Convergence of online k-means

Sanjoy Dasgupta Gaurav Mahajan Geelon So

November 22, 2021

Abstract

We prove global convergence for a general class of k-means algorithms performed over streaming data from a distribution—the centers asymptotically converge to the set of stationary points of the k-means objective function. To do so, we show that online k-means over a distribution can be interpreted as stochastic gradient descent with a stochastic learning rate schedule. Then, we prove convergence by extending techniques used in optimization literature to handle settings where center-specific learning rates may depend on the past trajectory of the centers.

1 Introduction

Lloyd's algorithm (Lloyd, 1982) is a popular iterative procedure for k-means clustering on a finite dataset in $\mathbb{R}^d$. At each step, the algorithm proposes k centers, say $W_1, \dots, W_k \in \mathbb{R}^d$. Each data point is then mapped to its closest center, partitioning the dataset into k clusters. The update simply sets each center to the mean of its corresponding cluster data. Because each step requires a pass over the whole dataset, large-scale data and streaming settings often use online variants of k-means, which compute updates on single data points.

Consider *online k-means algorithms* with updates that (i) receive a data point X, (ii) find the closest center W_{i^*} among $W_1, \dots, W_k$, and then (iii) update W_{i^*} using X. The long-term behavior of this procedure is unknown when it is applied to a never-ending stream of data points that is drawn from an underlying distribution p on $\mathbb{R}^d$. This leads to the following question:

> *If $X^{(1)}, X^{(2)}, \dots$ come from an underlying data distribution p, do these forms of online k-means algorithms converge to local optima of the k-means cost function f on p?*

1.1 A motivating example for analysis

Bottou and Bengio (1995) defines an example of an online k-means algorithm used in practice, which we call *naive online Lloyd's*. It simply sets each center W_i to the mean of all its previous updates, which can be computed in a streaming fashion. It does so by maintaining a counter N_i for the number of times each center has been updated so far. If W_i is the center closest to the next data point X, the update is:

$$W_i \leftarrow W_i - \frac{1}{N_i + 1}\big(W_i - X\big) \qquad \text{and} \qquad N_i \leftarrow N_i + 1$$

Generalizing, we consider a broader family of *online k-means* algorithms whose update is of the form:

$$W_i \leftarrow W_i - H_i \cdot \big(W_i - X_i\big)$$

where $H_i \in [0, 1]$ is a center-specific stochastic learning rate that may depend arbitrarily on the past and X_i is a random data point drawn from p, conditioned on having W_i as the closest center. Notice that W_i has a simple geometric meaning: it is a convex combination of all its previous updates.

*Dept. of Computer Science and Engineering, University of California, San Diego. {dasgupta, gmahajan, agso}@eng.ucsd.edu.

Algorithm *online k-means*
Initialize: k arbitrary distinct centers $W^{(0)} \in \mathbb{R}^{k \times d}$ from the support of p

1. **for** $n = 0, 1, 2, \ldots,$
2. **do** sample data points $X_i^{(n+1)} \sim p\big|_{V_i(W^{(n)})}$ for $i = 1, \ldots, k$
3. update center $W_i^{(n+1)} \leftarrow W_i^{(n)} - H_i^{(n+1)} \cdot \big(W_i^{(n)} - X_i^{(n+1)}\big)$

ALGORITHM. The class of online k-means algorithms analyzed in this work. This framework allows for more than one center to be updated per round; to recover the online setting with single-center updates, set the k-tuple of stochastic learning rates $H^{(n+1)}$ to be supported only on the ith center with probability equal to the mass of its Voronoi cell $V_i(W^{(n)})$. See Section 2 for notation.

1.2 Main results

Our main contributions: we (i) define a framework that covers a large class of online k-means algorithm, (ii) prove that algorithms in this family precisely perform stochastic gradient descent on the k-means cost function, (iii) prove asymptotic convergence to the set of stationary points of f, and (iv) extend Lloyd's algorithm to the online setting and prove its convergence.

(i) Online *k*-means framework Despite its simplicity, naive Lloyd's has eluded analysis. It is difficult to analyze because the update at each step depends on the center and the whole history of the trajectory of the algorithm; existing works replace the $\frac{1}{N_i+1}$ learning rate by a uniform-across-centers and deterministic learning rate $\frac{1}{n}$, where n is the number of elapsed iterations (Tang and Monteleoni, 2017). Introducing this general class of *online k-means* algorithms isolates the issue of *center-specific trajectory-dependent* learning rates, allowing us to prove convergence by placing mild conditions on them.

(ii) Analysis through SGD We show that algorithms in this family precisely perform stochastic gradient descent (SGD) on the k-means cost function f in Lemma 2.2. While known for k-means over finite datasets (Bottou and Bengio, 1995), the result does not trivially extend to distributions—the core difference is that there are finitely many ways to cluster a finite dataset, but infinitely many ways to cluster $\mathbb{R}^d$.

(iii) Convergence analysis Our framework is so broad that it even allows adversarially chosen learning rates. As an extreme example, an adversary can set the learning rate of a center W_i to zero; the iterates will never converge to a stationary point. Therefore, to prove convergence, we need to impose additional conditions. The key property that we shall require for convergence is that if a center W_i is far from its cluster mean—the mean of its Voronoi cell—then with constant probability, it shall be updated at a rate that is not too much slower than the rest of the centers. Theorem 4.2 provides the conditions for convergence.

(iv) Online Lloyd's algorithm It turns out that naive Lloyd's is particularly difficult to analyze. It is poorly conditioned in the sense that nothing seems to prevent iterates from making rare but large jumps—it is unclear whether it is not impossible to prove convergence for naive Lloyd's. Furthermore, naive Lloyd's may differ significantly from the original Lloyd's algorithm. In the original algorithm, centers are updated to the mean of the current clusters. But in naive Lloyd's, centers are set to the mean of all previous updates. But this mean-of-all-previous-updates does not generally well-approximate the mean of the current cluster because the underlying clusters drift about throughout the whole algorithm.

Instead, we start from the interpretation of Lloyd's algorithm as preconditioned gradient descent. Then, we define the *online Lloyd's algorithm* as a stochastic analog, which concurrently keeps an estimate of the preconditioner. We prove the consistency of our estimator to the Lloyd preconditioner in Appendix D. For the following theorem, we say that a k-tuple of centers $w \in \mathbb{R}^{k \times d}$ is *degenerate* if at least two of the centers coincide, $w_i = w_j$ for some $i \neq j$. The following theorem is an informal restatement of Theorem 4.3.

Theorem (informal)**.** Let p be a continuous density with bounded support on $\mathbb{R}^d$, and let f be its k-means cost function. Suppose that the set of stationary points $\{\nabla f = 0\}$ has no degenerate limit points. Then, the online Lloyd's algorithm globally converges to the set of stationary points:

$$\limsup_{n\to\infty} \inf_{w\in\{\nabla f=0\}} \|W^{(n)} - w\| = 0.$$

1.3 Related works

Bottou and Bengio (1995) show that Lloyd's algorithm for finite datasets can be viewed as gradient descent. Tang and Monteleoni (2017) use this to prove convergence of online k-means with uniform and deterministic learning rates, attaining rates of convergence. It would be of interest to prove rates in the distributional setting as well. However, one immediate difference is that the set of stationary points in the former setting is finite—hence isolated. There may be uncountably many stationary points in the distributional setting.

For our analysis of SGD, we make use of common frameworks to prove convergence (Bertsekas and Tsitsiklis, 2000; Li and Orabona, 2019). However, much of the general theory covers only uniform learning rates. Our work introduces a technique that might apply to prove convergence for more general SGD-based algorithms with non-uniform learning rates.

In our analysis of the k-means cost function f, we show that it admits a family of tangent quadratic upper bounds (Appendix A). This suggests that k-means over distributions, as in the finite setting, fits into the *majorization-minimization* (MM) scheme described by Mairal (2015). It would be of interest to generalize our work to iterative or online MM algorithms as in Cappé and Moulines (2009); Karimi et al. (2019).

2 Preliminaries

Let p be a density on $\mathbb{R}^d$ with bounded second moment. Notice that because p is a density, any Lebesgue measure zero set also has zero probability mass. We denote a tuple of k centers or prototypes in $\mathbb{R}^d$ by $w = \big(w_1, \dots, w_k\big) \in \mathbb{R}^{k\times d}$. Define the following:

- $V(w) = \big(V_1(w), \dots, V_k(w)\big)$ is the induced Voronoi partitioning[1] of $\mathbb{R}^d$ by w

$$V_i(w) = \left\{x \in \mathbb{R}^d : w_i \in \arg\min_{w_j} \|w_j - x\|\right\}$$

- $P(w) = \big(P_1(w), \dots, P_k(w)\big)$ is the probability mass of each of the Voronoi partitions

$$P_i(w) = \int_{V_i(w)} p(x)\, dx$$

- $M(w) = \big(M_1(w), \dots, M_k(w)\big)$ is the mean/center of mass of each of the Voronoi partitions

$$M_i(w) = \frac{1}{P_i(w)} \int_{V_i(w)} x\, p(x)\, dx.$$

While these functions are defined for all $w \in \mathbb{R}^{k\times d}$, we will be able to restrict our analysis to the set of non-degenerate tuples, where none of the centers coincides:

$$\mathcal{D} := \{w \in \mathbb{R}^{k\times d} : w_i \neq w_j,\ \forall i \neq j\}.$$

Later on, we will restrict the support of p to a closed ball $B(0, R)$ centered at the origin of radius R in $\mathbb{R}^d$. Let $\mathcal{D}_R$ be the set of non-degenerate tuples in $B(0, R)$:

$$\mathcal{D}_R := \{w \in \mathcal{D} : w_i \in B(0, R),\ \forall i \in [k]\}.$$

As last bits of notation, given a Borel set $S \subset \mathbb{R}^d$ with positive probability mass $p(S) > 0$, let $p\big|_S$ denote the distribution obtained by restricting p onto S. Finally, we will also let $[k]$ denote the set $\{1, \dots, k\}$.

[1]Strictly speaking, $V(w)$ does not partition $\mathbb{R}^{k\times d}$ because adjacent partitions $V_i(w)$ and $V_j(w)$ share boundary points. However, boundary points form a measure zero set, so we will encounter no problems.

2.1 The k-means problem

The k-means objective is to minimize:

$$f(w) := \frac{1}{2} \sum_{i \in [k]} \int_{V_i(w)} \|w_i - x\|^2 \, p(x) \, dx. \tag{1}$$

While the k-means cost is non-convex, we show that it is smooth on $\mathcal{D}$. Therefore, we aim for convergence to stationary points—the iterates $W^{(n)}$ approach the set of stationary points $\{\nabla f = 0\}$ in the limit as $n \to \infty$.

Definition 2.1 (Global convergence)**.** Let $\mathcal{D}$ be a domain and $f : \mathcal{D} \to \mathbb{R}$ be differentiable. We say that a sequence of points $(w^{(n)})_{n=0}^{\infty}$ in $\mathcal{D}$ *globally converges to stationary points of* f if all limit points of $(w^{(n)})_{n=0}^{\infty}$ are contained in the set of stationary points $\{\nabla f = 0\}$,

$$\bigcap_{n \geq 0} \overline{(w^{(n')})_{n' \geq n}} \subset \{\nabla f = 0\}.$$

Computing the derivative of the k-means cost (1) with respect to w is relatively involved because the both the domain of integration and the integrand depend on w.

Lemma 2.2 (Gradient of k-means objective)**.** *Let p be a density on $\mathbb{R}^d$ with $\mathbb{E}_{X \sim p}\left[\|X\|^2\right] < \infty$. Let f be the k-means objective (1). Then f is continuously differentiable on $\mathcal{D}$, where:*

$$\nabla_{w_i} f(w) = P_i(w) \cdot \big(w_i - M_i(w)\big). \tag{2}$$

Proof sketch. A change in f due to a small perturbation at w to $w + \varepsilon$ can be broken down into two parts. First, for points $x \in V_i(w) \cap V_i(w + \varepsilon)$ that remain within the ith Voronoi region, the accumulated change in cost is due to shifting the ith center,

$$\frac{1}{2} \int_{V_i(w) \cap V_i(w+\varepsilon)} \left(\|w_i + \varepsilon_i - x\|^2 - \|w_i - x\|^2\right) p(x) \, dx.$$

Note that in the limit as ε approaches 0, the domain of integration is the points in the interior of $V_i(w)$.

Second, for points $x \in V_i(w) \cap V_j(w + \varepsilon)$ that switch from the ith to the jth Voronoi region, the change in cost is due to switching regions. Note that in the limit as ε approaches 0, these points are on the boundary $V_i(w) \cap V_j(w)$. But as points on the boundary $V_i(w) \cap V_j(w)$ are equally distanced from either the ith and jth centers, this second term overall contributes nothing to the first-order change in f. It turns out that the derivative of f can be computed by treating the domains of integration as fixed. By dominated convergence, we can move the derivative past the integral:

$$\begin{aligned} \nabla_{w_i} f(w) &= \frac{1}{2} \int_{V_i(w)} \nabla_{w_i} \|w_i - x\|^2 \, p(x) \, dx \\ &= \int_{V_i(w)} (w_i - x) \, p(x) \, dx. \end{aligned}$$

Substituting the definition of P_i and M_i completes the proof. Appendix A makes this argument rigorous. □

3 Convergence of cost

Notice that for the k-means objective to be finite, p must have bounded second moment. But for our convergence analysis, we will make a stronger assumption:

Assumption 3.1. Assume p has bounded support, i.e. $\Pr_{X \sim p}(\|X\| > R) = 0$ for some $R > 0$.

We also make assumptions on the learning rates used in Algorithm *online k-means*. Let $X^{(n)}$ and $H^{(n)}$ denote the tuples $\left(X_1^{(n)}, \ldots, X_k^{(n)}\right)$ and $\left(H_1^{(n)}, \ldots, H_k^{(n)}\right)$, respectively.

Assumption 3.2. Define $\left(\mathcal{F}_n\right)_{n=0}^{\infty}$ to be the natural filtration associated to the *online k-means* algorithm. In particular, let $\mathcal{F}_0 = \sigma\left(W^{(0)}\right)$ be the σ-algebra generated by $W^{(0)}$. Define $\mathcal{F}_n = \sigma\left(X^{(n)}, H^{(n)}, \mathcal{F}_{n-1}\right)$ as the σ-algebra containing all information up to iteration n. We assume that:

(1) If $P_i(W^{(n)}) = 0$, then $H_i^{(n+1)} = 0$ almost surely.

(2) $H^{(n+1)}$ and $X^{(n+1)}$ are conditionally independent given $\mathcal{F}_n$.

(3) $0 \leq H_i^{(n+1)} \leq 1$.

When $P_i(W^{(n)}) = 0$, it may not be sensible to draw from $V_i(W^{(n)})$. The first assumption avoids this situation; we require ${H_i}^{(n+1)} = 0$ a.s. and let ${X_i}^{(n+1)}$ be arbitrary—it is not used to update the centers. The second assumption ensures that $H^{(n+1)}$ does not depend on $X^{(n+1)}$. The final assumption that $H_i^{(n+1)} \leq 1$ is not strictly necessary, but is simplifying for analysis and is intuitively natural: the update is a convex combination of the previous center and new data point. This makes it particularly easy to see that $W^{(n)}$ remains in $\mathcal{D}$, the set of non-degenerate tuples almost surely:

Lemma 3.1. *Let $0 \leq H_i^{(n+1)} \leq 1$. Then $W^{(n)} \in \mathcal{D}$ is non-degenerate for all $n \in \mathbb{N}$ almost surely.*

With Assumption 3.1, we also have that all centers and updates takes place in the closed ball $B(0, R) \subset \mathbb{R}^d$ of radius R almost surely. As this is a region with diameter $2R$, the amount that each center moves can be controlled by bounding the learning rates:

Lemma 3.2. *Suppose $\Pr_{X \sim p}(\|X\| > R) = 0$. Let $H_i^{(n)} \leq 1$ for all $n \in \mathbb{N}$ and $i \in [k]$. If $n > m$, then:*

$$\|W^{(m)} - W^{(n)}\| \leq 2R \cdot \sum_{i \in [k]} \sum_{m \leq n' < n} H_i^{(n'+1)} \quad \text{a.s.}$$

We do not lose much for making this assumption, and at any rate, we shall require the learning rate to converge to zero in our convergence analysis. From now on, we implicitly make Assumptions 3.1 and 3.2.

3.1 Convergence analysis

The first result we give is the asymptotic convergence of the k-means cost, $f(W^{(n)})$, which will be based on a supermartingale argument commonly used in proving the convergence of stochastic gradient descent. Recall that a bounded and monotonically decreasing real-valued sequence converges to a real value. This remains true in the random setting. A supermartingale is a noisy sequence that in expectation decreases monotonically. Provided that the noise can be controlled, then the *martingale convergence theorem* shows that such a stochastic sequence will converge to some real-valued random variable, e.g. see Durrett (2019).

Proposition 3.3 (Convergence of cost)**.** *Let $\left(W^{(n)}\right)_{n=0}^{\infty}$ be a sequence generated by the online k-means algorithm. If the following series converges:*

$$\sum_{n=1}^{\infty} \sum_{i \in [k]} \left(H_i^{(n)}\right)^2 < \infty \quad \text{a.s.},$$

then there is an $\mathbb{R}$-valued random variable f^* such that $f(W^{(n)})$ converges to f^* almost surely.

