Random Moments for Sketched Mixture Learning

Nicolas Keriven, Rémi Gribonval, Gilles Blanchard, Yann Traonmilin

Random Moments for Sketched Mixture Learning

Nicolas Keriven, Rémi Gribonval, Yann Traonmilin
INRIA Rennes-Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes, France
Email: firstname.lastname@inria.fr

Gilles Blanchard
Mathematics Institute, University of Potsdam
14476 Potsdam, Germany
Email: gilles.blanchard@math.uni-potsdam.de

***Abstract*—We present a method to solve large-scale mixture learning tasks from a *sketch* of the data, formed by random generalized empirical moments. We give empirical and theoretical results on k-means and Gaussian Mixture Model estimation problems.**

I. Introduction

Consider samples $z_i \in \mathbb{R}^d$, $1 \leq i \leq n$, drawn *i.i.d.* from a distribution π. Given a class of *hypotheses* $\mathcal{H}$ and a *loss function* $\ell : \mathbb{R}^d \times \mathcal{H} \rightarrow \mathbb{R}$, statistical learning consists in finding the hypothesis $h^* \in \mathcal{H}$ that minimizes the *expected risk* $\mathcal{R}(h) = \mathbb{E}_\pi \ell(z, h)$. Since the distribution π is not directly available, usual learning procedures minimize the empirical risk instead: $\hat{\mathcal{R}}_n(h) = \sum_i \ell(z_i, h)/n$.

This traditional approach is however challenged when samples z are high-dimensional (large d) or in great number (large n). The first case has been dealt with using random projections [1] or feature selection [2], while the second gave birth to online learning [3] or coresets [4]. We advocate here that when n is large, some learning tasks can be done using only a collection of generalized empirical moments, referred to as *sketch*, as a (highly) compressed representation of the database. A simple example is Principal Component Analysis (PCA), which can be done with only the empirical covariance. Such sketches can be computed online, and/or in a distributed/parallel manner, and do not require the database to be stored on one single device.

We present here a method to perform k-means or Gaussian Mixture Model (GMM) estimation with identity covariance from a sketch formed by a (weighted) random sampling of the characteristic function. Such inverse problems bear similarities with sparse recovery in continuous spaces [5]. Define the sketching operator:

$$\mathcal{A}\pi = \frac{1}{\sqrt{m}} \left[\mathbb{E}_{z \sim \pi} \exp(-i\omega_j^T z)/c_{\omega_j} \right]_{j=1}^m \tag{1}$$

where $c_{\omega_j} > 0$ are some weights and frequencies $\omega_j \in \mathbb{R}^d$ are drawn *i.i.d.* from a weighted Gaussian distribution $\Lambda(\omega) \propto c_\omega^2 \mathcal{N}(0, \sigma^2 \mathbf{I})$. The empirical sketch used in practice is denoted $\mathbf{y} = \frac{1}{n\sqrt{m}} \left[\sum_{i=1}^n \exp(-i\omega_j^T z_i)/c_{\omega_j} \right]_{j=1}^m$.

II. Main Results

We now present our main results on k-means and GMM estimation. In each case, c_ω and σ^2 are appropriately chosen and not detailed in this abstract. Leveraging tools from kernel embeddings of distributions [6] and Random Fourier features [7], our analysis is inspired by Compressive Sensing results [8], [9], adapted to the proposed infinite-dimensional framework.

A. k-means

In the k-means problem, each hypothesis is a set of centroids $h = \{\mathbf{c}_1, ..., \mathbf{c}_k\}$ and the loss function is $\ell(z, h) = \min_l \|z - \mathbf{c}_l\|_2^2$.
Assumptions. We restrict to a family of hypotheses where centroids are 2ε-separated from each other and contained in a ball of radius M, and denote $\mathcal{H}_{k,\varepsilon,M}$ the corresponding class of hypotheses.
Result. Denote $h^* \in \mathcal{H}_{k,\varepsilon,M}$ the true minimizer of the expected risk $\mathcal{R}$ and $\hat{h}$ the hypothesis recovered from the sketch by

$$\hat{h} = \underset{h \in \mathcal{H}_{k,\varepsilon,M}}{\operatorname{argmin}} \min_{\alpha_1, ..., \alpha_k} \left\| \mathbf{y} - \mathcal{A}\left(\sum_{l=1}^k \alpha_l \delta_{\mathbf{c}_l} \right) \right\|_2 \tag{2}$$

where $\alpha_l \geq 0$ and $\sum_{l=1}^k \alpha_l = 1$.

