Gradient Based Clustering

Aleksandar Armacki[†], Dragana Bajovic[‡], Dusan Jakovetic[‡], Soummya Kar[†]

Carnegie Mellon University[†], University of Novi Sad[‡]

Thirty-ninth International Conference on Machine Learning,
Baltimore, Maryland, USA,
17. - 23. July, 2022

Work of DB and DJ funded by H2020, GA 957337 and 871518. Work of AA and SK funded by NSF, GA CNS-1837607.

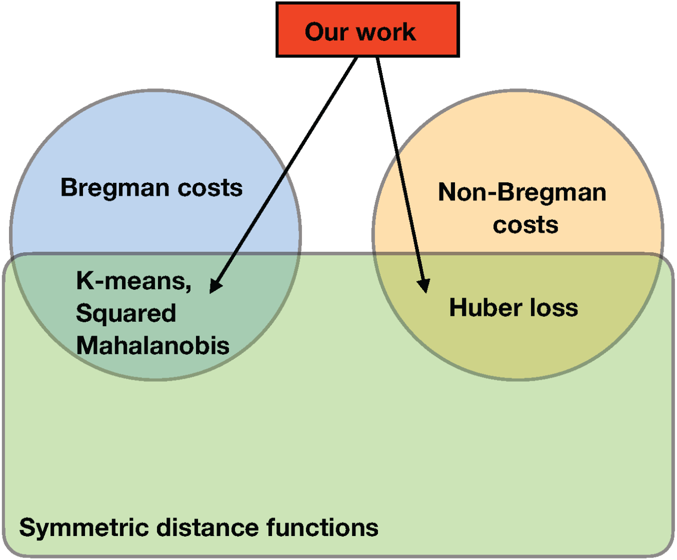
Introduction

- Clustering is a well-studied problem, e.g., Lloyd (1982), Banerjee et al. (2005), Pediredla and Seelamantula (2011)
- Prior works use specific cost functions and design tailored solvers
  - Banerjee et al. (2005) design an approach specific for Bregman costs
  - Pediredla and Seelamantula (2011) design an approach specific for Huber loss
- In Armacki et al. (2022), we propose a generic gradient-based approach to clustering
- Our approach is applicable to a wide array of costs, e.g., a large class of symmetric Bregman costs as well as non-Bregman costs, like Huber loss

Electrical & Computer ENGINEERING

Contributions

- We propose a gradient-based update rule, applicable to a wide range of costs
- We provide general convergence guarantees, independent of the choice of cost or distance functions
- We decouple the distance and cost functions, allowing for development of novel clustering algorithms
- Compared to Banerjee et al. (2005), our approach extends beyond Bregman costs
- Compared to other non-Bregman methods, e.g., Pediredla and Seelamantula (2011), our approach provides strong convergence guarantees to appropriately defined fixed points


Our work
Bregman costs
Non-Bregman costs
K-means, Squared Mahalanobis
Huber loss
Symmetric distance functions

Problem Formulation

- Input:
  - $g : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}_+$ - symmetric distance function
    - Example: $g(x, y) = \|x - y\|$
    - Example: $g(x, y) = \sqrt{(x - y)A(x - y)}$, for any $A \succ 0$
  - $K \in \mathbb{N}$ - desired number of clusters
  - $\mathcal{D} \subset \mathbb{R}^d$ - (finite) dataset
  - $p_y \in (0, 1)$ - weight assigned to point $y \in \mathcal{D}$

Electrical & Computer ENGINEERING

Problem Formulation - Cont'd

- General clustering problem:

$$\min_{x \in \mathbb{R}^{Kd}, C \in \mathcal{C}} J(x, C) = \sum_{k \in [K]} \sum_{y \in C(k)} p_y f(x(k), y) \qquad \text{(GC)}$$

  - $f : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}_+$ - cost function, such that, for all $x, y, z \in \mathbb{R}^d, \quad g(x,y) \leq g(z,y) \implies f(x,y) \leq f(z,y)$
    - Example: $f(x,y) = g(x,y)^2$
    - Example: $f(x,y) = \text{Huber loss}\,(g(x,y))$
  - $x(k) \in \mathbb{R}^d$ - center estimate for the k-th cluster
  - $C(k) \in \mathcal{C}$ - k-th cluster
  - $\mathcal{C}$ - the space of all K-partitions of $\mathcal{D}$, i.e., for any $C \in \mathcal{C}$, we have

$$|C| \leq K, \quad C(k) \cap C(j) = \emptyset, \text{ for } k \neq j, \quad \cup_{k=1}^{|C|} C(k) = \mathcal{D}$$

Electrical & Computer ENGINEERING

Proposed Method

- We propose a two step iterative algorithm to solve (GC)
- The method performs the following steps, in each iteration $t = 0, 1, \dots$:
  1. *Cluster assignment*: for all $y \in \mathcal{D}$, find $k \in [K]$, such that
  $$g(x_t(k), y) \leq g(x_t(j), y),\ \forall j \neq k, \quad (1)$$
  and assign the point y to $C_{t+1}(k)$.
  2. *Center update*: for all $k \in [K]$, perform
  $$x_{t+1}(k) = x_t(k) - \alpha \sum_{y \in C_{t+1}(k)} \nabla_{x_t} f(x_t(k), y), \quad (2)$$
  where $\alpha > 0$ is a fixed step-size.

