SPARSE QUANTIZED SPECTRAL CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Given a large data matrix, sparsifying, quantizing, and/or performing other entry-wise nonlinear operations can have numerous benefits, ranging from speeding up iterative algorithms for core numerical linear algebra problems to providing nonlinear filters to design state-of-the-art neural network models. Here, we exploit tools from random matrix theory to make precise statements about how the eigenspectrum of a matrix changes under such nonlinear transformations. In particular, we show that very little change occurs in the informative eigenstructure even under drastic sparsification/quantization, and consequently, that very little downstream performance loss occurs with very aggressively sparsified or quantized spectral clustering. We illustrate how these results depend on the nonlinearity, we characterize a phase transition beyond which spectral clustering becomes possible, and we show when such nonlinear transformations can introduce spurious non-informative eigenvectors.

1 INTRODUCTION

Sparsifying, quantizing, and/or performing other entry-wise nonlinear operations on large matrices can have many benefits. Historically, this has been used to develop iterative algorithms for core numerical linear algebra problems (Achlioptas & McSherry, 2007). More recently, this has been used to design better neural network models (Srivastava et al., 2014). A concrete example, amenable to theoretical analysis and ubiquitous in practice, is provided by spectral clustering, which can be solved by computing the eigenvectors of $\mathbf{X}^\mathsf{T}\mathbf{X}$, for $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ a large data matrix (Von Luxburg, 2007; Mahoney, 2016). When the amount of data n is large, the Gram "kernel" matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ can be enormous, impractical even to form and leading to computationally unaffordable algorithms. For instance, Lanczos iteration that operates through repeated matrix-vector multiplication suffers from an $O(n^2)$ complexity (Golub & Van Loan, 1996) and quickly becomes burdensome.

One approach to overcoming this limitation is simple subsampling: dividing $\mathbf{X}$ into subsamples of size εn, for some $\varepsilon \in (0, 1)$, on which one performs parallel computation, and then recombining. This leads to computational gain, but at the cost of degraded performance, since each data point $\mathbf{x}_i$ looses the cumulative effect of comparing to the *whole* dataset. An alternative cost-reduction procedure consists in *uniformly* randomly "zeroing-out" entries from the whole matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$, resulting in a sparse matrix with only an ε fraction of non-zero entries. For spectral clustering, by focusing on the eigenspectrum of the "zeroed-out" matrix, Zarrouk et al. (2020) showed that the same computational gain can be achieved at the cost of a much less degraded performance: for n/p rather large, almost no degradation is observed down to very small values of ε (e.g., $\varepsilon \approx 2\%$ for $n/p \gtrsim 100$).

Previous efforts showed that it is often advantageous to perform sparsification/quantization in a *data-dependent non-uniform* manner, rather than uniformly (Achlioptas & McSherry, 2007; Drineas & Zouzias, 2011). The focus there, however, is on (the non-asymptotic bound of) the approximation error between the original and the sparsified/quantized matrices. This, however, does *not* provide a direct access to the actual performance for spectral clustering or other downstream tasks of interest, e.g., since the top eigenvectors are known to exhibit a phase transition phenomenon (Baik et al., 2005; Saade et al., 2014). That is, they can behave very differently from those of the original matrix, even if the matrix after treatment is close in operator/Frobenius norm to the original matrix.

Here, we focus on the precise characterization of the eigenstructure of $\mathbf{X}^\mathsf{T}\mathbf{X}$ after entry-wise nonlinear transformation such as sparsification or quantization, in the large n, p regime, by performing simultaneously *data-dependent non-uniform sparsification and/or quantization* (down to

binarization). We consider a simple mixture data model with $\mathbf{x} \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$ and let $\mathbf{K} \equiv f(\mathbf{X}^\mathsf{T}\mathbf{X}/\sqrt{p})/\sqrt{p}$, where f is an entry-wise thresholding/quantization operator (thereby zeroing-out/quantizing entries of $\mathbf{X}^\mathsf{T}\mathbf{X}$); and we prove that this leads to significantly improved performances, with the same computational cost, in spectral clustering as uniform sparsification, but for a much reduced cost in storage induced by quantization. The only (non-negligible) additional cost arises from the extra need for evaluating each entry of $\mathbf{X}^\mathsf{T}\mathbf{X}$. Our main technical contribution (of independent interest, e.g., for those interested in entry-wise nonlinear transformations of feature matrices) consists in using random matrix theory (RMT) to derive the large n, p asymptotics of the eigenspectrum of $\mathbf{K}$ for a wide range of functions f, and then comparing to previously-established results for uniform subsampling and sparsification in (Zarrouk et al., 2020). Simulations on real-world data further collaborate our findings. Our main contributions are the following.

1. We derive the limiting eigenvalue distribution of $\mathbf{K}$ as $n, p \to \infty$ (Theorem 1), and we identify:
   (a) the existence of non-informative isolated eigenvectors of $\mathbf{K}$ for some nonlinear f (Corollary 1);
   (b) in the absence of such eigenvectors, a phase transition in the dominant eigenvalue-eigenvector $(\hat{\lambda}, \hat{\mathbf{v}})$ pair (Corollary 2): if the signal-to-noise ratio (SNR) $\|\boldsymbol{\mu}\|^2$ exceeds a threshold γ, then $\hat{\lambda}$ becomes isolated and $\hat{\mathbf{v}}$ contains data class-structure information exploitable for clustering; if not, then $\hat{\mathbf{v}}$ contains only noise and is asymptotically *orthogonal* to the class-label vector.
2. Letting f be a sparsification, quantization, or binarization operator, we show:
   (a) a selective sparsification operator such that $\mathbf{X}^\mathsf{T}\mathbf{X}$ can be drastically sparsified with very little degradation in clustering performance (Proposition 1 and Section 4.2), significantly outperforms the random uniform sparsification in (Zarrouk et al., 2020);
   (b) for a given matrix storage budget (i.e., fixed number of bits to store $\mathbf{K}$), an optimal design of the quantization/binarization operators (Proposition 2 and Section 4.3), the performances of which are compared against the original $\mathbf{X}^\mathsf{T}\mathbf{X}$ and its sparsified but not quantized version.

