Compressive k-Means with Differential Privacy

V. Schellekens[1], A. Chatalic[2], F. Houssiau[3], Y.-A. de Montjoye[3], L. Jacques[1] and R. Gribonval[2] ([1] UCLouvain / [2] Univ Rennes, Inria, CNRS, IRISA / [3] Imperial College London)

Context: sketched learning

A framework to learn from **compressed datasets** [2], made of two steps:

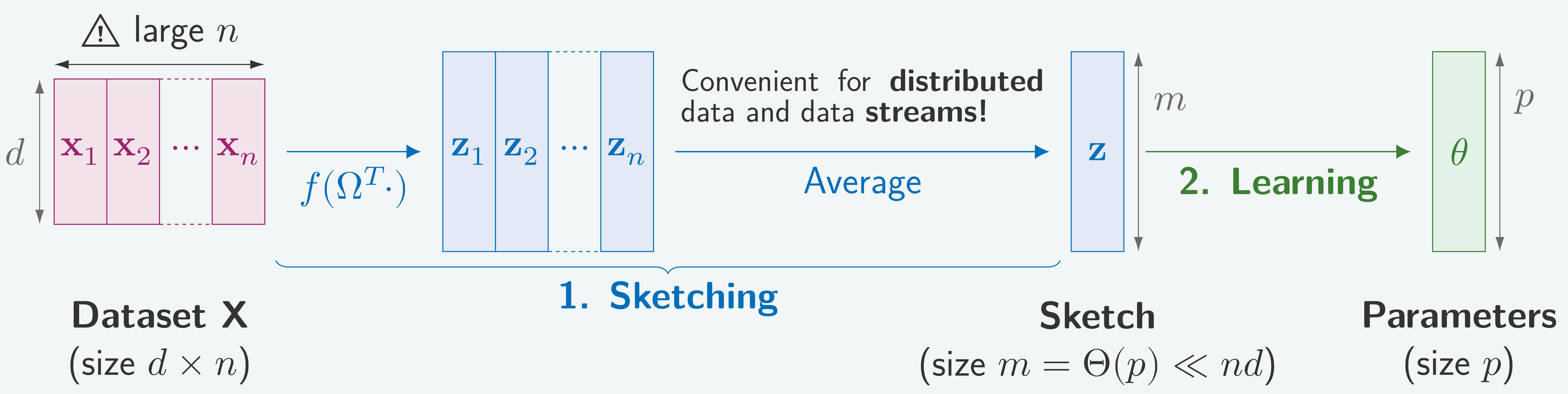
- Sketching = compressing the whole dataset into a single vector of generalized random moments:

$$\mathbf{z} \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \text{ with } \mathbf{z}_i \triangleq f(\Omega^T \mathbf{x}_i) \qquad (1)$$

(Ω: random matrix of m d-dimensional frequency vectors; f: pointwise nonlinearity (e.g. complex exponential or quantization))

- Learning = solving an inverse problem. For example for *compressive k-means*:

