Lloyd's K-Means Clustering Algorithm Is Frank-Wolfe in Disguise

Michael Pokojovy*, J. Marcus Jobe** and Simon Lacoste-Julien†

*Department of Mathematics & Statistics and School of Data Science, Old Dominion University, Norfolk, Virginia

**Department of Information Systems and Analytics, Miami University, Oxford, Ohio

†Canada CIFAR AI Chair, Department of Computer Science and Operations Research & Mila, Université de Montréal, Montréal (QC) Canada

jobejm@miamioh.edu

Introduction

Lloyd's K-means algorithm, also known as naïve K-means, is a widely used *ad hoc* optimization heuristic, designed to minimize the sum of squared errors (SSE) across all K-partitions of a dataset via iterative cluster refinement. We establish a novel connection between Lloyd's algorithm and the Frank-Wolfe (FW) algorithm, a prominent first-order method for projection-free optimization. We demonstrate that Lloyd's algorithm is a special case of FW. Leveraging recent advances in FW methods for concave objectives, we derive a non-asymptotic $\mathcal{O}(1/t)$ convergence rate to a local minimum of the SSE objective. To account for empty clusters, an outcome possible under Lloyd's greedy assignment, we develop an FW variant for semismooth objectives while retaining the same convergence rate that is solely controlled by the initial SSE value. We illustrate our findings with a simulation study for spherical Gaussian mixtures and a real-world image segmentation data.

Contributions: Main contributions are summarized as follows:

- We show that the Lloyd's algorithm with greedy cluster assignment is a special case of the FW algorithm with unit step-size for a semismooth concave objective over a polyhedral set. With empty clusters excluded from the feasible set, the objective becomes smooth. This observation allows us to leverage recent convergence results for concave FW to directly establish a non-asymptotic $\mathcal{O}(1/t)$ convergence rate (Yurtsever and Sra 2022).
- To accommodate for the practically relevant possibility of empty clusters, we develop an FW algorithm for general (semismooth) concave objectives with a new FW gap based on Clarke subdifferential. The same non-asymptotic $\mathcal{O}(1/t)$ convergence is recovered that solely depends on the initial global suboptimality (uniformly in sample size n, space dimension d and number of clusters K) and does not involve coreset considerations.

Lloyd's K-means Clustering

Consider a dataset $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$. We seek to minimize the sum of squares errors (SSE), a.k.a. within-cluster sum of squares (WCSS),

$$\mathrm{SSE}(\mathcal{C}) \equiv \sum_{k:C_k\neq\emptyset} \sum_{\boldsymbol{x}\in C_k} \big\|\boldsymbol{x} - \bar{\boldsymbol{x}}_k\big\|^2 \tag{1}$$

with $\bar{\boldsymbol{x}}_k = \frac{1}{|C_k|} \sum_{\boldsymbol{x}\in C_k} \boldsymbol{x}$ over all K-partitions $\mathcal{C} = (C_1, C_2, \ldots, C_K)$ of $\mathcal{D}$.

Algorithm 1 (Lloyd's K-means algorithm).

