Performance-complexity Trade-off in Large Dimensional Spectral Clustering

STA 290 Seminar, Dept. Stats, UC Davis

Zhenyu Liao

with Romain Couillet@Grenoble-Alpes and Michael Mahoney@UC Berkeley

ICSI and Department of Statistics University of California, Berkeley, USA

January 22, 2021

Outline

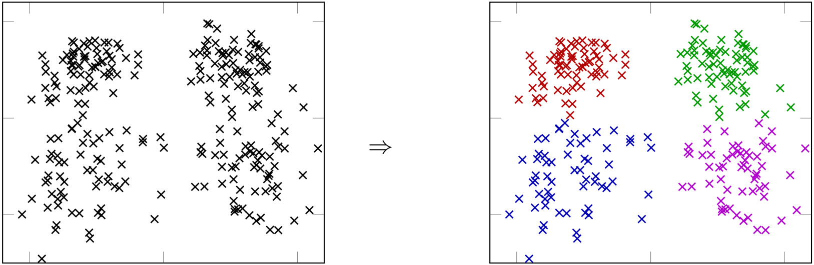
1. Introduction
2. Main Results
   - Model and problem setting
   - Uniform sparsification
   - Non-uniform sparsification, quantization, and binarization
3. Conclusion

Motivation: computationally efficient machine learning

- **Big Data**: number of data n and dimension p both large, thousands or millions
- ImageNet dataset (`http://www.image-net.org/`): in average $p = 0.2$ million pixels of in total $n = 14$ million high-resolution images
- Computational challenge: time and/or space complexity at least $O(n^2)$, unaffordable for Internet of Things (IoT) low-power devices
- **Idea**: compress machine learning models (e.g., sketching, quantized or binarized neural networks), with non-trivial performance-complexity trade-off
- **Objective**: theoretical understanding of performance-complexity trade-off, optimal design, how they depend on the data
- **Example**: unsupervised (kernel) spectral clustering

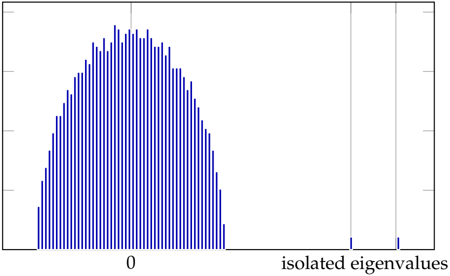
Remainder on clustering

- **Clustering**: unsurprised learning method to find possible groups/clusters from the data, with no pre-existing labels
- 2D example:

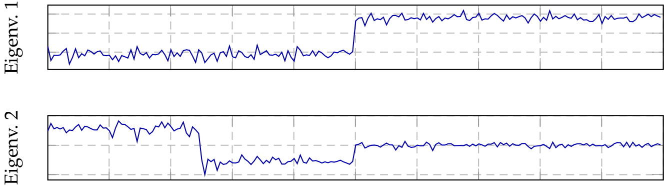


Reminder on kernel spectral clustering

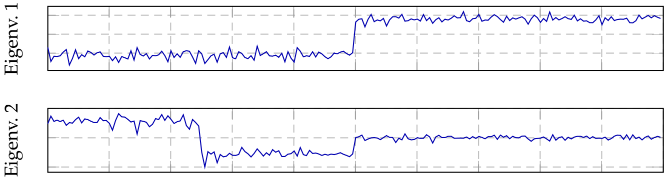
Two-step clustering of n data points based on kernel matrix $\mathbf{K} = \{f(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$:



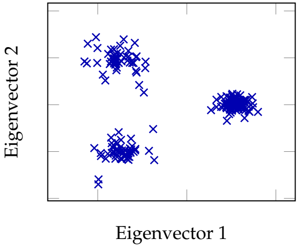
$\Downarrow$ **Top eigenvectors** $\Downarrow$



Reminder on kernel spectral clustering



$\Downarrow$ **K-dimensional representation** $\Downarrow$



$\Downarrow$

EM or k-means clustering.

Computational challenge in spectral clustering

- kernel/similarity matrix $\mathbf{K} = \{f(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$: pairwise comparison of n data points
- retrieve the top eigenvectors of $\mathbf{K} \in \mathbb{R}^{n \times n}$ with e.g., power method: suffer from an $O(n^2)$ complexity
- **Idea**: sparsifying, quantizing, and even binarizing: gain in both time and space!
- **Key object**: eigenspectrum of the "compressed" kernel matrix, in particular, statistics of top eigenvectors!

