Sketched Clustering via Hybrid GAMP

Evan Byrne and Philip Schniter
Dept. ECE, The Ohio State University, Columbus, OH, USA
Email: {byrne.133,schniter.1}@osu.edu

Antoine Chatalic and Rémi Gribonval
Univ. Rennes, Inria, CNRS, IRISA, France
Email: {antoine.chatalic,remi.gribonval}@irisa.fr

Given a dataset $\boldsymbol{X} \triangleq [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] \in \mathbb{R}^{N\times T}$ comprising T samples of dimension N, the standard clustering problem is to find K centroids $\boldsymbol{C} \triangleq [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K] \in \mathbb{R}^{N\times K}$ that minimize the sum of squared errors (SSE)

$$\mathrm{SSE}(\boldsymbol{X}, \boldsymbol{C}) \triangleq \frac{1}{T}\sum_{t=1}^{T} \min_k \|\boldsymbol{x}_t - \boldsymbol{c}_k\|_2^2. \tag{1}$$

Finding the optimal $\boldsymbol{C}$ is NP-hard. Thus, many heuristics have been proposed, like *k-means++* [1]. The computational complexity of k-means++ scales as $O(TKNI)$, with I the number of iterations, which is impractical for large T.

In *sketched clustering* [2]–[4], the dataset $\boldsymbol{X}$ is first sketched down to a vector $\boldsymbol{y}$ with $M = O(KN)$ components, from which the centroids $\boldsymbol{C}$ are subsequently extracted. In the typical case that $K \ll T$, the sketch consumes much less memory than the original dataset. Also, if the sketch can be performed efficiently, then—since the complexity of centroid-extraction is invariant to T—sketched clustering may be more efficient than direct clustering methods when T is large.

In this work, we focus on sketches of the type proposed by Keriven et al. in [2,3], which use $\boldsymbol{y} = [y_1, \ldots, y_M]^{\mathsf{T}}$ with

$$y_m = \frac{1}{T}\sum_{t=1}^{T} \exp(\mathrm{j}\boldsymbol{w}_m^{\mathsf{T}}\boldsymbol{x}_t) \tag{2}$$

and randomly generated $\boldsymbol{W} \triangleq [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M]^{\mathsf{T}} \in \mathbb{R}^{M\times N}$. Note that y_m in (2) can be interpreted as a sample of the empirical characteristic function, i.e.,

$$\phi(\boldsymbol{w}_m) = \int_{\mathbb{R}^N} p(\boldsymbol{x}) \exp(\mathrm{j}\boldsymbol{w}_m^{\mathsf{T}}\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{3}$$

under the empirical distribution $p(\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T}\delta(\boldsymbol{x} - \boldsymbol{x}_t)$, with Dirac $\delta(\cdot)$. Here, each $\boldsymbol{w}_m$ can be interpreted as a multidimensional frequency sample. The process of sketching $\boldsymbol{X}$ down to $\boldsymbol{y}$ via (2) costs $O(TMN)$ operations, but it can be performed efficiently in an online and/or distributed manner.

To recover the centroids $\boldsymbol{C}$ from $\boldsymbol{y}$, the state-of-the-art algorithm is *compressed learning via orthogonal matching pursuit with replacement* (CL-OMPR) [2,3]. It aims to solve

$$\arg\min_{\boldsymbol{C}} \min_{\boldsymbol{\alpha}:\mathbf{1}^{\mathsf{T}}\boldsymbol{\alpha}=1} \sum_{m=1}^{M} \left| y_m - \sum_{k=1}^{K} \alpha_k \exp(\mathrm{j}\boldsymbol{w}_m^{\mathsf{T}}\boldsymbol{c}_k) \right|^2 \tag{4}$$

using a greedy heuristic inspired by the OMP algorithm popular in compressed sensing. With sketch length $M \geq 10KN$, CL-OMPR typically recovers centroids of similar or better quality to those attained with k-means++. One may wonder, however, whether it is possible to recover accurate centroids with sketch lengths closer to the counting bound $M = KN$. Also, since CL-OMPR's computational complexity is $O(MNK^2)$, one may wonder whether it is possible to recover accurate centroids with computational complexity $O(MNK)$.

