Compressive k-Means with Differential Privacy

V. Schellekens*, A. Chatalic†, F. Houssiau‡, Y.-A. de Montjoye‡, L. Jacques* and R. Gribonval†
*ICTEAM/ELEN, UCLouvain †Univ Rennes, Inria, CNRS, IRISA ‡Imperial College London

***Abstract*—In the compressive learning framework, one harshly compresses a whole training dataset into a single vector of generalized random moments, the *sketch*, from which a learning task can subsequently be performed. We prove that this loss of information can be leveraged to design a differentially private mechanism, and study empirically the privacy-utility tradeoff for the k-means clustering problem.**

I. Introduction

The size and availability of datasets has increased in the last decades, and calls for machine learning methods able to process such collections efficiently, while protecting the privacy of data providers. In the compressive learning framework [1], the dataset is compressed into a vector of generalized random moments (the *sketch*), from which the desired model can be learned with reduced resources. We propose a mechanism based on this approach to learn from noisy dataset moments with provable privacy guarantees (differential privacy).

II. Differentially-Private Sketches

For a dataset $\mathcal{X} \triangleq \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$, the k-means problem consists in finding k cluster centroids $\mathcal{C} \triangleq \{\boldsymbol{c}_j \in \mathbb{R}^d\}_{j=1}^k$ minimizing the sum of squared errors (SSE):

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \text{SSE}(\mathcal{X}, \mathcal{C}) \triangleq \arg\min_{\mathcal{C}} \sum_{\boldsymbol{x}_i \in \mathcal{X}} \min_{\boldsymbol{c}_j \in \mathcal{C}} \|\boldsymbol{x}_i - \boldsymbol{c}_j\|_2^2. \quad (1)$$

In compressive learning [2], the dataset $\mathcal{X}$ is compressed into a sketch $\boldsymbol{z} \in \mathbb{C}^m$ of generalized moments as follows:

$$\boldsymbol{z} \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_{\boldsymbol{x}_i}, \text{ with } \boldsymbol{z}_{\boldsymbol{x}_i} \triangleq f(\Omega^T \boldsymbol{x}_i), \quad (2)$$

where $\Omega = [\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_m] \in \mathbb{R}^{d\times m}$ is a matrix of random frequency vectors drawn *i.i.d.* according to a well-chosen probability distribution [3], and $f : \mathbb{R} \mapsto \mathbb{C}$ is a (pointwise) nonlinear signature function, here assumed bounded and $2\pi-$periodic. This nonlinearity is typically the complex exponential $f_e : t \mapsto \exp(\mathrm{i}t)$ (i.e., $\boldsymbol{z}_{\boldsymbol{x}_i}$ are Random Fourier Features [4]), but we also consider one-bit universal quantization $f_q : t \mapsto \text{sign}(\cos(t)) + \mathrm{i}\,\text{sign}(\sin(t))$ [5], producing quantized sketch contributions [6]. One heuristic for k-means consists in finding k centroids whose sketch (computed using f_e) best matches the empirical sketch $\boldsymbol{z}$ [2].

Assuming the samples $\boldsymbol{x}_i$ represent sensitive information (e.g., medical records), our aim is to guarantee "privacy" for the providers of these data. Many definitions of privacy exist: for this work, we rely on the standard and widely used *differential privacy* [7]. Informally, it ensures that the output of a machine learning algorithm does not depend too much on the presence of one individual in the dataset.