Let the seeds $\boldsymbol{\mu}_1^{(-1)}, \ldots, \boldsymbol{\mu}_K^{(-1)} \in \mathbb{R}^d$ be given.
Initialize:
$C_k^{(0)} := \big\{\boldsymbol{x}_i \,\big|\, \|\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(-1)}\| \le \|\boldsymbol{x}_i - \boldsymbol{\mu}_{k'}^{(-1)}\| \text{ for } 1 \le k' \le K\big\}$,
$\boldsymbol{\mu}_k^{(0)} := \begin{cases} \frac{1}{|C_k^{(0)}|} \sum\limits_{\boldsymbol{x}_i\in C_k^{(0)}} \boldsymbol{x}_i, & \text{if } C_k^{(0)} \neq \emptyset, \\ \boldsymbol{\mu}_k^{(0)} := \boldsymbol{\mu}_k^{(-1)}, & \text{otherwise.} \end{cases}$
for $t = 0, \ldots, T-1$ **do**
  for $k = 1, \ldots, K$ **do**
    Let $C_k^{(t+1)} := \big\{\boldsymbol{x}_i \,\big|\, \|\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t)}\| \le \|\boldsymbol{x}_i - \boldsymbol{\mu}_{k'}^{(t)}\| \text{ for } 1 \le k' \le K\big\}$.
    Compute
$$\boldsymbol{\mu}_k^{(t+1)} := \begin{cases} \frac{1}{|C_k^{(t+1)}|} \sum_{\boldsymbol{x}_i\in C_k^{(t+1)}} \boldsymbol{x}_i, & \text{if } C_k^{(t+1)} \neq \emptyset, \\ \boldsymbol{\mu}_k^{(t)}, & \text{otherwise.} \end{cases}$$
    Let $\Delta\mathrm{SSE}_t := \mathrm{SSE}(\mathcal{C}^{(t)}) - \mathrm{SSE}(\mathcal{C}^{(t+1)})$.
    if $\Delta\mathrm{SSE}_t \le \epsilon$ **then**
      return $C_1^{(t)}, \ldots, C_K^{(t)}$ (and $\boldsymbol{\mu}_1^{(t)}, \ldots, \boldsymbol{\mu}_K^{(t)}$)
    end if
  end for
end for
return $C_1^{(T)}, \ldots, C_K^{(T)}$ (and $\boldsymbol{\mu}_1^{(T)}, \ldots, \boldsymbol{\mu}_K^{(T)}$).

One-Hot Encoding and Extension

The K-means SSE objective reads as

$$f(\boldsymbol{w}) = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} \|\boldsymbol{x}_i - \boldsymbol{x}_k^{\boldsymbol{w}}\|^2 \tag{2}$$

$$\text{with } \boldsymbol{x}_k^{\boldsymbol{w}} := \frac{1}{n_k^{\boldsymbol{w}}} \sum_{i=1}^{n} w_{ik} \boldsymbol{x}_i \quad \text{and} \quad n_k^{\boldsymbol{w}} := \sum_{i=1}^{n} w_{ik}$$

using the usual 'one-hot' encoding $\boldsymbol{w} \in \{0,1\}^{n\times K}$ such that $w_{ik} = 1$ if the i-th data point $\boldsymbol{x}_i \in \mathbb{R}^d$ belongs to the k-th cluster.

The SSE objective $f(w)$ can be continuously extended to $\mathcal{M} = \mathrm{conv}(\mathcal{M}_0)$ with

$$\mathcal{M}_0 = \Big\{\boldsymbol{w} \in \{0,1\}^{n\times K} \,\Big|\, \sum_{k=1}^{K} w_{ik} = 1, i = 1, \ldots, n\Big\} \tag{3}$$

by letting $\sum_{i=1}^{n} w_{ik} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k^{\boldsymbol{w}}\|^2 := 0$ for $n_k^{\boldsymbol{w}} = 0$.

Frank-Wolfe Algorithm for Semismooth Concave Objectives

Despite being continuous, the Lloyd's SSE objective fails to be smooth in $\mathcal{M}$. Therefore, we extend the recent results of Yurtsever and Sra 2022 to semismooth concave objectives.

Let $\mathcal{M} \subset \mathbb{R}^p$ be a non-empty convex compact set. We consider a general constrained optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{M}} f(\boldsymbol{x}), \tag{4}$$

where f is still concave, but only semismooth. Let

$$\mathrm{LMO}_{\mathcal{M}}(\boldsymbol{r}) := \arg\min_{\boldsymbol{s}\in\mathcal{M}} \langle \boldsymbol{s}, \boldsymbol{r}\rangle \tag{5}$$

denote the linear minimization oracle (LMO) at $\boldsymbol{r} \in \mathbb{R}^p$.

