A Calculus Based Proof of Finite Convergence for the Lloyd–Forgy k–Means Algorithm

Hanna Bawardi

December 23, 2025

1 Notation and Objective

Let $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ $(n \ge k \ge 1)$ and write

$$\mu = (\mu_1, \dots, \mu_k) \in (\mathbb{R}^d)^k, \qquad a : X \to \{1, \dots, k\}.$$

The *squared–error functional*

$$\Phi(a, \mu) := \sum_{i=1}^{n} \big\| x_i - \mu_{a(x_i)} \big\|_2^2 \tag{1}$$

is the quantity minimised by Lloyd's algorithm.

Conventions. Throughout, assignments are assumed to be *deterministically tie–broken.* Clusters that become empty during an iteration are either excluded from the analysis or re–initialised so that all centroids appearing in $\Phi(a, \mu)$ correspond to nonempty clusters. Under this convention, all centroid updates are well defined.

2 Gradient Structure for Fixed Assignments

Lemma 1 (Stationary point of Φ in μ)**.** *For any fixed assignment a the function $f_a : (\mathbb{R}^d)^k \to \mathbb{R}$, $f_a(\mu) = \Phi(a, \mu)$ is continuously differentiable and convex, and is strictly convex in the coordinates corresponding to nonempty clusters. Its unique critical point is $\mu^\star = C(a) := (c_{a,1}, \dots, c_{a,k})$, where $c_{a,j} = \frac{1}{|a^{-1}(j)|} \sum_{x_i \in a^{-1}(j)} x_i$.*

Proof. For $j \in \{1, \dots, k\}$ define $S_j := a^{-1}(j)$. Then

$$f_a(\mu) = \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_i - \mu_j\|_2^2.$$

Because each summand is a quadratic form in μ_j, $\nabla_{\mu_j} f_a(\mu) = 2|S_j|\big(\mu_j - c_{a,j}\big)$. Setting the gradient to zero yields $\mu_j = c_{a,j}$ for every j. The Hessian block in coordinates μ_j is $2|S_j|\mathbf{I}_d \succeq 0$; since at least k clusters are non-empty the Hessian is positive definite on the product of the corresponding coordinate subspaces, establishing strict convexity and uniqueness of the minimiser. $\square$

3 Descent Identities via the Gradient

Two algebraic facts, both derivable by expanding (1), underpin the calculus-flavoured proof.

Lemma 2 (Centroid update decreases Φ)**.** *For fixed a, writing $\tilde{\mu} := C(a)$,*

$$\Phi\big(a, \mu\big) - \Phi\big(a, \tilde{\mu}\big) \;=\; \sum_{j=1}^{k} |a^{-1}(j)|\, \|\mu_j - \tilde{\mu}_j\|_2^2 \;\geq\; 0,$$

with equality iff $\mu = \tilde{\mu}$.

Proof. Expand $\|x - \mu_j\|_2^2$ as $\|x - \tilde{\mu}_j\|_2^2 + 2\langle x - \tilde{\mu}_j, \tilde{\mu}_j - \mu_j\rangle + \|\mu_j - \tilde{\mu}_j\|_2^2$ and sum $x \in S_j$. Because $\sum_{x \in S_j}(x - \tilde{\mu}_j) = 0$, the middle term vanishes, giving the claimed identity. □

Lemma 3 (Best-response assignment)**.** *Let $A(\mu)$ assign each x_i to the* closest *current centre, breaking ties deterministically. Then for all assignments a, $\Phi\big(A(\mu), \mu\big) \leq \Phi(a, \mu)$, with strict inequality if $a \neq A(\mu)$.*

Proof. Immediate from the definition of $A(\mu)$ since each summand in (1) takes the minimal possible value. □

4 A Block Coordinate–Descent View

Define the *Lloyd operator*

$$T(a, \mu) := \big(A(\mu),\, C(A(\mu))\big).$$

Combining Lemmas 2 and 3,

$$\begin{aligned}
\Phi\big(T(a, \mu)\big) &= \Phi\big(A(\mu), C(A(\mu))\big) \\
&\leq \Phi\big(A(\mu), \mu\big) && \text{(Lemma 2, } a = A(\mu)\text{)} \\
&< \Phi\big(a, \mu\big) && \text{if } a \neq A(\mu) \text{ or } \mu \neq C(a).
\end{aligned}$$

5 Finite Convergence via Descent and Finite Assignments

Theorem 4 (Finite termination)**.** *Starting from any $(a^{(0)}, \mu^{(0)})$ the sequence $(a^{(t)}, \mu^{(t)})_{t \geq 0} := T^{\,t}(a^{(0)}, \mu^{(0)})$ terminates after finitely many steps at $(a^*, \mu^*) = T(a^*, \mu^*)$, i.e. a block–coordinate stationary point satisfying*

$$a^* = A(\mu^*), \qquad \nabla_\mu \Phi(a^*, \mu^*) = \mathbf{0}.$$

Proof. By strict descent, $\big(\Phi(a^{(t)}, \mu^{(t)})\big)_{t \geq 0}$ is strictly decreasing until a fixed point is reached. Because X is finite the number of distinct assignments is $k^n < \infty$; each assignment determines a *unique* optimal centroid tuple by Lemma 2. Hence only finitely many distinct values of Φ are attainable, so strict descent can occur at most that many times. When descent ceases we have $a^* = A(\mu^*)$ (otherwise the assignment step would still decrease Φ) and $\mu^* = C(a^*)$ (otherwise the centroid step would). The gradient condition follows from Lemma 1 (§3). □

References

[1] S. P. Lloyd, Least squares quantization in PCM, *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.