Definition 1 (Differential Privacy)**.** *Let $\sim$ be the neighboring relation between datasets that differ by at most one record ($\mathcal{X} \sim \mathcal{X}' \Leftrightarrow (|\mathcal{X}| = |\mathcal{X}'|$ and $|(\mathcal{X} \cup \mathcal{X}') \setminus (\mathcal{X} \cap \mathcal{X}')| \leq 2)$). A randomized algorithm F is said to achieve differential privacy with parameter $\epsilon > 0$ (noted $\epsilon-$DP) if for any measurable set S of the co-domain of F:*

$$\forall \mathcal{X}, \mathcal{X}' \text{ s.t. } \mathcal{X} \sim \mathcal{X}' : \quad \mathbb{P}\left[F(\mathcal{X}) \in S\right] \leq e^{\epsilon}\mathbb{P}\left[F(\mathcal{X}') \in S\right].$$

In order to satisfy this definition, we produce a *scrambled sketch* by adding Laplace noise to the usual sketching process (2); the individual sketches $\boldsymbol{z}_{\boldsymbol{x}_i}$ are also subsampled to reduce the computational cost, i.e. only some of the m entries are computed for each sample $\boldsymbol{x}_i$ (cf. Figure 2). Formally, the scrambled sketch $\boldsymbol{s}_{\mathcal{X}}$ using $r \in [\![1, m]\!]$ (i.e., $r \in \mathbb{N} : 1 \leq i \leq m$) measurements per record, is defined as

$$\boldsymbol{s}_{\mathcal{X}} \triangleq \frac{1}{\alpha_r n}\sum_{i=1}^n (\boldsymbol{z}_{\boldsymbol{x}_i} \odot \boldsymbol{b}_{\boldsymbol{x}_i}) + \frac{1}{\sqrt{\alpha_r m n}}\boldsymbol{\xi}, \quad (3)$$

where $(\boldsymbol{b}_{\boldsymbol{x}_i})_{1\leq i\leq n}$ are uniformly drawn random binary masks with r nonzero entries, $\xi_j \overset{\text{iid}}{\sim} \mathcal{L}(\sigma_\xi/2) + \mathrm{i}\mathcal{L}(\sigma_\xi/2)$ for all $j \in [\![1, m]\!]$, $\alpha_r \triangleq r/m$ is called the *subsampling parameter*, and $\odot$ is the pointwise multiplication. We can now state our main result.

Theorem 1. *The local sketching mechanism (3)—with r measurements per input sample and noise standard deviation $\sigma_\xi = \frac{2c_f\sqrt{rm}}{\sqrt{n\epsilon}}$, where $c_f = 2\max_t(|\Re f(t)| + |\Im f(t)|)$ depends on the non-linearity f (e.g., $c_{f_e} = 2\sqrt{2}$, $c_{f_q} = 4$)—achieves $\epsilon-$DP.*

Proof: The proof is a direct generalization of [8, Prop. 1], with a general nonlinearity $f(\cdot)$ instead of the complex exponential $\exp(\mathrm{i}\cdot)$. ∎

This means that the scrambled sketch $\boldsymbol{s}_{\mathcal{X}}$ can be publicly released while protecting the ϵ-differential privacy of the records in $\mathcal{X}$. In a distributed context, all data provider compute scrambled sketches that can later be further averaged, as depicted in the attack model described in Figure4. The result will still be private by composition properties of differential privacy [9, Section 2.4.2].

III. Experiments and perspectives

When the privacy guarantee strengthens (i.e. $\epsilon \to 0$), the learning performance is expected to degrade, as $\boldsymbol{s}_{\mathcal{X}}$ becomes a poorer estimate for $\boldsymbol{z}$. Figure1 shows this privacy-utility tradeoff, for the k-means scenario, using two different sketch sizes $m = 10kd$ and $m = 100kd$ (experimental protocol described in the caption). Results are also provided for the standard DPLloyd [10], and its variant with improved initialization [11]. For a given sketch size, quantization degrades only slightly the results, and subsampling with $r = 1$ measurements (instead of m) has no significant impact for considered parameters ($n = 10^7$).