For spectral clustering, the surprisingly small performance drop, accompanied by a huge reduction in computational cost, contributes to improved algorithms for large problems. More generally, our proposed analysis sheds light on the effect of entry-wise nonlinear transformations on the eigenspectra of data/feature matrices. Thus, looking forward (and perhaps more importantly, given the use of nonlinear transformations in designing modern neural network models as well as the recent interest in applying RMT to neural network analyses (Li & Nguyen, 2018; Seddik et al., 2018; Jacot et al., 2019)), we expect that our analysis opens the door to improved analysis of computationally efficient methods for large dimensional machine learning and neural network models more generally.

2 System model and preliminaries

Basic setup. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be independently drawn (non-necessarily uniformly) from:

$$\mathcal{C}_1 : \mathbf{x}_i = -\boldsymbol{\mu} + \mathbf{z}_i, \quad \mathcal{C}_2 : \mathbf{x}_i = +\boldsymbol{\mu} + \mathbf{z}_i \tag{1}$$

with $\mathbf{z}_i \in \mathbb{R}^p$ having i.i.d. zero-mean, unit-variance, κ-kurtosis, sub-exponential entries, $\boldsymbol{\mu} \in \mathbb{R}^p$ such that $\|\boldsymbol{\mu}\|^2 \to \rho \geq 0$ as $p \to \infty$, and $\mathbf{v} \in \{\pm 1\}^n$ with $[\mathbf{v}]_i = -1$ for $\mathbf{x}_i \in \mathcal{C}_1$ and $+1$ for $\mathbf{x}_i \in \mathcal{C}_2$.[1] The data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ can be compactly written as $\mathbf{X} = \mathbf{Z} + \boldsymbol{\mu}\mathbf{v}^\mathsf{T}$ for $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ so that $\|\mathbf{v}\| = \sqrt{n}$ and both $\mathbf{Z}, \boldsymbol{\mu}\mathbf{v}^\mathsf{T}$ have operator norm of order $O(\sqrt{p})$ in the $n \sim p$ regime. In this setting, the Gram (or *linear* kernel) matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ achieves optimal clustering performance on the mixture model (1); see Remark 1 below. However, it consists of a dense $n \times n$ matrix, which becomes quickly expensive to store or to perform computation on, as n increases.

Thus, we consider instead the following *entry-wise nonlinear* transformation of $\mathbf{X}^\mathsf{T}\mathbf{X}$:

$$\mathbf{K} = \left\{\delta_{i \neq j} f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}\right\}_{i,j=1}^n \tag{2}$$

for $f : \mathbb{R} \to \mathbb{R}$ satisfying some regularity conditions (see Assumption 1 below), where $\delta_{i \neq j}$ equals 1 for $i \neq j$ and equals 0 otherwise. The diagonal elements $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_i/\sqrt{p})$ (**i**) bring no additional information for clustering and (**ii**) do not scale properly for p large ($\mathbf{x}_i^\mathsf{T}\mathbf{x}_i/\sqrt{p} = O(\sqrt{p})$). Thus, following (El Karoui, 2010; Cheng & Singer, 2013), they are discarded.

[1] The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices.

Most of our technical results hold for rather generic functions f, but we are particularly interested in f with nontrivial numerical properties (e.g., promoting quantization and sparsity):

$$\textbf{Sparsification:} \quad f_1(t) = t \cdot 1_{|t|>\sqrt{2}s} \tag{3}$$

$$\textbf{Quantization:} \quad f_2(t) = 2^{2-M}(\lfloor t \cdot 2^{M-2}/\sqrt{2}s \rfloor + 1/2) \cdot 1_{|t|\le\sqrt{2}s} + \mathrm{sign}(t) \cdot 1_{|t|>\sqrt{2}s} \tag{4}$$

$$\textbf{Binarization:} \quad f_3(t) = \mathrm{sign}(t) \cdot 1_{|t|>\sqrt{2}s}\,. \tag{5}$$

Here, $s \ge 0$ is some *truncation threshold*, and $M \ge 2$ is a number of information bits.[2] The visual representations of these fs are given in Figure 1-(left). For f_3, taking $s \to 0$ leads to the sign function $\mathrm{sign}(t)$. In terms of storage, the quantization f_2 consumes $2^{M-2}+1$ bits per non-zero entry, while the binarization f_3 takes values in $\{\pm 1, 0\}$ and thus consumes 1 bit per non-zero entry.

Random matrix theory. To provide a precise description of the eigenspectrum of $\mathbf{K}$ for the non-linear f of interest, to be used in the context of spectral clustering, we will provide a large dimensional characterization for the *resolvent* of $\mathbf{K}$, defined for $z \in \mathbb{C} \setminus \mathbb{R}^+$ as

$$\mathbf{Q}(z) \equiv (\mathbf{K} - z\mathbf{I}_n)^{-1}. \tag{6}$$

This quantity, which plays a central role in RMT (Bai & Silverstein, 2010), will be used in two primary ways. First, the normalized trace $\frac{1}{n}\operatorname{tr}\mathbf{Q}(z)$ is the so-called *Stieltjes transform* of the eigenvalue distribution of $\mathbf{K}$, and the large n, p eigenvalue distribution of $\mathbf{K}$ will be used to characterize the phase transition beyond which spectral clustering becomes theoretically possible (Corollary 2). Second, for $(\hat{\lambda}, \hat{\mathbf{v}})$, an eigenvalue-eigenvector pair of $\mathbf{K}$, and $\mathbf{a} \in \mathbb{R}^n$, a deterministic vector, by Cauchy's integral formula, the "angle" between $\hat{\mathbf{v}}$ and $\mathbf{a}$ is given by $|\hat{\mathbf{v}}^\mathsf{T}\mathbf{a}|^2 = -\frac{1}{2\pi\imath}\oint_{\Gamma(\hat{\lambda})} \mathbf{a}^\mathsf{T}\mathbf{Q}(z)\mathbf{a}\,dz$, where $\Gamma(\hat{\lambda})$ is a positively oriented contour surrounding $\hat{\lambda}$ only. Letting $\mathbf{a} = \mathbf{v}$, this will be exploited to characterize the spectral clustering error rate (Proposition 1).