System model

Data: two-class signal-plus-noise mixture

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

$$\mathcal{C}_1 : \mathbf{x}_i \sim \mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_p), \quad \mathcal{C}_2 : \mathbf{x}_i \sim \mathcal{N}(+\boldsymbol{\mu}, \mathbf{I}_p). \tag{1}$$

We have $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] = \mathbf{Z} + \boldsymbol{\mu}\mathbf{v}^\mathsf{T}$ for Gaussian $\mathbf{Z} \in \mathbb{R}^{p \times n}$, $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\mathbf{v} \in \{\pm 1\}^n$.

Large dimensional asymptotics

As $n, p \to \infty$ with $p/n \to c \in (0, \infty)$ and signal-to-noise ratio (SNR) $\|\boldsymbol{\mu}\|^2 \to \rho \geq 0$.

Previous work:

- Gram (kernel) matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$, extensively studied in random matrix theory
- (limiting) eigenvalue distribution: the Marčenko-Pastur law [MP67]
- spiked model and phase transition of top eigenvalue-eigenvector [BBP05]

[1] Vladimir A Marčenko and Leonid Andreevich Pastur. "Distribution of eigenvalues for some sets of random matrices". In: *Mathematics of the USSR-Sbornik* 1.4 (1967), p. 457

[2] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *The Annals of Probability* 33.5 (2005), pp. 1643–1697

Previous work

- for $\|\boldsymbol{\mu}\| = 0$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, eigenvalue distribution of $\mathbf{X}^\mathsf{T}\mathbf{X}/p$ converges to the Marčenko–Pastur law

$$\mu(dx) = (1-c)^+\delta(x) + \frac{1}{2\pi x}\sqrt{(x-a)^+(b-x)^+}dx$$

where $a = (1 - 1/\sqrt{c})^2$, $b = (1 + 1/\sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$.

- for $\|\boldsymbol{\mu}\| > 0$, depending on SNR $\rho = \lim \|\boldsymbol{\mu}\|^2$, one *isolated* eigenvalue may "jump" out of the Marčenko–Pastur bulk, with associated eigenvector aligned to $\mathbf{v}$!

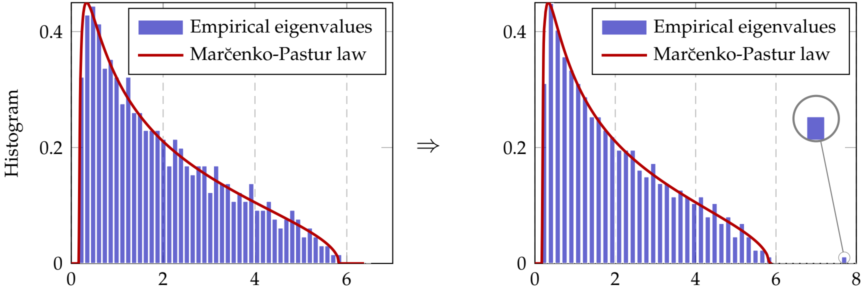


Figure: Eigenvalues of $\mathbf{X}^\mathsf{T}\mathbf{X}/p$ versus the Marčenko-Pastur law, $p = 512$, $n = 1\,024$, with $\rho = 0$ **(left)** and $\rho = 2$ **(right)**.

Uniform sparsification: method

Objective: "compress" linear Gram matrix $\mathbf{X}^\mathsf{T}\mathbf{X} \in \mathbb{R}^{n\times n}$.

Uniform sparsification

Setting uniformly a proportion $1-\varepsilon$ entries to zero with a symmetric Bernoulli mask $\mathbf{B} \in \{0,1\}^{n\times n}$

$$\mathbf{K} = \frac{1}{p}\mathbf{X}^\mathsf{T}\mathbf{X} \odot \mathbf{B}, \quad \mathbf{B}_{ij} \sim \mathrm{Bern}(\varepsilon) \text{ for } 1 \leq i < j \leq n \tag{2}$$

with $\odot$ the (entry-wise) Hadamard product, $[\mathbf{B}]_{ji} = [\mathbf{B}]_{ij}$ and $[\mathbf{B}]_{ii} = b \in \{0,1\}$.