Suppose the function f possesses a (set-valued) Clarke subdifferential $\partial f(\boldsymbol{x})$ at any $\boldsymbol{x} \in \mathcal{M}$. Further, the set $\partial f(\boldsymbol{x}) \neq \emptyset$ is compact and convex for any $\boldsymbol{x} \in \mathcal{M}$. Algorithm 2 below protocols our semismooth FW with unit step-size. Note that for smooth objectives, it automatically reduces to the smooth version since $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$.

Algorithm 2 (Semismooth FW for concave functions).

Let $\boldsymbol{x}^{(0)} \in \mathcal{M}$.
for $t = 0, \ldots, T-1$ **do**
  Choose arbitrary $\boldsymbol{r}^{(t)} \in \partial f(\boldsymbol{x}^{(t)})$.
  Compute $\boldsymbol{s}^{(t)} := \mathrm{LMO}_{\mathcal{M}}\big(\boldsymbol{r}^{(t)}\big)$.
  Define the FW update direction $\boldsymbol{d}_t := \boldsymbol{s}^{(t)} - \boldsymbol{x}^{(t)}$.
  Compute the FW gap $g_t := \big\langle \boldsymbol{d}_t, -\boldsymbol{r}^{(t)}\big\rangle$.
  if $g_t \le \epsilon$ **then**
    return $\boldsymbol{x}^{(t)}$
  end if
  Select step-size $\gamma_t = 1$.
  Update $\boldsymbol{x}^{(t+1)} := \boldsymbol{x}^{(t)} + \gamma_t \boldsymbol{d}_t \equiv \boldsymbol{s}^{(t)}$.
end for
return $\boldsymbol{x}^{(T)}$.

Before discussing non-asymptotic convergence of Algorithm 2, we introduce a suitable generalization of the FW gap to be used as a stationarity measure.

Cnt'd

Recall that $\boldsymbol{x} \in \mathcal{M}$ is referred to as Clarke-stationary if

$$0 = \max_{\boldsymbol{y}\in\mathcal{M}} \langle \boldsymbol{y} - \boldsymbol{x}, -\boldsymbol{r}\rangle = 0 \tag{6}$$

for at least one $\boldsymbol{r} \in \partial f(\boldsymbol{x})$. Similar to Liu et al. 2024, we introduce the Clarke version of FW gap

$$g_{\mathrm{C}}(\boldsymbol{x}) := \min_{\boldsymbol{r}\in\partial f(\boldsymbol{x})} \max_{\boldsymbol{y}\in\mathcal{M}} \langle \boldsymbol{y} - \boldsymbol{x}, -\boldsymbol{r}\rangle. \tag{7}$$

By compactness, the latter expression is well-defined and finite. Further, for any $\boldsymbol{r} \in \partial f(\boldsymbol{x})$,

$$\max_{\boldsymbol{y}\in\mathcal{M}} \langle \boldsymbol{y} - \boldsymbol{x}, -\boldsymbol{r}\rangle \ge \langle \boldsymbol{y} - \boldsymbol{x}, -\boldsymbol{r}\rangle|_{\boldsymbol{y}=\boldsymbol{x}} = 0.$$

Thus, we trivially have $g_{\mathrm{C}}(\boldsymbol{x}) \ge 0$. By Equation (6), $g_{\mathrm{C}}(\boldsymbol{x}) = 0$ if and only if $\boldsymbol{x}$ is Clarke stationary.

Similar to Yurtsever and Sra 2022, we can show:

Theorem 1 (Semismooth FW on concave objectives). Consider running the FW Algorithm 2 with unit step-size for the optimization problem in Equation (4). Then the minimal FW gap $\tilde{g}_t := \min_{0\le\tau\le t} g_\tau$ encountered by the iterates after t iterations satisfies:

$$\tilde{g}_t \le \frac{h_0}{t+1} \quad \text{for } 0 \le t \le T-1 \tag{8}$$

where $h_0 := f(\boldsymbol{x}^{(0)}) - \min_{\boldsymbol{x}\in\mathcal{M}} f(\boldsymbol{x})$ is the initial global suboptimality.

