Sparse Quantized Spectral Clustering

Ninth International Conference on Learning Representations (ICLR 2021)

Zhenyu Liao

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

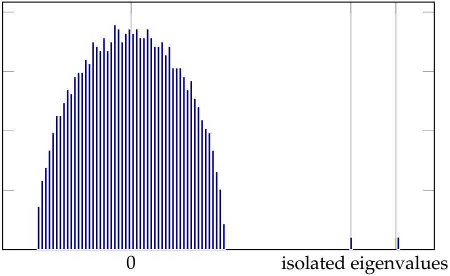
ICSI and Department of Statistics University of California, Berkeley, USA

Motivation: computationally efficient machine learning

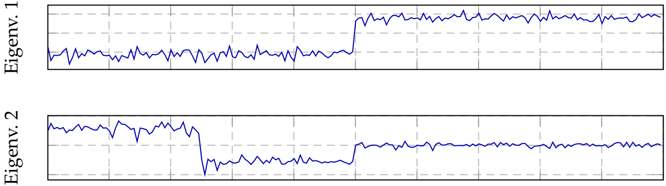
- **Big Data**: number of data n and dimension p both large, thousands or millions
- ImageNet dataset (`http://www.image-net.org/`): in average $p = 0.2$ million pixels of in total $n = 14$ million high-resolution images
- Computational challenge: time and/or space complexity at least $O(n^2)$, unaffordable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

Reminder on kernel spectral clustering

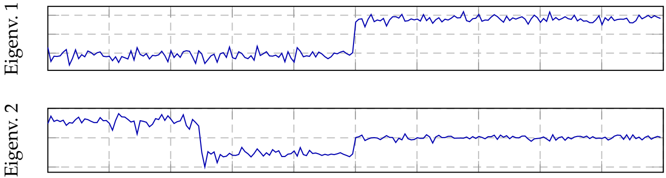
Two-step clustering of n data points based on kernel matrix $\mathbf{K} = \{f(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$:



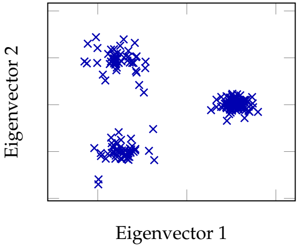
$\Downarrow$ **Top eigenvectors** $\Downarrow$



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:

- Dense 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

"Compressed" spectral clustering: method

Compression as nonlinear transformation

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

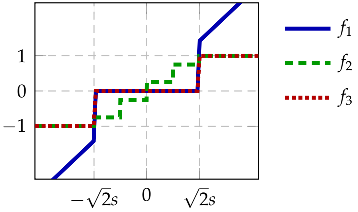
$$\mathbf{K} = \left\{ f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p} \right\}_{i,j=1}^n \tag{2}$$

with

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

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

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



Tuning parameters:

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

"Compressed" spectral clustering: performance analysis

Notations

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

$$a_0 = \mathbb{E}[f(\xi)] = 0, \quad \mathbf{a_1} = \mathbb{E}[\xi f(\xi)], \quad \mathbf{a_2} = \mathbb{E}[\xi^2 f(\xi)]/\sqrt{2}, \quad \boldsymbol{\nu} = \mathbb{E}[f^2(\xi)] \geq a_1^2 + a_2^2. \tag{3}$$

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

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

$$z = -\frac{1}{m(z)} - \frac{\nu - a_1{}^2}{c} m(z) - \frac{a_1{}^2 m(z)}{c + a_1 m(z)}. \tag{4}$$

"Compressed" spectral clustering: attention!

Theorem (Informative spike and a phase transition)

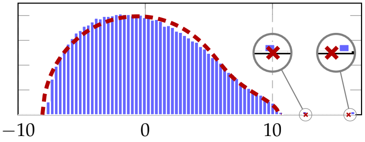
For $a_1 > 0$ *and* $\mathbf{a_2 = 0}$, *similarly define* $F(x) = x^4 + 2x^3 + \left(1 - \frac{c\nu}{a_1{}^2}\right) x^2 - 2cx - c$ *and* $G(x) = \frac{a_1}{c}(1+x) + \frac{a_1}{x} + \frac{\nu - a_1{}^2}{a_1}\frac{1}{1+x}$ *and let* γ *be the largest real solution to* $F(\gamma) = 0$. *Then,*

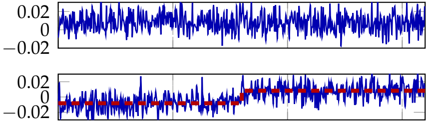
$$\hat{\lambda} \to \lambda = \begin{cases} G(\rho), & \rho > \gamma \\ G(\gamma), & \rho \leq \gamma \end{cases}, \quad \frac{1}{n}|\hat{\mathbf{v}}^\mathsf{T}\mathbf{v}|^2 \to \alpha = \begin{cases} \frac{F(\rho)}{\rho(1+\rho)^3}, & \rho > \gamma \\ 0, & \rho \leq \gamma \end{cases} \tag{5}$$

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

Remark (Spurious non-informative spikes)

If $\mathbf{a_2 \neq 0}$, 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.