From a technical perspective, unlike linear random matrix models, $\mathbf{K}$ (and thus $\mathbf{Q}(z)$) involves non-linear dependence between its entries. To break this difficulty, following the ideas of (Cheng & Singer, 2013), we exploit the fact that, by the central limit theorem, $\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p} \to \mathcal{N}(0,1)$ in distribution as $p \to \infty$. As such, up to $\boldsymbol{\mu}\mathbf{v}^\mathsf{T}$, which is treated separately with a perturbation argument, the $[\mathbf{K}]_{ij}$s asymptotically behave like a family of *dependent* standard Gaussian variables to which f is applied. Expanding f in a series of *orthogonal polynomials* with respect to the Gaussian measure allows for "unwrapping" this dependence. A few words on the theory of orthogonal polynomials (Abramowitz & Stegun, 1965; Andrews et al., 2000) are thus convenient to pursue our analysis.

Orthogonal polynomial framework. For a probability measure μ, let $\{P_l(x), l \ge 0\}$ be the orthonormal polynomials with respect to $\langle f, g\rangle \equiv \int fg d\mu$ obtained by the Gram-Schmidt procedure on the monomials $\{1, x, x^2, \ldots\}$, such that $P_0(x) = 1$, P_l is of degree l and $\langle P_{l_1}, P_{l_2}\rangle = \delta_{l_1 - l_2}$. By the Riesz-Fischer theorem (Rudin, 1964, Theorem 11.43), for any function $f \in L^2(\mu)$, the set of square-integrable functions with respect to $\langle\cdot,\cdot\rangle$, one can formally expand f as

$$f(x) \sim \textstyle\sum_{l=0}^\infty a_l P_l(x), \quad a_l = \int f(x)P_l(x)\mu(dx) \tag{7}$$

where "$f \sim \sum_{l=0}^\infty a_l P_l$" indicates that $\|f - \sum_{l=0}^L a_l P_l\| \to 0$ as $L \to \infty$ with $\|f\|^2 = \langle f, f\rangle$.

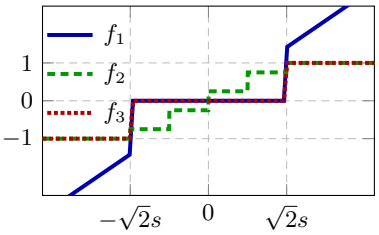
To investigate the asymptotic behavior of $\mathbf{K}$ as $n, p \to \infty$, we make the following assumption on f.

Assumption 1. *Let $\xi_p = \mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p}$ and $\{P_{l,p}(x),\, l \ge 0\}$ be the orthonormal polynomials with respect to the measure μ_p of ξ_p. For $f \in L^2(\mu_p)$, $f(x) \sim \sum_{l=0}^\infty a_{l,p}P_{l,p}(x)$ with $a_{l,p}$ in (7) and **(i)** $\sum_{l=0}^\infty a_{l,p}P_{l,p}(x)\mu_p(dx)$ converges in $L^2(\mu_p)$ to $f(x)$ uniformly over large p and **(ii)** as $p \to \infty$, $\sum_{l=1}^\infty a_{l,p}^2 \to \nu \ge 0$ and, for $l = 0, 1, 2$, $a_{l,p} \to a_l$ converges with $a_0 = 0$.*

Since $\xi_p \to \mathcal{N}(0,1)$, the parameters a_0, a_1, a_2 and ν are simply moments of the standard Gaussian measure involving f. More precisely, for $\xi \sim \mathcal{N}(0,1)$,

$$a_0 = \mathbb{E}[f(\xi)], \quad a_1 = \mathbb{E}[\xi f(\xi)], \quad \sqrt{2}a_2 = \mathbb{E}[\xi^2 f(\xi)] - a_0, \quad \nu = \mathbb{E}[f^2(\xi) - a_0^2] \ge a_1^2 + a_2^2. \tag{8}$$

[2] Also, here, $\lfloor\cdot\rfloor$ denotes the floor function, while $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ and $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$ denotes the error function and complementary error function, respectively.



f	a_1	ν
f_1	$\mathrm{erfc}(s) + 2se^{-s^2}/\sqrt{\pi}$	$\mathrm{erfc}(s) + \frac{2}{\sqrt{\pi}} se^{-s^2}$
f_2	$\sqrt{\frac{2}{\pi}} \cdot 2^{1-M}(1 + e^{-s^2} + \sum_{k=1}^{2^{M-2}-1} 2e^{-\frac{k^2 s^2}{4^{M-2}}})$	$1 - \frac{2^M-1}{4^{M-1}} \mathrm{erf}(s) - \sum_{k=1}^{2^{M-2}-1} \frac{k\,\mathrm{erf}(ks \cdot 2^{2-M})}{2^{2M-5}}$
f_3	$e^{-s^2}\sqrt{2/\pi}$	$\mathrm{erfc}(s)$

Figure 1: Visual representations of different functions f defined in (3)-(5) **(left)** and their associated parameters a_1 and ν **(right)** ($a_2 = 0$ for each of these fs) defined in Assumption 1.

Imposing the condition $a_0 = 0$ simply discards the *non-informative* rank-one matrix $a_0 \mathbf{1}_n \mathbf{1}_n^\mathsf{T}/\sqrt{p}$ from $\mathbf{K}$. The three parameters (a_1, a_2, ν) are of crucial significance in determining the spectral behavior of $\mathbf{K}$ (see Theorem 1 below).The sparse f_1, quantized f_2, and binary f_3 of our primary interest all satisfy Assumption 1, with the corresponding $a_2 = 0$ (as we shall see in Corollary 1 below, this is important for the spectral clustering use case) and a_1, ν given in Figure 1-(right). With those introductory elements at hand, we are in position to present our main technical results.

3 Main Technical Results

Our main technical result, from which our performance-complexity trade-off analysis will follow, provides an asymptotic *deterministic equivalent* $\bar{\mathbf{Q}}(z)$ for the *random resolvent* $\mathbf{Q}$, defined in (6). (A deterministic equivalent is a deterministic matrix $\bar{\mathbf{Q}}(z)$ such that, for any deterministic sequence of matrices $\mathbf{A}_n \in \mathbb{R}^{n \times n}$ and vectors $\mathbf{a}_n, \mathbf{b}_n \in \mathbb{R}^n$ of bounded (spectral and Euclidean) norms, $\frac{1}{n} \operatorname{tr} \mathbf{A}_n(\mathbf{Q}(z) - \bar{\mathbf{Q}}(z)) \to 0$ and $\mathbf{a}_n^\mathsf{T}(\mathbf{Q}(z) - \bar{\mathbf{Q}}(z))\mathbf{b}_n \to 0$ almost surely as $n, p \to \infty$. We denote this relation $\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z)$.) This is given in the following theorem. The proof is in Appendix A.1.