Lloyd's K-Means as Frank-Wolfe for Concave Objective

Lemma 1 characterizes f as a concave function in the interior of $\mathcal{M} = \mathrm{conv}(\mathcal{M}_0)$ (cf. Equation (3)). Since f is continuous in $\mathcal{M}$, due to Bauer's maximum principle, f attains a global minimum at one of the extreme points $\mathcal{M}_0$. This also applies to any local minimum that either must be an extreme point itself or be "tied" with one.

Lemma 1. For any $\boldsymbol{w} \in \mathcal{M}$ with $n_k^{\boldsymbol{w}} > 0$ for all k, ∇f and $\nabla^2 f$ can be expressed as

$$\partial_{w_{ik}} f(\boldsymbol{w}) = \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k^{\boldsymbol{w}}\|^2 \quad \text{and}$$
$$\partial_{w_{ik}} \partial_{w_{jl}} f(\boldsymbol{w}) = -\frac{2}{n_k^{\boldsymbol{w}}} \cdot \mathbb{1}_{\{k=l\}} \cdot (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k^{\boldsymbol{w}}) \cdot (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_k^{\boldsymbol{w}}).$$

Moreover, the Hessian $\nabla^2 f$ is negative semidefinite.

Unfortunately, ∇f can not be continuously extended to $\mathcal{M}$ so that the smooth FW algorithm cannot be applied to f over $\mathcal{M}$. However, the following theorem holds for a semismooth extension of f.

Theorem 2. Let $\boldsymbol{w}^{(t)}$, $0 \le t \le T$, be produced by Lloyd's K-means Algorithm 1. Then the membership vector $\boldsymbol{w}^{(t)}$ coincides with the the t-th iterate of the FW Algorithm 2 applied to minimizing f over $\mathcal{M}^{\mathrm{adm}}$ such that

$$\tilde{g}_t = \min_{0\le\tau\le t} g_\tau = \min_{0\le\tau\le t} \Delta\mathrm{SSE}_\tau \le \frac{h_0}{t+1}$$

for $t = 0, \ldots, T-1$, where $h_0 = f(\boldsymbol{x}^{(0)}) - \min_{\boldsymbol{x}\in\mathcal{M}} f(\boldsymbol{x})$ and $\Delta\mathrm{SSE}_t = f\big(\boldsymbol{w}^{(t)}\big) - f\big(\boldsymbol{w}^{(t+1)}\big)$.

Simulation Study

We illustrate our theoretical results with a simulation study. Since the iteration becomes stationary in finite time, we focus on the *worst-case* performance over randomly selected seeds – for a fixed dataset or over multiple datasets sampled from a given population. Letting g_t^{WC} denote the worst-case (i.e., largest) FW gap (over all replications) at iteration t, the convergence rate of g_t^{WC} was estimated using OLS applied to the regression model $\log(g_t^{\mathrm{WC}}) = \beta_0 + \beta_1 \log(t) + \varepsilon$ with i.i.d. normal errors ε.

Preasymptotic regime was defined as bottom $2/3$ of the $\log(t)$ range. Theoretical values were assumed $\beta_1 = -1$ and $\beta_0 = (n-1)\,\mathrm{tr}(\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is estimated using sample covariance. Synthetic data were simulated from the Gaussian mixture $\sum_{k=1}^{K} \pi_k \varphi_k(\boldsymbol{x}|\boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{I})$ with $\pi_k = \frac{1}{K}$ and $\sigma_k = 1$ (cf. `make_blobs()` of Python's `scikit-learn`). For each (n, d, K) with $n = 500, 1000, 5000$, $d = 2, 5, 10$ and $K = 5, 10, 20, 50$, we performed $10{,}000$ Monte Carlo (MC) simulations to estimate g_t^{WC}.