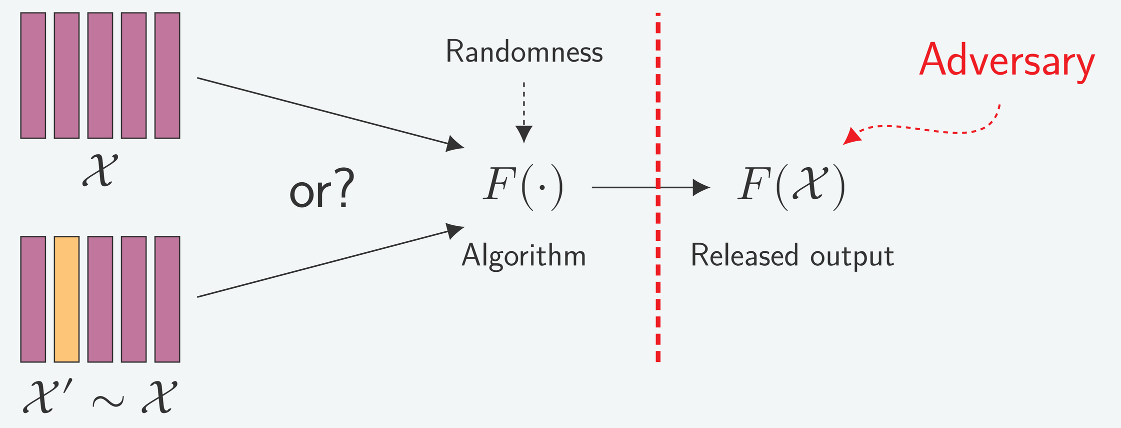
$$\mathcal{C}^* \in \arg\min_{\{\mathbf{c}_j\}_{j=1}^k} \|\mathbf{z} - \textstyle\sum_{j=1}^k \alpha_j f(\Omega^T \mathbf{c}_j)\|_2. \qquad (2)$$



Learning tasks that can (currently) be solved with sketching: k-means clustering, GMM fitting, PCA.

Privacy formalism: Differential Privacy (DP)

Question: how to protect sensitive datasets when learning?



An answer: Differential Privacy [1], random algorithm F has ϵ−DP if

$$\mathbb{P}[F(\mathcal{X}) \in S] \leq \exp(\epsilon) \cdot \mathbb{P}[F(\mathcal{X}') \in S], \quad \forall S, \forall \mathcal{X} \sim \mathcal{X}', \qquad (3)$$

where ϵ is the privacy parameter ($\epsilon \downarrow$ means privacy $\uparrow$).

- 👍 Widely studied and accepted
- 👍 Strong, robust guarantee
- 👍 Easy to implement
- 👎 Very conservative
- 👎 Selection of privacy parameter ϵ?

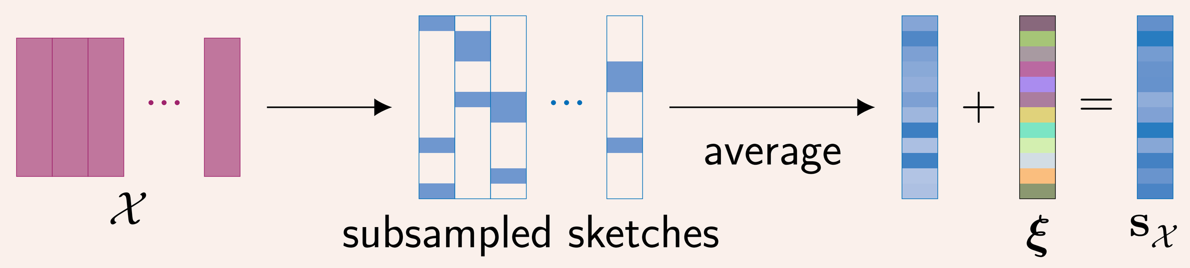
Contribution: private sketching for (e.g.) k-means clustering

Goal: leverage the information loss induced by sketching in formal privacy guarantees.