Theorem 1 (Deterministic equivalent)**.** *Let $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, let $\mathbf{Q}(z)$ be defined in (6), and let $\Im[\cdot]$ denote the imaginary part of a complex number. Then, under Assumption 1,*

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_n - \mathbf{V}\mathbf{\Lambda}(z)\mathbf{V}^\mathsf{T}, \quad \mathbf{\Lambda}(z) = \begin{bmatrix} \Theta(z)m^2(z) & \Theta(z)\Omega(z)\frac{\mathbf{v}^\mathsf{T}\mathbf{1}_n}{n}m(z) \\ \Theta(z)\Omega(z)\frac{\mathbf{v}^\mathsf{T}\mathbf{1}_n}{n}m(z) & \Theta(z)\Omega^2(z)\frac{(\mathbf{v}^\mathsf{T}\mathbf{1}_n)^2}{n^2} - \Omega(z) \end{bmatrix},$$

with $\sqrt{n}\mathbf{V} = [\mathbf{v}, \mathbf{1}_n]$, $\Omega(z) = \frac{a_2^2(\kappa-1)m^3(z)}{2c^2 - a_2^2(\kappa-1)m^2(z)}$, $\Theta(z) = \frac{a_1\|\boldsymbol{\mu}\|^2}{c + a_1 m(z)(1+\|\boldsymbol{\mu}\|^2) + a_1\|\boldsymbol{\mu}\|^2\Omega(z)(\mathbf{v}^\mathsf{T}\mathbf{1}_n)^2/n^2}$, for κ the kurtosis of the entries of $\mathbf{Z}$, and $m(z)$ the unique solution, such that $\Im[m(z)] \cdot \Im[z] \geq 0$, to

$$-\frac{1}{m(z)} = z + \frac{a_1^2 m(z)}{c + a_1 m(z)} + \frac{\nu - a_1^2}{c} m(z). \tag{9}$$

As a consequence of Theorem 1, the *empirical spectral measure* $\omega_n = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(\mathbf{K})}$ of $\mathbf{K}$ has a deterministic limit ω as $n, p \to \infty$, uniquely defined through its Stieltjes transform $m(z) \equiv \int (t-z)^{-1}\omega(dt)$ as the solution to (9).[3] This limiting measure ω does *not* depend on the law of the independent entries of $\mathbf{Z}$, so long that they are sub-exponential, with zero mean and unit variance. In particular, taking $a_1 = 0$ in (9) gives the (rescaled) Wigner semi-circle law ω (Wigner, 1955), and taking $\nu = a_1^2$ (i.e., $a_l = 0$ for $l \geq 2$) gives the Marčenko-Pastur law ω (Marcenko & Pastur, 1967). See Remark 2 in Appendix A.2 for more discussions on this point.

Non-informative spikes. Going beyond just the limiting spectral measure of $\mathbf{K}$, Theorem 1 also shows that *isolated eigenvalues* (often referred to as *spikes*) may be found in the eigenspectrum of $\mathbf{K}$ at very specific locations. Such spikes and their associated eigenvectors are typically thought to provide information on the data class-structure (and they do when f is linear). However, when f is nonlinear, this is not always the case: it is possible that *not all* these spikes are "useful" in the sense of being informative about the class structure. This is made precise in the following, where we assume $\boldsymbol{\mu} = \mathbf{0}$, i.e., that there is no class structure. The proof is in Appendix A.2.

Corollary 1 (Non-informative spikes)**.** *Assume $\boldsymbol{\mu} = \mathbf{0}$, $\kappa \neq 1$ and $a_2 \neq 0$, so that in the notations of Theorem 1, $\Theta(z) = 0$ and $\bar{\mathbf{Q}}(z) = m(z) + \Omega(z)\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ for $\Omega(z)$ defined in Theorem 1. Then, if $x_\pm \equiv \pm\frac{1}{a_2}\sqrt{\frac{2}{\kappa-1}}$ satisfies $a_1 x_\pm \neq \pm 1$ and $(\nu - a_1^2)x_\pm^2 + \frac{a_1^2 x_\pm^2}{(1+a_1 x_\pm)^2} < 1/c$, two eigenvalues*

[3] For $m(z)$ the Stieltjes transform of a measure ω, ω can be obtained via $\omega([a,b]) = \lim_{\epsilon \downarrow 0} \frac{1}{\pi}\int_a^b \Im[m(x + \imath\epsilon)]dx$ for all $a < b$ continuity points of ω.

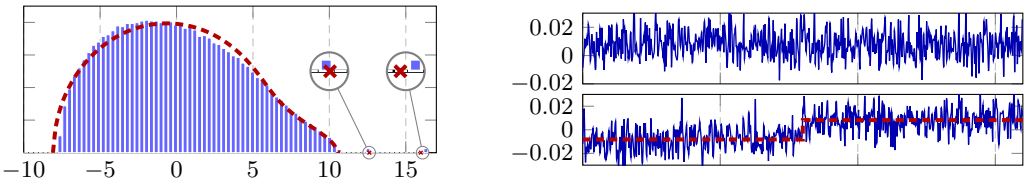


Figure 2: **(Left)** Histogram of eigenvalues of $\mathbf{K}$ (**blue**) versus the limiting spectrum and spikes (**red**). **(Right)** Eigenvectors of the largest **(top)** and second largest **(bottom)** eigenvalues of $\mathbf{K}$ (**blue**), versus the rescaled class label $\alpha\mathbf{v}/\sqrt{n}$ (**red**, from Corollary 2). $f(t) = \sin(t) - 3\cos(t) + 3/\sqrt{e}$, $p = 800$, $n = 6\,400$, $\boldsymbol{\mu} = 1.1 \cdot \mathbf{1}_p/\sqrt{p}$, $\mathbf{v} = [-\mathbf{1}_{n/2};\ \mathbf{1}_{n/2}]$ on Student-t data with $\kappa = 5$.

of $\mathbf{K}$ *converge to* $z_\pm = -\frac{1}{cx_\pm} - \frac{a_1^2 x_\pm}{1+a_1 x_\pm} - (\nu - a_1^2)x_\pm$ *away from the support of* ω*. If instead* $x_\pm = \pm 1/a_1$ *for* $a_1 \neq 0$*, then a single eigenvalue of* $\mathbf{K}$ *isolates with limit* $z = -\frac{\nu}{a_1} - \frac{a_1(2-c)}{2c}$.