Proof sketch. Lemma A.3 shows that for every $w \in \mathcal{D}$, there is a quadratic upper bound tangent to f at w:

$$f(w') \leq f(w) + \nabla f(w)^\top (w' - w) + \frac{1}{2}\|w' - w\|^2.$$

Set $w' = W^{(n+1)}$ and $w = W^{(n)}$ to bound the amount f decreases each step of the algorithm. Also recall:

$$W_i^{(n+1)} - W_i^{(n)} = -H_i^{(n+1)} \cdot \left(W_i^{(n)} - X_i^{(n+1)}\right).$$

Thus, in expectation, the algorithm takes a step in the direction of the negative gradient:

$$\nabla_{w_i} f(W^{(n)}) = P_i(W^{(n)}) \cdot \big(W_i^{(n)} - M_i(W^{(n)})\big).$$

Suppose that we are able to neglect the quadratic term $\frac{1}{2}\|w' - w\|^2$ of the upper bound. Then this shows that $f(W^{(n)})$ is a supermartingale; f decreases in expectation each iteration since it takes a step in the direction of the negative gradient. In order to apply the martingale convergence theorem, we need to know not only that the expected value decreases, but that the total amount of noise is bounded. Indeed, the noise at each step can be bounded by Lemma 3.2. In particular,

$$\operatorname{Var}\big(\|W_i^{(n+1)} - W_i^{(n)}\|\big) \le \sum_{i\in[k]} \mathbb{E}\left[\big\|W_i^{(n+1)} - W_i^{(n)}\big\|^2\right] \le 4R^2 \sum_{i\in[k]} \big(H_i^{(n+1)}\big)^2.$$

Thus the total noise of the process is finite:

$$\sum_{n\in\mathbb{N}} \operatorname{Var}\big(\|W^{(n+1)} - W^{(n)}\|\big) < \infty \qquad \text{a.s.}$$

Martingale convergence shows that $f(W^{(n)})$ converges.

Of course, we cannot just drop the quadratic term in the upper bound. But notice that the quadratic terms form convergent series; each term is dominated by terms of a convergent series:

$$\frac{1}{2}\big\|W^{(n+1)} - W^{(n)}\big\|^2 \le 2R^2 \sum_{i\in[k]} \big(H_i^{(n+1)}\big)^2.$$

So, except for a convergent series, $f(W^{(n)})$ is a supermartingale. Martingale convergence applies here too.

Appendix B fills in the technical details. □

4 Convergence of iterates

Our main result is the convergence of the k-means iterates $W^{(n)}$ to the set of stationary point $\{\nabla f = 0\}$. For this, we need some additional assumption besides Assumptions 3.1 and 3.2. Lemmas 3.1 and 3.2 show that the iterates $W^{(n)}$ are non-degenerate and bounded, where $W^{(n)} \in \mathcal{D}_R \subset B(0, R)$. From now on, we restrict all sets to the (topological) subspace $\mathcal{D}_R \subset \mathbb{R}^{k\times d}$. For example, when we write $\{\nabla f = 0\}$, we implicitly mean $\{\nabla f = 0\} \cap \mathcal{D}_R$. We assume:

Assumption 4.1. The set $\{\nabla f = 0\}$ of stationary points is compact in $\mathcal{D}_R$.

Geometrically, this means that $\{\nabla f = 0\}$ has no degenerate limit point (see Lemma 4.1's proof). This lets us prove that $W^{(n)}$ converges to $\{\nabla f = 0\}$ by showing that $\nabla f(W^{(n)})$ converges to zero. Formally, we apply Lemma C.4, which relies on (i) the continuity of ∇f, and (ii) the existence of a compact subset of $\mathcal{D}_R$ containing $\{\nabla f = 0\}$. Indeed, a consequence is that level sets $\{\|\nabla f\| \le \varepsilon\}$ are compact for small enough ε.

Lemma 4.1. *Let $\{\nabla f = 0\}$ be compact in $\mathcal{D}_R$. There exists $\varepsilon_0 > 0$ so that if $\varepsilon \in [0, \varepsilon_0]$, the sets $\{\|\nabla f\| \le \varepsilon\}$ and $\{\|\nabla_{w_i} f\| \le \varepsilon\}$ are compact in $\mathcal{D}_R$ for $i \in [k]$.*

From here on, we additionally make Assumption 4.1.

4.1 Convergence analysis

The key idea is to link the convergence of $\nabla f(W^{(n)}) \to 0$ to that of $f(W^{(n)}) \to f^*$. Broadly speaking, we need conditions on the learning rate so that whenever $\|\nabla f(W^{(n)})\|$ is larger than ε, then the cost is likely to decrease by at least δ. Then, Borel-Cantelli (Lemma C.5) would show that the cost decreases by a large amount infinitely often if the gradient is large infinitely often. But as the cost converges, this cannot happen.

Were we in the noiseless gradient descent setting, we could—in a single step—turn a large gradient into a large decrease in cost. We simply ensure that the iterates move in a direction that significantly decreases f

by lower bounding the learning rates for centers $i \in [k]$ with large gradients $\nabla_{w_i} f$. However, in our setting, the true gradient direction is obscured by the presence of noise, and so lower bounding the learning rates by a constant no longer guarantees a decrease in cost.

Instead, in the stochastic gradient descent setting, we can—over many steps—turn a *region* with large gradients into a large decrease in cost, with high probability. Since our learning rate decays to zero, eventually, whenever we enter a region with large gradients, we remain in that region for sufficiently many iterations so as to average out the noise and recover the underlying signal to move along the negative gradient direction. As in the noiseless setting, we want the choice of learning rate to not disproportionately dampen learning for centers $i \in [k]$ with large gradients $\nabla_{w_i} f$. In other words, the iterates should not leave this region of large gradients before having accumulated enough learning on those centers with large gradients.

As a motivating adversarial example, suppose we decrease the learning rates for centers with large gradients while increasing the rates for centers with small gradients. Then, the iterates may escape the region with large gradients by moving orthogonally to the gradient descent direction, not significantly decreasing the function value. And as an extreme example, if we never update the ith center, so $H_i^{(n+1)} \equiv 0$ for all n, then the cost $f(W^{(n)})$ may converge while the gradients $\|\nabla f(W^{(n)})\|$ may never converge to zero.

To preclude these examples, we need to be able to control the learning rates whenever the gradient $\|\nabla_{w_i} f(W^{(m)})\|$ becomes large for center $i \in [k]$ at iteration m. But because the convergence of cost analysis required the learning rates to decay to zero (Proposition 3.3), to accumulate the same amount of learning, we will need to control the learning rates over increasingly many iterations as $m \to \infty$. In particular, if the gradient becomes large at iteration m, we shall aim to control the learning rates during the interval between m and a future horizon $T(m)$, for some appropriately chosen function $T : \mathbb{N} \to \mathbb{N}$.

We impose two types of conditions on the learning rate. The first condition allows us to ensure that the iterates remain within the region of large gradients during the interval from m to $T(m)$. Recall Lemma 3.2 showed that we can bound the total displacement between $W^{(m)}$ and $W^{(T(m))}$ by bounding the accumulated learning rates; the first condition is of the form:

$$\sum_{j \in [k]} \sum_{m \le n < T(m)} H_j^{(n+1)} < r.$$

The second condition ensures that we accumulate enough learning for the center $i \in [k]$ with large gradients, so that the cost decreases by a constant between iterations m and $T(m)$,

$$\sum_{m \le n < T(m)} H_i^{(n+1)} > s.$$

In fact, it is enough that these conditions hold with constant probability.

Theorem 4.2 (Convergence of iterates)**.** *Let $W^{(n)}$ and $H^{(n+1)}$ be as in Proposition 3.3. Suppose that for all $i \in [k]$, $\varepsilon > 0$, and sufficiently small $r > 0$, there exists $T : \mathbb{N} \to \mathbb{N}$, $m_0 \in \mathbb{N}$, and some $s, c > 0$ so that:*

$$\Pr\left(\sum_{j \in [k]} \sum_{m \le n < T(m)} H_j^{(n+1)} < r \quad \text{and} \sum_{m \le n < T(m)} H_i^{(n+1)} > s \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| > \varepsilon \right) > c,$$

for any $m > m_0$. Then, $W^{(n)}$ globally converges to stationary points of the k-means cost f almost surely.

Proof sketch. We show for any $\varepsilon > 0$ and $i \in [k]$, the gradient $\|\nabla_{w_i} f(W^{(n)})\|$ is greater than ε only finitely often. Suppose that for all $w \in \mathcal{D}_R$ with large gradient $\|\nabla_{w_i} f(w)\| > \varepsilon$, there is a region around w also with large gradients. In particular, that $\|\nabla_{w_i} f(w')\| > \frac{\varepsilon}{2}$ whenever $\|w - w'\| < R_0$ for some R_0 depending on ε.

Let $r_0 = R_0/2R$ and choose $r < r_0$. Then, for all $m \le n < T(m)$, we obtain:

$$\|W^{(m)} - W^{(n)}\| < R_0,$$

whenever the first event in the probability in the theorem statement holds. This follows from Lemma 3.2, which bounds this displacement through the learning rate bound r_0. In short, with constant probability during this interval, the ith gradient remains large, $\|\nabla_{w_i} f(W^{(n)})\| > \frac{\varepsilon}{2}$.

Since the ith gradient remains large, the cost is likely to decrease significantly as long as the ith center is updated enough times. Lemma B.1 relates the learning rate to the decrease in cost, so that if the second event in the probability statement holds, then the cost decreases by at least:

$$\sum_{m \le n < T(m)} H_i^{(n+1)} \|\nabla_{w_i} f(W^{(n)})\|^2 > \frac{s\varepsilon^2}{4},$$

if we may neglect the noise terms of Lemma B.1. And so if $\|\nabla_{w_i} f(W^{(n)})\| > \varepsilon$ infinitely often, then the cost must decrease by a constant infinitely often by Borel-Cantelli, contradicting the convergence of the cost.

The proof is only slightly more complicated because of noise, but not significantly so as the noise forms a convergent series, by Lemma B.2. As a result, there is a random time M after which a deterministic cost decrease of $s\varepsilon^2/4$ will completely dominate any increase of cost due to the noise term.

The remaining technical complication is our implicit assumption earlier that $\|\nabla_{w_i} f\|$ is uniformly continuous on $\mathcal{D}_R$, which would allow us to find a constant R_0 given ε that holds for all $\|\nabla_{w_i} f(w)\| > \varepsilon$. Though $w \mapsto \|\nabla_{w_i} f(w)\|$ is continuous on $\mathcal{D}_R$, it is not guaranteed to be uniformly continuous since $\mathcal{D}_R$ is not compact. Our solution is to apply Lemma 4.1 to find an $\varepsilon_0 > 0$ for which $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ is compact, so:

$$\mathcal{D}_R = \{\|\nabla_{w_i} f\| \le \varepsilon_0\} \cup \{\|\nabla_{w_i} f\| > \varepsilon_0\}. \tag{3}$$

The first subset is compact, so uniform continuity holds. On the second subset, we rule out the possibilities that (i) the iterates eventually remain in this set, and (ii) the iterates exit and re-enter this set infinitely often. Case (i) is impossible by an argument akin to our earlier one on $[\varepsilon, \varepsilon_0]$. In fact, here we can bypass finding R_0 altogether since the iterates are guaranteed to remain in a region with large gradients forever. Case (ii) is impossible because this forces the iterates to enter $[\varepsilon, \varepsilon_0]$ infinitely often once the learning rate has become sufficiently small. Thus for all $\varepsilon > 0$, the iterates eventually never return to $\{\|\nabla_{w_i} f\| > \varepsilon\}$. □

We prove a slightly more general theorem in Theorem C.1, which makes the decomposition (3) explicit. We introduce separate conditions on the learning rate on $\{\|\nabla_{w_i} f\| \in [\varepsilon, \varepsilon_0]\}$ and $\{\|\nabla_{w_i} f\| > \varepsilon_0\}$; that way, we can remove the learning rate upper bound condition where we do not need it.

4.2 Online Lloyd's algorithm

Let us return now to extending Lloyd's algorithm into the online setting. Lloyd's algorithm is defined to iteratively set each center w_i to $M_i(w)$,

$$w_i \leftarrow M_i(w).$$

Since $\nabla_{w_i} f(w) = P_i(w) \cdot \big(w_i - M_i(w)\big)$, the Lloyd update is precisely *preconditioned gradient descent*, where the preconditioner at w for the ith center is $P_i(w)^{-1}$.

$$w_i \leftarrow w_i - \frac{1}{P_i(w)} \nabla_{w_i} f(w).$$

A noiseless gradient descent algorithm is often converted into a stochastic one by introducing decaying learning rates, which allows the noise to average out over time (Bottou, 1998). Applying the common decay rate of n^{-1}, we might aim for an update that, in expectation, looks like:

$$w_i \leftarrow w_i - \frac{1}{nP_i(w)} \nabla_{w_i} f(w).$$

As a first pass at designing the stochastic version of Lloyd's algorithm, suppose that we had access directly to p. Then, we might consider the following idealized learning rate achieving the above expected update:

$$H_{\text{ideal},i}^{(n+1)} \leftarrow \frac{\mathbb{1}\{I^{(n+1)} = i\}}{nP_i(W^{(n)})},$$

where $I^{(n+1)}$ is the index of the updated center; it is drawn from the distribution $P(W^{(n)})$ over $[k]$. Naive Lloyd's algorithm described in Section 1.1 could then be considered as a rough approximation to this idealized learning rate. To compare, recall the naive Lloyd's update:

$$H^{(n+1)}_{\text{naive},i} \leftarrow \frac{\mathbb{1}\{I^{(n+1)} = i\}}{N_i^{(n)} + 1},$$

where $N_i^{(n)}$ is the update count for the ith center. If $P(W^{(n)})$ does not vary much over time, then $N_i^{(n)} + 1$ is approximately equal to $nP_i(W^{(n)})$. In other words, the naive Lloyd's update appears to approximate the idealized learning rate by applying a stochastic preconditioner computed on:

$$\widehat{P}_i^{(n)} = \frac{N_i^{(n)}}{n}.$$

We call this algorithm naive as it assumes that $\widehat{P}_i^{(n)}$ is a reasonable estimator of ${P_i}^{(n)} := P_i(W^{(n)})$. Because the Voronoi partitions drift throughout the whole algorithm, this assumption is generally false.

However, the main issue with naive Lloyd's is not its naiveté. Rather, the issue lies with the idealized preconditioner $P_i(w)^{-1}$, which is poorly conditioned—it can become arbitrarily large. While this is not a problem in the noiseless setting, in the stochastic setting, the learning rate cannot become unbounded.

As a second pass, let us consider a second idealized online Lloyd's, where we introduce an upper bounding rate of t_n^{-1} for some positive sequence t_n. Set the learning rate:

$$H^{(n+1)}_{\text{ideal}',i} \leftarrow \frac{\mathbb{1}\{I^{(n+1)} = i\}}{\max\left\{nP_i^{(n)},\, t_n\right\}},$$

where we've refrained from preconditioning when $P_i(w) < t_n/n$. Note that if $t_n = o(n)$, then we can expect the set of points on which the algorithm will precondition to grow. Of course, this is idealized since we generally do not have direct access to p with which to compute $P_i(w)$.

This motivates what we call the *online Lloyd's algorithm*, which simultaneously constructs an estimator $\widehat{P}_i^{(n)}$ of ${P_i}^{(n)}$ based on the empirical rate at which the ith center has recently been updated in the past s_n steps, for some sequence s_n. We define its learning rate by:

$$H_i^{(n+1)} \leftarrow \frac{\mathbb{1}\{I^{(n+1)} = i\}}{\max\left\{n\widehat{P}_i^{(n)},\, t_n\right\}} \qquad \text{and} \qquad \widehat{P}_j^{(n)} = \frac{1}{s_n} \sum_{n_\circ \le n' < n} \mathbb{1}\{I^{(n'+1)} = j\}, \tag{4}$$

where we denote $n_\circ = n - s_n$, and where s_n and t_n are non-decreasing sequences in $\mathbb{N}$. On the one hand, in order to obtain a low-bias estimator, we need $s_n = o(n)$ so that updates in the distant past are forgotten. But on the other, we would like the estimator to concentrate as n goes to infinity, so we also require $s_n \uparrow \infty$.

By introducing conditions on s_n and t_n, we show that $\widehat{P}_i^{(n)}$ is a consistent estimator of ${P_i}^{(n)}$ in Appendix D. Furthermore, we obtain the following convergence theorem, as a corollary of Theorem C.1.

Theorem 4.3 (Convergence of iterates, online Lloyd's)**.** *Let $H_j^{(n)}$ and $\widehat{P}_j^{(n)}$ as in (4). Let s_n and t_n be non-decreasing sequences satisfying:*

$$\lim_{n\to\infty} \frac{n^{2/3}\log n}{s_n} = \lim_{n\to\infty} \frac{s_n \log s_n}{t_n} = \lim_{n\to\infty} \frac{t_n}{n} = 0.$$

If p is continuous, then $W^{(n)}$ globally converges to stationary points of the k-means cost f almost surely.

Proof sketch. To apply Theorem 4.2, we need to show that with constant probability:

$$\sum_{j\in[k]} \sum_{m\le n<T(m)} H_j^{(n+1)} < r \qquad \text{and} \qquad \sum_{m\le n<T(m)} H_i^{(n+1)} > s,$$

whenever $\|\nabla_{w_i} f(W^{(m)})\| \ge \varepsilon$. We achieve this by proving tighter bounds *in expectation*:

$$\sum_{j\in[k]} \sum_{m\le n<T(m)} \mathbb{E}\left[H_j^{(n+1)} \,\middle|\, \mathcal{F}_n\right] < r/100 \qquad \text{and} \qquad \sum_{m\le n<T(m)} \mathbb{E}\left[H_i^{(n+1)} \,\middle|\, \mathcal{F}_n\right] > 100s,$$

which can be converted into bounds *in probability* by the Markov's and Azuma-Hoeffding's inequalities.

Notice that the conditional expectation is:

$$\mathbb{E}\left[H_j^{(n+1)} \,\middle|\, \mathcal{F}_n\right] = \frac{P_j^{(n)}}{\max\{n\widehat{P}_j^{(n)}, t_n\}}.$$

Assume $P_j^{(n)}$ is not too small and that $\widehat{P}_j^{(n)} = P_j^{(n)}$, so the estimator is perfect. Then in fact:

$$\mathbb{E}\left[H_j^{(n+1)} \,\middle|\, \mathcal{F}_n\right] = \frac{1}{n}.$$

Define the map $T_r : \mathbb{N} \to \mathbb{N}$ so that $T_r(m)$ is the unique natural number so that:

$$\sum_{m \le n < T_r(m)} \frac{1}{n} \le r < \sum_{m \le n \le T_r(m)} \frac{1}{n}.$$

We obtain the desired conditional expectation bounds if we set $T \equiv T_{Cr}$ with $C = 1/100k$ and $s = Cr/100$.

