{\sqrt{n}} \mathbf{D}_q^{-1} \mathbf{X} \mathbf{D}_q^{-1} + \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\mathsf{T}$ where:

- The eigenvectors of the deterministic low matrix $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ contain the class canonical vectors $\mathbf{J}$ meaning that when the eigenvalues of the former matrix are sufficiently large, the isolated eigenvectors of $\tilde{\mathbf{L}}$ are correlated to $\mathbf{J}$.
- The matrix $\mathbf{\Lambda}$ contains the matrix of affinities between classes $\mathbf{M}$.

As it is common in spiked random matrix analysis, a complete study of $\tilde{\mathbf{L}}$ was performed *i*) The evaluation of the phase transition beyond which spectral community detection is possible. *ii*) Study of the isolated eigenvectors; in particular the evaluation of the average level of each plateau of the eigenvectors step functions and the average fluctuations around each plateau using the approaches described in Section 2. The different results can be found in our article[3].

The anticipated performances of spectral community detection were then evaluated using the aforementionned class-wise means and variances, in a 2−class graph generated using the DC-SBM. As in Figure 5, the empirical and theoretical correct clustering rates for $\mathbf{M} = \delta \mathbf{I}_2$ and varying δ, with μ the uniform distribution in $[.2, .8]$ and $\mu = \delta_{.5}$, a perfect match is obtained between theory and practice. An important outcome from this study is that the limiting spectrum of

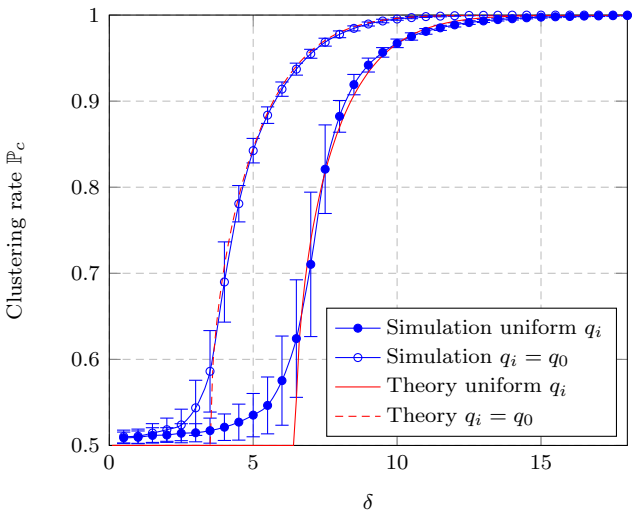


Figure 5: Performance of community detection, for q_i uniformly distributed in $[.2, .8]$, $\mathbf{M} = \delta \mathbf{I}_2$, $c_1 = c_2 = \frac{1}{2}$, and for $q_i = q_0 = .5$. Simulations for $n = 2000$.

the matrix $\mathbf{L}$ is larger for more spread out measures μ and this prevents the appearance of spiked eigenvalues. This is illustrated in Figure 5 where the phase transition point beyond which clustering is feasible is seen to be shifted to larger values of δ for the uniform distribution. The normalization of the modularity matrix $\mathbf{A} - \frac{\mathbf{q}\mathbf{q}^\mathsf{T}}{\frac{1}{n}\mathbf{q}^\mathsf{T}\mathbf{1}_n}$ by $\mathbf{D}^{-1}$ might be the reason why the main spectrum of $\mathbf{L}$ is more spread out. A natural tradeoff should thus be done between the correction of the eigenvectors biases and the avoidance of spectrum spread. One could for instance consider the unormalized modularity matrix $\mathbf{A} - \frac{\mathbf{q}\mathbf{q}^\mathsf{T}}{\frac{1}{n}\mathbf{q}^\mathsf{T}\mathbf{1}_n}$, and perform spectral clustering on its isolated eigenvectors pre-multiplied by $\mathbf{D}^{-1}$. This will, at the same time, correct the biases of the eigenvectors and reduce the spectrum spread.