Electrical & Computer ENGINEERING

Main Results

Definition

A pair $(x_\star, C_\star)$ is a fixed point of (1)-(2) if

1. *Optimal clusters*: for all $k \in [K]$ and $y \in C_\star(k)$, we have $g(x_\star(k), y) \leq g(x_\star(j), y)$
2. *Optimal centers*: $\nabla_x J(x_\star, C_\star) = 0$

Theorem

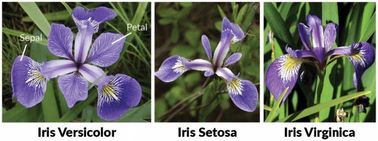
For the step-size choice $\alpha < \frac{2}{L}$ and any initialization $x_0 \in \mathbb{R}^{Kd}$, the sequence of points (x_t, C_t), generated by (1)-(2), converges to a fixed point.

Electrical & Computer ENGINEERING

Numerical Results - Data

- We evaluate the performance of the gradient based clustering methods on two real datasets, MNIST and Iris
- For MNIST, we chose $K = 7$ clusters, corresponding to the first seven digits, with $n = 500$ samples per digit
- For Iris, we use the whole dataset, i.e., $K = 3$ clusters, corresponding to different Iris flowers, with $n = 50$ samples per flower

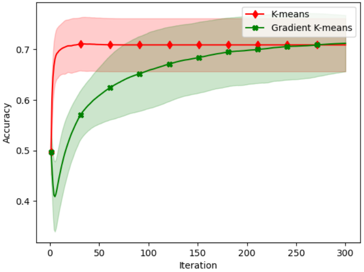
MNIST digits



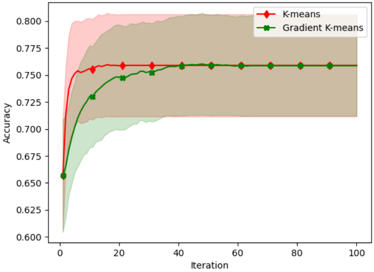
Iris flowers. Credit: gadictos.com

Numerical Results - Noiseless

- We use the standard K-means cost with Euclidean distance, i.e., $f(x,y) = \|x-y\|^2$
- Benchmark: Lloyd's algorithm Lloyd (1982), Banerjee et al. (2005)



K-means on MNIST data, averaged across 20 runs



K-means on Iris data, averaged across 20 runs

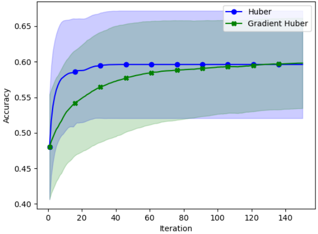
Electrical & Computer ENGINEERING

Numerical Results - Noisy

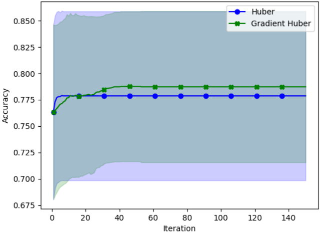
- We add zero mean Gaussian noise to $p = 20\%$ of data points, with variance $\sigma^2 = 2$
- We use the Huber loss cost with Euclidean distance, i.e.,

$$f(x,y) = \begin{cases} \frac{\|x-y\|^2}{2}, & \|x-y\| \leq \delta, \\ \delta\|x-y\| - \frac{\delta^2}{2}, & \|x-y\| > \delta \end{cases}$$

- Benchmark: Huber loss clustering from Pediredla and Seelamantula (2011)



Huber loss on MNIST data, averaged across 20 runs



Huber loss on Iris data, averaged across 20 runs

Conclusion

- We propose a general gradient-based method for clustering
- The method encompasses a wide range of functions, such as a class of Bregman divergences and Huber loss
- The method provably converges to a properly defined fixed point, with arbitrary initialization
- Numerical results on real data show the method is competitive, in comparison to existing methods

References

A. Armacki, D. Bajovic, D. Jakovetic, and S. Kar. Gradient based clustering. *arXiv preprint arXiv:2202.00720*, 2022.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6(58): 1705–1749, 2005. URL http://jmlr.org/papers/v6/banerjee05b.html.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

A. K. Pediredla and C. S. Seelamantula. A Huber-loss-driven clustering technique and its application to robust cell detection in confocal microscopy images. In *2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 501–506, 2011.