If $m \geq \mathcal{O}\left(k^2 d^3 \texttt{polylog}(k, d) \log(1/\rho \cdot M/\varepsilon)\right)$, then with joint probability $1 - \rho$ on the drawing of z_i and ω_j it holds that

$$\mathcal{R}(\hat{h}) \lesssim \mathcal{R}(h^*) + \mathcal{O}\left(\sqrt{kd^2/n}\right). \tag{3}$$

B. Gaussian mixture with identity covariance

In the GMM learning problem, a hypothesis is a set of means and weights $h = \{\mu_1, ..., \mu_k, \alpha_1, ..., \alpha_k\}$ and the loss function is $\ell(z, h) = -\log \pi_h(z)$, where $\pi_h = \sum_{l=1}^k \alpha_l \mathcal{N}(\mu_l, \mathbf{I})$ is a GMM.
Assumptions. We restrict to a class of hypotheses where means are separated from each other and contained in a ball of radius M, and denote $\mathcal{H}_{k,M}$ the corresponding class of hypotheses. Unlike k-means, the separation between means cannot be as small as desired, and there is a trade-off between the required separation and the required number of measurements m. A few values are given in Table I.
Result. Denote $h^* \in \mathcal{H}_{k,M}$ the true minimizer of the expected risk $\mathcal{R}$ and $\hat{h}$ the hypothesis recovered from the sketch by solving

$$\hat{h} = \underset{h \in \mathcal{H}_{k,M}}{\operatorname{argmin}} \|\mathbf{y} - \mathcal{A}\pi_h\|_2. \tag{4}$$

If the number of measurements m is large enough (see Tab. I), with joint probability $1 - \rho$ on the drawing of z_i and ω_j it holds that

$$\mathcal{R}(\hat{h}) - \mathcal{R}(h^*) \lesssim \inf_{h \in \mathcal{H}_{k,M}} \|\pi - \pi_h\|_{\mathrm{TV}} + \mathcal{O}\left(\sqrt{1/n}\right) \tag{5}$$

where the $\mathcal{O}$ hides some dependencies in k, d (roughly behaving like m in Tab. I). The bound also involves the best approximation of π by a GMM for the TV norm (L^1 norm for densities).

III. Experimental Results

The optimization problems (2) and (4) are non-convex and seem hard to solve exactly. Heuristically, a greedy algorithm inspired by sparse recovery referred to as Compressive Learning OMP (CLOMP) [10]–[12] has been previously shown to perform well. We compare a Matlab implementation of CLOMP available at [13] with Matlab's `kmeans` function and VLFeat's [14] `gmm` function.

In Fig. 1, the sketched approach is seen to lead to tremendous savings in time of execution and memory consumption when the number of items n is large, while achieving the same precision as the corresponding traditional approach for a limited number of measurements $m \approx \mathcal{O}(kd)$. Fig. 2 further confirms that $m \approx \mathcal{O}(kd)$ is empirically sufficient, hence the theoretical guarantees for $m \gtrsim \mathcal{O}\left(k^2 d^2\right)$ are probably pessimistic.

Further work will combine the sketching technique with dimensionality-reduction methods to treat *both* large d and n.

TABLE I

TRADE-OFF BETWEEN REQUIRED SEPARATION OF MEANS AND NUMBER OF MEASUREMENTS IN THE GMM LEARNING PROBLEM.

Separation of means	Number of measurements
$\mathcal{O}\left(\sqrt{d \log k}\right)$	$m \geq \mathcal{O}\left(k^2 d^2 \mathtt{polylog}(k,d) \log(M/\rho)\right)$
$\mathcal{O}\left(\sqrt{d + \log k}\right)$	$m \geq \mathcal{O}\left(k^3 d^2 \mathtt{polylog}(k,d) \log(M/\rho)\right)$
$\mathcal{O}\left(\sqrt{\log k}\right)$	$m \geq \mathcal{O}\left(k^2 d^2 e^d \mathtt{polylog}(k,d) \log(M/\rho)\right)$

ACKNOWLEDGMENT

This work was supported in part by the European Research Council, PLEASE project (ERC-StG- 2011-277906).

REFERENCES

[1] C. Boutsidis, A. Zouzias, and P. Drineas, "Random Projections for k-means Clustering," in *Advances in Neural Information and Processing Systems (NIPS)*, 2010, pp. 298–306.

[2] J. Altschuler, A. Bhaskara, Gang, Fu, V. Mirrokni, A. Rostamizadeh, and M. Zadimoghaddam, "Greedy Column Subset Selection: New Bounds and Distributed Algorithms," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

[3] O. Cappé and E. Moulines, "Online EM Algorithm for Latent Data Models," *Journal of the Royal Statistical Society*, vol. 71, no. 3, pp. 593–613, 2009.