2.4 Echo-state Neural Networks (within Task 2.3)

The contribution to be presented in this section concerns our first steps into the performance characterization of large dimensional neural networks. The overall line of strategy consists in starting from the study of simple networks (single layer, randomly connected, with linear activations) down to increasingly more elaborate networks (multiple layers, recurrent, with non-linear activations, using backpropagation of the error, etc.).

There exist three main barriers to break into the characterizations of neural networks using random matrix tools. The first easy one, which we shall address next, is the possibly recurrent nature of these networks; this is especially the case for handling time series. The second difficulty is the non-linearity of the activation functions of neural networks. This aspect is currently under investigation and first results are appearing, which shall not be presented presently. The last challenge consists in handling the learning by back-propagation of the error; there the difficulty lies in that the neural network performance is strongly data dependent.

Presently, our interest is to handle the first aspect of neural networks: their (possibly) recursive nature. As such, avoiding the difficulty of non-linear activations and back-propagation of the error, we shall focus on a "simple" family of neural networks, known as *echo-state networks*, ESN for short. Figure 6 depicts an instance of the ESN under study here. To define an ESN, we assume here an n-node network with connectivity matrix $W \in \mathbb{R}^{n \times n}$ such that $\|W\| < 1$,

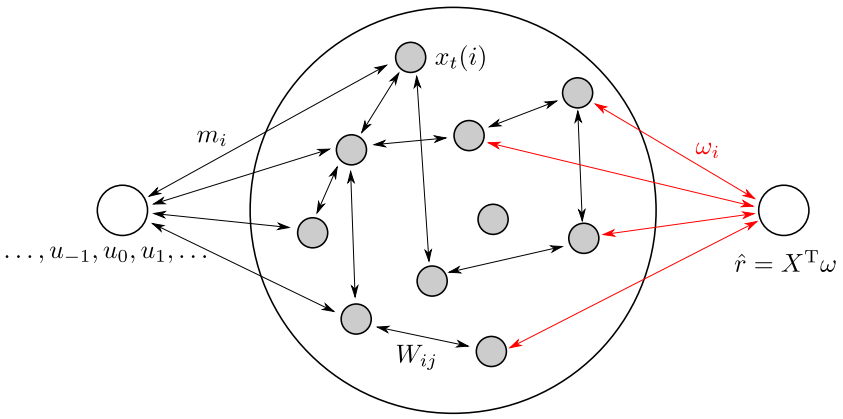


Figure 6: Echo-state neural network.

source-to-reservoir vector $m \in \mathbb{R}^n$, states $x_t \in \mathbb{R}^n$, $t = -\infty, \ldots, \infty$, and internal noise $\eta\varepsilon_t \sim \mathcal{N}(0, \eta^2 I_n)$, fed by a scalar input $u_t \in \mathbb{R}$. The state evolution equation follows:

$$x_{t+1} = Wx_t + mu_{t+1} + \eta\varepsilon_{t+1}. \tag{2.5}$$

The specificity of ESN's is that the weight matrix W as well as the input layer m is *not trained*. Only the network-to-sink layer, that connects the states x_t to a desired output node, is trained. In the training phase, one wishes to map an input sequence $u = [u_0, \ldots, u_{T-1}]^{\mathsf{T}}$ to a corresponding known output sequence $r = [r_0, \ldots, r_{T-1}]^{\mathsf{T}}$. To this end, we shall enforce the reservoir-to-sink connections of the network, gathered into a vector $\omega \in \mathbb{R}^n$ and depicted in color in Figure 6, so to minimize the quadratic reconstruction error

$$E_\eta(u, r) \equiv \frac{1}{T} \left\| X^{\mathsf{T}}\omega - r \right\|^2.$$

The solution to this classical problem is to take ω to be the least-square regressor

$$\omega \equiv \begin{cases} (XX^{\mathsf{T}})^{-1}Xr & , \ T > n \\ X(X^{\mathsf{T}}X)^{-1}r & , \ T \leq n. \end{cases} \tag{2.6}$$

with $X = [x_1, \ldots, x_T] \in \mathbb{R}^{n \times T}$. The matrix X is inherently random because, in the first place, the additional noise ε_t is random and, possibly also, the connectivity matrix W may be random as well. At first, we shall only account for the randomness in ε_t.