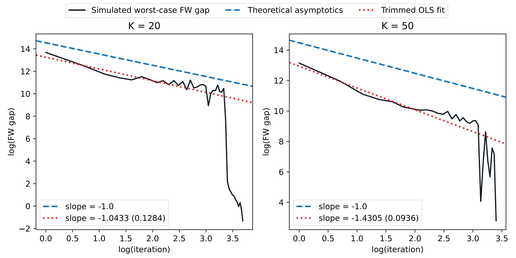

Fig. 1: Worst-case convergence rate on synthetic "blobs" data for $n = 5{,}000$ and $d = 10$.

Figure 1 displays the worst-case FW gap plot vs t. Applying an upper-sided Student's t-test of $H_0 : \beta_1 = -1$ vs $H_1 : \beta_1 > -1$ across all scenarios considered, the p-values never dropped below $1-(1-0.05)^{1/36} \approx 0.0014$, which is the Šidák-adjusted value to maintain a family-wise 0.05 test size, the composite null hypothesis failed to be rejected, corroborating that the worst-case preasymptotic rate was never slower than $\mathcal{O}(1/t)$.

Example: Image Segmentation Dataset

Consider the image segmentation dataset from UCI Machine Learning Repository. Pooling training and test data, each of $2{,}310$ instances is classified as one of the seven classes: `BRICKFACE`, `CEMENT`, `FOLIAGE`, `GRASS`, `PATH`, `SKY` and `WINDOW`. Keeping only continuous features (columns 6 through 19), we end up with 14 variables ($d = 14$). Pokojovy and Jobe 2022 empirically demonstrated that the 7-component mixture is non-Gaussian. An MC simulation of size $50{,}000$ was performed for $K = 7$. The results are displayed in Figure 2. The slope estimate is reported along with the standard error (in parentheses). Based on an upper-sided Student's t-test with $\hat{\beta}_1 = -1.2451$ and $\mathrm{se}(\hat{\beta}_1) = 0.1228$, the null hypothesis $H_0 : \beta_1 = -1$ failed to be rejected vs $H_1 : \beta_1 > -1$ at $\alpha = 0.05$, corroborating an $\mathcal{O}(1/t)$ worst-case rate.

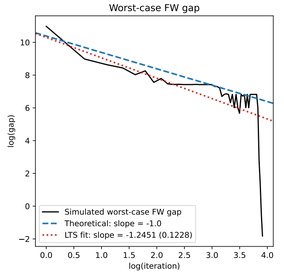

Fig. 2: Worst-case convergence rate.

Summary and Conclusions

We emphasize the importance of non-asymptotic convergence guarantees for Lloyd's K-means – or any other machine learning algorithm for that matter – that hold uniformly with respect to the sample size, space dimension, hyperparameters, initial seeding or optimization "intricacies" like the curvature constant or geometric aspects of the domain. Building upon the work of (Yurtsever and Sra 2022) in the context of convex-concave procedure, we proved that Lloyd's algorithm is a special case of FW. In doing so, we not only were able to perfectly carry over their results to Lloyd's K-means clustering but developed an FW variant for general semismooth concave objectives and proved an $\mathcal{O}(1/t)$ convergence rate. Owing to concavity, we could employ the less technical Clarke's instead of Goldstein's subdifferential prevalent in the nonsmooth literature allowing for better transparency and broader accessibility.

References

Liu, Z., C. Chen, L. Luo, and B.K.H. Low. 2024. "Zeroth-Order Methods for Constrained Nonconvex Nonsmooth Stochastic Optimization." In *Proc of the 41st ICML*, 235:30842–30872. PMLR.

Pokojovy, M., and J. M. Jobe. 2022. "A robust deterministic affine-equivariant algorithm for multivariate location and scatter." *Computational Statistics & Data Analysis* 172 (107475): 1–24.

Yurtsever, Alp, and Suvrit Sra. 2022. "CCCP is Frank-Wolfe in Disguise." In *Advances in Neural Information Processing Systems (NeurIPS)*, 35:29989–30001.