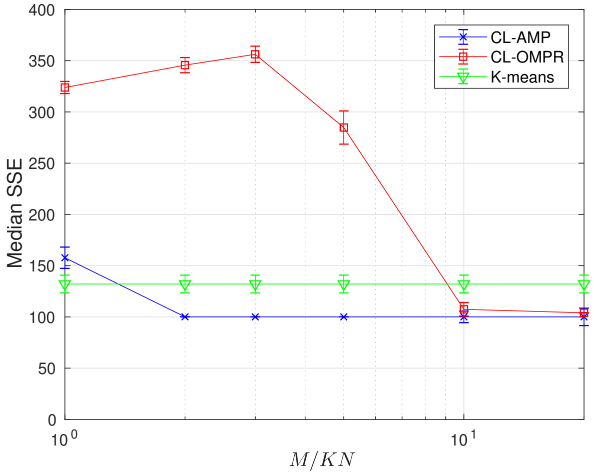
In answer to these questions, we propose the *compressive learning via approximate message passing* (CL-AMP) algorithm [5], which has computational complexity $O(MNK)$. Numerical experiments show that CL-AMP accurately recovers centroids from sketches of length $M = 2KN$, an improvement over CL-OMPR. Also, experiments show that CL-AMP recovers centroids faster and more accurately than k-means++ for large T.

CL-AMP treats centroid recovery as a high-dimensional inference problem, based on the Gaussian mixture model

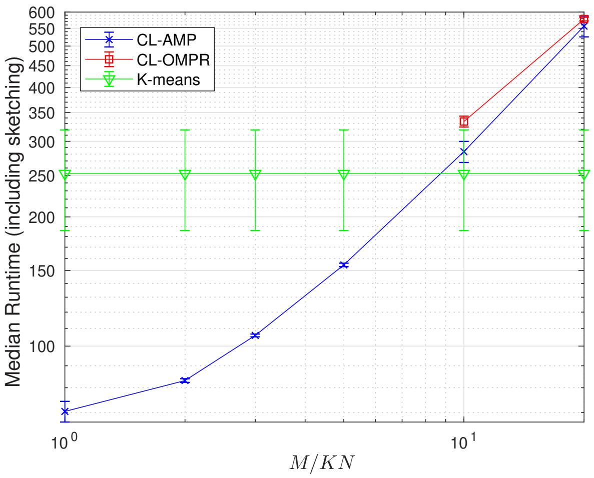
$$\boldsymbol{x}_t \sim \sum_{k=1}^{K} \alpha_k \mathcal{N}(\boldsymbol{c}_k, \boldsymbol{\Phi}_k), \tag{5}$$

where α_k and covariances $\boldsymbol{\Phi}_k$ are treated as deterministic unknown parameters. In particular, CL-AMP computes an approximation to the MMSE estimate $\widehat{\boldsymbol{C}} = \mathbb{E}\{\boldsymbol{C} \,|\, \boldsymbol{y}\}$, where the expectation is taken over the posterior density $p(\boldsymbol{C}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{C})p(\boldsymbol{C})$. The form of the sketch in (2) implies that $p(\boldsymbol{y}|\boldsymbol{C}) = \prod_{m=1}^{M} p_{\mathsf{y}|\mathbf{z}}(y_m|\boldsymbol{w}_m^{\mathsf{T}}\boldsymbol{C})$, which can be recognized as a generalized linear model (GLM) on the random linear transform outputs $\boldsymbol{w}_m^{\mathsf{T}}\boldsymbol{C}$. As such, sketched clustering is ripe for the application of the *simplified hybrid generalized AMP* (SHyGAMP) algorithm from [6], which is a generalization of the GAMP algorithm [7]. As described in [5], the likelihood depends on $\boldsymbol{\Phi}_k$ through $\boldsymbol{w}_m^{\mathsf{T}}\boldsymbol{\Phi}_k\boldsymbol{w}_m$, which concentrates to an m-invariant value "τ_k" in the high dimensional limit. The EM-GAMP algorithm can then be used to estimate $\{\alpha_k, \tau_k\}$.

The full details of CL-AMP are given in [5]. Here we show just a few numerical results with synthetic clusters $\boldsymbol{c}_k$. All results represent the median over 10 trials, and runtime is not shown whenever SSE is $> 1.5\times$ that of k-means++. Figures 1a and 1b show SSE (1) and runtime vs. sketch length M. We see that CL-AMP allows shorter sketch-length M than CL-OMPR, and yields better SSE and runtime than k-means++ when $M \in [2, 5]$. Figures 2a, 2b, and 2c show SSE, runtime with sketching, and runtime without sketching, respectively, vs. sample size T. We see that CL-AMP yields better SSE than CL-OMPR and k-means++ for all tested T, and that CL-AMP runs faster than CL-OMPR and k-means++ for large T.