We assumed (i) $P_j^{(n)}$ is not too small and (i) $\widehat{P}_j^{(n)}$ is perfect; neither is true in general. The first caveat is easier to deal with; the fear is that the centers with low update probabilities may sporadically contribute a large amount $1/t_n$, when compared to the $1/n$ of the other centers. However, our conditions on s_n and t_n ensure that centers with small Voronoi masses also have little effect on the overall behavior. For the second, we show that the estimator is highly concentrated around its true mean (Lemma D.10).

To sketch why $\widehat{P}_j^{(n)}$ concentrates about $P_j^{(n)}$, we have:

$$\widehat{P}_j^{(n)} = \frac{1}{s_n} \sum_{n_\circ \le n' < n} \mathbb{1}\{I^{(n'+1)} = j\} \qquad \text{and} \qquad P_j^{(n')} = \mathbb{E}\left[\mathbb{1}\{I^{(n'+1)} = j\} \,\middle|\, \mathcal{F}_{n'}\right].$$

Azuma-Hoeffding's shows that $\widehat{P}_j^{(n)}$ concentrates:

$$\Pr\left(\left|\widehat{P}_j^{(n)} - \frac{1}{s_n} \sum_{n_\circ \le n' < n} P_j^{(n')}\right| \ge a \,\middle|\, \mathcal{F}_{n_\circ}\right) \le 2\exp\left(-\frac{1}{2} s_n a^2\right).$$

We just need to show that $P_j^{(n')}$ remains close to $P_j^{(n)}$ over this interval. If P_j were L-Lipschitz, then:

$$\left|P_j^{(n')} - P_j^{(n)}\right| \le 2RL \cdot \sum_{j'' \in [k]} \sum_{n_\circ \le n'' < n} H_{j''}^{(n''+1)},$$

for all $n_\circ \le n' < n$, where conditions on s_n and t_n forces the right-hand side to go to zero (Lemma D.9).

This would be the proof if P_j were globally Lipschitz on $\mathcal{D}_R$. But this is not generally the case. For example, consider the 2-means problem on $\mathbb{R}^2$. When the two centers are very close together, then the Voronoi cells they produce are much more sensitive to small perturbations than when they are far apart. However, it is the case that when p is continuous, then $P_j : \mathcal{D}_R \to [0,1]$ is locally Lipschitz (Lemma D.11), so that P_j is Lipschitz on compact subsets K of $\mathcal{D}_R$.

A slight complication arises in each step of this proof because cannot directly condition on the iterates remaining in K; conditioning on a future event destroys the martingale property required by Azuma-Hoeffding's. To overcome this issue, we construct a core-set $K_\circ \subset K$ so that if $W^{(n_\circ)}$ is initially in $K_\circ$ and the accumulated learning rate is bounded:

$$W^{(n_\circ)} \in K_\circ \quad \text{and} \quad \sum_{j \in [k]} \sum_{n_\circ \le n' < n} H_j^{(n'+1)} < r_\circ,$$

then $W^{(n')}$ remains in K almost surely throughout the interval $n_\circ \le n' \le n$. As this construction splits the analysis into two cases (iterates in and outside K), in the actual proof, this theorem follows from the more general Theorem C.1, which allows us to break the analysis down into these two cases.

Appendix D makes this argument rigorous. □

Acknowledgements

Thanks to Yian Ma, Zhi Wang, and reviewers for helpful discussions that improved this work.

References

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 2000.

Léon Bottou. Online algorithms and stochastic approximations. In *Online Learning and Neural Networks*. Cambridge University Press, 1998.

Léon Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems*, 1995.

Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2009.

Rick Durrett. *Probability: theory and examples*. Cambridge University Press, 2019.

Pavel Grinfeld. *Introduction to tensor analysis and the calculus of moving surfaces*. Springer, 2013.

Belhal Karimi, Hoi-To Wai, Eric Moulines, and Marc Lavielle. On the global convergence of (fast) incremental expectation maximization methods. *arXiv preprint arXiv:1910.12521*, 2019.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 1982.

Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 2015.

Cheng Tang and Claire Monteleoni. Convergence rate of stochastic k-means. In *Artificial Intelligence and Statistics*, 2017.

Contents of the appendix

- **Appendix A: analysis of the k-means cost f**
  - Proposition A.1 shows that f has a family of quadratic upper bounds
  - Lemma 2.2 computes the gradient of f
  - Lemma A.3 combines the previous two results to give an analytic upper bound on f
- **Appendix B: convergence analysis for cost $f(W^{(n)})$**
  - Lemma B.1 uses Lemma A.3 to upper bound $f(W^{(n+1)}) - f(W^{(n)})$
  - Lemma B.2 proves that noise terms in Lemma B.1 forms a convergent series
  - Proposition 3.3 proves that the cost $f(W^{(n)})$ converges almost surely
- **Appendix C: convergence analysis for iterates $W^{(n)}$**
  - Lemma 3.1 shows that we can restrict the analysis to non-degenerate tuples of centers $\mathcal{D}$
  - Lemma 3.2 shows that we can bound $\|W^{(m)} - W^{(n)}\|$ by bounding the learning rates instead
  - Theorem C.1 is our general convergence result for iterates
    - Lemma C.4 shows that convergence of $W^{(n)}$ can be passed to convergence of $\nabla f(W^{(n)}) \to 0$
  - Theorem 4.2 is a simplified version of Theorem C.1 seen in the main body
    - Lemma C.2 and Lemma 4.1 shows that Theorem 4.2 satisfies the conditions in Theorem C.1
- **Appendix D: analysis of the online Lloyd's algorithm**
  - Theorem 4.3 proves the convergence of online Lloyd's algorithm through Theorem C.1
    - Lemma D.4 shows half of condition (A1) of Theorem C.1: total learning rates upper bound
    - Lemma D.6 shows other half of (A1): learning rates lower bound for centers with large gradients
      - Both Lemma D.4 and Lemma D.6 make use of the concentration of $\widehat{P}_j^{(n)}$ shown in Lemma D.9 and Lemma D.10, which requires that the maps P_j are locally Lipschitz, shown in Lemma D.11
      - They also make use of purely analytic lemmas Lemma D.12 and Corollary D.13
    - Lemma D.7 shows condition (A2) of Theorem C.1
  - Lemma D.9 shows that the accumulated learning rate over the window of s_n steps goes to zero
  - Lemma D.10 shows that the estimator $\widehat{P}_j^{(n)}$ is consistent with non-asymptotic concentration rates
  - Lemma D.11 proves that P_j is locally Lipschitz, an assumption of Lemma D.10
  - Lemma D.12 and Corollary D.13 are auxiliary technical lemmas to analyze $T(m)$

A Analysis of the k-means cost

In order to analyze the k-means algorithm through the lens of gradient descent, we need to be able to prove smoothness properties and calculate the gradient of the k-means objective,

$$f(w) := \frac{1}{2} \sum_{i \in [k]} \int_{V_i(w)} \|w_i - x\|^2 \, p(x) \, dx. \tag{1}$$

But because the domains and integrands depend on w simultaneously, taking the gradient is not so straightforward. To simplify analysis, we can fix the Voronoi partition with respect to some tuple of centers $w' \in \mathbb{R}^{k \times d}$ in order to define a family of surrogate objectives parametrized by w',

$$g(w; w') := \frac{1}{2} \sum_{i \in [k]} \int_{V_i(w')} \|w_i - x\|^2 \, p(x) \, dx. \tag{5}$$

A.1 Family of upper bounds of the cost

It turns out that $\{g(\,\cdot\,; w') : w' \in \mathbb{R}^{k \times d}\}$ forms a family of convex, quadratic upper bounds of f. That $g(\,\cdot\,; w')$ is convex and quadratic is easy to see, since it is a sum of convex combinations of convex quadratic functions. The following proposition shows that g dominates f.

Proposition A.1. *Let p be a density on $\mathbb{R}^d$ with bounded second moment. Let f be the k-means objective (1) and g be the k-means surrogate (5). Then for all $w, w' \in \mathbb{R}^{k \times d}$,*

$$f(w) \le g(w; w').$$

Proof. Notice that by definition, $f(w) = g(w; w)$. We claim that:

$$f(w) = g(w; w) \le g(w; w'). \tag{6}$$

To see this, note that g is an integral accumulating $\|w_i - x\|^2$ when x is in the ith partition. The integral only decreases if x moves into its Voronoi partition, $j^* = \arg\min \|w_j - x\|^2$,

$$\begin{aligned} g(w; w') &= \frac{1}{2} \int_{\mathbb{R}^{k \times d}} \sum_{i \in [k]} \|w_i - x\|^2 \cdot \mathbb{1}_{V_i(w')}(x) \, p(x) \, dx \\ &\ge \frac{1}{2} \int_{\mathbb{R}^{k \times d}} \min_{i \in [k]} \|w_i - x\|^2 \, p(x) \, dx \\ &= g(w; w), \end{aligned}$$

where $\mathbb{1}_{V_i(w')}(x)$ is the indicator of $V_i(w')$. The inequality holds as the first integrand dominates the second. □

A.2 Gradient of the cost

While taking the derivative of f is nontrivial, taking the derivative of g is much easier; by dominated convergence, the derivative with respect to w is:

$$\begin{aligned} \nabla_{w_i} g(w; w') &= \int_{V_i(w')} (w_i - x) \, p(x) \, dx \\ &= P_i(w') \cdot \big(w_i - M_i(w')\big). \end{aligned} \tag{7}$$

We provide two proofs computing the gradient of f: (i) an elementary proof based on a local approximation $f(w+\varepsilon) = g(w+\varepsilon; w) + \mathrm{error}_w(\varepsilon)$, and (ii) a short proof using results from differential geometry. Our target:

Lemma 2.2 (Gradient of k-means objective)**.** *Let p be a density on $\mathbb{R}^d$ with $\mathbb{E}_{X \sim p}\left[\|X\|^2\right] < \infty$. Let f be the k-means objective (1). Then f is continuously differentiable on $\mathcal{D}$, where:*

$$\nabla_{w_i} f(w) = P_i(w) \cdot \big(w_i - M_i(w)\big). \tag{2}$$

Given the form of the gradient (2), it is straightforward to show continuity:

Proof of continuous gradient. Assuming (2), the following holds:

$$\begin{aligned}
&\|\nabla_{w_i} f(w+\varepsilon) - \nabla_{w_i} f(w)\| \\
&\qquad = \Big\| \int_{V_i(w+\varepsilon)} (w_i + \varepsilon_i - x)\, p(x)\, dx - \int_{V_i(w)} (w_i - x)\, p(x)\, dx \Big\| \\
&\qquad \le \int_{V_i(w+\varepsilon)\cap V_i(w)} \|\varepsilon_i\|\, p(x)\, dx + 2 \int_{V_i(w+\varepsilon)\Delta V_i(w)} \big(\|w_i - x\| + \|\varepsilon_i\|\big)\, p(x)\, dx,
\end{aligned}$$

where Δ is the symmetric difference operator. Taking as $\|\varepsilon\| \to 0$, the limit of both integrals converge to zero by dominated convergence, proving continuity. □

We now show (2) through two different approaches.

A.2.1 An elementary proof

If ε is a small perturbation, we can write $f(w+\varepsilon) = g(w+\varepsilon; w) + \text{error}_w(\varepsilon)$, where the error is accumulated over points near the boundaries of the partitions $V_i(w)$. In the proof, we show that $\text{error}_w(\varepsilon) = o\big(\|\varepsilon\|\big)$, so only the $g(w+\varepsilon; w)$ term contributes to the derivative of $f(w)$.

Proof of Lemma 2.2. Fix $w, h \in \mathbb{R}^{k\times d}$ where $\|h\| = 1$. We proceed by computing the directional derivative $D_h f(w)$ along h. Notice that:

$$f(w) = \frac{1}{2} \sum_{i\in[k]} \int_{V_i} \|w_i - x\|^2\, p(x)\, dx \quad \text{and} \quad f(w+th) = \frac{1}{2} \sum_{i\in[k]} \int_{V_i^t} \|w_i + th_i - x\|^2\, p(x)\, dx,$$

where we let V_i and V_i^t respectively abbreviate $V_i(w)$ and $V_i(w+th)$ for $t > 0$. Notice that:

$$V_i^t \;=\; \big[V_i \;\cup\; (V_i^t \setminus V_i)\big] \;\setminus\; (V_i \setminus V_i^t),$$

where (i) V_i and $(V_i^t \setminus V_i)$ are disjoint and (ii) V_i contains $(V_i \setminus V_i^t)$. It follows that:

$$\begin{aligned}
f(w+th) = \frac{1}{2} &\sum_{i\in[k]} \int_{V_i} \|w_i + th_i - x\|^2\, p(x)\, dx \\
&+ \frac{1}{2} \sum_{i\in[k]} \left(\int_{V_i^t \setminus V_i} \|w_i + th_i - x\|^2\, p(x)\, dx \;-\; \int_{V_i \setminus V_i^t} \|w_i + th_i - x\|^2\, p(x)\, dx \right).
\end{aligned}$$

We claim that this second line is $o(t)$. Assuming this for now, it follows that the derivative is:

$$\begin{aligned}
D_h f(w) &= \lim_{t\to 0} \frac{1}{t} \big(f(w+th) - f(w)\big) \\
&= \lim_{t\to 0} \frac{1}{t} \left(\frac{1}{2} \sum_{i\in[k]} \int_{V_i} \big(\|w_i + th_i - x\|^2 - \|w_i - x\|^2\big)\, p(x)\, dx \right) + \lim_{t\to 0} \frac{o(t)}{t} = D_h g(w; w),
\end{aligned}$$

where the derivative $D_h g(w; w)$ is taken only over the first argument (i.e. the partition V_i is fixed). But this implies that $\nabla_w f(w) = \nabla_w g(w; w')$ when $w' = w$. Applying (7), we obtain:

$$\nabla_{w_i} f(w) = P_i(w) \cdot \big(w_i - M_i(w)\big).$$

To finish the proof, we need to show that the above second line is $o(t)$. We claim it is equal to:

$$\frac{1}{2} \sum_{i\neq j} \int_{V_{j\to i}^t} \big(\, \|w_i + th_i - x\|^2 - \|w_j + th_j - x\|^2 \,\big)\, p(x)\, dx, \tag{8}$$

where the domain of the integral $V_{j\to i}^t = \big(V_i^t \setminus V_i\big) \cap \big(V_j \setminus V_j^t\big)$ is the set of points $x \in \mathbb{R}^d$ that originally began in the jth partition $V_j(w)$ but after the small perturbation, ended up in the ith partition $V_i(w + th)$. Indeed, $V_i^t \setminus V_i$ is the disjoint union $\bigcup_j V_{j\to i}^t$, since every point in V_i^t not originally in V_i must have come from some other partition V_j. Similarly, $V_j \setminus V_j^t = \bigcup_i V_{j\to i}^t$.

Intuitively, if x swaps partitions due to a small perturbation, then $\|w_i - x\|^2 \sim \|w_j - x\|^2$. In fact, we will show that the magnitude of $\|w_i + th_i - x\|^2 - \|w_j + th_j - x\|^2$ is $O(t\|x\|)$ for $x \in V_{j\to i}^t$. First, to simplify notation, let:

$$\alpha = \frac{w_i - w_j}{2} \quad \beta = \frac{w_i + w_j}{2} \quad \gamma = \frac{h_i - h_j}{2} \quad \delta = \frac{h_i + h_j}{2}.$$

By the polarization identity, we have:

$$\begin{aligned} \|w_i - x\|^2 - \|w_j - x\|^2 &= 4\big\langle \alpha, \beta - x \big\rangle \\ \|w_i + th_i - x\|^2 - \|w_j + th_j - x\|^2 &= 4\big\langle \alpha + t\gamma, \beta + t\delta - x \big\rangle. \end{aligned}$$

Furthermore, because $x \in V_{j\to i}^t$, we have the inequalities:

$$\|w_j - x\|^2 \le \|w_i - x\|^2 \qquad \text{and} \qquad \|w_i + th_i - x\|^2 \le \|w_j + th_j - x\|^2.$$

Plugging in the equality from polarization, we obtain:

$$\underbrace{\|w_i + th_i - x\|^2 - \|w_j + th_j - x\|^2}_{\le 0} = \underbrace{\|w_i - x\|^2 - \|w_j - x\|^2}_{\ge 0} + 4t\big(\langle \alpha, \delta \rangle + \langle \gamma, \beta - x \rangle + t\langle \gamma, \delta \rangle\big),$$

which implies that $\big|\|w_i + th_i - x\|^2 - \|w_j + th_j - x\|^2\big| \le Ct(\|x\| + 1)$ when $t < 1$, where C is a constant that may depend on w. It follows that:

$$\left| \int_{V_{j\to i}^t} \big(\|w_i + th_i - x\|^2 - \|w_j + th_j - x\|^2 \big)\, p(x)\, dx \right| \le \int_{V_{j\to i}^t} Ct(\|x\| + 1)\, p(x)\, dx.$$

As $\mathop{\mathbb{E}}_{p}\big[\|X\|^2\big] < \infty$, we know p also has bounded first moment. By dominated convergence:

$$\lim_{t\to 0} \frac{1}{t} \int_{V_{j\to i}^t} Ct(\|x\| + 1)\, p(x)\, dx = 0,$$

since $V_{j\to i}^t$ decreases to some measure zero subset of $V_i \cap V_j$. And so, (8) is $o(t)$. □

A.2.2 A short proof using Leibniz integral rule

Suppose we are given an integral where both its domain and integrand are time-varying. Then the derivative of that integral is given by Leibniz rule:

$$\frac{d}{dt} \int_{a(t)}^{b(t)} F(x,t)\, dx = \int_{a(t)}^{b(t)} \frac{\partial F(x,t)}{\partial t}\, dx + \Big(b'(t) F\big(b(t), t\big) - a'(t) F\big(a(t), t\big) \Big).$$

In particular, we break down the time derivative into two pieces: (i) the accumulated time derivative at each point in the domain, and (ii) the weighted velocity at the boundary at which the domain is expanding or contracting. In higher dimensions, the time derivative of a volume integral can be decomposed into the same two pieces.