To such an ω are associated an $E_\eta(u; r)$ which it is convenient to see here as

$$E_\eta(u, r) = \begin{cases} \lim_{\gamma \downarrow 0} \gamma \frac{1}{T} r^{\mathsf{T}} \tilde{Q}_\gamma r & , \ T > n \\ 0 & , \ T \leq n \end{cases} \tag{2.7}$$

where $\tilde{Q}_\gamma = (\frac{1}{T} X^{\mathsf{T}}X + \gamma I_T)^{-1}$.

To characterize $E_\eta(u, r)$, we first need a deterministic equivalent for $\tilde{Q}_\gamma$. This is provided in the following result.

Theorem 4 (Deterministic Equivalent) *Denote $A = MU$ with $M = [m, Wm, \ldots, W^{T-1}m]$ and $U = T^{-\frac{1}{2}}\{u_{j-i}\}_{i,j=0}^{T-1}$. Then under mild technical assumptions, for $\gamma > 0$, and with $Q_\gamma = (\frac{1}{T}XX^{\mathsf{T}} + \gamma I_n)^{-1}$ and $\tilde{Q}_\gamma = (\frac{1}{T}X^{\mathsf{T}}X + \gamma I_T)^{-1}$, we have that, as $n \to \infty$,*

$$Q_\gamma \leftrightarrow \bar{Q}_\gamma \equiv \frac{1}{\gamma}\left(I_n + \eta^2 \tilde{R}_\gamma + \frac{1}{\gamma} A \left(I_T + \eta^2 R_\gamma \right)^{-1} A^{\mathsf{T}} \right)^{-1}$$
$$\tilde{Q}_\gamma \leftrightarrow \bar{\tilde{Q}}_\gamma \equiv \frac{1}{\gamma}\left(I_T + \eta^2 R_\gamma + \frac{1}{\gamma} A^{\mathsf{T}} \left(I_n + \eta^2 \tilde{R}_\gamma \right)^{-1} A \right)^{-1}$$

where $R_\gamma \in \mathbb{R}^{T\times T}$ *and* $\tilde{R}_\gamma \in \mathbb{R}^{n\times n}$ *are solutions to*

$$R_\gamma = \left\{ \frac{1}{T} \operatorname{tr}\left(S_{i-j}\bar{Q}_\gamma\right) \right\}_{i,j=1}^{T}$$
$$\tilde{R}_\gamma = \sum_{q=-\infty}^{\infty} \frac{1}{T} \operatorname{tr}\left(J^q \bar{\tilde{Q}}_\gamma\right) S_q$$

with $[J^q]_{ij} \equiv \boldsymbol{\delta}_{i+q,j}$, $S_q \equiv \sum_{k\geq 0} W^{k+(-q)^+}(W^{k+q^+})^\mathsf{T}$ *(and* $(x)^+ = \max(x,0)$*).*

Applying Theorem 4 in the limit where $\gamma \to 0$, we have in particular the following first limiting performance result.