Note that for a fixed sketch size m, and as shown in Figure3, one useful quantity to explain the clustering error (utility) is the signal-to-noise ratio, whose definition (SNR) and expression for the non-quantized case (SNR_e) are

$$\text{SNR} \triangleq \frac{\|\boldsymbol{z}\|^2}{\sum_{j=1}^m \text{Var}((\boldsymbol{s}_{\mathcal{X}})_j)} \text{ and } \text{SNR}_e = \frac{\alpha_r n\|\boldsymbol{z}\|^2}{1 - \alpha_r\|\boldsymbol{z}\|^2 + \sigma_\xi^2}, \quad (4)$$

where $\alpha_r \triangleq r/m$ and $\boldsymbol{z}$ denotes the true sketch, i.e. the expectation w.r.t. the true data distribution.

It is not yet known if this privacy-preserving sketching mechanism is optimal (e.g., leading to the best privacy-utility tradeoff), or what kind of mechanism (if it exists) might be so. We also leave for future work the extension of private compressive learning for other tasks than k-means.

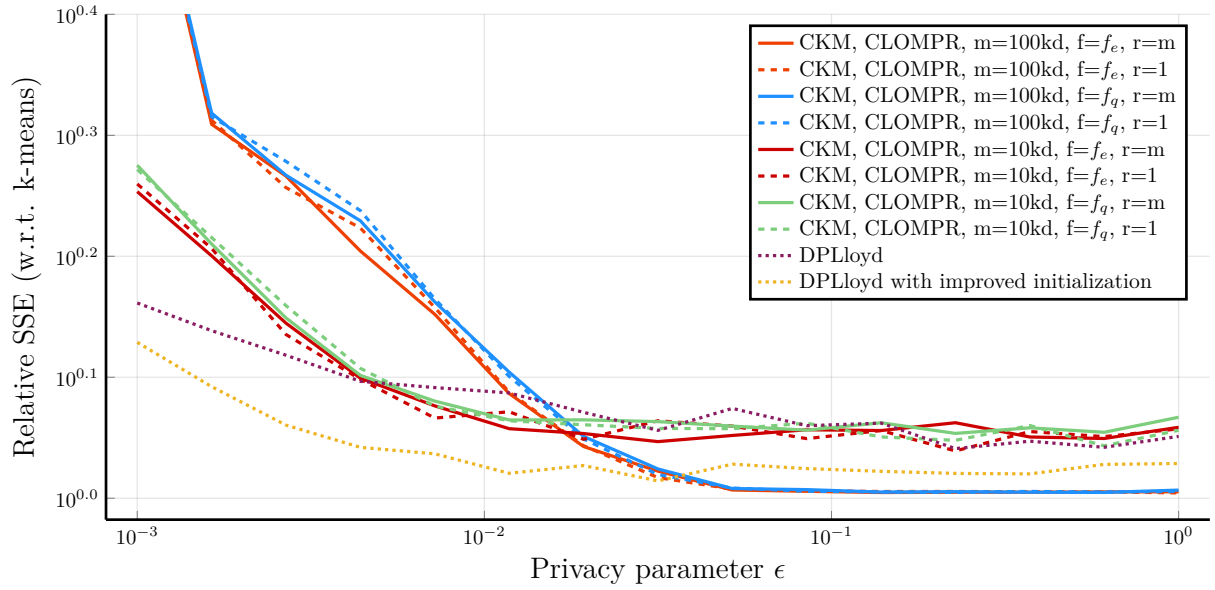

Fig. 1. Privacy-utility tradeoff (best viewed in colors). CKM stands for Compressive k-means (our approach). Improved initialization for DPLloyd refers to the approach proposed in [11]. Parameters: $k = d = 10$, $n = 10^7$. Data drawn according to Gaussian mixtures with k gaussians of covariances $\boldsymbol{I}_d$, and centers $(\boldsymbol{\mu}_i)_{1\leq i\leq k} \sim \mathcal{N}(0, (2.5k^{1/d})^2\boldsymbol{I}_d)$. Medians over 50 trials, all methods have similar variances (hidden for readability) except for smaller values ofϵ.