Corollary 1 (combined with Theorem 1) says that, while the limiting spectrum ω is *universal* with respect to the distribution of the entries of $\mathbf{Z}$, the existence and position of the non-informative spikes are *not universal* and depend on the kurtosis κ of the distribution. (See Figure 12 in Appendix A.2 for an example.) This is far from a mathematical curiosity. Given how nonlinear transformations are used in machine learning practice, being aware of the existence of *spurious non-informative spikes* in the eigenspectrum of $\mathbf{K}$ (well separated from the bulk of eigenvalues, but that correspond to random noise instead of signal), as a function of properties of the nonlinear f, is of fundamental importance for downstream tasks. For example, for spectral clustering, their associated eigenvectors may be mistaken as informative ones by spectral clustering algorithms, even in the complete absence of classes. This is confirmed by Figure 2 where two isolated eigenvalues (on the right side of the bulk) are observed, with only the second largest one corresponds to an eigenvector that contains class-label information. For further discussions on this point, see Appendix A.2 and A.3.

Informative spikes. From Theorem 1, we see that the eigenspectrum of $\mathbf{K}$ depends on f only via a_1, a_2, and ν. In particular, a_1 and ν determine the limiting spectral measure ω. From Corollary 1, we see that a_2 contributes by **(i)** introducing (at most two) non-informative spikes and **(ii)** reducing the ratio a_1/ν (since $\nu = \sum_{i\geq 1} a_i^2$), thereby necessarily *enlarging* the support of ω (see Remark 2 in Appendix A.2). Taking $a_2 = 0$ reduces the length of the support of ω and, as such, maximizes the "chance" of the appearance of an informative spike (the eigenvector of which is positively correlated with the label vector $\mathbf{v}$). See Remark 1 below for a more precise statement.

In particular, by taking $a_2 = 0$ and $a_1 \neq 0$, we obtain only informative spikes, and we can characterize a phase transition depending on the SNR ρ. The proof of the following is in Appendix A.3.

Corollary 2 (Informative spike and a phase transition)**.** *For* $a_1 > 0$ *and* $a_2 = 0$*, let*

$$F(x) = x^4 + 2x^3 + \left(1 - \frac{c\nu}{a_1^2}\right)x^2 - 2cx - c, \quad G(x) = \frac{a_1}{c}(1+x) + \frac{a_1}{x} + \frac{\nu - a_1^2}{a_1}\frac{1}{1+x}, \tag{10}$$

and let γ *be the largest real solution to* $F(\gamma) = 0$*. Then, under Assumption 1, we have* $\sqrt{c} \leq \gamma \leq \sqrt{c\nu}/a_1$*, and the largest eigenpair* $(\hat{\lambda}, \hat{\mathbf{v}})$ *of* $\mathbf{K}$ *satisfies*

$$\hat{\lambda} \to \lambda = \begin{cases} G(\rho) & \rho > \gamma, \\ G(\gamma) & \rho \leq \gamma; \end{cases} \quad \frac{|\hat{\mathbf{v}}^\mathsf{T}\mathbf{v}|^2}{n} \to \alpha = \begin{cases} \frac{F(\rho)}{\rho(1+\rho)^3} & \rho > \gamma, \\ 0 & \rho \leq \gamma; \end{cases} \tag{11}$$

almost surely as $n, p \to \infty$*,* $p/n \to c \in (0, \infty)$*, where we recall* $\rho \equiv \lim_{p\to\infty} \|\boldsymbol{\mu}\|^2$ *and* $\|\mathbf{v}\| = \sqrt{n}$.

Without loss of generality, we discuss only the case $a_1 > 0$.[4] For $a_1 > 0$, both $F(x)$ and $G(x)$ are increasing functions on $x \in (\gamma, \infty)$. Then, as expected, both λ and α increase with the SNR ρ. Moreover, the phase transition point γ is an increasing function of c and of ν/a_1^2. As such, the optimal choice of f in the sense of the smallest phase transition point is the linear function $f(t) = t$ with $a_1 = \nu = 1$ and $\gamma = \sqrt{c}$. This recovers the classical random matrix result (Baik et al., 2005).

4 Clustering performance of sparse and quantized operators

We start, in Section 4.1, by providing a sharp asymptotic characterization of the clustering error rate and demonstrating the optimality of the linear function under (1). Then, in Section 4.2, we

[4] Otherwise we could consider $-\mathbf{K}$ instead of $\mathbf{K}$ and the largest eigenvalue becomes the smallest one.

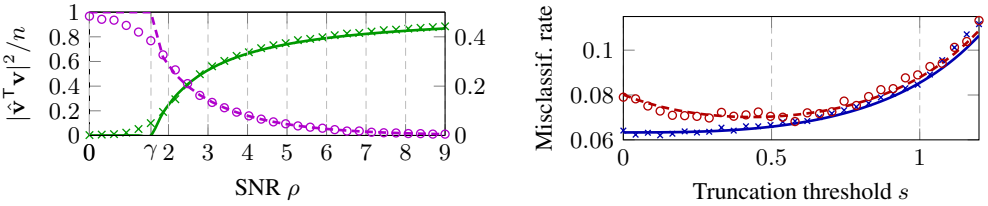


Figure 3: **(Left)** Alignment $|\hat{\mathbf{v}}^\mathsf{T}\mathbf{v}|^2/n$ (**green crosses**) and misclassification rate (**purple circles**) in markers versus their limiting behaviors in lines, for $f(t) = \mathrm{sign}(t)$. **(Right)** Misclassification rate versus s of sparse f_1 (**blue crosses**) and binary f_3 (**red circles**) with $\rho = 4$. Here, $p = 512$, $n = 256$, $\boldsymbol{\mu} \propto \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{v} = [-\mathbf{1}_{n/2};\ \mathbf{1}_{n/2}]$ on Gaussian data, and we averaged over 250 runs.

discuss the advantageous performance of the proposed selective sparsification approach (with f_1) versus the uniform or subsampling approach studied previously in (Zarrouk et al., 2020). Finally, in Section 4.3, we derive the optimal truncation threshold s_{opt}, for both quantized f_2 and binary f_3, so as to achieve an optimal performance-complexity trade-off for a given storage budget.