But generalizing Leibniz rule to higher dimensions, we need some notation. Let $\Omega(t) \subset \mathbb{R}^n$ be a smoothly time-varying differentiable n-manifold for $t \in (-\tau, \tau)$ with boundary $S(t) = \partial\Omega(t)$. That is, there is a domain $U \subset \mathbb{R}^n$ and a continuously differentiable map $\phi : (-\tau, \tau) \times U \to \mathbb{R}^n$ where ϕ_t is a diffeomorphism of U onto $\Omega(t)$.

We write $x = x(t, u) = \phi(t, u)$ and $v = \partial x / \partial t$. For points $x \in S(t)$ on the boundary, denote the surface normal by $N = N(x)$. The *surface velocity* at $x \in S(t)$ is defined as $C(x) = N^\top v$, which is an invariant in the sense that it is coordinate-independent (Grinfeld, 2013).

Theorem A.2 (General Leibniz rule, Grinfeld (2013))**.** *Let $\Omega(t) \subset \mathbb{R}^n$ be a smoothly time-varying smooth n-manifold with boundary $S(t)$ over times $t \in (-\tau, \tau)$. Let $C(t, x)$ be the surface velocity of the point $x \in S(t)$ at time t. If $F : (-\tau, \tau) \times \mathbb{R}^n \to \mathbb{R}^m$ is smooth, then for $t \in (-\tau, \tau)$:*

$$\frac{d}{dt} \int_{\Omega(t)} F \, d\Omega = \int_{\Omega(t)} \frac{\partial F}{\partial t} \, d\Omega + \int_{S(t)} F \, C dS.$$

As a result, if we consider the directional derivative of our objective f in the direction of h,

$$\frac{d}{dt} f(w + th) = \frac{1}{2} \sum_{i \in [k]} \frac{d}{dt} \int_{V_i(w+th)} \|w_i + th_i - x\|^2 p(x) \, dx, \tag{9}$$

each of the integrals will split into two: (i) the integrals computing the accumulated rate of change, and (ii) those computing weighted surface velocities at the boundaries of the Voronoi partition. The first exactly coincides with $\frac{d}{dt} g(w+th; w)$. The second terms vanish since the weighted surface velocities at the boundaries of two partitions exactly cancel each other out. Formally, we have:

Proof of Lemma 2.2. Fix $w, h \in \mathbb{R}^{k \times d}$ where h is unit. The directional derivative $D_h f(w)$ is given by (9) evaluated at time $t = 0$. Applying the general Leibniz rule yields:

$$D_h f(w) = \sum_{i \in [k]} h_i^\top \int_{V_i(w)} (w_i - x) p(x) \, dx + \frac{1}{2} \sum_{i \in [k]} \int_{\partial V_i(w)} \|w_i - x\|^2 p(x) \, C_i dS, \tag{10}$$

where $C_i(x)$ is the surface velocity of a point $x \in \partial V_i(w)$ at time 0. Notice that if $x \in \partial V_i(w)$, then it is also contained in exactly one other boundary, $x \in \partial V_j(w)$. On the one hand, the weight of the integrands are equal $\|w_i - x\|^2 p(x) = \|w_j - x\|^2 p(x)$. But on the other, the surface velocities of x of $\partial V_i \cap \partial V_j$ are equal and opposite, $C_i(x) = -C_j(x)$, since it is an invariant. Therefore,

$$\int_{\partial_i V(w) \cap \partial_j V(w)} \left(\|w_i - x\|^2 p(x) \, C_i \right) + \left(\|w_j - x\|^2 p(x) \, C_j \right) dS = 0.$$

Thus, the second set of integrals in (10) vanishes. By the chain rule, $D_h f(w) = h^\top \nabla f(w)$, so:

$$\nabla_{w_i} f(w) = \int_{V_i(w)} (w_i - x) p(x) \, dx. \qquad \square$$

A.3 An analytic upper bound of the cost

Because $g(\,\cdot\,; w')$ is quadratic, we can describe the upper bound analytically using our computation of ∇f.

Lemma A.3 (Quadratic upper bound)**.** *Let p a density on $\mathbb{R}^d$ have bounded second moment. If f is the k-means objective (1), then for all $w, w^+ \in \mathbb{R}^{k \times d}$*

$$f(w^+) \leq f(w) + \left\langle \nabla f(w), w^+ - w \right\rangle + \frac{1}{2} (w^+ - w)^\top \mathbf{H} \, (w^+ - w), \tag{11}$$

where we let $\mathbf{H} = \mathbf{H}_w \, g(w; w)$ be the Hessian of $g(\,\cdot\,; w)$. In particular,

$$f(w^+) \leq f(w) + \left\langle \nabla f(w), w^+ - w \right\rangle + \frac{1}{2} \|w^+ - w\|^2. \tag{12}$$

Proof. Let $w, w^+ \in \mathbb{R}^{k \times d}$. Recall that $f(w^+)$ is upper bounded by $g(w^+; w)$ by Proposition A.1. Because $g(\,\cdot\,; w)$ is quadratic, it is equal to its second-order Taylor expansion. Then, we have:

$$f(w^+) \leq g(w^+; w) = g(w; w) + \left\langle \nabla_w g(w; w), w^+ - w \right\rangle + \frac{1}{2} (w^+ - w)^\top \mathbf{H} \, (w^+ - w).$$

The first assertion (11) follows because $f(w) = g(w; w)$ and $\nabla f(w) = \nabla_w \, g(w; w)$.

Notice that the Hessian $\mathbf{H}$ is constant since g is quadratic:

$$\mathbf{H}_w \, g(w^+; w)_{\,ij} = \begin{cases} P_i(w)\, \mathbf{I}_{d\times d} & i = j \\ 0 & i \neq j, \end{cases} \tag{13}$$

where $\mathbf{I}_{d\times d}$ is the d-dimensional identity matrix. Because $P(w)$ is a probability vector, the spectral norm is bounded, $\|\mathbf{H}_{w^+} \, g(w^+; w)\|_* = \max_{i\in[k]} P_i(w) \leq 1$. From this, (12) immediately follows. □

B Convergence of cost

Lemma B.1. *Let f be the k-means cost and $(W^{(n)}, H^{(n)}, X^{(n)})_{t=1}^{\infty}$ as in the online k-means algorithm. Then:*

$$f(W^{(n+1)}) \leq f(W^{(n)}) - A_{n+1} + N_{n+1}, \tag{14}$$

where A_{n+1} is the exact gradient descent term and $N_{n+1} = -B_{n+1} + C_{n+1}$ is the noise term:

- $A_{n+1} = \sum_{i\in[k]} H_i^{(n+1)} P_i^{-1}(W^{(n)}) \|\nabla_{w_i} f(W^{(n)})\|^2$
- $B_{n+1} = \sum_{i\in[k]} H_i^{(n+1)} \nabla_{w_i} f(W^{(n)})^\top \big(M_i(W^{(n)}) - X_i^{(n+1)}\big)$
- $C_{n+1} = \frac{1}{2} \sum_{i\in[k]} \big(H_i^{(n+1)}\big)^2 \|W_i^{(n)} - X_i^{(n+1)}\|^2.$

In particular, $(A_n)_{n=1}^{\infty}$ and $(C_n)_{n=1}^{\infty}$ are nonnegative sequences; and, since $H^{(n+1)}$ and $X^{(n+1)}$ are conditionally independent given $\mathcal{F}_n$, $(B_n)_{n=1}^{\infty}$ is a martingale difference sequence.

Proof. We simply rewrite the update in the following form:

$$\begin{aligned} W_i^{(n+1)} &= W_i^{(n)} - H_i^{(n+1)} \Big(W_i^{(n)} - M_i(W^{(n)}) + M_i(W^{(n)}) - X_i^{(n+1)} \Big) \\ &= W_i^{(n)} - H_i^{(n+1)} P_i(W^{(n)})^{-1} \nabla_{w_i} f(W^{(n)}) - H_i^{(n+1)} \Big(M_i(W^{(n)}) - X_i^{(n+1)} \Big), \end{aligned} \tag{15}$$

where the second line follows from Lemma 2.2 showing $\nabla_{w_i} f(w) = P_i(w)\big(w_i - M_i(w)\big)$. Recall the quadratic upper bound in Lemma A.3, reproduced here:

$$f(W^{(n+1)}) \leq f(W^{(n)}) + \langle \nabla f(W^{(n)}), W^{(n+1)} - W^{(n)} \rangle + \frac{1}{2} \|W^{(n+1)} - W^{(n)}\|^2 \tag{12}$$

Combining (12) and (15) immediately yields (14). □

Lemma B.2. *Let $(N_n)_{n=1}^{\infty}$ as in Lemma B.1. Suppose that:*

$$\sum_{n=1}^{\infty} \sum_{i\in[k]} \big(H_i^{(n)}\big)^2 < \infty \quad a.s.$$

Then the series $\sum_{n=1}^{\infty} N_n = -\sum_{n=1}^{\infty} B_n + \sum_{n=1}^{\infty} C_n < \infty$ converges almost surely.

Proof. Assumption 3.1 and 3.2 imply that the all iterates $W_i^{(n)}$ and updates $X_i{}^{(n+1)}$ remain in the closed ball $B(0, R)$. Recall from Lemma 2.2 that $\nabla_{w_i} f(w) = P_i(w) \cdot \big(w - M_i(w)\big)$. Thus:

$$\big|\nabla_{w_i} f(W^{(n)})^\top \big(M_i(W^{(n)}) - X_i^{(n+1)}\big)\big| < 4R^2 \quad \text{and} \quad \|W_i^{(n)} - X_i^{(n+1)}\|^2 < 4R^2.$$

The series $\sum B_n$ converges almost surely by martingale convergence, Theorem B.3, which we may apply as:

$$\sum_{n=1}^{\infty} \mathbb{E}[B_n^2] \leq 16R^4 \cdot \sum_{n=1}^{\infty} \sum_{i\in[k]} \big(H_i^{(n)}\big)^2 < \infty.$$

The series $\sum C_n$ converges almost surely since it is dominated by a convergent series:

$$\sum_{n=1}^{\infty} C_n \leq 2R^2 \cdot \sum_{n=1}^{\infty} \sum_{i\in[k]} \big(H_i^{(n)}\big)^2 < \infty.$$

□

First we show that the cost $f(W^{(n)})$ converges. Notice that if the noise term N_{n+1} in Lemma B.1 did not contain the nonnegative C_{n+1} term, then $f(W^{(n)})$ is seen to be a supermartingale bounded below since $f \geq 0$. Then, the convergence of $f(W^{(n)})$ would immediately follow from the martingale convergence theorem. But in reality, $f(W^{(n)})$ is just almost a supermartingale. Still, we can obtain convergence since $\sum C_n < \infty$ converges almost surely, from Lemma B.2. This next lemma proves this formally.

Proposition 3.3 (Convergence of cost)**.** *Let $\left(W^{(n)}\right)_{n=0}^{\infty}$ be a sequence generated by the online k-means algorithm. If the following series converges:*

$$\sum_{n=1}^{\infty} \sum_{i \in [k]} \left(H_i^{(n)}\right)^2 < \infty \quad \text{a.s.},$$

then there is an $\mathbb{R}$-valued random variable f^* such that $f(W^{(n)})$ converges to f^* almost surely.

Proof. Let $(M_n)_{n=1}^{\infty}$ be defined by:

$$M_{n+1} = f(W^{(n+1)}) - \sum_{\tau=0}^{n} C_{\tau+1}.$$

Lemma B.1 shows that M_n is an $\mathcal{F}_n$-supermartingale:

$$\begin{aligned}
\mathbb{E}\left[M_{n+1} \,\middle|\, \mathcal{F}_n\right] &\leq f(W^{(n)}) - \mathbb{E}\left[A_{n+1} + B_{n+1} - C_{n+1} \,\middle|\, \mathcal{F}_n\right] - \mathbb{E}\left[\sum_{\tau=0}^{n-1} C_{\tau+1} + C_{n+1} \,\middle|\, \mathcal{F}_n\right] \\
&= f(W^{(n)}) - \sum_{\tau=0}^{n-1} C_{\tau+1} - \mathbb{E}[A_{n+1} \,|\, \mathcal{F}_n] \leq M_n,
\end{aligned}$$

where we used the fact that $(A_n)_{n=1}^{\infty}$ is nonnegative and $(B_n)_{t=0}^{\infty}$ is an $\mathcal{F}_n$-martingale. Furthermore, $(C_n)_{n=1}^{\infty}$ is nonnegative and Lemma B.2 shows that $\sum C_n < \infty$ converges, the supermartingale is bounded below:

$$-\infty < -\sum_{t=1}^{\infty} C_n \leq M_n.$$

By the martingale convergence theorem, both $(M_n)_{n=1}^{\infty}$ and $f(W^{(n)})$ converge almost surely. □

Theorem B.3 (Martingale convergence theorem, Durrett (2019))**.** *Let $(M_n)_{n\in\mathbb{N}}$ be a (sub)martingale with:*

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left[M_n^+\right] < \infty,$$

where $M_n^+ := \max\{0, M_n\}$. Then as $n \to \infty$, M_n converges a.s. to a limit M with $\mathbb{E}[|M|] < \infty$.

Remark B.4 (Specific forms of martingale convergence)**.** We use two specific forms of Theorem B.3 in our proofs. The first applies to martingale difference sequences $(B_n)_{n\in\mathbb{N}}$. Let $M_n = \sum_{m=1}^{n} B_n$. Since the terms in a martingale difference sequence are orthogonal, we have for all $n \in \mathbb{N}$:

$$\mathbb{E}\left[M_n^+\right]^2 \leq \mathbb{E}\left[\left(\sum_{m=1}^{n} B_m\right)^2\right] = \sum_{m=1}^{n} \mathbb{E}\left[B_m^2\right].$$

It follows that the condition $\sum_{n=1}^{\infty} \mathbb{E}[B_n^2] < \infty$ implies $\sup_{n\in\mathbb{N}} \mathbb{E}\left[M_n^+\right] < \infty$.

The other applies to lower bounded supermartingales $(M_n)_{n\in\mathbb{N}}$. As $(-M_n)_{n\in\mathbb{N}}$ is then an upper bounded submartingale, it converges to some $-M$. We may apply martingale convergence: let $c \in \mathbb{R}$ be a lower bound such that $M_n > c$ almost surely. Then $-M_n$ is a submartingale with $\sup_{n\in\mathbb{N}} \mathbb{E}\left[-M_n^+\right] \leq \max\{0, -c\} < \infty$.

C Convergence of iterates

Lemma 3.1. *Let $0 \le H_i^{(n+1)} \le 1$. Then $W^{(n)} \in \mathcal{D}$ is non-degenerate for all $n \in \mathbb{N}$ almost surely.*

Proof. By assumption, $W^{(0)} \in \mathcal{D}$. It suffices to show by induction that $W^{(n+1)}$ is non-degenerate almost surely if $W^{(n)}$ is non-degenerate. Notice that ${W_i}^{(n+1)} \in V_i(W^{(n)})$ since it is a convex combination of points in the Voronoi region $V_i(W^{(n)})$. Therefore, the only way for two initially distinct centers ${W_i}^{(n+1)}$ and ${W_j}^{(n+1)}$ to possibly meet is if the updates ${X_i}^{(n+1)}$ and ${X_j}^{(n+1)}$ are come from the boundary of their Voronoi regions. But because the boundary is a measure zero set, this occurs almost never. □

Lemma 3.2. *Suppose $\Pr_{X\sim p}(\|X\| > R) = 0$. Let $H_i^{(n)} \le 1$ for all $n \in \mathbb{N}$ and $i \in [k]$. If $n > m$, then:*

$$\|W^{(m)} - W^{(n)}\| \le 2R \cdot \sum_{i\in[k]} \sum_{m\le n'<n} H_i^{(n'+1)} \quad \text{a.s.}$$

Proof. We claim that $W_i^{(n)} \in B(0,R)$ for all $n \in \mathbb{N}_0$. This is shown true by induction. Assumption 3.1 states that p is supported only in $B(0,R)$. Since $W_i^{(0)}$ comes from the support of p, this claim holds for $n = 0$. If $W_i^{(n)} \in B(0,R)$, then $W_i^{(n+1)}$ is a convex combination of points in $B(0,R)$ almost surely:

$$W_i^{(n+1)} = \left(1 - H_i^{(n+1)}\right) \cdot W_i^{(n)} + H_i^{(n+1)} \cdot X_i^{(n+1)},$$

since $H_i^{(n+1)} \in [0,1]$ and $X_i^{(n+1)} \sim p$.

As a result of this, we can upper bound the displacement:

$$\begin{aligned}
\|W^{(m)} - W^{(n)}\| &\overset{(i)}{\le} \sum_{j\in[k]} \left\|W_j^{(m)} - W_j^{(n)}\right\| \\
&\overset{(ii)}{\le} \sum_{j\in[k]} \sum_{m\le n'<n} \left\|H_j^{(n'+1)} \cdot \left(W_i^{(n')} - X_i^{(n'+1)}\right)\right\| \\
&\overset{(iii)}{\le} 2R \cdot \sum_{j\in[k]} \sum_{m\le n'<n} H_j^{(n'+1)}, \qquad (16)
\end{aligned}$$

where (i) follows from Minkowski's inequality, (ii) follows from triangle inequality, and (iii) follows from our initial claim since $W_i^{(n')}, X_i^{(n'+1)} \in B(0,R)$ almost surely, so that $\|W_i^{(n')} - X_i^{(n'+1)}\| < 2R$. □

Theorem C.1 (Convergence of iterates, generalized)**.** *Let $(W^{(n)})_{n=0}^{\infty}$ and $H^{(n+1)}$ be as in Proposition 3.3. Suppose there exists $\varepsilon_0 > 0$ such that $\{\|\nabla f\| \le \varepsilon_0\}$ is compact, and for all $i \in [k]$,*

(A1) *For any $\varepsilon \in (0, \varepsilon_0)$, there is an $r_0 \equiv r_0(\varepsilon) > 0$ so that if $r \in (0, r_0)$, then there exist $T : \mathbb{N} \to \mathbb{N}$, $m_0 \in \mathbb{N}$, and $s, c > 0$, which may all depend on ε and r, so that:*

$$\Pr\left(\sum_{j\in[k]} \sum_{m\le n<T(m)} H_j^{(n+1)} < r \quad \text{and} \quad \sum_{m\le n<T(m)} H_i^{(n+1)} > s \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \in [\varepsilon, \varepsilon_0)\right) > c,$$

for any $m > m_0$, and,

(A2) *If $\liminf_{n\to\infty} \|\nabla_{w_i} f(W^{(n)})\| \ge \varepsilon_0$, then $\sum_{n\in\mathbb{N}} H_i^{(n)} = \infty$ almost surely.*

Then, $W^{(n)}$ globally converges to stationary points of f almost surely, where f is the k-means cost (1).