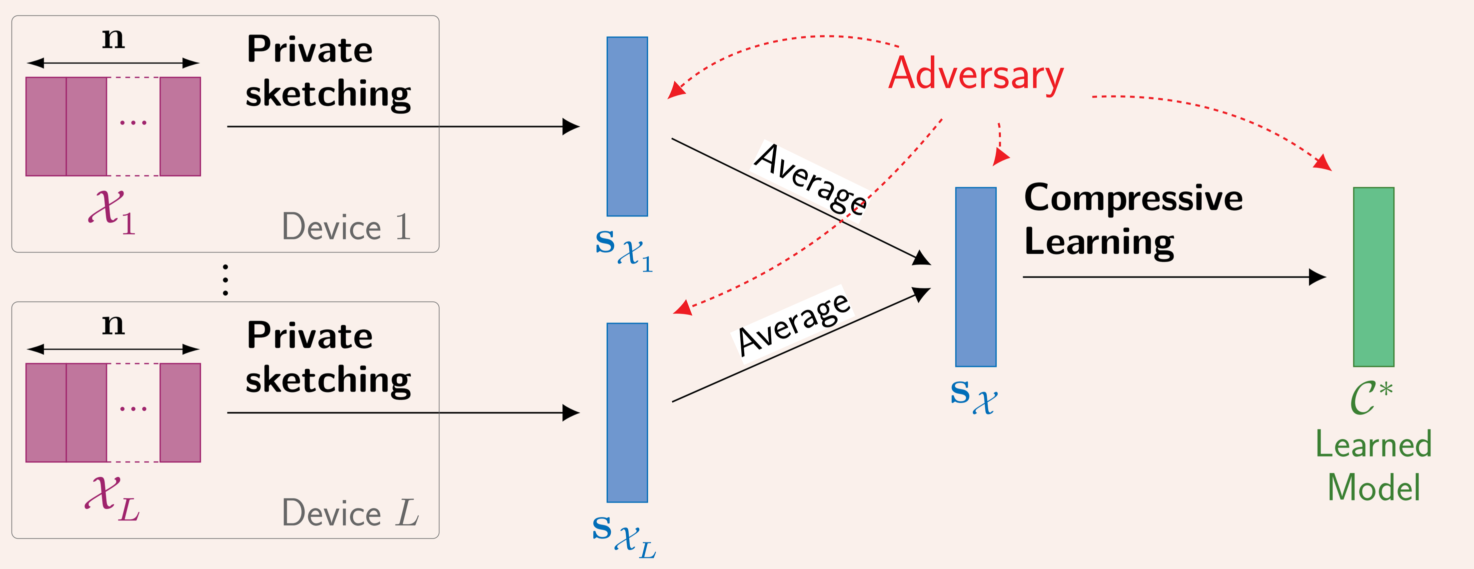
Idea: construct private sketch $\mathbf{s}_\mathcal{X}$ using two privacy-inducing elements:

- *subsample* individual sketches $\mathbf{z}_i$ with binary masks $\mathbf{b}_i$ (keeps r entries);
- *add Laplacian noise* $\xi \sim \mathcal{L}(\frac{\sigma_\xi}{\sqrt{2}})$ on top of the average:

$$\mathbf{s}_\mathcal{X} \triangleq \left(\frac{1}{\alpha_r n}\sum_{i=1}^{n} \mathbf{z}_i \odot \mathbf{b}_i\right) + \frac{1}{\sqrt{\alpha_r m n}}\xi. \qquad (4)$$



The noisy sketch $\mathbf{s}_\mathcal{X}$ can be released publicly without harming the privacy of users in $\mathcal{X}$.



Main result

The noisy sketching mechanism (4) with r measurements per input sample and noise standard deviation $\sigma_\xi = \frac{2 c_f \sqrt{r m}}{\sqrt{n}\,\epsilon}$ **achieves ϵ-differential privacy.** (c_f depends on the non-linearity, e.g. $c_f = 2\sqrt{2}$ for the complex exponential.)

Experimental results: solving Compressive k-Means (CKM) on the private sketch $\mathbf{s}_\mathcal{X}$

Problem: k-means clustering

Input: $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$ a set of n d-dimensional points.
Output: k centroids $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_k\} \subset \mathbb{R}^d$ minimizing the sum of squared errors:

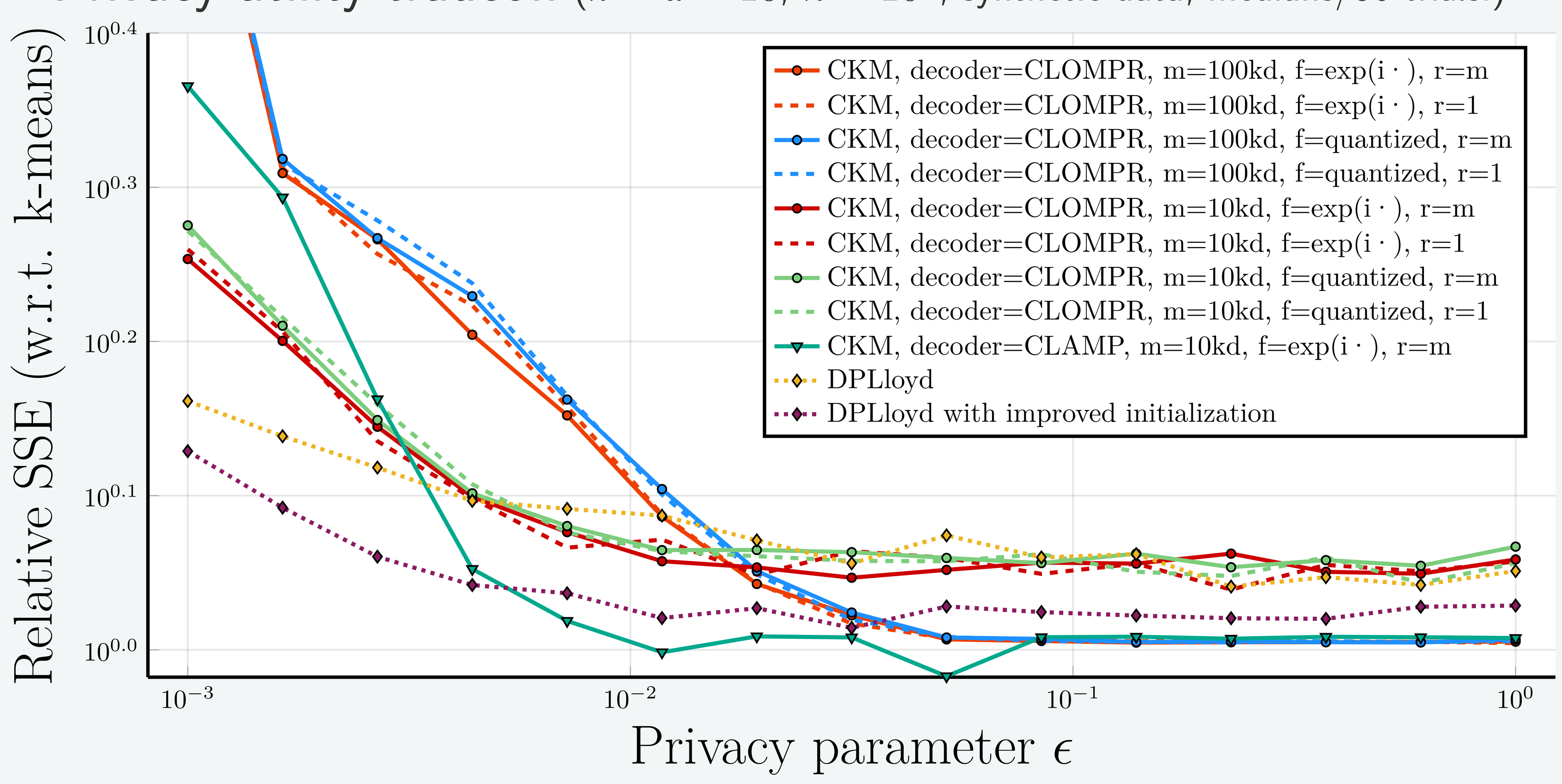
$$\text{SSE}(\mathcal{X}, \mathcal{C}) = \textstyle\sum_{i=1}^n \min_j \|\mathbf{x}_i - \mathbf{c}_j\|^2.$$

Note: We are learning $p = kd$ parameters; In practice we need $m \approx kd$ to get good clustering results with compressive k-means [3].