4.1 Performance of spectral clustering

The technical results in Section 3 provide conditions under which $\mathbf{K}$ admits an informative eigenvector $\hat{\mathbf{v}}$ that is non-trivially correlated with the class label vector $\mathbf{v}$ (and thus that is exploitable for spectral clustering) in the $n, p \to \infty$ limit. Since the exact (limiting) alignment $|\mathbf{v}^\mathsf{T}\hat{\mathbf{v}}|$ is known, along with an additional argument on the normal fluctuations of $\hat{\mathbf{v}}$, we have the following result for the performance of the spectral clustering method. The proof is in Appendix A.4.

Proposition 1 (Performance of spectral clustering)**.** *Let Assumption 1 hold, let $a_1 > 0, a_2 = 0$, and let $\hat{\mathcal{C}}_i = \mathrm{sign}([\hat{\mathbf{v}}]_i)$ be the estimate of the underlying class $\mathcal{C}_i$ of the datum $\mathbf{x}_i$, with the convention $\hat{\mathbf{v}}^\mathsf{T}\mathbf{v} \geq 0$, for $\hat{\mathbf{v}}$ the top eigenvector of $\mathbf{K}$. Then, the misclassification rate satisfies $\frac{1}{n}\sum_{i=1}^n \delta_{\hat{\mathcal{C}}_i \neq \mathcal{C}_i} \to \frac{1}{2}\,\mathrm{erfc}(\sqrt{\alpha/(2-2\alpha)})$, almost surely, as $n, p \to \infty$, for $\alpha \in [0, 1)$ defined in (11).*

Despite being asymptotic results valid in the $n, p \to \infty$ limit, the results of Proposition 1 and Corollary 2 closely match empirical results for n, p in the hundreds. This is illustrated in Figure 3. Proposition 1 further confirms that the misclassification rate, being a decreasing function of α, increases with ν/a_1^2 (for c and ρ fixed). This leads to the following remark.

Remark 1 (Optimality of linear function)**.** *Since both the phase transition point γ and the misclassification rate grow with ν/a_1^2, the linear function $f(t) = t$ with the minimal $\nu/a_1^2 = 1$ is* optimal *in the sense of (**i**) achieving the* smallest *SNR ρ or the* largest *ratio $c = \lim p/n$ (i.e., the fewest samples n) necessary to observe an informative (isolated) eigenvector, and (**ii**) upon existence of such an isolated eigenvector, reaching the* lowest *classification error rate.*

According to Remark 1, any f with $\nu/a_1^2 > 1$ induces performance degeneration (compared to the optimal linear function). However, by choosing f to be one of the functions f_1, f_2, f_3 defined in (3)-(5), one may trade clustering performance optimality for reduced storage size and computational time. Figure 4 displays the decay in clustering performance and the gain in storage size of the sparse f_1, quantized f_2, and binary f_3, when compared to the optimal but "dense" linear function. As $s \to 0$, both performance and storage size under f_1 naturally approach those of the linear function. This is unlike f_2 or f_3, which approach the sign function. For $s \gg 1$, the performance under sparse f_1 becomes comparable to that of binary f_3 (which is significantly worse than quantized but non-sparse f_2) but for a larger storage cost. In particular, using f_2 or f_3 in the setting of Figure 4, one can reduce the storage size by a factor of 32 or 64, at the price of a performance drop less than 1%.

4.2 Comparison to uniform sparsification and subsampling

From Figure 4, we see that the classification error and storage gain of the sparse f_1 increase monotonically, as the truncation threshold s grows. For f_1, the number of non-zero entries of $\mathbf{K}$ is approximately $\mathrm{erfc}(s)n^2$ with truncation threshold s. Thus, the *sparsity level* $\varepsilon_{\mathrm{selec}} = \mathrm{erfc}(s) \in [0, 1]$ can be defined and compared to uniform sparsification or subsampling approaches.

Recall (from the introduction) that the cost of spectral clustering may be reduced by subsampling the whole dataset in $1/\varepsilon_{\mathrm{sub}}$ chunks of $n\varepsilon_{\mathrm{sub}}$ data vectors each. Alternatively, as investigated recently (Zarrouk et al., 2020), the cost can be reduced by uniformly zeroing-out $\mathbf{X}^\mathsf{T}\mathbf{X}$ with a symmetric

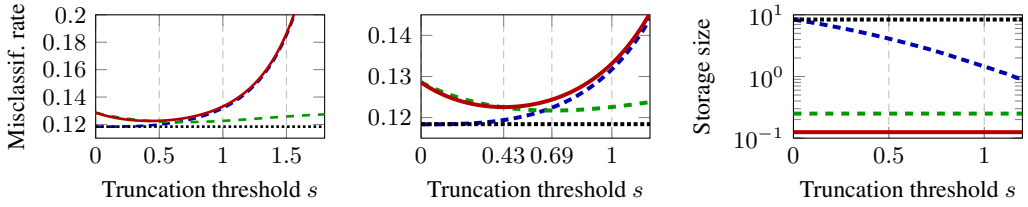


Figure 4: Clustering performance (**left**, a zoom-in in **middle**) and storage size (MB) **(right)** of f_1 (**blue**), f_2 with $M = 2$ (**green**), f_3 (**red**), and linear $f(t) = t$ (**black**), versus the truncation threshold s, for SNR $\rho = 2$, $c = 1/2$ and $n = 10^3$, with 64 bits per entry for non-quantized matrices.

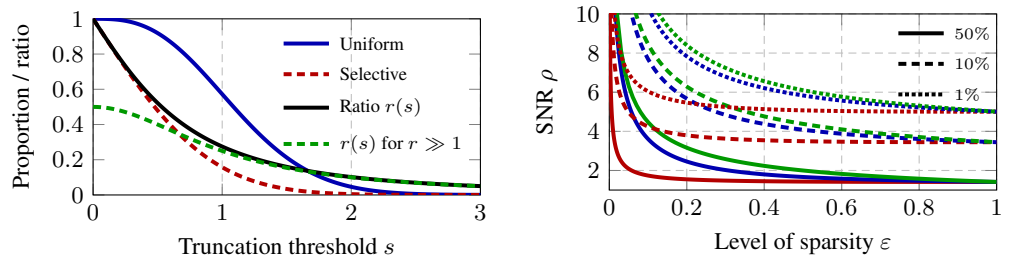


Figure 5: **(Left)** Proportion of non-zero entries with uniform versus selective sparsification f_1 and their ratio, as a function of the truncation threshold s. **(Right)** Comparison of $1\%, 10\%$ classification error and phase transition (i.e., 50% error) curves between subsampling (**green**), uniform (**blue**) and selective sparsification f_1 (**red**), as a function of sparsity level ε and SNR ρ, for $c = 2$.