$\Rightarrow$ Evaluate clustering performance of $\mathbf{K}$ via eigenspectrum study: limiting eigenvalue distribution, statistics of the top eigenvalue-eigenvector pair.

Key object: resolvent matrix $\mathbf{Q}(z) = (\mathbf{K} - z\mathbf{I}_n)^{-1}$ for $z \in \mathbb{C}$ not an eigenvalue of $\mathbf{K}$.

- $\frac{1}{n}\operatorname{tr}\mathbf{Q}(z)$ is the *Stieltjes transform* of the eigenvalue distribution of $\mathbf{K}$
- used to characterize the phase transition (of isolated eigenvalue-eigenvector) beyond which spectral clustering becomes theoretically possible
- for $(\hat{\lambda}, \hat{\mathbf{v}})$ an eigenpair of $\mathbf{K}$ and label vector $\mathbf{v} \in \mathbb{R}^n$, by Cauchy's integral formula, the "angle": $|\hat{\mathbf{v}}^\mathsf{T}\mathbf{v}|^2 = -\frac{1}{2\pi\imath}\oint_{\Gamma(\hat{\lambda})} \mathbf{v}^\mathsf{T}\mathbf{Q}(z)\mathbf{v}\,dz$, for $\Gamma(\hat{\lambda})$ positively circling $\hat{\lambda}$

Uniform sparsification: performance analysis

Theorem (Limiting spectral measure)

As $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, the empirical spectral measure $\omega_{\mathbf{K}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{K})}$ of $\mathbf{K}$ converges to a deterministic limit ω, uniquely defined through its Stieltjes transform $m(z) = \int (t-z)^{-1} \omega(dt)$ solution to

$$z = b - \frac{1}{m(z)} - \frac{\varepsilon}{c} m(z) + \frac{\varepsilon^3 m^2(z)}{c(c + \varepsilon m(z))}. \tag{3}$$

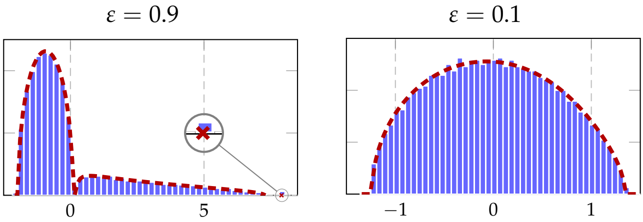
Theorem (Isolated eigenpair and a phase transition)

Define $F(x) = x^4 + 2x^3 + \left(1 - \frac{c}{\varepsilon}\right) x^2 - 2cx - c$, $G(x) = b + \frac{\varepsilon}{c}(1+x) + \frac{1}{1+x} + \frac{\varepsilon}{x(1+x)}$ and let γ be the largest real solution to $F(\gamma) = 0$. Then, the largest eigenpair $(\hat{\lambda}, \hat{\mathbf{v}})$ of $\mathbf{K}$ satisfies

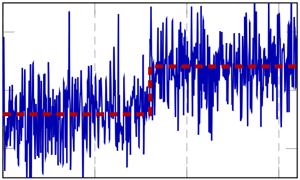
$$\hat{\lambda} \to \lambda = \begin{cases} G(\rho), & \rho > \gamma \\ G(\gamma), & \rho \leq \gamma \end{cases}, \quad \frac{1}{n} |\hat{\mathbf{v}}^\mathsf{T} \mathbf{v}|^2 \to \alpha = \begin{cases} \frac{F(\rho)}{\rho(1+\rho)^3}, & \rho > \gamma \\ 0, & \rho \leq \gamma \end{cases} \tag{4}$$

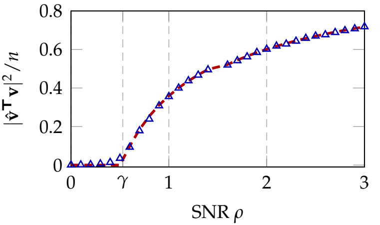
as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, for SNR $\rho = \lim \|\boldsymbol{\mu}\|^2$.

Uniform sparsification: implications



Remark: becomes the Marčenko–Pastur law (of $\mathbf{X}^\mathsf{T}\mathbf{X}/p$) as $\varepsilon \to 1$ and semicircle law as $\varepsilon \to 0$, a "mixed" of behavior in the sense of *free additive convolution* [Voi86].