Proposition 1 (Training MSE) *Under the same mild assumptions as above, let* $r \in \mathbb{R}^T$ *be a vector of Euclidean norm* $O(\sqrt{T})$*. Then, with* $E_\eta(u,r)$ *defined in* (2.7)*, as* $n \to \infty$*,*

$$E_\eta(u,r) \leftrightarrow \begin{cases} \frac{1}{T} r^\mathsf{T} \tilde{\mathcal{Q}} r & ,\ c<1 \\ 0 & ,\ c>1. \end{cases}$$

where, for $c<1$*,*

$$\tilde{\mathcal{Q}} \equiv \left(I_T + \mathcal{R} + \frac{1}{\eta^2} A^\mathsf{T} \tilde{\mathcal{R}}^{-1} A \right)^{-1}$$

and $\mathcal{R}$*,* $\tilde{\mathcal{R}}$ *are solutions to*[1]

$$\mathcal{R} = c\left\{ \frac{1}{n} \operatorname{tr}\left(S_{i-j}\tilde{\mathcal{R}}^{-1}\right) \right\}_{i,j=1}^{T}$$
$$\tilde{\mathcal{R}} = \sum_{q=-\infty}^{\infty} \frac{1}{T} \operatorname{tr}\left(J^q (I_T + \mathcal{R})^{-1}\right) S_q.$$

Although seemingly not simple, note that, by writing

$$A^\mathsf{T} \tilde{\mathcal{R}}^{-1} A = U^\mathsf{T} \left\{ m^\mathsf{T} (W^{i-1})^\mathsf{T} \tilde{\mathcal{R}}^{-1} W^{j-1} m \right\}_{i,j=1}^{T} U$$

the matrix $\tilde{\mathcal{Q}}$ involved in the asymptotic expression for $E_\eta(u,r)$ clearly features independently:

- the input data matrix U composed in columns of the successive delayed versions of the vector $T^{-\frac{1}{2}}[u_{-(T-1)}, \ldots, u_{T-1}]^\mathsf{T}$;
- the network structuring matrices $\mathcal{R}$ and $(W^{i-1})^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^{j-1}$;
- the factor η^{-2}, not present in $\mathcal{R}, \tilde{\mathcal{R}}$, which trades off the need for *regularizing* the ill-conditioned matrix $A^\mathsf{T}\tilde{\mathcal{R}}^{-1}A$ (through the matrix M in A) and the need to increase the weight of the information-carrying matrix $A^\mathsf{T}\tilde{\mathcal{R}}^{-1}A$ (through the matrix U in A).

Note in particular that, since $\|W\| < 1$, the matrix $\{m^\mathsf{T}(W^{i-1})^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^{j-1}m\}_{i,j=1}^T$ has an exponentially decaying profile down the rows and columns (essentially decaying with $i+j$). As such, all but the first few columns of $\tilde{\mathcal{R}}^{-\frac{1}{2}}MU$ vanish as n, T grow large, providing us with a first testimony of the ESN short term memory specificity, since only the first columns of U (i.e., the first delays of $\{u_t\}$) are accounted for. The matrix $\tilde{\mathcal{R}}^{-\frac{1}{2}}M$ then plays the important role of tuning the memory decay.

For W random, Theorem 4 further simplifies and we in particular obtain the following short-hand formulation for W a random orthogonally invariant orthogonal matrix (often called a real Haar matrix).

[1] $\mathcal{R}$ and $\tilde{\mathcal{R}}$ are rigorously the limits of R_γ and $\gamma\tilde{R}_\gamma$ from Theorem 4, respectively, as $\gamma \downarrow 0$.

Corollary 1 (Haar W, $n/T < 1$) *Let $W = \sigma Z$ with Z random real Haar and m be independent of W with $\|m\| = 1$. Then, under the same mild assumptions as above and with $n/T < 1$,*

$$E_\eta(u, r) \leftrightarrow (1 - c)\frac{1}{T} r^\mathsf{T} Q r$$

where $Q = (I_T + \frac{1}{\eta^2} U^\mathsf{T} D U)^{-1}$ with $D \in \mathbb{R}^{T \times T}$ the diagonal matrix

$$D_{ii} = (1 - \sigma^2)\sigma^{2(i-1)}.$$

We clearly see through Corollary 1 the impact of σ which weighs through D_{ii} the successive delay vectors $U_{\cdot,i}$ starting from $i = 1$ for zero delay. This is reminiscent of the works [20] where the diagonal elements of D were understood qualitatively as a *memory curve*, with the property that $\sum_{i \geq 1} D_{ii} = 1$, so that the ESN *allocates* a total unit amount of memorization capabilities across the successively delayed versions of u.

