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

$$A\pi = \frac{1}{\sqrt{m}} \left[ \sum_{j=1}^{m} \exp(-i \omega_j^T z_i) / \sqrt{c_{\omega_j}} \right]^{m}_{i=1}$$

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 $\Lambda(\omega) \propto c_{\omega_j}^2 \exp(-\omega_j^2 \sigma^2)$. The empirical sketch used in practice is denoted $y = \frac{1}{n/\sqrt{m}} \left[ \sum_{i=1}^{n} \exp(-i \omega_j^T z_i) / \sqrt{c_{\omega_j}} \right]^{m}_{j=1}$.

II. MAIN RESULTS

We now present our main results on $k$-means and GMM estimation. In each case, $c_{\omega}$ and $\sigma^2$ 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 $h = \{c_1, \ldots, c_k\}$ and the loss function is $\ell(z, h) = \min_{\ell} \|z - c_{\ell}\|^2$. Assumptions. We restrict to a family of hypotheses where centroids are $2\varepsilon$-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

$$\hat{h} = \arg\min_{h \in \mathcal{H}_{k,\varepsilon,M}} \min_{\alpha_1, \ldots, \alpha_k} \|y - A \left( \sum_{i=1}^{k} \alpha_i \delta_{c_i} \right) \|^2_2$$

where $\alpha_i \geq 0$ and $\sum_{i=1}^{k} \alpha_i = 1$.

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

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

B. Gaussian mixture with identity covariance

In the GMM learning problem, a hypothesis is a set of means and weights $h = \{\mu_1, \ldots, \mu_k, \alpha_1, \ldots, \alpha_k\}$ and the loss function is $\ell(z, h) = -\log \pi_k(z)$, where $\pi_k = \sum_{i=1}^{k} \pi_i \mathcal{N}(\mu_i, I)$ 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

$$\hat{h} = \arg\min_{h \in \mathcal{H}_{k,M}} \|y - A\pi_h\|^2_2.$$

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

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

where the $O$ hides some dependencies in $k, d$ (roughly behaving like $m$ in Tab. I). The bound also involves the best approximation of $\pi$ 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 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 $m \approx O(kd)$. Fig. 2 further confirms that $m \approx O(kd)$ is empirically sufficient, hence the theoretical guarantees for $m \gtrsim O(k^2d^2)$ are probably pessimistic.

Further work will combine the sketching technique with dimensionality-reduction methods to treat both large $d$ and $n$. 
<table>
<thead>
<tr>
<th>Separation of means</th>
<th>Number of measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{O}(\sqrt{\log k})$</td>
<td>$m \geq \mathcal{O}(k^2d^2 \text{polylog}(k, d) \log(M/\rho))$</td>
</tr>
<tr>
<td>$\mathcal{O}(\sqrt{d \log k})$</td>
<td>$m \geq \mathcal{O}(k^3d^2 \text{polylog}(k, d) \log(M/\rho))$</td>
</tr>
<tr>
<td>$\mathcal{O}(\log k)$</td>
<td>$m \geq \mathcal{O}(k^2d^2e^d \text{polylog}(k, d) \log(M/\rho))$</td>
</tr>
</tbody>
</table>

**ACKNOWLEDGMENT**

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

**REFERENCES**


---

**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 $d = 10$. 

---

**TABLE I**

<table>
<thead>
<tr>
<th>Trade-off between required separation of means and number of measurements in the GMM learning problem.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separation of means</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>$\mathcal{O}(\sqrt{\log k})$</td>
</tr>
<tr>
<td>$\mathcal{O}(\sqrt{d \log k})$</td>
</tr>
<tr>
<td>$\mathcal{O}(\log k)$</td>
</tr>
</tbody>
</table>