[1] Dan Voiculescu. "Addition of certain non-commuting random variables". In: *Journal of Functional Analysis* 66.3 (1986), pp. 323–346

Non-uniform "compressed" spectral clustering: method

Intuition: can we do better by treating the entries in a non-uniform manner?

Non-uniform compression

Entry-wise *nonlinear* transformation of $\mathbf{X}^\mathsf{T}\mathbf{X}$:

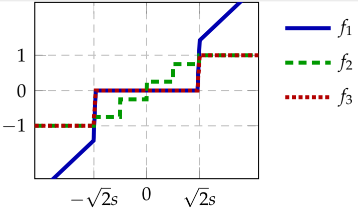
$$\mathbf{K} = \left\{ f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p} \right\}_{i,j=1}^{n} \tag{5}$$

with

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

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

Binarization: $f_3(t) = \mathrm{sign}(t) \cdot 1_{|t|>\sqrt{2}s}$



Tuning parameters:

- truncation threshold $s > 0$
- number of information bits M

Compressed kernel matrix: intuition

Object of interest

Entry-wise *nonlinear* transformation of $\mathbf{X}^\mathsf{T}\mathbf{X}$:

$$\mathbf{K} = \left\{ f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p} \right\}_{i,j=1}^{n} \tag{6}$$

Recall $\mathbf{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$ with $\|\boldsymbol{\mu}\| = O(1)$, so $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p} \to \mathcal{N}(0,1)$ in law as $p \to \infty$.

$$\boxed{\sqrt{p}[\mathbf{K}]_{ij} \simeq f(\mathcal{N}(0,1)).}$$

Notations

For each f and $\xi \sim \mathcal{N}(0,1)$, define the (generalized) moments

$$a_0 = \mathbb{E}[f(\xi)] = 0, \quad a_1 = \mathbb{E}[\xi f(\xi)], \quad \sqrt{2}a_2 = \mathbb{E}[\xi^2 f(\xi)], \quad \nu = \mathbb{E}[f^2(\xi)] \geq a_1^2 + a_2^2. \tag{7}$$

"Compressed" spectral clustering: performance analysis

For each f and $\xi \sim \mathcal{N}(0,1)$, define the (generalized) moments

$$a_0 = \mathbb{E}[f(\xi)] = 0, \quad a_1 = \mathbb{E}[\xi f(\xi)], \quad \sqrt{2} a_2 = \mathbb{E}[\xi^2 f(\xi)], \quad \nu = \mathbb{E}[f^2(\xi)] \geq a_1^2 + a_2^2. \tag{8}$$

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

with $\mathbf{a_2 = 0}$, $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$, $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$ error/comple. error function.

Theorem (Limiting spectral measure)

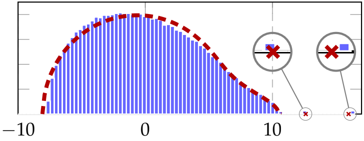
As $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, the empirical spectral measure $\omega_{\mathbf{K}} = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(\mathbf{K})}$ of $\mathbf{K}$ converges to a deterministic limit ω, uniquely defined through its Stieltjes transform $m(z) = \int (t-z)^{-1} \omega(dt)$ solution to

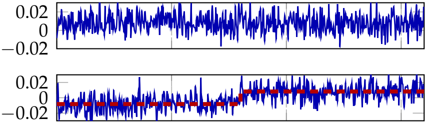
$$z = -\frac{1}{m(z)} - \frac{\nu - a_1^2}{c} m(z) - \frac{a_1^2 m(z)}{c + a_1 m(z)}. \tag{9}$$

"Compressed" spectral clustering: attention!

Remark (Spurious non-informative spikes)

If $\mathbf{a_2} \neq \mathbf{0}$, then there may be *up to two* **non-informative** eigenvalues (with eigenvectors containing only random noise) on the *left or right* of the main bulk.