Proof. We claim that for all $\varepsilon > 0$, the iterates $W^{(n)}$ eventually never return to the set $\{\|\nabla f\| > \varepsilon\}$ almost surely. If so, then the sequence of gradients converges to zero $\|\nabla f(W^{(n)})\| \to 0$ almost surely, since the claim holds simultaneously for any countable sequence of $\varepsilon_j \downarrow 0$. Because the map $w \mapsto \|\nabla f(w)\|$ is continuous and the iterates eventually remain in a compact region $\{\|\nabla f\| \le \varepsilon_0\}$, the convergence of gradients $\|\nabla f(W^{(n)})\|$ to zero implies the almost sure convergence of the iterates $W^{(n)}$ to the set of stationary points (Lemma C.4):

$$\limsup_{n\to\infty} \inf_{\{w:\nabla f(w)=0\}} \|W^{(n)} - w\| = 0 \quad \text{a.s.}$$

The claim remains to be proven: that the iterates $W^{(n)}$ eventually never return to the set $\{\|\nabla f\| > \varepsilon\}$ for all $\varepsilon > 0$. Note that $\|\nabla f(w)\|$ is upper bounded by $\sum_{i\in[k]} \|\nabla_{w_i} f(w)\|$, so it suffices to consider each center individually and show that the iterates eventually never return to the set $\{\|\nabla_{w_i} f\| > \frac{\varepsilon}{k}\}$. And so, we show that for all $i \in [k]$ and $\varepsilon > 0$, if $\|\nabla_{w_i} f(W^{(n)})\| > \varepsilon$ infinitely often, then $f(W^{(n)})$ does not converge. As this would contradict Proposition 3.3, we indeed have $\|\nabla_{w_i} f(W^{(n)})\| > \varepsilon$ finitely often almost surely.

We first consider the case $\varepsilon < \varepsilon_0$ and show that the iterates eventually never return to the set $\{\|\nabla_{w_i} f\| \in [\varepsilon, \varepsilon_0)\}$. Fix $i \in [k]$ and any $\varepsilon \in (0, \varepsilon_0)$. Note that $\{\|\nabla_{w_i} f\| \le \frac{\varepsilon}{2}\}$ and $\{\|\nabla_{w_i} f\| \in [\varepsilon, \varepsilon_0]\}$ are disjoint compact sets; because $w \mapsto \|\nabla f(w)\|$ is continuous, they are closed subsets of the compact set $\{\|\nabla f\| \le \varepsilon_0\}$. And so, these two sets are bounded away from each other by some distance $R_0 > 0$. Without loss of generality, we may assume that r_0 given in (A1) satisfies $r_0 \le R_0/2R$. Fix any $r \in (0, r_0)$.

We may now apply condition (A1) to control the behavior of the iterates for a non-negligible number of iterations upon entering the set $\{\|\nabla_{w_i} f\| \in [\varepsilon, \varepsilon_0)\}$. In particular, given (ε, r), let (T, m_0, s, c) be chosen so that (A1) holds. Suppose at time $m > m_0$, the iterate $W^{(m)}$ enters this set. For parsimony, call the two events within the first probability in the theorem statement Ξ_1 and Ξ_2,

$$\Xi_1 = \left\{ \sum_{j\in[k]} \sum_{m\le n<T(m)} H_j^{(n+1)} < r \right\} \qquad \text{and} \qquad \Xi_2 = \left\{ \sum_{m\le n<T(m)} H_i^{(n+1)} > s \right\}.$$

We show that given m is sufficiently large, if both of these events hold, then the iterates will remain in the set $\{\|\nabla_{w_i} f\| > \frac{\varepsilon}{2}\}$ for a sufficient amount of time to decrease $f(W^{(n)})$ by a constant amount. This will allow us to apply Borel-Cantelli to show that $f(W^{(n)})$ does not converge. For the first claim, recall that Lemma 3.2 bounds the distance iterates travel away from $W^{(m)}$ via the summed learning rates:

$$\sum_{j\in[k]} \sum_{m\le n<T(m)} H_j^{(n+1)} < r \qquad \Longrightarrow \qquad \sup_{m\le n<T(m)} \|W^{(m)} - W^{(n)}\| < 2R \cdot r < R_0.$$

Consequently, Ξ_1 implies that $\|\nabla_{w_i} f(W^{(n)})\| > \frac{\varepsilon}{2}$ on the interval $m \le n < T(m)$. The second event Ξ_2 can be used to show that $f(W^{(n)})$ decreases a constant amount on this interval. Lemma B.1 shows for all $n \in \mathbb{N}$,

$$f\big(W^{(n)}\big) \le f\big(W^{(m)}\big) - \sum_{m\le n'<n} A_{n'+1} + \sum_{m\le n'<n} N_{n'+1}.$$

Lemma B.2 shows that $\sum_{n=0}^{\infty} N_{n+1}$ converges almost surely, so that $\sum_{m\le n'<n} N_{n'+1} < \delta$ when m is sufficiently large. More precisely, for any $\delta > 0$, there almost surely exists an $\bar{\mathbb{N}}$-random variable M_δ so that:

$$\left| \sum_{n'\ge M_\delta} N_{n'+1} \right| < \delta/2.$$

In particular, $\sum_{m\le n'<n} N_{n'+1} < \delta$ holds for all $m > M_\delta$. Let $m' = T(m) - 1$. Then:

$$\begin{aligned}
f\big(W^{(m')}\big) &\overset{(i)}{\le} f\big(W^{(m)}\big) - \sum_{m\le n<m'} \sum_{j\in[k]} H_j^{(n+1)} P_j^{-1}(W^{(n)}) \|\nabla_{w_j} f(W^{(n)})\|^2 + \delta \\
&\overset{(ii)}{\le} f(W^{(m)}) - \frac{\varepsilon^2}{4} \sum_{m\le n<m'} H_i^{(n+1)} + \delta \\
&\overset{(iii)}{\le} f(W^{(m)}) - \frac{s\varepsilon^2}{4} + \delta \\
&\overset{(iv)}{<} f(W^{(m)}) - \delta,
\end{aligned}$$

where (i) substitutes in the expression for A_{n+1}, (ii) drops the summation over other centers except for i, and $\|\nabla_{w_i} f(W^{(n)})\| > \frac{\varepsilon}{2}$ holds if the first event occurs, (iii) follows if Ξ_2 occurs, and (iv) sets $\delta < s\varepsilon^2/8$.

We have thus shown that if condition (A1) holds, $m > \max\{m_0, M_\delta\}$, and $\delta < s\varepsilon^2/8$, then:

$$\Pr\left(f(W^{(n)}) < f(W^{(m)}) - \delta \text{ for some } m \le n < T(m) \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \in [\varepsilon, \varepsilon_0)\right) > c. \tag{17}$$

That is, if $\|\nabla_{w_i} f(W^{(m)})\|$ is large at iteration m, then with positive probability, within a bounded amount of time, $f(W^{(n)})$ will decrease by a constant amount δ. But we also know that $f(W^{(n)})$ converges almost surely to some f^*, by Proposition 3.3, and so decreases by δ only a finite number of times. We claim that by Borel-Cantelli, Lemma C.5, the event $\|\nabla_{w_i} f(W^{(n)})\| \in [\varepsilon, \varepsilon_0)$ must also occur only finitely often.

Assume for the sake of contradiction that $\|\nabla_{w_i} f(W^{(n)})\| \in [\varepsilon, \varepsilon_0)$ infinitely often. Then we can define the infinite sequence of stopping times:

$$\tau_0 = 0 \qquad \text{and} \qquad \tau_{j+1} = \inf\{n > \tau_j + T(\tau_j) : \|\nabla_{w_i} f(W^{(n)})\| \in [\varepsilon, \varepsilon_0)\}.$$

Then (17) states that when $\tau_j > \max\{m_0, M\}$,

$$\Pr\left(f(W^{(n)}) < f(W^{(\tau_j)}) - \delta \text{ for some } \tau_j \le n < T(\tau_j) \,\middle|\, \mathcal{F}_{\tau_j}\right) > c',$$

where the event in the probability is $\mathcal{F}_{\tau_{j+1}}$-measurable. Borel-Cantelli, Lemma C.5, then implies that $f(W^{(n)})$ decreases by a constant amount δ infinitely often, which contradicts the convergence of $f(W^{(n)})$.

To finish the proof, we need to show that the iterates eventually never return to the set $\{\|\nabla_{w_i} f\| \ge \varepsilon_0\}$. We do this by ruling out (i) after some iteration m, the iterates never leave this set, and (ii) the iterates exit and re-enter this set infinitely often. The first case is impossible, for then condition (A2) implies that $\sum_{n\in\mathbb{N}} H_i^{(n)} = \infty$ almost surely. By Lemma B.1, this leads to an unbounded decrease in cost,

$$\liminf_{N\to\infty} f(W^{(N)}) \le \lim_{N\to\infty} \left(f(W^{(m)}) - \varepsilon_0^2 \sum_{m\le n<N} H_i^{(n+1)} + \delta\right) = -\infty.$$

The second case is also impossible; when the learning rates become sufficiently small, each time the iterates leave $\{\|\nabla_{w_i} f\| \ge \varepsilon_0\}$, they must enter $\{\|\nabla_{w_i} f\| \in [\varepsilon, \varepsilon_0)\}$. Thus, the iterates eventually never return to $\{\|\nabla_{w_i} f\| \ge \varepsilon_0\}$. This shows that for all $\varepsilon > 0$, we almost surely have $\|\nabla_{w_i} f(W^{(n)})\| > \varepsilon$ finitely often. □

We now prove Theorem 4.2, the simplified version of Theorem C.1 seen in the main body of the paper (reproduced below the next lemma). While simpler, it imposes a stronger condition on the learning rate:

Lemma C.2. *Let $i \in [k]$ and $\varepsilon > 0$. Suppose there exists $T : \mathbb{N} \to \mathbb{N}$, $m_0 \in \mathbb{N}$, and $s, c > 0$,*

$$\Pr\left(\sum_{m\le n<T(m)} H_i^{(n+1)} > s \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| > \varepsilon\right) > c,$$

for all $m > m_0$. Then $\liminf_{n\to\infty} \|\nabla_{w_i} f(W^{(n)})\| \ge \varepsilon$ implies $\sum_{n\in\mathbb{N}} H_i^{(n)} = \infty$ almost surely.

Proof. Suppose that there is an $\mathbb{N}$-random variable M such that if $m > M$, then $\|\nabla_{w_i} f(W^{(m)})\| \ge \varepsilon$. That is, the limit infimum condition holds. By assumption, we have:

$$\Pr\left(\sum_{m\le n<T(m)} H_i^{(n+1)} > s \,\middle|\, \mathcal{F}_m, m > \max\{m_0, M\}\right) > c.$$

The Borel-Cantelli lemma (Lemma C.5) shows that there are infinitely many (non-overlapping) intervals $m_j \le n < T_r(m_j)$ on which the sum of $H_i^{(n+1)}$ is at least s, and so the total sum is infinite almost surely. □

Lemma 4.1. *Let $\{\nabla f = 0\}$ be compact in $\mathcal{D}_R$. There exists $\varepsilon_0 > 0$ so that if $\varepsilon \in [0, \varepsilon_0]$, the sets $\{\|\nabla f\| \le \varepsilon\}$ and $\{\|\nabla_{w_i} f\| \le \varepsilon\}$ are compact in $\mathcal{D}_R$ for $i \in [k]$.*

Proof. Because the inclusion map $\iota : \mathcal{D}_R \to \mathbb{R}^{k\times d}$ is continuous, if $\{\nabla f = 0\}$ is compact in $\mathcal{D}_R$, then it is compact in $\mathbb{R}^{k\times d}$. On the other hand, the set of degenerate points $Z := \mathbb{R}^{k\times d} \setminus \mathcal{D}$ is closed in $\mathbb{R}^{k\times d}$, for it is the union of closed sets A_{ij} for $i \neq j$ defined by:

$$A_{ij} := \{\|w_i - w_j\| = 0\}.$$

Recall that if a closed set and a compact set in a metric space are disjoint, then they are separated by some positive distance $\alpha > 0$. So, as $\{\nabla f = 0\}$ and Z are disjoint, no limit point of $\{\nabla f = 0\}$ is degenerate.

And, because ∇f is continuous on $\mathcal{D}$, this implies that that there exists $\varepsilon_0 > 0$ such that $\{\|\nabla f\| \le \varepsilon_0\}$ is compact. In particular, the $\alpha/2$-expansion of $\{\nabla f = 0\}$ is compact in $\mathcal{D}$, where the $\alpha/2$-expansion is the set of points a distance less than or equal to $\alpha/2$ from a stationary point. Additionally, its boundary is compact and separated from $\{\nabla f = 0\}$, so $w \mapsto \|\nabla f(w)\|$ attains a minimum $2\varepsilon_0 > 0$ on it. It follows by continuity that $\{\|\nabla f\| \le \varepsilon\}$ for any $\varepsilon \in [0, \varepsilon_0]$ is a closed set contained in the $\alpha/2$-expansion, hence compact. □

The following lemma is used later in Appendix D, using the same argument to show compactness:

Lemma C.3. *Let $\varepsilon_0 > 0$ be given so that the set $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ is compact. Fix $0 \le \varepsilon \le \varepsilon' \le \varepsilon_0$. Then the level set $K := \{\|\nabla_{w_i} f\| \in [\varepsilon, \varepsilon']\}$ is a nonempty compact set.*

Proof. By Lemma 2.2, the map $\phi : w \mapsto \|\nabla_{w_i} f(w)\|$ is continuous. Since $K = \phi^{-1}([\varepsilon, \varepsilon'])$ is the inverse of a closed set, it is a closed subset of $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$, hence compact. Furthermore, K is nonempty; if this were not the case, then we claim that $\{\|\nabla_{w_i} f\| \le \varepsilon\} = \mathcal{D}_R$. But this cannot be as $\mathcal{D}_R$ is not compact.

As for the claim, note that the map $w \mapsto \|\nabla_{w_i} f(w)\|$ is continuous and that the set $\{\|\nabla_{w_i} f\| = 0\}$ is nonempty. So if there were some point $w \in \mathcal{D}_R$ with $\|\nabla_{w_i} f(w)\| > \varepsilon$, then the intermediate value theorem implies that there is some other point w' with $\|\nabla_{w_i} f(w')\| = \varepsilon$, which violates our assumption. □

Theorem 4.2 (Convergence of iterates)**.** Let $W^{(n)}$ and $H^{(n+1)}$ be as in Proposition 3.3. Suppose that for all $i \in [k]$, $\varepsilon > 0$, and sufficiently small $r > 0$, there exists $T : \mathbb{N} \to \mathbb{N}$, $m_0 \in \mathbb{N}$, and some $s, c > 0$ so that:

$$\Pr\left(\sum_{j\in[k]} \sum_{m\le n<T(m)} H_j^{(n+1)} < r \quad \text{and} \sum_{m\le n<T(m)} H_i^{(n+1)} > s \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| > \varepsilon\right) > c,$$

for any $m > m_0$. Then, $W^{(n)}$ globally converges to stationary points of the k-means cost f almost surely.

Proof. By Lemma 4.1, there exists ε_0 such that $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ is compact for all $i \in [k]$. The conclusion follows from verifying the conditions of Theorem C.1. Condition (A1) is assumed. Condition (A2) follows from Lemma C.2, in which we set the ε parameter to ε_0. □

Lemma C.4. *Let (K, d) be a compact metric space and $h : K \to \mathbb{R}_{\ge 0}$ continuous. Define its zero set as $Z = \{x \in K : h(x) = 0\}$. For all $\varepsilon > 0$, there exists $\delta > 0$ such that $h(x) < \delta$ implies $d(x, Z) < \varepsilon$.*

Proof by contradiction. Suppose there exists some sequence x_n that remains bounded away from Z, so that $d(x_n, Z) \ge \varepsilon$, but $h(x_n)$ converges to zero. Then, by compactness, there is a convergent subsequence $x_{n_k} \to x$. By continuity, $h(x) = 0$, so that $x \in Z$. This is a contradiction; all the x_n are ε-bounded away from Z. □

Lemma C.5 (Second Borel-Cantelli lemma, Durrett (2019))**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\mathcal{F}_n$, $n \ge 0$ be a filtration with $\mathcal{F}_0 = \{\varnothing, \Omega\}$ and let B_n, $n \ge 1$ a sequence of events with $B_n \in \mathcal{F}_n$. Then:*

$$\{B_n \text{ occurs infinitely often}\} = \left\{\sum_{n=1}^{\infty} P(B_n \,|\, \mathcal{F}_{n-1}) = \infty\right\}.$$

D Analysis of the online Lloyd's algorithm

In this section, we prove the convergence for the online Lloyd's learning rate reproduced here:

$$H_j^{(n+1)} = \frac{\mathbb{1}\{I^{(n+1)} = i\}}{\max\{n\widehat{P}_j^{(n)}, t_n\}} \qquad \text{and} \qquad \widehat{P}_j^{(n)} = \frac{1}{s_n} \sum_{n_\circ \le n' < n} \mathbb{1}\{I^{(n'+1)} = j\}. \tag{4}$$

where we denote $n_\circ = n - s_n$, and where s_n and t_n are non-decreasing sequences. Denote $P_j^{(n)} := P_j(W^{(n)})$.