(a) SSE vs. M



(b) Runtime (including sketching) vs. M

Fig. 1: Performance vs. sketch length M for $K = 10$ clusters, dimension $N = 100$, and $T = 10^7$ training samples.

References

[1] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. Symp. Discrete Alg. (SODA)*, 2007, pp. 1027–1035.

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

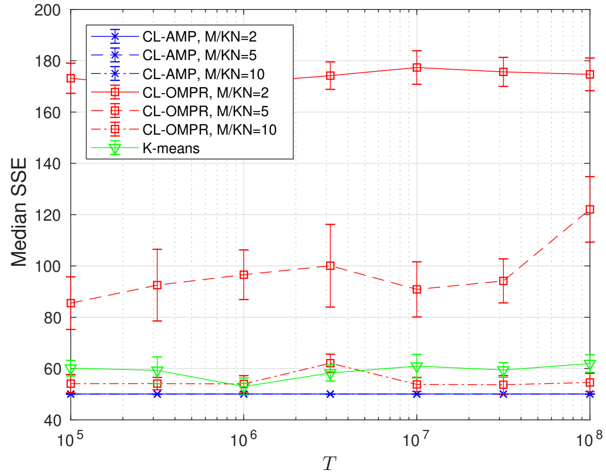
[3] N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval, "Compressive K-means," in *Proc. IEEE Int. Conf. Acoust. Speech & Signal Process.*, 2017, pp. 6369–6373.

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

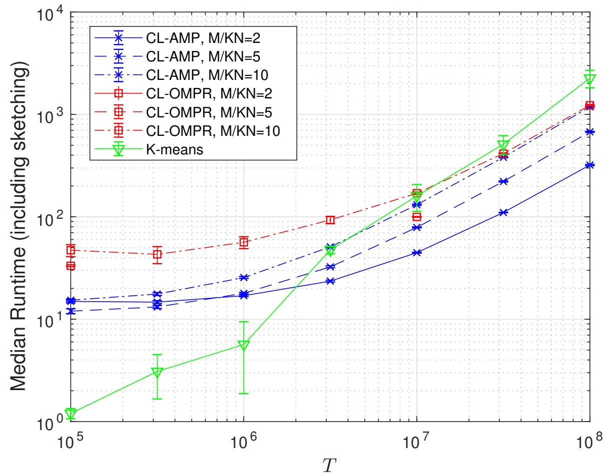
[5] E. Byrne, A. Chatalic, R. Gribonval, and P. Schniter, "Sketched clustering via hybrid approximate message passing," *arXiv:1712.02849v2*, 2019.

[6] E. M. Byrne and P. Schniter, "Sparse multinomial logistic regression via approximate message passing," *IEEE Trans. Signal Process.*, vol. 64, no. 21, pp. 5485–5498, 2016.

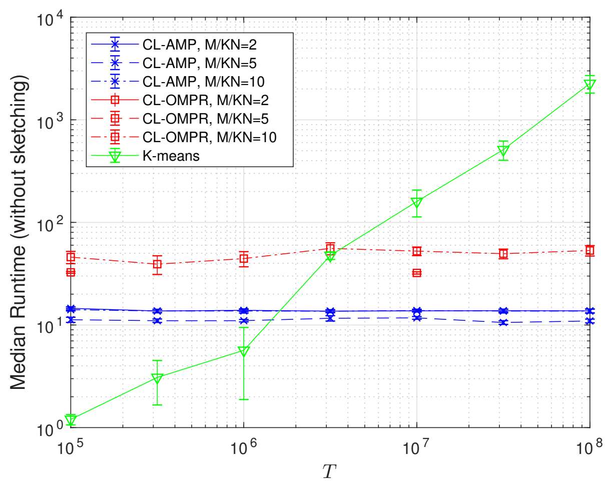
[7] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inform. Thy.*, Aug. 2011, pp. 2168–2172, (full version at *arXiv:1010.5141*).



(a) SSE vs. T



(b) Runtime (including sketching) vs. T



(c) Runtime (without sketching) vs. T

Fig. 2: Performance vs. training size T for $K = 10$ classes, dimension $N = 50$, and sketch size $M \in \{2, 5, 10\} \times KN$.