Theorem (Informative spike and a phase transition)

For $a_1 > 0$ and $a_2 = 0$, similarly define $F(x) = x^4 + 2x^3 + \left(1 - \frac{c\nu}{a_1^2}\right) x^2 - 2cx - c$ *and* $G(x) = \frac{a_1}{c}(1 + x) + \frac{a_1}{x} + \frac{\nu - a_1^2}{a_1}\frac{1}{1+x}$ *and let γ be the largest real solution to $F(\gamma) = 0$. Then,*

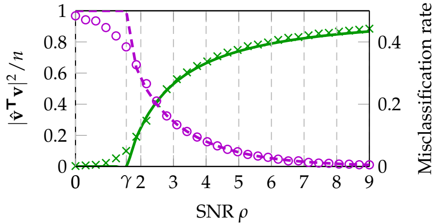
$$\hat{\lambda} \to \lambda = \begin{cases} G(\rho), & \rho > \gamma \\ G(\gamma), & \rho \leq \gamma \end{cases}, \quad \frac{1}{n}|\hat{\mathbf{v}}^\mathsf{T}\mathbf{v}|^2 \to \alpha = \begin{cases} \frac{F(\rho)}{\rho(1+\rho)^3}, & \rho > \gamma \\ 0, & \rho \leq \gamma \end{cases} \tag{10}$$

as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, for SNR $\rho = \lim \|\boldsymbol{\mu}\|^2$.

"Compressed" spectral clustering: practical implications

Corollary (Performance of spectral clustering)

Let $a_1 > 0, a_2 = 0$, *and let* $\hat{\mathcal{C}}_i = \operatorname{sign}([\hat{\mathbf{v}}]_i)$ *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* $\frac{1}{n}\sum_{i=1}^n \delta_{\hat{\mathcal{C}}_i \neq \mathcal{C}_i} \to \frac{1}{2}\operatorname{erfc}(\sqrt{\alpha/(2-2\alpha)})$, *as* $n, p \to \infty$, *for* α *the limit of the eigenvector alignment* $\frac{1}{n}|\hat{\mathbf{v}}^\mathsf{T}\mathbf{v}|^2$.



Remark (Optimality of linear $f(t) = t$)

Both phase transition point γ and misclassification rate grow with ν/a_1^2, the linear $f(t) = t$ with minimal $\nu/a_1^2 = 1$ is *optimal* in: **(i)** *smallest* SNR ρ or *largest* ratio p/n to observe a spike, and **(ii)** upon existence, reaching *lowest* classification error rate.

Uniform versus non-uniform sparsification

Comparison between uniform (Bernoulli) sparsification and "selective" non-uniform sparsification $f_1(t) = t \cdot 1_{|t|>\sqrt{2}s}$. **Same** performance with **different** *level of sparsity*:

$$\varepsilon_{\text{unif}} = \operatorname{erfc}(s) + 2se^{-s^2}/\sqrt{\pi} > \operatorname{erfc}(s) = \varepsilon_{\text{selec}} \tag{11}$$

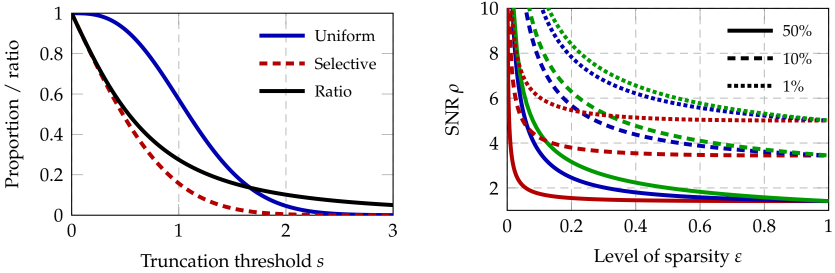
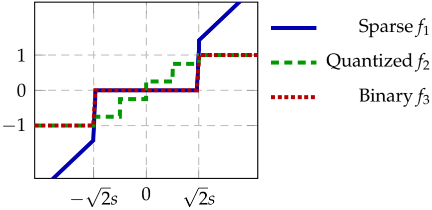


Figure: **(Left)** Proportion of non-zero entries with uniform versus selective sparsification f_1 and their ratio, as a function of the truncation threshold s. **(Right)** Comparison of 1%, 10% error and phase transition (i.e., 50% error) curves between subsampling (**green**), uniform (**blue**) and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



Tuning parameters:

- truncation threshold $s > 0$
- number of information bits M

Performance depends on f **only** via $\nu/a_1^2 \Rightarrow$ Convex in s for quantized f_2 and binary f_3!