Theorem 4.3 (Convergence of iterates, online Lloyd's)**.** *Let $H_j^{(n)}$ and $\widehat{P}_j^{(n)}$ as in (4). Let s_n and t_n be non-decreasing sequences satisfying:*

$$\lim_{n\to\infty} \frac{n^{2/3}\log n}{s_n} = \lim_{n\to\infty} \frac{s_n \log s_n}{t_n} = \lim_{n\to\infty} \frac{t_n}{n} = 0.$$

If p is continuous, then $W^{(n)}$ globally converges to stationary points of the k-means cost f almost surely.

Remark D.1 (Existence of s_n and t_n)**.** Note that it is fairly easy to construct sequences s_n and t_n satisfying the condition of Theorem 4.3. In particular, let $s_n = n^\alpha$ and $t_n = n^\beta$, where $2/3 < \alpha < \beta < 1$. For example $s_n = n^{3/4}$ and $t_n = n^{5/6}$.

We show convergence by verifying conditions (A1) and (A2) of Theorem C.1. The bulk of our effort is spent on the first condition. Here is a brief guide to the objects in this analysis. Recall the form of (A1):

$$\Pr\left(\sum_{j\in[k]} \sum_{m\le n<T(m)} H_j^{(n+1)} < r \quad \text{and} \quad \sum_{m\le n<T(m)} H_i^{(n+1)} > s \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \in [\varepsilon, \varepsilon_0) \right) > c.$$

Since $H_j^{(n+1)}$ depends on the estimator $\widehat{P}_j^{(n)}$, there are two time units of analysis: (i) many short intervals of length s_n from $n_\circ$ to n used to compute the estimators, and (ii) the much longer interval from m to $T(m)$ over which we aim to bound the behavior of the accumulated learning rates.

It turns out that our ability to control $\widehat{P}_j^{(n)}$ depends on how well-behaved the maps $P_j : \mathcal{D}_R \to [0,1]$ are on a neighborhood of the trajectory of the iterates during the short intervals. The main issue is that the P_j's are not nice everywhere on $\mathcal{D}_R$. All is not lost though, for (A1) requires these bounds only when $\|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0$. Hope remains if $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ lies in some region K of $\mathcal{D}_R$ on which the maps P_j are well-behaved. Indeed, we shall be able to find such a K onto which we can restrict our analysis. But we cannot simply condition on a future event that the trajectories remain in K, since many of the tools we use from martingale analysis break if we do so. To handle this, let us define the notion of a *core set*.

Definition D.2 (r-core set)**.** Given $S \subset \mathcal{D}_R$, we say that $S_\circ$ is an *r-core set* of S if for all $m, n \in \mathbb{N}$,

$$W^{(m)} \in S_\circ \quad \text{and} \quad \sum_{j\in[k]} \sum_{m\le n'<n} H_j^{(n'+1)} < r \qquad \Longrightarrow \qquad \forall m \le n' \le n, \quad W^{(n')} \in S \quad \text{a.s.} \tag{18}$$

In other words, if we are presently in an r-core set $W^{(m)} \in S_\circ$, then we are guaranteed to remain in S so long as the accumulated learning rate does not exceed r.

Remark D.3. Recall from Lemma 3.2 that the displacement in iterates is bounded by the accumulated learning rate by an additional factor of $2R$. It follows that $S_\circ$ is an r-core set of S whenever (i) $S_\circ$ is contained in S, and (ii) $S_\circ$ is separated from the boundary ∂S by a distance of $2R{\cdot}r$. Here, $\partial S := \text{closure}(S)\backslash\text{interior}(S)$.

If we find an $r_\circ$-core set $K_\circ$ of K, we can ensure that iterates remain in K from times $n_\circ$ through n whenever $W^{(n_\circ)}$ begins in $K_\circ$ and the accumulated learning rates do not exceed $r_\circ$. It turns out that we will eventually always be able to upper bound the accumulated learning rate by $r_\circ$; in fact, Lemma D.9 shows that the accumulated learning rate over this short interval $n_\circ$ to n converges to zero.

This allows us to analyze $\widehat{P}_j^{(n)}$. For example, if $W^{(n_\circ)} \in K_\circ$, Lemma D.10 applies Azuma-Hoeffding's to show that the estimator is consistent. In fact, it is concentrated with high probability:

$$\Pr\left(\left|\widehat{P}_j^{(n)} - P_j^{(n)}\right| \lesssim a_n \,\middle|\, \mathcal{F}_{n_\circ},\ W^{(n_\circ)} \in K_\circ\right) > 1 - \frac{1}{n}. \tag{19}$$

where $a_n \to 0$ is a sequence depending on s_n and t_n that converges to zero.

So far, our discussion has focused on the analyses over the short intervals. But, we also have to bound the behavior of the learning rates over the long interval from m to $T(m)$. Here, we run into the same issue: at time m, we cannot condition on the future event that the iterates remain in $K_\circ$, which we need to control the individual learning rates. We need to be able to be able to choose $K_\circ$ and K so that $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ is an r_0-core set of $K_\circ$. This is in fact possible; we obtain a sequence of core sets seen in Figure 1.

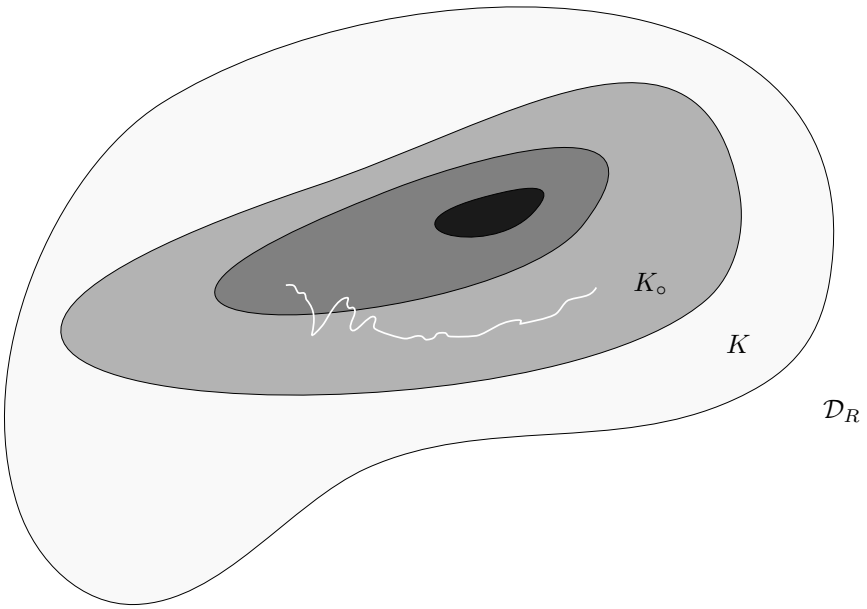

Figure 1: A sequence of subsets: $\{\nabla f = 0\} \subset \{\|\nabla_{w_i} f\| \le \varepsilon_0\} \subset K_\circ \subset K \subset \mathcal{D}_R$. The page is $\mathcal{D}_R$. As P_j is not well-behaved over all of $\mathcal{D}_R$, we construct a compact subset K (*light gray*) over which the maps P_j are L-Lipschitz. K contains an $r_\circ$-core set $K_\circ$ (*gray*), allowing Lemma D.10 to control the behavior of estimators over short intervals of length s_n when $W^{(n_\circ)} \in K_\circ$. To control the learning rates over the long interval m to $T(m)$, we chose $K_\circ$ and K so that $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ (*dark gray*) is an r_0-core set of $K_\circ$. We show in Lemma D.4 that when iterates start within this set, then they do not exit $K_\circ$ with constant probability during the long interval. The white squiggly line depicts the trajectory of such a sequence of iterates. Notice that $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ contains the set of stationary points (*black*).

D.1 Proof of Theorem 4.3

Fix $\varepsilon_0 > 0$ so that $K := \{\|\nabla_{w_i} f\| \le 3\varepsilon_0\}$ is compact; such an ε_0 exists by Lemma 4.1. Because each of the $P_j : \mathcal{D}_R \to [0,1]$ is locally Lipschitz by Lemma D.11, there exists a constant $L > 0$ so that they are all L-Lipschitz on K. Put $K_\circ := \{\|\nabla_{w_i} f\| \le 2\varepsilon_0\}$. Then $K_\circ$ is bounded away from $\partial K := \{\nabla_{w_i} f\| = 3\varepsilon_0\}$, since both are disjoint, non-empty, and compact sets, by Lemma C.3. Thus, Remark D.3 implies that $K_\circ$ is an $r_\circ$-core set of K for some $r_\circ > 0$. Similarly, $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ is an r_0-core set of $K_\circ$ for some $r_0 > 0$.

We also define the sequence:

$$a_n := c \cdot \left(\frac{1}{t_{n_\circ}} + \frac{s_n \log s_n}{n}\right) \qquad \text{and} \qquad c := \max\{1, 256kRL\}. \tag{20}$$

For any $r > 0$, define the function $T_r : \mathbb{N} \to \mathbb{N}$ so that $T_r(m)$ is the unique natural number so that:

$$\sum_{m \le n < T_r(m)} \frac{1}{n} \le r < \sum_{m \le n \le T_r(m)} \frac{1}{n}. \tag{21}$$

Because $\{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ is an r_0-core set of $K_\circ$, the following lemma shows that we can choose T so that eventually, whenever $W^{(m)} \in \{\|\nabla_{w_i} f\| \le \varepsilon_0\}$, then iterates will remain in $K_\circ$ for the whole duration from m through $T(m)$ with constant probability. In fact, it more generally verifies the first half of condition (A1).

Lemma D.4. *Let $H_j^{(n)}$ and $\widehat{P}_j^{(n)}$ be defined as in (4), and let s_n and t_n be non-decreasing sequences in $\mathbb{N}$. Let $\varepsilon_0, K, K_\circ, L, r_\circ, r_0, a_n, c$ be given as above. Let $0 < r < \min\{\ln 2, r_0\}$. Assume that there exists $m_0 \in \mathbb{N}$ such that for all $n \ge m_0$, the following hold:*

(a) $4n^{2/3}(\log 2n)^{1/3} \le s_n \le \frac{1}{2}n - 1$

(b) $a_n < \min\{cr_\circ/16k, t_n/n\}$

(c) $s_{2n}/t_n < r/12k$.

Then, the function $T \equiv T_{Cr}$ where $C = 1/18k$ satisfies for all $m \ge m_0$:

$$\Pr\left(\sum_{j\in[k]} \sum_{m \le n < T(m)} H_j^{(n+1)} \ge r \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0\right) \le \frac{1}{3}.$$

Proof of Lemma D.4. The essence of the proof will be to apply Markov's inequality by bounding the expectation of the summed learning rates. Since $H_j^{(n+1)}$ is of the form:

$$H_j^{(n+1)} = \frac{\mathbb{1}\{I^{(n+1)} = j\}}{\max\{n\widehat{P}_j^{(n)}, t_n\}},$$

we upper bound it via the concentration result (19) of Lemma D.10, which lower bounds $\widehat{P}_j^{(n)}$ when iterates have not strayed out of $K_\circ$ by time $n_\circ$. Then, we apply Markov's to a related stopped process that sets learning rates to zero once iterates exit $K_\circ$. Let $(Z_n)_{n>m}$ be the accumulated learning rates from time m,

$$Z_n = \sum_{j\in[k]} \sum_{m \le n' < n} H_j^{(n'+1)}.$$

Define τ as the exit time from $K_\circ$ and $(Z_{n\wedge\tau})_{n>m}$ be the stopped process:

$$\tau := \min_{n \ge m} \{W^{(n)} \notin K_\circ\} \qquad \text{and} \qquad Z_{n\wedge\tau} = \sum_{j\in[k]} \sum_{m \le n' < n\wedge\tau} H_j^{(n'+1)}, \tag{22}$$

where $n \wedge \tau := \min\{n, \tau\}$. We conditioned on $W^{(m)} \in \{\|\nabla_{w_i} f\| \le \varepsilon_0\}$ to be initially in an r_0-core set of $K_\circ$. So, the iterates will remain in $K_\circ$ through iteration n if $Z_n < r_0$. We claim that if $r < r_0$, then the events $\{Z_n < r\}$ and $\{Z_{n\wedge\tau} < r\}$ are equal. Indeed, we have that if $Z_n < r < r_0$, then $Z_{n\wedge\tau} = Z_n$. And because the accumulated learning rate when the process stops Z_τ must be at least r_0, we also have that if $Z_{n\wedge\tau} < r < r_0$, then the process has not stopped yet, so that $Z_{n\wedge\tau} = Z_n$. Thus:

$$\Pr\left(Z_n \ge r \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0\right) = \Pr\left(Z_{n\wedge\tau} \ge r \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0\right).$$

We now bound the right-hand side by bounding the expected value of $Z_{T(m)\wedge\tau}$ and applying Markov's. The expected value of $Z_{T(m)\wedge\tau}$ can be bounded by considering each term within the summation (22) individually:

$$\begin{aligned}
&\mathbb{E}\left[Z_{T(m)\wedge\tau} \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0\right] \\
&\qquad \le \sum_{j\in[k]} \sum_{m \le n < T(m)} \mathbb{E}\left[H_j^{(n+1)} \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0, \tau > n_\circ\right].
\end{aligned}$$

Consider two intervals: (1) a warm-up interval $m \le n \le m + s_{T(m)}$ during which $n_\circ < m$ is possible, and (2) the tail interval $m + s_{T(m)} < n < T(m) \wedge \tau$ during which $n_\circ \ge m$ holds. For interval (1), we use the coarse bound $H_j^{(n+1)} \le t_m^{-1}$. In interval (2), for any center $j \in [k]$ and iteration $m + s_{T(m)} \le n < T(m)$, we have:

$$\begin{aligned}
&\mathbb{E}\left[H_j^{(n+1)} \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0, \tau > n_\circ\right] \\
&\quad \overset{(i)}{=} \mathbb{E}\left[\mathbb{E}\left[H_j^{(n+1)} \,\middle|\, \mathcal{F}_n\right] \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0, \tau > n_\circ\right] \\
&\quad \overset{(ii)}{=} \mathbb{E}\left[\frac{P_j^{(n)}}{\max\{n\widehat{P}_j^{(n)}, t_n\}} \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0, \tau > n_\circ\right] \\
&\quad \overset{(iii)}{\le} \frac{3}{n}.
\end{aligned}$$

where (i) follows from the tower law for conditional expectations, (ii) from plugging in the form of $H_j^{(n+1)}$, and (iii) we must prove. But assuming this to be the case, we then have the upper bound:

$$\mathbb{E}\left[Z_{T(m)\wedge\tau} \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \le \varepsilon_0\right] \le \frac{k(s_{T(m)}+1)}{t_m} + \sum_{m \le n < T(m)} \frac{3k}{n} \quad \le \quad \frac{r}{3},$$

where the last inequality holds because m_0 is sufficiently large so that $s_{2m}/t_m < r/12k$ in condition (c) holds, and because $T \equiv T_{Cr}$, where $C = 1/18k$. In particular, we assumed that $Cr < \ln 2$, so Corollary D.13 shows that $T(m) \le 2m$; as s_n is non-decreasing, $s_{T(m)} + 1 \le 2s_{2m}$. Thus, the first term is upper bounded by $r/6$. The second term is less than $r/6$ by the definition of T. The lemma follows from Markov's inequality.

Only inequality (iii) above is left. Since $n_\circ \ge m$, by the tower law again, it suffices to show:

$$\mathbb{E}\left[\frac{P_j^{(n)}}{\max\{n\widehat{P}_j^{(n)}, t_n\}} \,\middle|\, \mathcal{F}_{n_\circ},\ \tau > n_\circ\right] \le \frac{3}{n}.$$

Note that because $\tau > n_\circ$, we have $W^{(n_\circ)} \in K_\circ$. This along with conditions (a) and (b) shows that Lemma D.9 and Lemma D.10 may be applied; we use them as follows. Consider two cases separately:

$$\{P_j^{(n_\circ)} \le a_n\} \qquad \text{and} \qquad \{P_j^{(n_\circ)} > a_n\}.$$

Lemma D.9 shows that $P_j^{(n_\circ)}$ and $P_j^{(n)}$ are almost surely within $a_n/8$ of each other. Thus in the first case:

$$\mathbb{E}\left[\frac{P_j^{(n)}}{\max\{n\widehat{P}_j^{(n)}, t_n\}} \,\middle|\, \mathcal{F}_{n_\circ},\ \tau > n_\circ\ P_j^{(n_\circ)} \le a_n\right] \le \frac{9}{8}\frac{a_n}{t_n} \le \frac{3}{n},$$

where the last inequality holds because we assumed that $a_n/t_n \le \frac{1}{n}$.

In the second case, Lemma D.10 shows that $P_j^{(n)}/\widehat{P}_j^{(n)} < 2$ with probability at least $1 - \frac{1}{n}$. Because $H_j^{(n+1)} \le 1$, the failure mode contributes at most $\frac{1}{n}$ to the expectation:

$$\mathbb{E}\left[\frac{P_j^{(n)}}{\max\{n\widehat{P}_j^{(n)}, t_n\}} \,\middle|\, \mathcal{F}_{n_\circ},\ \tau > n_\circ,\ P_j^{(n_\circ)} > a_n\right] \le \frac{2}{n} + \frac{1}{n} \le \frac{3}{n}. \qquad \square$$

Remark D.5. As we mentioned earlier, when r, T and m_0 are defined as in Lemma D.4, this result shows that the iterates from m through $T(m)$ remain in $K_\circ$ with probability at least $2/3$ when $m \ge m_0$. This is because we conditioned on $W^{(m)} \in \{\|\nabla_{w_i} f\| \le \varepsilon_0\}$, which is an r_0-core set of $K_\circ$, and we assumed $r < r_0$.

Introducing a few more conditions, which we highlight in blue, leads to all of (A1).