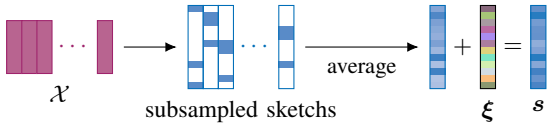

Fig. 2. Sketching with subsampling and additive noise on the sketch.

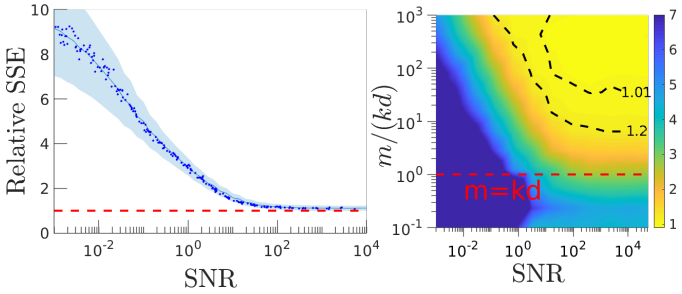

Fig. 3. (Left) Correlation between relative (w.r.t. k-means with 3 replicates) SSE and SNR for different values of α_r and σ_ξ, using $m = 10kd$, $f = f_e$. Medians of 40 trials, blue area shows the standard deviation. (Right) SSE as a function of SNR and m/kd, using $n = 10^5$, interpolated from a 12x12 grid, means of 40 trials, $k = d = 10$.

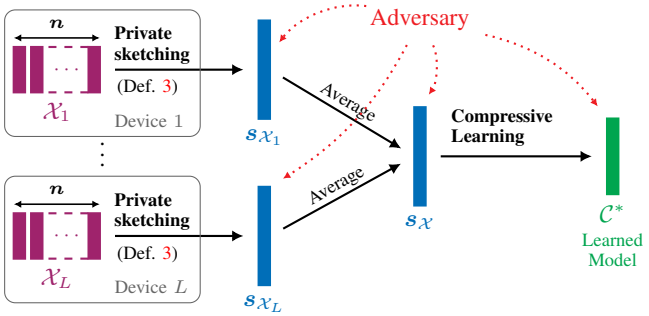

Fig. 4. Our attack model: L devices should protect the privacy of their local datasets $(\mathcal{X}_l)_{1\leq l\leq L}$ while allowing an algorithm to learn a model from it (in our case, the centroids $\mathcal{C}^*$). The (public) matrix of frequencies Ω is used for both "private sketching" and "compressive learning". All devices publish their scrambled sketches $\boldsymbol{s}_{\mathcal{X}_l}$, which are combined into the global sketch $\boldsymbol{s}_{\mathcal{X}}$.

References

[1] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin, "Compressive statistical learning with random feature moments," *arXiv preprint arXiv:1706.07180*, 2017.

[2] N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval, "Compressive K-means," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 6369–6373. [Online]. Available: https://hal.inria.fr/hal-01386077/document

[3] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez, "Sketching for large-scale learning of mixture models," *Information and Inference: A Journal of the IMA*, vol. 7, no. 3, pp. 447–508, 2017.

[4] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.

[5] P. T. Boufounos, S. Rane, and H. Mansour, "Representation and coding of signal geometry," *Information and Inference: A Journal of the IMA*, vol. 6, no. 4, pp. 349–388, 2017.

[6] V. Schellekens and L. Jacques, "Quantized compressive k-means," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1211–1215. [Online]. Available: http://arxiv.org/abs/1804.10109

[7] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[8] V. Schellekens, A. Chatalic, F. Houssiau, Y.-A. de Montjoye, L. Jacques, and R. Gribonval, "Differentially Private Compressive K-means," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[9] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, pp. 19–30.

[10] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: the SuLQ framework," in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2005, pp. 128–138.

[11] D. Su, J. Cao, N. Li, E. Bertino, and H. Jin, "Differentially private k-means clustering," in *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*. ACM, pp. 26–37.