Conclusions:

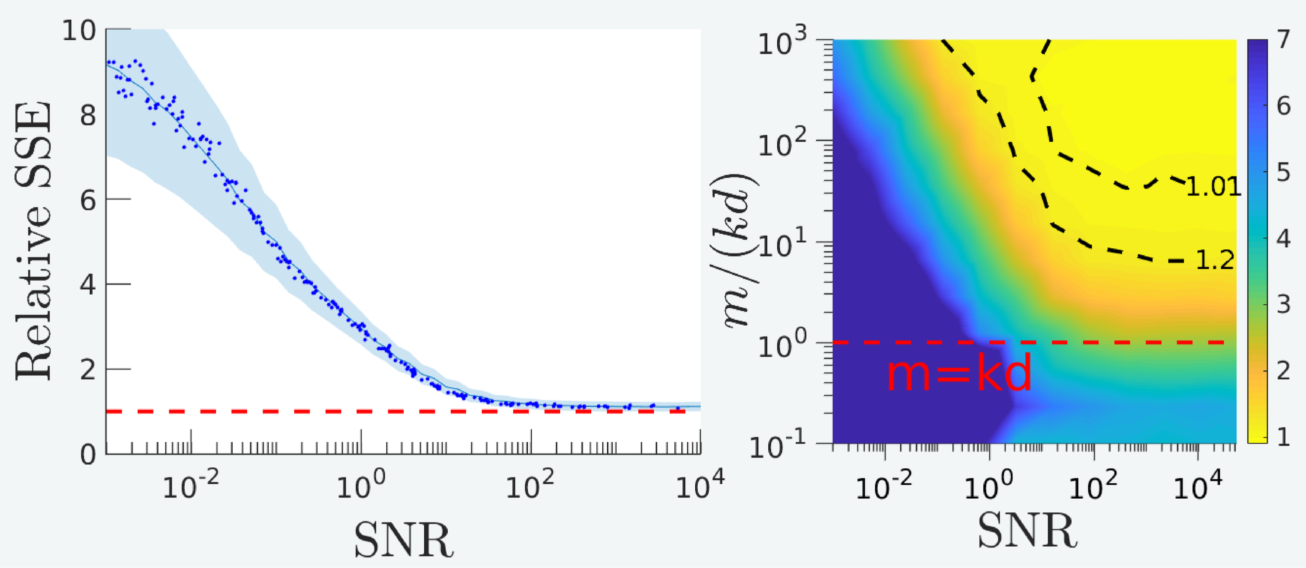
- A **generic** differentially private method, yielding privacy-utility tradeoffs similar to problem-specific techniques.
- Quantization does not degrade the results much.
- Subsampling reduces the time complexity without changing the tradeoff.

Privacy-utility tradeoff ($k = d = 10$, $n = 10^7$, synthetic data, medians/50 trials.)



Quantifying utility with the signal-to-noise ratio

$$\text{SNR} \triangleq \frac{\|\mathbf{z}\|^2}{\sum_{j=1}^m \text{Var}((\mathbf{s}_\mathcal{X})_j)} = \frac{\alpha_r n \|\mathbf{z}\|^2}{\underbrace{1 - \alpha_r \|\mathbf{z}\|^2}_{\text{subsampling}} + \underbrace{32\frac{\alpha_r m^2}{n\varepsilon^2}}_{\text{noise}}} \text{ where } \alpha_r \triangleq \frac{r}{m}.$$



Perspectives

- The bounds are actually tight (a bit trickier to show).
- Guarantees for PCA as well.
- Extension to other learning tasks.

References

See [4] for full paper and proof.

[1] Cynthia Dwork. "Differential privacy: A survey of results". 2008.
[2] Rémi Gribonval et al. "Compressive statistical learning with random feature moments". 2017.
[3] Nicolas Keriven et al. "Compressive K-means". Mar. 5, 2017.
[4] Vincent Schellekens et al. "Differentially Private Compressive K-Means". May 2019.