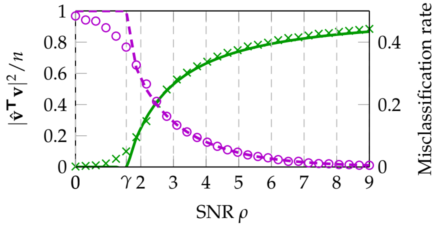
"Compressed" spectral clustering: practical implications

Corollary (Performance of spectral clustering)

Let $a_1 > 0, a_2 = 0$, and $\hat{\mathcal{C}}_i = \operatorname{sign}([\hat{\mathbf{v}}]_i)$ be the estimate of the underlying class $\mathcal{C}_i$ of the datum $\mathbf{x}_i$, with $\hat{\mathbf{v}}^\mathsf{T}\mathbf{v} \geq 0$ 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 eigenvector alignment $\frac{1}{n}|\hat{\mathbf{v}}^\mathsf{T}\mathbf{v}|^2$.



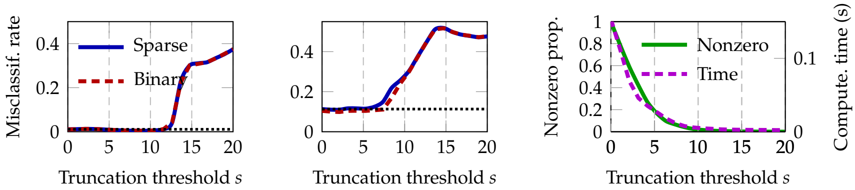


Figure: Clustering performance (**left** and **middle**), proportion of nonzero entries and computational time of the top eigenvector for f_3 (**right**), on the **MNIST** dataset: digits $(0, 1)$ (**left**) and $(5, 6)$ (**middle** and **right**) with $n = 2\,048$ and performance of the linear function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

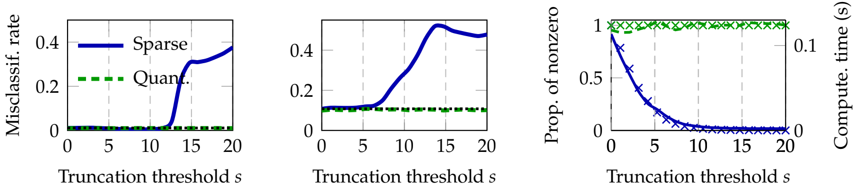


Figure: Clustering performance (**left** and **middle**), proportion of nonzero entries, and computational time of the top eigenvector (**right**, in markers) of sparse f_1 and quantized f_2 with $M = 2$, on the **MNIST** dataset.

Experiments on real-world data

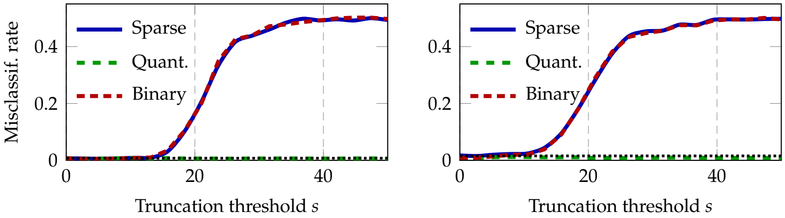


Figure: Clustering performance of sparse f_1, quantized f_2 (with $M = 2$) and binary f_3 as a function of the truncation threshold s on *GoogLeNet* features of the **ImageNet** datasets: (**left**) class "pizza" versus "daisy" and (**right**) class "hamburger" versus "coffee", for $n = 1\,024$ and performance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

Conclusion and take-away message

Take-away message:

- theoretical analysis of **performance-complexity trade-offs** in computationally efficient machine learning methods
- compare with [Zar+20]: non-uniform treatment significantly outperforms uniform (sparsification) scheme
- spurious non-informative eigenvectors may appear if not properly done!

References:

- Zhenyu Liao, Romain Couillet, and Michael W. Mahoney. "Sparse Quantized Spectral Clustering". In: *International Conference on Learning Representations*. 2021.

talk at **Poster Session 10**, and `https://zhenyu-liao.github.io/` for more info!

Thank you!