[4] D. Feldman and M. Langberg, "A unified framework for approximating and clustering data," *Proceedings of the forty-third annual ACM symposium on Theory of computing*, no. 46109, pp. 569–578, 2011.

[5] E. J. Candès and C. Fernandez-Granda, "Super-resolution from noisy data," *Journal of Fourier Analysis and Applications*, vol. 19, no. 6, pp. 1229–1254, 2013.

[6] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert Space Embedding for Distributions," in *International Conference on Algorithmic Learning Theory*, 2007, pp. 13–31.

[7] A. Rahimi and B. Recht, "Random Features for Large Scale Kernel Machines," *Advances in Neural Information Processing Systems (NIPS)*, no. 1, pp. 1–8, 2007.

[8] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[9] A. Bourrier, M. E. Davies, and T. Peleg, "Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems," *IEEE Transactions on Information Theory*, vol. 60, no. 12, pp. 7928–7946, 2014.

[10] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérèz, "Sketching for Large-Scale Learning of Mixture Models," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2016.

[11] N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval, "Compressive K-means," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2017.

[12] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérèz, "Sketching for Large-Scale Learning of Mixture Models," *arXiv preprint arXiv:1606.02838*, pp. 1–50, 2016.

[13] N. Keriven, N. Tremblay, and R. Gribonval, "SketchMLbox : a Matlab toolbox for large-scale learning of mixture models," 2016. [Online]. Available: sketchml.gforge.inria.fr

[14] A. Vedaldi and B. Fulkerson, "VLFeat - An open and portable library of computer vision algorithms," Tech. Rep., 2010. [Online]. Available: www.vlfeat.org

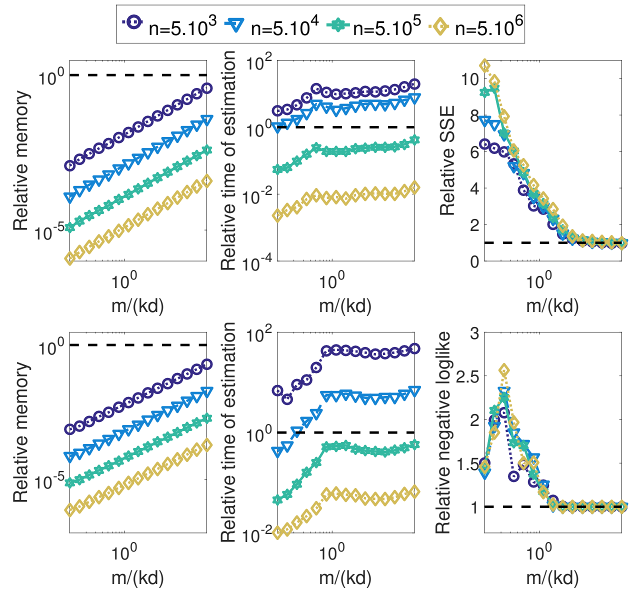

Fig. 1. Relative memory consumption (left), time of estimation (center) and precision (right) for compressive k-means (top) and GMM estimation (bottom) with $k = 10$ components in dimension $d = 10$, compared to Matlab's `kmeans` and VLFeat's `gmm` functions (dotted black lines).

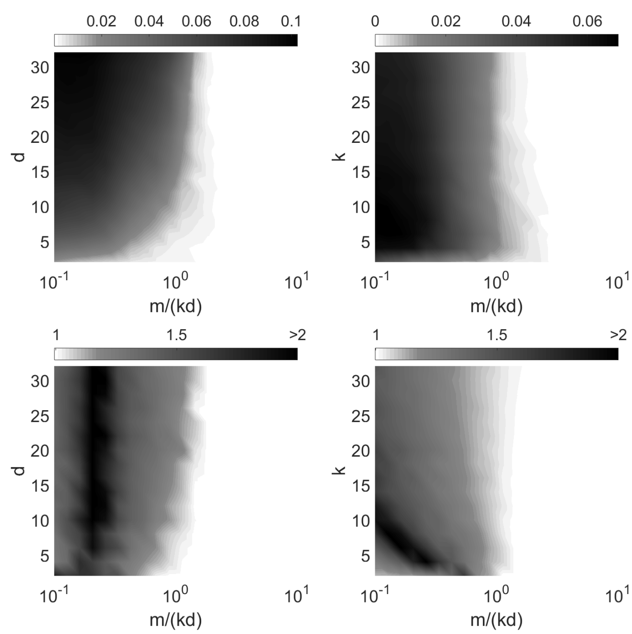

Fig. 2. Relative precision for k-means (top) and GMM estimation (bottom) with respect to the relative number of measurements $m/(kd)$. On the left $k = 10$ and on the right $d = 10$.