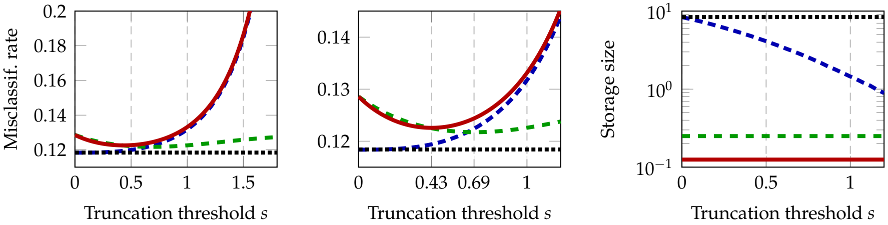


Figure: Clustering performance (**left**, a zoom-in in **middle**) and storage size (MB) **(right)** of f_1 (**blue**), f_2 with $M = 2$ (**green**), f_3 (**red**), and linear $f(t) = t$ (**black**), versus the truncation threshold s, for SNR $\rho = 2$, $c = 1/2$ and $n = 10^3$, with 64 bits per entry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

Experiments on real-world data

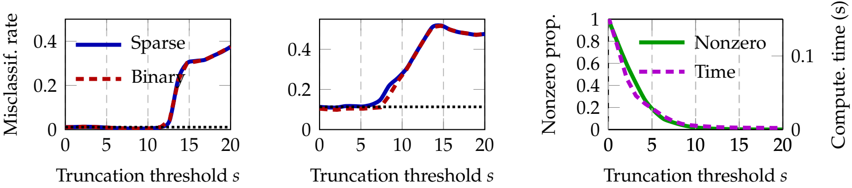


Figure: Clustering performance (**left** and **middle**), proportion of nonzero entries and computational time of the top eigenvector for f_3 (**right**), on the MNIST dataset: digits $(0, 1)$ (**left**) and $(5, 6)$ (**middle** and **right**) with $n = 2\,048$ and performance of the linear function in **black**.

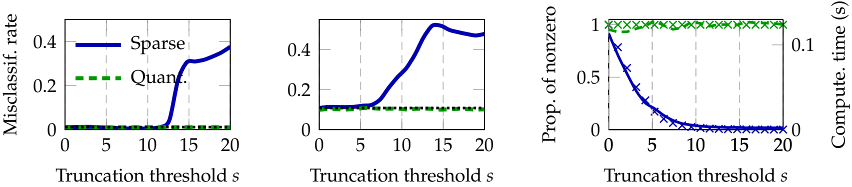


Figure: Clustering performance (**left** and **middle**), proportion of nonzero entries, and computational time of the top eigenvector (**right**, in markers) of sparse f_1 and quantized f_2 with $M = 2$, on the MNIST dataset.

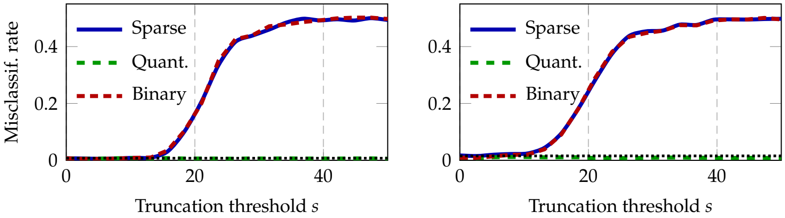


Figure: Clustering performance of sparse f_1, quantized f_2 (with $M = 2$) and binary f_3 as a function of the truncation threshold s on *GoogLeNet* features of the ImageNet datasets: (**left**) class "pizza" versus "daisy" and (**right**) class "hamburger" versus "coffee", for $n = 1\,024$ and performance of the linear function in **black**. Results averaged over 10 runs.

Take-away message:

- theoretical analysis of **performance-complexity trade-offs** in computationally efficient machine learning methods
- non-uniform treatment significantly outperforms uniform (sparsification) scheme
- spurious non-informative eigenvectors may appear if not properly done!

Future work:

- more generic model, e.g., K-class $\mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, $a \in \{1, \ldots, K\}$
- nonlinear transformation in modern ML, e.g., neural nets

References:

- Tayeb Zarrouk et al. "Performance-complexity trade-off in large dimensional statistics". In: *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2020, pp. 1–6
- Zhenyu Liao, Romain Couillet, and Michael W Mahoney. "Sparse quantized spectral clustering". In: *arXiv preprint arXiv:2010.01376* (2020). Accepted for publication, Proc. of the 2021 ICLR Conference.

and my homepage `https://zhenyu-liao.github.io/` for more information!

Thank you!