Similarly, one can further develop relations to characterize the performance of echo-state networks on test datasets. Since the results take more involved forms while not bringing significant additional insights, we elude their detailing here. A complete exposition of the results are available in the submitted article [4]. In practice, the derived theoretical performances are shown to be good matches of the performances of finite dimensional ESN's. This is depicted through an example in Figure 7 for the special case of the so-called Mackey–Glass model prediction task (here $r_t = u_{t+1}$). Observe that, as n, T (and $\hat{T}$ the testing duration) grow large, the deterministic limiting approximations become extremely tight.

3 On-going and Future Activities

Referring back to the timeline and description of work packages and tasks, along with previous discussions, all tasks within the project are well engaged, and most particularly tasks falling within WP1. As an overview summary of the on-going and future activities of WP1:

- Significant advances have been made within Task 1.1 into the understanding of kernel matrices of the *radial type*, i.e., with $\kappa(x, y) = f(\|x - y\|^2)$, and for Gaussian input data. While clearly not a major ambition in the following year of the project (but possibly of the last year), one extension of these results are to consider the (in fact simpler) *outer-product* kernels, i.e., with $\kappa(x, y) = f(x^\mathsf{T} y)$. Another extension concerns the case of more "impulsive" or "heavy-tailed" input data, such as elliptical models of input data. Such datasets dramatically change the developments exposed earlier, as the concentration of all norm differences $\|x_i - x_j\|^2$ to a certain limit no longer holds. Other generalizations concern some shift in the growth regime, such as: increasing the number of classes with n, changing the growth rate between p and n, etc.

- The study of structured spiked models within Task 1.2 is also well underway and has been used significantly in the various applications within WP2. Further investigations in this direction concern deeper analysis of the entry distribution of eigenvectors arising from inhomogeneous datasets (as in particular for community detection in graphs with inhomogeneous degree distributions).

- Task 1.3, now converted to a wider purpose study of mathematical specificities of neural networks, has covered the study of some recursive models of matrices. The next challenging work within this task is to investigate matrices with non-linear entries, largely used in neural networks. Then, if time allows, a study of the performance of neural networks involving back-propagation of the error will be made; the strong difficulty related to this question has to do with the fact that gradient descent methods are being used to update in succession the system parameters, said parameters being function of all the deterministic data and the original random parameters.

As for WP2, the on-going and future investigations can be gathered as follows:

- As an extension of the work [1, 2], and mostly based on the key Theorem 1, two studies are currently being undertaken:

  - *kernel semi-supervised learning*: this study concerns the case where the dataset is divided into a subset of *labelled* data and a subset of *unlabelled* data. Many kernel methods consist in studying a functional of the kernel matrix K defined earlier, now subdivided into four submatrices: K_{uu}, K_{ul}, K_{lu}, K_{ll} depending on