Lemma D.6. *Let $H_j^{(n)}$ and $\widehat{P}_j^{(n)}$ be defined as in (4), and let s_n and t_n be non-decreasing sequences in $\mathbb{N}$. Let $\varepsilon_0, K, K_\circ, L, r_\circ, r_0, a_n, c$ be given as above. Let $\varepsilon \in (0, \varepsilon_0)$ and also let $0 < r < \min\{\ln 2, r_0, \varepsilon/8R^2L, \frac{1}{C}\ln\frac{7}{6}\}$. Set $s = Cr/8$. Assume that there exists $m_0 \in \mathbb{N}$ such that for all $n \geq m_0$, the following hold:*

(a) $4n^{2/3}(\log 2n)^{1/3} \leq s_n \leq \min\{\frac{1}{2}n - 1, \frac{1}{2}(e^{Cr/2} - 1)n - e^{Cr/2}\}$

(b) $a_n < \min\{cr_\circ/16k, t_n/n, \varepsilon/4R\}$

(c) $s_{2n}/t_n < r/12k$.

(d) $e^{Cr/2}\sqrt{2n\ln 6}/s < t_n < n\varepsilon/8R$.

Then, the function $T \equiv T_{Cr}$ where $C = 1/18k$ satisfies for all $m \geq m_0$:

$$\Pr\left(\sum_{j\in[k]}\sum_{m\leq n<T(m)} H_j^{(n+1)} < r \quad \text{and} \quad \sum_{m\leq n<T(m)} H_i^{(n+1)} > s \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \in [\varepsilon, \varepsilon_0)\right) > \frac{1}{3}.$$

Proof of Lemma D.6. The learning rate $H_i^{(n+1)}$ has conditional expectation:

$$\mathbb{E}\left[H_i^{(n+1)} \,\middle|\, \mathcal{F}_n\right] = \mathbb{E}\left[\frac{\mathbb{1}\{I^{(n+1)} = i\}}{\max\{n\widehat{P}_i^{(n)}, t_n\}} \,\middle|\, \mathcal{F}_n\right] = \frac{P_i^{(n)}}{\max\{n\widehat{P}_i^{(n)}, t_n\}}.$$

Thus, the sequence $H_i^{(n+1)} - \mathbb{E}\left[H_i^{(n+1)} \,\middle|\, \mathcal{F}_n\right]$ is a martingale difference sequence with bounded increments:

$$\left|\frac{\mathbb{1}\{I^{(n+1)} = i\} - P_i^{(n)}}{\max\{n\widehat{P}_i^{(n)}, t_n\}}\right| \leq \frac{1}{t_n}.$$

Define μ and ν as follows:

$$\mu := \sum_{m\leq n<T(m)} \frac{P_i^{(n)}}{\max\{n\widehat{P}_i^{(n)}, t_n\}} \qquad \text{and} \qquad \nu := \left(\sum_{m\leq n<T(m)} \frac{1}{t_n^2}\right)^{-1}.$$

Azuma-Hoeffding's implies that the accumulated learning rates for the ith center concentrates about μ,

$$\Pr\left(\sum_{m\leq n<T(m)} H_i^{(n+1)} > \mu - s \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \in [\varepsilon, \varepsilon_0)\right) > 1 - \exp\left(-\frac{1}{2}\nu s^2\right) > \frac{5}{6}, \tag{23}$$

where the last inequality follows from:

$$\nu = \left(\sum_{m\leq n<T(m)} \frac{1}{t_n^2}\right)^{-1} \overset{(i)}{\geq} \frac{1}{e^{Cr} - 1}\frac{t_m^2}{m} \overset{(ii)}{>} \frac{2\ln 6}{s^2},$$

since (i) $T(m) - m \leq (e^{Cr} - 1)m$ by Corollary D.13 and t_n is non-decreasing, and (ii) $t_n > e^{Cr/2}\sqrt{2n\ln 6}/s$ by assumption (d). We also claim that:

$$\Pr\left(\sum_{j\in[k]}\sum_{m\leq n<T(m)} H_j^{(n+1)} < r \quad \text{and} \quad \mu > 2s \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \in [\varepsilon, \varepsilon_0)\right) > \frac{1}{2}. \tag{24}$$

Assuming the claim, we obtain the desired result by combining (23) and (24) by a union bound.

Now, only the claim remains, but let's first reduce notation. Denote the two events in (24) by Ξ and E,

$$\Xi = \left\{\sum_{j\in[k]}\sum_{m\leq n<T(m)} H_j^{(n+1)} < r\right\} \qquad \text{and} \qquad \mathrm{E} := \left\{\sum_{m\leq n<T(m)} \frac{P_i^{(n)}}{\max\{n\widehat{P}_i^{(n)}, t_n\}} > 2s\right\},$$

and let $\mathrm{F} = \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \in [\varepsilon, \varepsilon_0)$ denote the conditioning; (24) states that $\Pr\left(\Xi \cap \mathrm{E} \,|\, \mathrm{F}\right) > 1/2$. From Lemma D.4, we have $\Pr(\Xi \,|\, \mathrm{F}) > 2/3$. Thus, we just need to show that the event $\mathrm{E}|\mathrm{F}$ also likely holds. This turns out to be the case if for nearly all iterations between m and $T(m)$, the conditional expectation satisfies:

$$\frac{P_i^{(n)}}{\max\{n\widehat{P}_i^{(n)}, t_n\}} > \frac{1}{2n}. \tag{25}$$

Again to reduce notation, denote the event in (25) by E_n. Then specifically, E occurs if all events E_n occur over times $m + s_{T(m)} \le n < T(m)$. This is due the definition of T, which implies the following:

$$\begin{aligned}
\sum_{m+s_{T(m)} \le n < T(m)} \frac{1}{2n} &\overset{(i)}{\ge} \frac{1}{2} \log \frac{T(m)}{m + s_{T(m)}} \\
&\overset{(ii)}{\ge} \frac{1}{2} \log \frac{e^{Cr}(m-1)}{m + s_{T(m)}} \\
&\overset{(iii)}{\ge} 2s,
\end{aligned}$$

where (i) follows from Lemma D.12, (ii) from Corollary D.13, and (iii) from setting $s = Cr/8$ and our choice of upper bound on s_n. In particular, because $r < \ln 2$, Corollary D.13 shows that $T(m) \le 2m$. Since s_n is non-decreasing, $s_{T(m)} \le s_{2m} \le (e^{Cr/2} - 1)m - e^{Cr/2}$. A little bit of algebra then verifies (iii).

The natural way to prove (24) would then be to union bound the probability of failure:

$$\Pr\left(\Xi^c \cup \mathrm{E}^c \,\middle|\, \mathrm{F}\right) \le \Pr\left(\Xi^c \,\middle|\, \mathrm{F}\right) + \sum_{m+s_{T(m)} \le n < T(m)} \Pr\left(\mathrm{E}_n^c \,\middle|\, \mathrm{F}\right).$$

This turns out to be too coarse of a bound for us; there is too much overcounting of certain outcomes in Ξ^c that are also contained in E_n^c. Instead, we use a finer union bound—since the first term $\Pr(\Xi^c \,|\, \mathrm{F})$ already accounts for the bad outcomes in Ξ^c, the remaining sum needs only measure the outcomes in $\mathrm{E}_n^c \cap \Xi$. In fact, we only loosen the bound if we measure outcomes in $\mathrm{E}_n^c \cap \Xi_n$, for a superset $\Xi_n \supset \Xi$. Thus:

$$\begin{aligned}
\Pr\left(\Xi^c \cup \mathrm{E}^c \,\middle|\, \mathrm{F}\right) &\le \Pr\left(\Xi^c \,\middle|\, \mathrm{F}\right) + \sum_{m+s_{T(m)} \le n < T(m)} \Pr\left(\mathrm{E}_n^c \cap \Xi \,\middle|\, \mathrm{F}\right) \\
&\le \Pr\left(\Xi^c \,\middle|\, \mathrm{F}\right) + \sum_{m+s_{T(m)} \le n < T(m)} \Pr\left(\mathrm{E}_n^c \cap \Xi_n \,\middle|\, \mathrm{F}\right) \\
&\le \Pr\left(\Xi^c \,\middle|\, \mathrm{F}\right) + \sum_{m+s_{T(m)} \le n < T(m)} \Pr\left(\mathrm{E}_n^c \,\middle|\, \mathrm{F}, \Xi_n\right),
\end{aligned} \tag{26}$$

where in the last step, we use the general fact that $\Pr(A \cap B) \le \Pr(A \,|\, B)$.

Lemma D.4 bounds the first term in (26) with $\Pr(\Xi^c \,|\, \mathrm{F}) \le 1/3$. For the others, notice that conditioned on F, the event Ξ, which bounds the accumulated learning rates by r, implies the $\mathcal{F}_{n_\circ}$-measurable events:

$$\Xi_n := \left\{W^{(n_\circ)} \in K_\circ \quad \text{and} \quad P_i^{(n_\circ)} \ge \frac{\varepsilon}{4R}\right\}.$$

This is because $r < \min\{r_0, \varepsilon/8R^2L\}$. If $\Xi \,|\, \mathrm{F}$ occurs, the bound $r < r_0$ implies that all iterates remain in $K_\circ$, as discussed in Remark D.5. Furthermore, since P_i is L-Lipschitz on $K_\circ$, we also have for all $m \le n \le T(m)$,

$$\left|P_i^{(n)} - P_i^{(m)}\right| \le 2RL \cdot \sum_{j \in [k]} \sum_{m \le n' < n} H_j^{(n'+1)} \le 2RLr.$$

Thus, $r < \varepsilon/8R^2L$ implies $2RLr < \varepsilon/4R$. So, $P_i^{(n)}$ is lower bounded because $P_i^{(m)} \ge \varepsilon/2R$. This comes from the gradient, $\nabla_{w_i} f(w) = P_i(w) \cdot \big(w_i - M_i(w)\big)$, and that the initial iterate satisfies $\|\nabla_{w_i} f(W^{(m)})\| \ge \varepsilon$.

We claim that $\Pr(\mathrm{E}_n^c \,|\, \mathrm{F}, \Xi_n) \le \frac{1}{n}$. If this is the case, then (26) implies (24):

$$\Pr\left(\Xi^c \cup \mathrm{E}^c \,\middle|\, \mathrm{F}\right) \le \frac{1}{3} + \sum_{m+s_{T(m)} \le n < T(m)} \frac{1}{n}$$
$$\overset{(i)}{\le} \frac{1}{3} + \frac{T(m)-m}{m} \overset{(ii)}{\le} \frac{1}{3} + \left(e^{Cr}-1\right) \overset{(iii)}{\le} \frac{1}{2},$$

where we use (i) $\frac{1}{m} \ge \frac{1}{n}$ on this interval, (ii) $T(m) - m \le (e^{Cr}-1)m$, and (iii) $r < \frac{1}{C}\ln\frac{7}{6}$.

Now, all that is left is to verify that $\Pr(\mathrm{E}_n^c \,|\, \mathrm{F}, \Xi_n)$ is bounded above:

$$\Pr\left(\frac{P_i^{(n)}}{\max\{n\widehat{P}_i^{(n)}, t_n\}} \le \frac{1}{2n} \,\middle|\, \mathcal{F}_m, \|\nabla_{w_i} f(W^{(m)})\| \in [\varepsilon, \varepsilon_0),\ W^{(n_\circ)} \in K_\circ,\ P_i^{(n_\circ)} \ge \frac{\varepsilon}{4R}\right) \le \frac{1}{n}.$$

This is true by Lemma D.10, which shows multiplicative concentration $\frac{1}{2}P_i^{(n)} < \widehat{P}_i^{(n)} < 2P_i^{(n)}$ with probability at least $1 - \frac{1}{n}$. This lemma applies because Ξ_n is $\mathcal{F}_{n_\circ}$-measurable with $n_\circ > m$, since we only consider iterations n between $m + s_{T(m)}$ and $T(m)$, and because $\varepsilon/4R > a_n$. Concentration implies E_n: the left tail bound shows that $\max\{n\widehat{P}_i^{(n)}, t_n\} = n\widehat{P}_i^{(n)}$ as $t_n < n\varepsilon/8R$ and $P_i^{(n)} \ge \varepsilon/4R$; the right tail shows $P_i^{(n)}/n\widehat{P}_i^{(n)} < 1/2n$. □

We now verify condition (A2) of Theorem C.1. The proof remixes techniques used to prove (A1). If $\|\nabla_{w_i} f\|$ is lower bounded by a constant, then so is P_i. And so $\widehat{P}_i^{(n)}$ will also be lower bounded by a constant with high probability. On average $H_i^{(n+1)}$ is on the order of n^{-1}, whose sum diverges.

Lemma D.7. *Let $H_j^{(n)}$ and $\widehat{P}_j^{(n)}$ be defined as in (4) and $\lim_{n\to\infty} \frac{\log n}{s_n} = \lim_{n\to\infty} \frac{t_n}{n} = 0$. Let $\varepsilon_0 > 0$. Then:*

$$\liminf_{n\to\infty} \|\nabla_{w_i} f(W^{(n)})\| \ge \varepsilon_0 \qquad \Longrightarrow \qquad \sum_{n\in\mathbb{N}} H_i^{(n)} = \infty \quad \text{a.s.}$$

Proof of Lemma D.7. We saw in the proof of Lemma D.6 that $P_i^{(n)} > \varepsilon_0/2R$ is implied by $\|\nabla_{w_i} f(W^{(n)})\| > \varepsilon$, since the gradient is $\nabla_{w_i} f(w) = P_i(w) \cdot \big(w_i - M_i(w)\big)$. Therefore, if the limit infimum condition holds, then there exists a random variable $N \in \mathbb{N}$ such that if $n > N$, then:

$$P_i^{(n)} \ge \frac{\varepsilon_0}{4R}.$$

That is, the probability of updating the center i eventually remains at least $\varepsilon_0/4R$. Thus, $\widehat{P}_i^{(n)} > \varepsilon_0/8R$ holds with high probability at any sufficiently large iteration. In particular, if n is large enough to satisfy $n_\circ > N$ and $s_n > \frac{128R^2}{\varepsilon_0^2}\ln n$, then Azuma-Hoeffding's implies (Lemma D.10 uses the same technique),

$$\Pr\left(\widehat{P}_j^{(n)} \le \frac{\varepsilon_0}{8R} \,\middle|\, \mathcal{F}_{n_\circ}\right) \le \Pr\left(\widehat{P}_j^{(n)} \le \frac{1}{s_n}\sum_{n_\circ \le n' < n} P_j^{(n')} - \frac{\varepsilon_0}{8R} \,\middle|\, \mathcal{F}_{n_\circ}\right) \le \exp\left(-\frac{s_n\varepsilon_0^2}{128R^2}\right) \le \frac{1}{n}.$$

This bound on $\widehat{P}_i^{(n)}$ implies one on $H_i^{(n+1)}$. In particular, if n is sufficiently large so that $t_n < n\varepsilon_0/8R$, then:

$$\Pr\left(H_i^{(n+1)} = \frac{P_i^{(n)}}{\max\{n\widehat{P}_i^{(n)}, t_n\}} = \frac{P_i^{(n)}}{n\widehat{P}_i^{(n)}} \ge \frac{1}{n}\frac{\varepsilon_0}{4R} \,\middle|\, \mathcal{F}_{n_\circ}\right) > 1 - \frac{1}{n},$$

where the inequality in gray comes from the lower bound $P_i^{(n)} \ge \varepsilon_0/4R$ and the upper bound $\widehat{P}_i^{(n)} = 1$. And so, we have that for n sufficiently large:

$$\Pr\left(\sum_{m \le n' < T_r(m)} H_i^{(n'+1)} > c \,\middle|\, \mathcal{F}_{n_\circ}\right) > 1 - r,$$

where we may take $r < 1$ and we set $c = r\varepsilon_0/4R$. This inequality follows directly from a union bound and the definition of T_r. Borel-Cantelli implies that the accumulated learning on the ith center increases by c infinitely often, which implies that $\sum_{n\in\mathbb{N}} H_i^{(n)}$ diverges. □

Under assumptions on s_n and t_n, the conditions of Lemma D.6 and Lemma D.7 are verified. Thus, conditions (A1) and (A2) of Theorem C.1 are satisfied, proving Theorem 4.3. ■

Remark D.8 (Generalized union bound)**.** The modified union bound used in Lemma D.6 may be of generic interest: let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $A, B, C \in \mathcal{F}$ be events such that $A^c \subset C$. Then:

$$P(A \cup B) \overset{(i)}{=} P(A) + P(B \cap A^c) \overset{(ii)}{\leq} P(A) + P(B \cap C) \overset{(iii)}{\leq} P(A) + P(B \,|\, C),$$

where (i) $A \cup B$ is the disjoint union $A \sqcup (B \cap A^c)$, (ii) $B \cap A^c \subset B \cap C$, and (iii) $P(B \cap C) \leq P(C)P(B \,|\, C)$. This is useful because $P(B \,|\, C)$ may in general be easier to bound than $P(B \,|\, A^c)$, as was our case.

D.2 Consistency and concentration of $\widehat{P}_j^{(n)}$

The estimator $\widehat{P}_j^{(n)}$ for $P_j^{(n)} := P_j(W^{(n)})$ is consistent, provided P_j is locally Lipschitz and (s_n, t_n) satisfies:

$$\frac{1}{t_n} \to 0 \qquad \text{and} \qquad \frac{s_n \log s_n}{n} \to 0, \qquad \text{as} \quad n \to \infty.$$

Specifically, we give non-asymptotic rates of concentration in Lemma D.10.

The estimator $\widehat{P}_j^{(n)}$ depends on the trajectory of the past s_n iterates up to that point. In particular, since $I^{(n'+1)}$ is drawn from $P(W^{(n')})$, Azuma-Hoeffding's shows that the estimator tends to concentrate around:

$$\frac{1}{s_n} \sum_{n_\circ \leq n' < n} P_j^{(n')}.$$

Therefore, $\widehat{P}_j^{(n)}$ concentrates around $P_j^{(n)}$, as long as $P_j^{(n')}$ does not vary too much over $n_\circ \leq n' \leq n$. The amount of variation can be bounded because the maps $P_j : \mathcal{D}_R \to [0,1]$ are locally Lipschitz (Lemma D.11). We just need to ensure that the iterates do not move too much—we achieve this by upper bounding the accumulated learning rates between $n_\circ$ and n, achieved by the next lemma: by bounding the learning rates, we can control the change in P_j whenever iterates stay within a region K on which P_j is L-Lipschitz.