mask matrix $\mathbf{B}$, with $\mathbf{B}_{ij} \sim \mathrm{Bern}(\varepsilon_{\mathrm{unif}})$ for $1 \leq i < j \leq n$ and $\mathbf{B}_{ii} = 0$. On average, a proportion $1 - \varepsilon_{\mathrm{unif}}$ of the entries of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is set to zero, so that $\varepsilon_{\mathrm{unif}} \in [0, 1]$ controls the sparsity level (and thus the storage size as well as computational time). Similar to our Corollary 2, the associated eigenvector alignment (and thus the clustering accuracy via Proposition 1) in both cases can be derived. Specifically, taking $\varepsilon_{\mathrm{unif}} = a_1^2/\nu$ in (Zarrouk et al., 2020, Theorem 3.2), we obtain the same $F(x)$ as in our Corollary 2 and therefore the same phase transition point γ and eigenvector alignment α. As for subsampling, its performance can be obtained by letting $a_1^2 = \nu$ and changing c into $c/\varepsilon_{\mathrm{sub}}$ in our Corollary 2. Consequently, the same clustering performance is achieved by either uniform or selective sparsification (with f_1) with $\varepsilon_{\mathrm{unif}} = a_1^2/\nu = \mathrm{erfc}(s) + 2se^{-s^2}/\sqrt{\pi} > \mathrm{erfc}(s) = \varepsilon_{\mathrm{selec}}$, and our proposed selective sparsification approach thus leads to *strictly sparser* matrices. Moreover, the ratio $r(s) = \mathrm{erfc}(s)/(\mathrm{erfc}(s) + 2se^{-s^2}/\sqrt{\pi})$ is a decreasing function of s and approximates as $r(s) \sim (1 + s^2)^{-1}/2$ for $s \gg 1$, meaning that the gain in storage size and computational time is more significant as the matrix becomes sparser. This is depicted in Figure 5-(left).

Fixing α in Corollary 2 to achieve a given clustering performance level (via Proposition 1), one may then retrieve "equi-performance" curves in the (ε, ρ)-plane, for uniform sparsification, selective sparsification, and subsampling. This is displayed in Figure 5-(right), showing that a dramatic performance gain is achieved by the proposed selective sparsification. Besides, here for $c = 2$, as much as 80% sparsity may be obtained with selective sparsification at constant SNR ρ, with virtually no performance loss (red curves are almost flat on $\varepsilon \in [0.2, 1]$). This fails to hold for uniform sparsification (Zarrouk et al. (2020) obtain such a result only when $c \lesssim 0.1$) or subsampling.

4.3 Optimally quantized and binarized matrices

From Figure 4, we see that the classification error of the quantized $f_2(M; s; t)$ and binarized $f_3(s; t)$ do *not* increase monotonically with the truncation threshold s. It can be shown (and visually confirmed in Figure 4) that, for a given $M \geq 2$, the ratio ν/a_1^2 of both f_2 and f_3 is convex in s and has a unique minimum. This leads to the following result, the proof of which is straightforward.

Proposition 2 (Optimal design of quantized and binarized functions)**.** *Under the assumptions and notations of Proposition 1, the classification error rate is minimized at $s = s_{\mathrm{opt}}$ with*

1. s_{opt} the unique solution to $a_1(s_{\mathrm{opt}})\nu'(s_{\mathrm{opt}}) = 2a_1'(s_{\mathrm{opt}})\nu(s_{\mathrm{opt}})$ for quantized f_2, with $a_1'(s)$ and $\nu'(s)$ the corresponding derivatives with respect to s in Figure 1-(right); and

2. $s_{\mathrm{opt}} = \exp(-s_{\mathrm{opt}}^2)/(2\sqrt{\pi}\,\mathrm{erfc}(s_{\mathrm{opt}})) \approx 0.43$ for binary f_3, with $\nu/a_1^2 \approx 1.24$ and $\varepsilon \approx 0.54$.

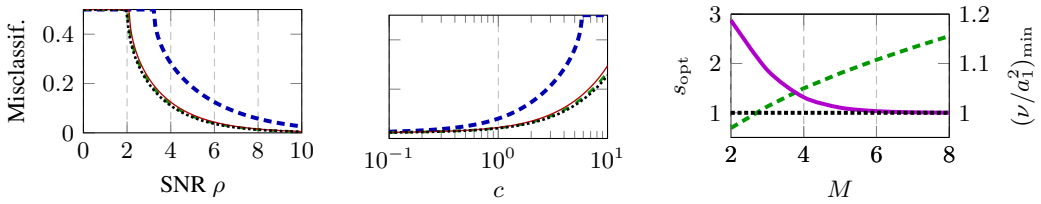


Figure 6: Performance of f_1 (**blue**), f_2 with $M = 3$ (**green**) and f_3 (**red**) of the same storage size, versus SNR for $c = 4$ (**left**) and versus c for SNR $\rho = 4$ (**middle**). **(Right)** Optimal threshold s_{opt} (**green**) and $(\nu/a_1^2)_{\min}$ (**purple**) of f_2 versus M. Curves for linear $f(t) = t$ are displayed in **black**.

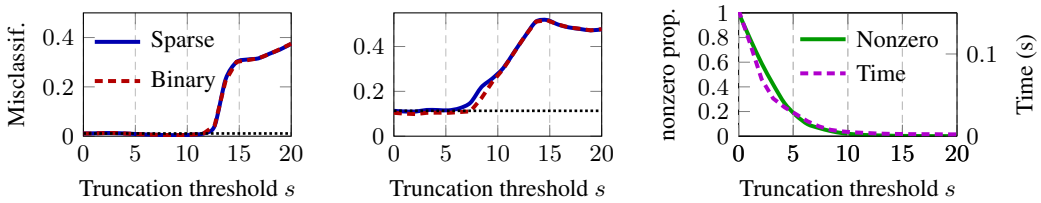


Figure 7: Clustering performance of sparse f_1 and binary f_3 (**left** and **middle**), proportion of nonzero entries and computational time of the top eigenvector for f_3 (**right**), as a function of the truncation threshold s on the MNIST dataset: digits $(0, 1)$ (**left**) and $(5, 6)$ (**middle** and **right**) with $n = 2\,048$ and performance of the linear function in **black**, Results averaged over 100 runs.