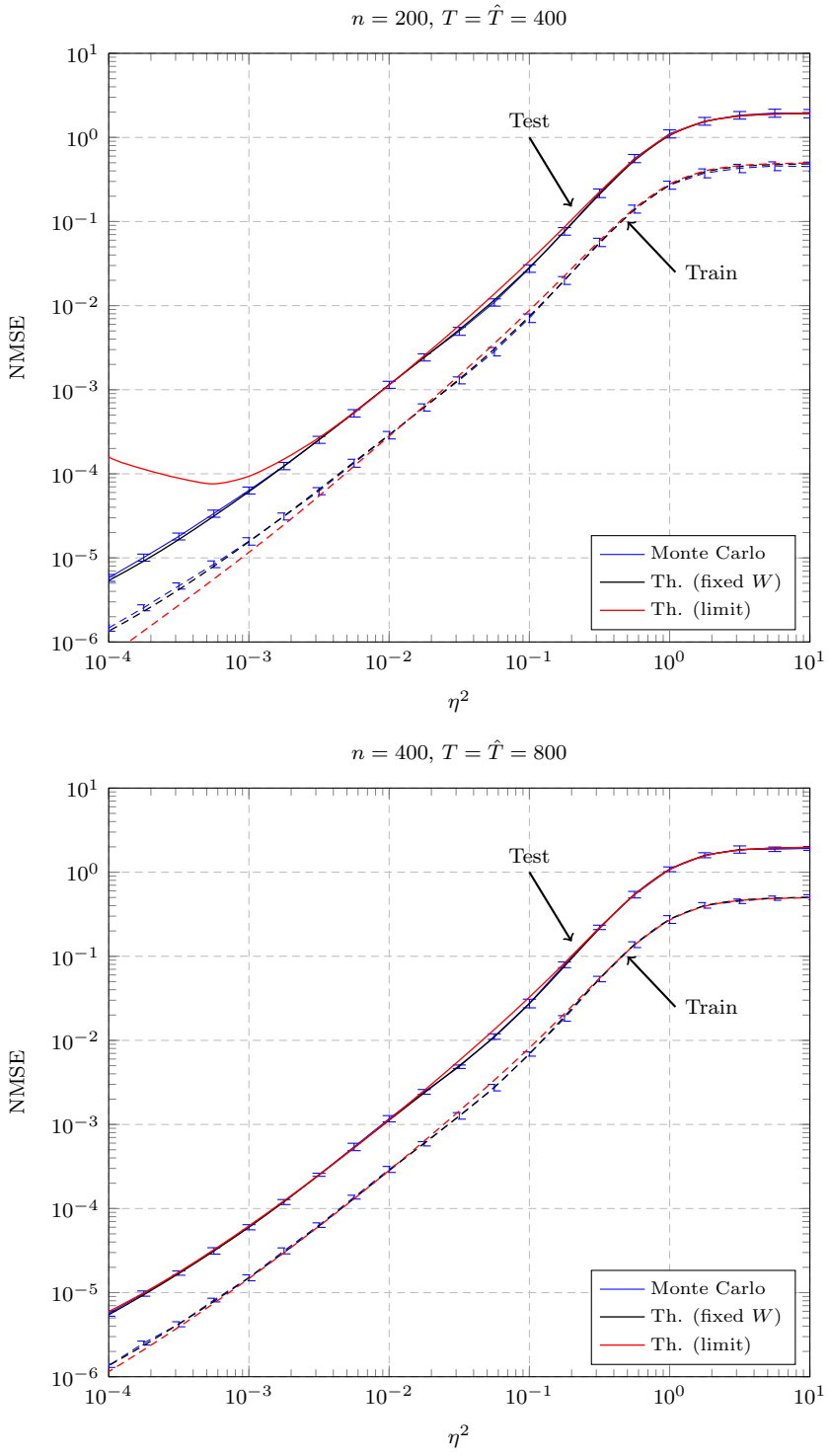


Figure 7: Training and testing (normalized) MSE for the Mackey Glass one-step prediction, W Haar, $n = 200$, $T = \hat{T} = 400$ (top) and $n = 400$, $T = \hat{T} = 800$ (bottom). Comparison between Monte Carlo simulations (Monte Carlo) and theory for fixed and random W.

their data entries being labelled or unlabelled. The performance of such semi-supervised methods are often described up to a tuning parameter that, so far, is only intuitively, qualitatively selected. Our objective is a sound performance analysis of these methods so to provide a quantitative understanding of the performances

and of the optimal hyperparameters tuning.