Lemma D.9. *Let $H_j^{(n)}$ and $\widehat{P}_j^{(n)}$ be defined as in (4). If $e \leq s_n + 1 \leq \frac{n}{2}$, then:*

$$\sum_{j\in[k]} \sum_{n_\circ \leq n' < n} H_j^{(n'+1)} \leq \frac{16k}{t_{n_\circ}} + \frac{16k s_n \log s_n}{n} \quad \text{a.s.}$$

Let $K \subset \mathcal{D}_R$ be given so that the restriction $P_j\big|_K : K \to [0,1]$ is L-Lipschitz. Conditioned on $W^{(n_\circ)}, \ldots, W^{(n)}$ remaining in K, then for all $n_\circ \leq n' \leq n$:

$$\left| P_j^{(n')} - P_j^{(n)} \right| \leq 32kRL \cdot \left(\frac{1}{t_{n_\circ}} + \frac{s_n \log s_n}{n} \right) \quad \text{a.s.} \tag{27}$$

Proof. Fix $j \in [k]$. The following chain of inequalities holds almost surely:

$$\begin{aligned} \sum_{n_\circ \leq n' < n} H_j^{(n'+1)} &\leq \sum_{n_\circ \leq n' < n} \frac{\mathbb{1}\{I^{(n'+1)} = j\}}{\max\{n' \cdot \widehat{P}_j^{(n')}, t_{n'}\}} \\ &\leq \frac{1}{t_{n_\circ}} + \sum_{n'=1}^{s_n - 1} \frac{1}{(n - s_n) \cdot \frac{n'}{s_n}} \leq \frac{1}{t_{n_\circ}} + \frac{16 s_n \log s_n}{n} \quad \text{a.s.} \end{aligned}$$

The first equality expands the definition of the learning rate (4). The next inequality comes the worst-case scenario where the jth center has had no recent updates (so that $\widehat{P}_j^{(n-s_n)} = 0$), and it is updated every single time following during this window of length s_n. We've re-indexed the sum by subtracting $n - s_n$ from the original index. For the first term, since $\widehat{P}_j{}^{(n-s_n)} = 0$, we use the bound $H_j^{(n_\circ+1)} \le t_{n_\circ}^{-1}$. And for the rest, we bound each $\widehat{P}_j^{(n')}$ by $\frac{1}{s_n}, \frac{2}{s_n}, \ldots, \frac{s_n-1}{s_n}$, respectively. This is the worst-case scenario, since delaying an update simply introduces a zero in the sum and shifts the rest of the bounds to the right. The final inequality upper bounds the partial sums of the harmonic series, and uses the assumption $e \le s_n \le \frac{n}{2}$.

For the second bound, Lemma 3.2 converts the learning rate bound to one over distance $\|W^{(n')} - W^{(n)}\|$, introducing a factor of $2R$. Then, Lipschitz continuity bounds $|P_j^{(n')} - P_j^{(n)}|$ by introducing a factor of L. □

The analysis to show that $\widehat{P}_j^{(n)}$ concentrates around $P_j^{(n)}$ would be quite straightforward if P_j were globally Lipschitz—then we could use Lemma D.9 to design conditions on s_n and t_n to force the accumulated learning rates to go to zero over periods of s_n,

$$\lim_{n\to\infty} \sum_{j\in[k]} \sum_{n_\circ \le n' < n} H_j^{(n'+1)} = 0.$$

Thus over a small interval s_n, the iterates would remain close together, and the bias of $\widehat{P}_j^{(n)}$ would be forced to zero in the limit. And if $s_n \uparrow \infty$, then Azuma-Hoeffding's would imply increasingly tight concentration.

Unfortunately, P_j is not generally globally Lipschitz; the local Lipschitz constant at $w \in \mathcal{D}_R$ depends on the distances between centers $\|w_j - w_{j'}\|$, so we need to perform our analysis on a subset $K \subset \mathcal{D}_R$ on which P_j is L-Lipschitz. While we need to know that the iterates $W^{(n_\circ)}, \ldots, W^{(n)}$ remain in K, this event is not generally contained in $\mathcal{F}_{n_\circ}$. Directly conditioning on it would introduce new dependencies that prevent us from applying Azuma-Hoeffding's. We can overcome this issue by conditioning on an $\mathcal{F}_{n_\circ}$-measurable event contained within this event instead: that $W^{(n_\circ)}$ is contained in $K_\circ$, some $r_\circ$-core set of K.

Lemma D.10 (Estimator concentration)**.** *Let $H_j^{(n)}$ and $\widehat{P}_j^{(n)}$ be defined as in (4). Let $K \subset \mathcal{D}_R$ be given so that the restriction $P_j\big|_K : K \to [0,1]$ is L-Lipschitz. Let $K_\circ$ be an $r_\circ$-core set of K. Let $c = \max\{1, 256kRL\}$ and $a_n = c \cdot \left(\frac{1}{t_{n_\circ}} + \frac{s_n \log s_n}{n}\right)$. If s_n satisfies $4n^{2/3}(\log 2n)^{1/3} \le s_n \le \frac{n}{2} - 1$ and $a_n < cr_\circ/16k$, then:*

$$\Pr\left(\left|\widehat{P}_j^{(n)} - P_j^{(n)}\right| < \frac{3}{8}a_n \,\middle|\, \mathcal{F}_{n_\circ},\ W^{(n_\circ)} \in K_\circ\right) > 1 - \frac{1}{n}.$$

In particular, a multiplicative bound holds:

$$\Pr\left(\frac{1}{2}P_j^{(n)} < \widehat{P}_j^{(n)} < 2P_j^{(n)} \,\middle|\, \mathcal{F}_{n_\circ},\ P_j^{(n_\circ)} > a_n,\ W^{(n_\circ)} \in K_\circ\right) > 1 - \frac{1}{n}.$$

Proof. The following sequence during the interval $n_\circ \le n' < n$ is a martingale difference sequence:

$$\mathbb{1}\{I^{(n'+1)} = j\} - P_j^{(n')}.$$

In fact, since the event $\{W^{(n_\circ)} \in K_\circ\}$ is $\mathcal{F}_{n_\circ}$-measurable, we can condition on it, and the sequence remains a martingale difference sequence. Then, $\widehat{P}_j^{(n)}$ is concentrated:

$$\Pr\left(\left|\widehat{P}_j^{(n)} - \frac{1}{s_n}\sum_{n_\circ \le n < n} P_j^{(n')}\right| \ge \frac{a_n}{4} \,\middle|\, \mathcal{F}_{n_\circ},\ W^{(n_\circ)} \in K_\circ\right) \overset{(i)}{\le} 2\exp\left(-\frac{s_n a_n^2}{32}\right) \overset{(ii)}{\le} 2\exp\left(-\frac{s_n^3}{64n^2}\right) \overset{(iii)}{\le} \frac{1}{n},$$

where (i) follows from Azuma-Hoeffding's, (ii) from $\frac{1}{32}a_n^2 \ge \frac{1}{64}s_n^2/n^2$ since $c \ge 1$, and (iii) from plugging in the lower bound on s_n in the theorem statement.

To complete the theorem, we need to relate $P_j^{(n')}$ to $P_j^{(n)}$, which we can do whenever the iterates remain in K. Indeed, we conditioned on $W^{(n_\circ)} \in K_\circ$, and we also have:

$$\sum_{j\in[k]} \sum_{n_\circ \le n' < n} H_j^{(n'+1)} \overset{(i)}{\le} \frac{16ka_n}{c} \overset{(ii)}{<} r_\circ \quad \text{a.s.},$$

where (i) is the first result of Lemma D.9, which we may apply since $4 \le s_n \le \frac{n}{2} - 1$, and (ii) we assumed that $a_n < cr_\circ/16k$. By the core-set property of $K_\circ$, the iterates remain in K during this interval on which P_j is L-Lipschitz. We can now apply the second result (27) of Lemma D.9, which shows that for all $n_\circ \le n' \le n$,

$$\left|P_j^{(n')} - P_j^{(n)}\right| \le 32kRL \cdot \left(\frac{1}{t_{n_\circ}} + \frac{s_n \log s_n}{n}\right) \le \frac{a_n}{8} \quad \text{a.s.}$$

By triangle inequality:

$$\left|\frac{1}{s_n} \sum_{n_\circ \le n' < n} P_j^{(n')} - P_j^{(n)}\right| \le \frac{a_n}{8} \quad \text{a.s.}$$

A further application of triangle inequality yields the desired additive concentration bound:

$$\Pr\left(\left|\widehat{P}_j^{(n)} - P_j^{(n)}\right| \ge \frac{3}{8}a_n \,\middle|\, \mathcal{F}_{n_\circ},\ W^{(n_\circ)} \in K_\circ\right) \le \frac{1}{n}.$$

If we further condition on the $\mathcal{F}_{n_\circ}$-measurable event $\{P_j^{(n_\circ)} > a_n\}$, then (27) implies $P_j^{(n)} > \frac{7}{8}a_n$, from which we also obtain the multiplicative bound. □

D.3 Local Lipschitzness of P_j

Lemma D.11 (P_j is locally Lipschitz)**.** *Let p be a density supported in the closed ball $B(0,R)$. If p is continuous on $B(0,R)$, then then the maps $P_j : \mathcal{D}_R \to [0,1]$ are locally Lipschitz.*

Proof. We first prove this in the setting where there are only two centers (i.e. $k = 2$), before generalizing. Given two tuples of centers $w, w' \in \mathcal{D}_R$. Then the difference $P_j(w) - P_j(w')$ is:

$$P_j(w) - P_j(w') = \int_{V_j(w)\setminus V_j(w')} p(x)\,dx - \int_{V_j(w')\setminus V_j(w)} p(x)\,dx.$$

Since p is continuous on the closed set $B(0,R)$, it attains a maximum $p_{\max} = \sup p(x) < \infty$. Let λ be the Lebesgue measure. It follows by triangle inequality that:

$$|P_j(w) - P_j(w')| \le p_{\max} \cdot \Big(\lambda\big(V_j(w) \setminus V_j(w')\big) + \lambda\big(V_j(w') \setminus V_j(w)\big)\Big).$$

Therefore, to prove that P_j is locally Lipschitz, we need to bound how much the jth Voronoi can grow/shrink when the two centers w are perturbed slightly to $w' \in \mathcal{D}_R$. As $w \in \mathcal{D}_R$, the two centers are separated $\|w_1 - w_2\| > 0$. We claim that if the perturbation $\|w - w'\|$ is a factor smaller than the separation, $\|w - w'\| \le \frac{1}{4}\|w_1 - w_2\|$, then the jth Voronoi region can only grow linearly with $\|w - w'\|$,

$$\lambda\big(V_j(w') \setminus V_j(w)\big) \le L_w \|w - w'\|,$$

for some $L_w > 0$. And, the same can be said for the other term, measuring how much the region can shrink. If this claim holds, then P_j is locally Lipschitz, where the local Lipschitz constant at w is $p_{\max} \cdot 2L_w$.

Fix $w \in \mathcal{D}_R$ and let $2r = \|w_1 - w_2\|$ be the separation of its two centers. By a change of coordinates, we may without loss of generality assume that:

$$w_1 = (-r, 0, \ldots, 0) \qquad \text{and} \qquad w_2 = (r, 0 \ldots, 0).$$

Therefore, the boundary of their Voronoi partitions is the hyperplane $\{x \in \mathbb{R}^d : x_1 = 0\}$. We now show that if the perturbed centers w' satisfy $\|w - w'\| = \varepsilon \le \frac{r}{2}$, then $V_1(w')$ is contained in the halfspace:

$$V_1(w') \subset \left\{x \in \mathbb{R}^d : x_1 \le \left(1 + \frac{2R}{r}\right)\varepsilon\right\},$$

from which local Lipschitzness follows:

$$\lambda\big(V_1(w') \setminus V_1(w)\big) \le (2R)^{d-1}\left(1 + \frac{2R}{r}\right) \cdot \|w - w'\|,$$

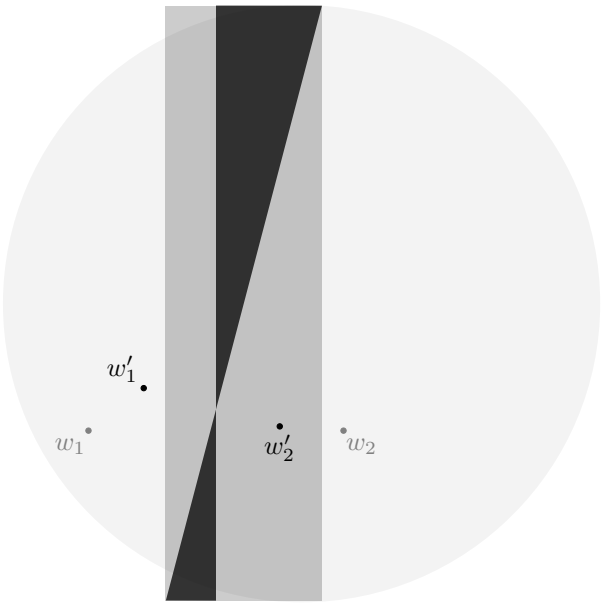

Figure 2: A two-dimensional projection of the 2-means problem in $\mathbb{R}^d$. The light gray disk represents the support of the distribution p, which has a diameter of $2R$. The initial tuple $w = (w_1, w_2)$ partitions the space along the vertical hyperplane. After a small perturbation to $w' = (w_1', w_2')$, a new Voronoi partition is induced, where the black region corresponds the symmetric difference $V_1(w)\Delta V_1(w') = V_2(w)\Delta V_2(w')$. The probability mass of this region can be upper bounded by the rectangular gray region whose width is $O(\|w - w'\|)$ and lengths in all other directions are $2R$.

since $V_1(w') \setminus V_1(w)$ is contained in the rectangular region where the last $d-1$ coordinates have length $2R$ and the first coordinate length $(1 + 2R/r)\varepsilon$. Figure 2 depicts this argument.

We show that $V_1(w')$ is contained in the above halfspace by upper bounding the first coordinate of points in $V_1(w')$. Note that the new boundary induced by w' is the hyperplane H intersecting $\frac{1}{2}(w_1' + w_2')$ defined by the normal vector $w_1' - w_2'$:

$$H := \frac{1}{2}(w_1' + w_2') + \big\{x \in \mathbb{R}^d : (w_1' - w_2')^\top x = 0\big\}.$$

Thus, $V_1(w')$ is to the left of H. The first term $\frac{1}{2}\|w_1' - w_2'\|$ contributes at most ε to the first coordinate of points in H, since $\|w - w'\| = \varepsilon$. Since after the change of coordinates, all points in $B(0, R)$ must now be at most a distance of $2R$ away from $\frac{1}{2}(w_1 + w_2')$, we just need to bound the first coordinate of points in:

$$\big\{x \in B(0, 2R) : (w_1' - w_2')^\top x = 0\big\}.$$

Let $w_1' - w_2' = (\alpha_1, \ldots, \alpha_d)$. Then if x in this set satisfies:

$$|x_1| = \left|\frac{\alpha_2 x_2 + \cdots + \alpha_d x_d}{\alpha_1}\right| \leq \frac{\|w_1' - w_2'\| \cdot \|x\|}{r},$$

by Cauchy-Schwarz and the fact that $|\alpha_1| \geq r$, which follows from the form of w and that the perturbation is less than $r/2$. That is, $|x_1| \leq 2R\varepsilon/r$. This shows that $V_1(w')$ is contained the above halfspace.

At this point, we have show the result for $k = 2$. The setting for general k is an easy extension. Let Δ be the symmetric difference. Then as before, we need to show:

$$\lambda\big(V_j(w)\Delta V_j(w')\big) \leq 2L_w\|w - w'\|,$$

for some $L_w > 0$ and w' in a neighborhood of w.

Given w and fixed j, consider a collection of $k-1$ induced 2-means problems constructed on $\widetilde{w}_\ell := (w_j, w_\ell)$ for $\ell \neq j$. Let $\widetilde{V}$ map the 2-center $\widetilde{w} \in \mathbb{R}^{2\times d}$ to its Voronoi partitions. Then:

$$V_j(w)\Delta V_j(w') \subset \bigcup_{\ell \neq j} \widetilde{V}_j(\widetilde{w}_\ell)\Delta\widetilde{V}_j(\widetilde{w}'_\ell).$$

It follows that we may reduce to the 2-center case, since:

$$\lambda\big(V_j(w)\Delta V_j(w')\big) \leq \sum_{\ell \neq j} \lambda\big(\widetilde{V}_j(\widetilde{w}_\ell)\Delta\widetilde{V}_j(\widetilde{w}'_\ell)\big) \qquad \square$$

D.4 Properties of $T(m)$

Recall we defined for $r > 0$, the function $T_r : \mathbb{N} \to \mathbb{N}$ so that $T_r(m)$ is the unique natural number so that:

$$\sum_{m \leq n < T_r(m)} \frac{1}{n} \leq r < \sum_{m \leq n \leq T_r(m)} \frac{1}{n}. \tag{21}$$

The following lemma and corollary give properties of T_r.

Lemma D.12. *Let $1 < m < m'$ be in $\mathbb{N}$. Then:*

$$\log \frac{m'}{m} \leq \sum_{m \leq n < m'} \frac{1}{n} \leq \log \frac{m'-1}{m-1}.$$

Proof.

$$\log \frac{m'}{m} \leq \int_m^{m'} \frac{1}{x}\, dx \leq \sum_{m \leq n < m'} \frac{1}{n} \leq \int_m^{m'} \frac{1}{x-1}\, dx = \log \frac{m'-1}{m-1}. \qquad \square$$

Corollary D.13. *Let $r > 0$. Let $\alpha := e^r - 1$ and set $T \equiv T_r$. Then:*

$$\alpha(m-1) \leq T(m) - m \leq \alpha m.$$

Proof. Combining Lemma D.12 with the definition of $T(m)$, we have:

$$\log \frac{T(m)}{m} \leq \sum_{m \leq n < T(m)} \frac{1}{n} \leq r < \sum_{m \leq n < T(m)+1} \frac{1}{n} \leq \log \frac{T(m)}{m-1}.$$

Rearranging yields the result. $\square$