Therefore, the optimal threshold s_{opt} for quantized f_2 or binary f_3 under (1) is *problem-independent*, as it depends neither on ρ not on c. In particular, note that **(i)** the binary $f_3(s_{\text{opt}}; \cdot)$ is *consistently* better than $f(t) = \text{sign}(t)$ for which $\nu/a_1^2 = \pi/2 \approx 1.57 > 1.24$; and **(ii)** the performance of quantized f_2 can be *worse*, though very slightly, than that of binary f_3 for small s, but significantly better for not-too-small s. These are visually confirmed in the left and middle displays of Figure 4.

As already observed in Figure 4-(right), a significant gain in storage size can be achieved by using f_2 or f_3, versus the performance-optimal but dense linear function, with virtually no performance loss. Figure 6 compares the performance of the optimally designed f_2 and f_3, to sparse f_1 that has *approximately* the same storage size.[5] A significant drop in classification error is observed by using quantization f_2 or binarization f_3 rather than sparsification f_1. Also, the performances of f_2 and f_3 are extremely close to the theoretical optimal (met by $f(t) = t$). This is further confirmed by Figure 6-(right) where, for the optimal f_2, the ratio ν/a_1^2 gets close to 1, for all $M \geq 5$.

Figure 7 next evaluates the clustering performance, the proportion of nonzero entries in $\mathbf{K}$, and the computational time of the top eigenvector, for sparse f_1 and binary f_3, versus linear $f(t) = t$, as a function of the truncation threshold s, on the popular MNIST datasets (LeCun et al., 1998). Depending on (the SNR ρ of) the task, up to 90% of the entries can be discarded almost "for free". Moreover, the curves of the binary f_3 appear strikingly close to those of the sparse f_1, showing the additional advantage of using the former to further reduce the storage size of $\mathbf{K}$. More empirical results on various datasets are provided in Appendix B to confirm our observations in Figure 7.

5 Concluding Remarks

We have evaluated performance-complexity trade-offs when sparsifying, quantizing, and binarizing a linear kernel matrix via a thresholding operator. Our main technical result characterizes the change in the eigenspectrum under these operations; and we have shown that, under an information-plus-noise mixture data, sparsification and quantization, when carefully employed, maintain the informative eigenstructure and incur almost negligible performance loss in spectral clustering. Empirical results on real data demonstrate that these conclusions hold far beyond the present statistical model.

Our results opens the door to theoretical investigations of a broad range of cost-efficient methods in machine learning, including subsampling techniques (Mensch et al., 2017), distributed optimization (Wang et al., 2018), randomized algorithms (Mahoney, 2011), and provides seminal keys to the understanding and improvement of the performance-complexity trade-off beyond kernels: to sparse, quantized, or even binary neural networks (Courbariaux et al., 2016) of growing interest today.

[5] We set the threshold s of f_1 such that $\text{erfc}(s) = 3/64$, so that the storage size of sparse f_1 (64 bits per entry) is the same as quantized f_2 with $M = 3$ (with 3 bits per entry), which is *three times* that of binary f_3.

REFERENCES

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

Dimitris Achlioptas and Frank McSherry. Fast computation of low-rank matrix approximations. *Journal of the ACM (JACM)*, 54(2):9–es, 2007.

George E Andrews, Richard Askey, and Ranjan Roy. *Special functions*, volume 71. Cambridge university press, 2000.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

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, 2005.

Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature, 2018.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to $+1$ or -1. *arXiv preprint arXiv:1602.02830*, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Yen Do and Van Vu. The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 2(03):1350005, 2013.

P. Drineas and A. Zouzias. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality. *Information Processing Letters*, 111:385–389, 2011.

Noureddine El Karoui. On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5):3191–3216, 2010.

Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1-2):27–85, 2019.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

Arthur Jacot, Franck Gabriel, and Clement Hongler. The asymptotic spectrum of the hessian of dnn throughout training. In *International Conference on Learning Representations*, 2019.

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)*, pp. 694–698. IEEE, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ping Li and Phan-Minh Nguyen. On random deep weight-tied autoencoders: Exact asymptotic analysis, phase transitions, and implications to training. In *International Conference on Learning Representations*, 2018.

Zhenyu Liao and Romain Couillet. Inner-product kernels are asymptotically equivalent to binary discrete kernels. *arXiv preprint arXiv:1909.06788*, 2019.

Anna Lytova, Leonid Pastur, et al. Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *The Annals of Probability*, 37(5):1778–1840, 2009.

M. W. Mahoney. Lecture notes on spectral graph methods. Technical report, 2016. Preprint: arXiv:1608.04845.

Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Vladimir A Marcenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Stochastic subsampling for factorizing huge matrices. *IEEE Transactions on Signal Processing*, 66(1):113–128, 2017.

Vinay Uday Prabhu. Kannada-mnist: A new handwritten digits dataset for the kannada language. *arXiv preprint arXiv:1908.01242*, 2019.

Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Alaa Saade, Florent Krzakala, and Lenka Zdeborová. Spectral clustering of graphs with the bethe hessian. In *Advances in Neural Information Processing Systems*, pp. 406–414, 2014.

Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. A kernel random matrix-based approach for sparse pca. In *International Conference on Learning Representations*, 2018.

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. *arXiv preprint arXiv:2001.08370*, 2020.

Jack W Silverstein and ZD Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.

Jack W Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15: 1929–1958, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Terence Tao, Van Vu, Manjunath Krishnapur, et al. Random matrices: Universality of esds and the circular law. *The Annals of Probability*, 38(5):2023–2065, 2010.

Dan Voiculescu. Addition of certain non-commuting random variables. *Journal of functional analysis*, 66(3):323–346, 1986.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Shusen Wang, Fred Roosta, Peng Xu, and Michael W Mahoney. Giant: Globally improved approximate newton method for distributed optimization. In *Advances in Neural Information Processing Systems*, pp. 2332–2342, 2018.

Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Tayeb Zarrouk, Romain Couillet, Florent Chatelain, and Nicolas Le Bihan. Performance-complexity trade-off in large dimensional statistics. In *2020 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2020.