  - *kernel support vector machines*: on the far end of the labelled versus unlabelled spectrum are (kernel) support vector machines, which assume the existence of a large set of labelled data, used to create "decision regions" for subsequent unlabelled data appearing one by one. Similar to the semi-supervised learning study, our objective is to study the asymptotic performance of support vector machines and derive optimal tuning of its hyperparameters.

- On the graph community detection side, an important new object of analysis (which came up in the recent literature of community detection) is the so-called Bethe Hessian matrix which, in case of very sparse networks, demonstrates performances reaching the so-far optimal (but complex and barely fathomed) belief propagation approach. The Bethe Hessian matrix is however only studied for homogeneous graph stochastic block models. We wish here to extend its analysis to inhomogeneous networks. One possible important difficulty is that, as it was designed to handle sparse networks, the extension of its analysis to the dense network under consideration in our own analyses might turn out more challenging than expected.

- A second aspect currently under study is the performance of simple but *non-linear* neural networks. To simplify the network, we take it to be a one-layer *extreme learning machine*, which consists of a first *fixed* (often randomly selected) connectivity matrix from the input to the neural network and of a network to sink connectivity matrix which is learned by ridge regression. At the neurons, a non-linear activation function is applied to the input data. Our objective is to understand the regression performance of such non-linear networks, using the anticipated results from Task 1.3.

References

[1] R. Couillet and F. Benaych-Georges, "Kernel spectral clustering of large dimensional data," *arXiv preprint arXiv:1510.03547*, 2016.

[2] F. Benaych-Georges and R. Couillet, "Spectral analysis of the gram matrix of mixture models," *ESAIM: Probability and Statistics*, 2016. [Online]. Available: http://dx.doi.org/10.1051/ps/2016007

[3] H. Tiomoko Ali and R. Couillet, "Performance analysis of spectral community detection in realistic graph models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*, 2016.

[4] R. Couillet, G. Wainrib, H. Sevi, and H. T. Ali, "The asymptotic performance of linear echo state neural networks," *(submitted to) Journal on Machine Learning Research*, 2016.

[5] R. Couillet and A. Kammoun, "Statistical subspace kernel spectral clustering of large dimensional data," *under patenting process*, 2016.

[6] R. Couillet, G. Wainrib, H. Sevi, and H. Tiomoko Ali, "A random matrix approach to echo-state neural networks," in *International Conference on Machine Learning (ICML)*, New York, USA, 2016.

[7] ——, "Training performance of echo state neural networks," in *IEEE Statistical Signal Processing Workshop (SSP)*, Palma de Mallorca, Spain, 2016.

[8] N. El Karoui, "The spectrum of kernel random matrices," *The Annals of Statistics*, vol. 38, no. 1, pp. 1–50, 2010.

[9] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, 2006.

[10] F. Chapon, R. Couillet, W. Hachem, and X. Mestre, "The outliers among the singular values of large rectangular random matrices with additive fixed rank deformation," *Markov Processes and Related Fields*, vol. 20, pp. 183–228, 2014.

[11] F. Benaych-Georges and R. R. Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[12] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. NY, USA: Cambridge University Press, 2011.

[13] W. Hachem, P. Loubaton, and J. Najim, "Deterministic equivalents for certain functionals of large random matrices," *Annals of Applied Probability*, vol. 17, no. 3, pp. 875–930, 2007.

[14] A. Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press*, vol. 14, pp. 849–856, 2001.

[15] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[16] U. Von Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering," *The Annals of Statistics*, pp. 555–586, 2008.

[17] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

[18] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, "Spectral redemption in clustering sparse networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20 935–20 940, 2013.

[19] A. Saade, F. Krzakala, and L. Zdeborová, "Spectral clustering of graphs with the bethe hessian," in *Advances in Neural Information Processing Systems*, 2014, pp. 406–414.

[20] S. Ganguli, D. Huh, and H. Sompolinsky, "Memory traces in dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 105, no. 48, pp. 18 970–18 975, 2008.