AXIS-ALIGNED PROJECTIVE CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study axis-aligned (j, k, z)-projective clustering of n points in $\mathbb{R}^d$, where the objective sums the zth powers of Euclidean distances to the nearest of k axis-aligned j-dimensional affine subspaces. For $j, k \geq 2$ and every fixed $z \geq 1$, we obtain total-sensitivity bounds of $(j + k)^{O_z(\min\{j,k\})} \log n$, independent of the ambient dimension and requiring no bounds on input coordinates. Sensitivity sampling then yields small weighted coresets, with ambient dimension entering only through the complexity of the query family. Complementary lower bounds show that logarithmic dependence on n and exponential dependence on $\min\{j, k\}$ are unavoidable, already in $\mathbb{R}^{j+1}$. We also establish $\widetilde{O}_z(\sqrt{kj/n})$ expected excess-risk bounds on the unit sphere for linear subspaces and affine subspaces within unit distance of the origin. This bound is optimal up to logarithmic factors for the linear class when $d \geq 2kj$ and n is sufficiently large relative to kj.

1 INTRODUCTION

Projective clustering extends classical center-based clustering by representing each cluster with a low-dimensional affine subspace. Owing to the considerable flexibility in its problem formulation, projective clustering has found broad applications across a wide range of machine learning tasks Lucic et al. (2018); Aggarwal & Yu (2000); Procopiuc (2010). However, solving such problems is generally computationally challenging, and the use of popular data reduction tools, such as coresets, often requires additional assumptions on the structure of the input data Tukan et al. (2022); Varadarajan & Xiao (2012a).

We focus on a special case known as axis-aligned projective clustering. Namely, for $I \subset [d]$ with $|I| = d - j$ and $a \in \mathbb{R}^{|I|}$, an axis-aligned j-dimensional affine subspace has the form

$$F = \{y \in \mathbb{R}^d : y_I = a\},$$

where y_I is the $|I|$-dimensional vector formed by the coordinates of y on I. Its distance from x is $d(x, F) = \|x_I - a\|_2$. Thus each cluster penalizes deviations in its own $d - j$ coordinates, linking the model to cluster-specific feature selection. This model is natural when coordinates have intrinsic meaning: different clusters may be characterized by different subsets of features, while the selected features remain interpretable. This connection motivated the projected-clustering model of Aggarwal et al. (1999), which was later generalized to arbitrarily oriented cluster-specific subspaces Aggarwal & Yu (2000).

Let $\mathcal{F}_{d,j,k}$ denote the family of all k-tuples $F = (F_1, \ldots, F_k)$ of axis-aligned j-dimensional affine subspaces of $\mathbb{R}^d$. Given a finite set $X \subseteq \mathbb{R}^d$ and a fixed $z \geq 1$, the (j, k, z)-projective clustering problem seeks $F \in \mathcal{F}_{d,j,k}$ that minimizes the clustering cost

$$\phi_F(X) = \sum_{x \in X} \phi_F(x), \qquad \text{where} \quad \phi_F(x) = \min_{1 \leq i \leq k} d(x, F_i)^z,$$

where $d(x, F_i)$ denotes the Euclidean distance between x and F_i.

We consider coresets, which have been a major tool for reducing the computational and storage costs for large datasets. For $\epsilon \in (0, 1)$, an ϵ-coreset is a pair (C, w), where $C \subseteq X$ and $w : C \to \mathbb{R}_{\geq 0}$, that simultaneously preserves the clustering cost for every collection $F = (F_1, \ldots, F_k)$ of axis-aligned j-dimensional affine subspaces:

$$(1 - \epsilon)\phi_F(X) \leq \phi_F(C) = \sum_{y \in C} w(y)\phi_F(y) \leq (1 + \epsilon)\phi_F(X)$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for all F. Such a coreset is called an L_z-coreset. The sensitivity of a point $x \in X$ is its maximum relative contribution to the clustering cost,

$$\sigma_X(x) = \sup_F \frac{\phi_F(x)}{\phi_F(X)},$$

where the supremum ranges over all collections $F = (F_1, \ldots, F_k)$ of axis-aligned j-dimensional affine subspaces. Throughout the paper, a ratio with zero denominator is interpreted as zero whenever its numerator is also zero. The total sensitivity is $S(X) = \sum_{x \in X} \sigma_X(x)$. Sensitivity sampling constructs a coreset by sampling points with probabilities proportional to upper bounds on their sensitivities. Consequently, an upper bound on $S(X)$, together with a complexity bound for the query family, yields an upper bound on the coreset size.

Existing coreset-size bounds for general projective clustering in the regime $j, k \geq 2$ require additional assumptions on the input. When the input points have polynomially bounded integral coordinates, Varadarajan & Xiao (2012b;a) obtained L_1-coresets by using L_∞-coresets[1] to bound the sensitivities. Under the same assumption, Tukan et al. (2022) subsequently improved the dependence of the L_∞-coreset size on d from exponential to polynomial, and used this result to obtain an explicit size bound for the L_2-coreset. For arbitrary real inputs, direct sensitivity bounds and coresets are known only in the special cases $j = 1$ (k-line clustering) and $k = 1$ (subspace approximation) (Varadarajan & Xiao, 2012a). Thus, no comparable small-coreset guarantee is known for arbitrary real inputs in the general $j, k \geq 2$ regime. Moreover, this gap cannot be closed by directly applying the existing L_∞ machinery: there are arbitrary real inputs for which every $1/2$-coreset for the L_∞ objective has size at least $n-2$, already when $j = 1$, $k = 2$ and $d = 2$ (Har-Peled, 2004). In order to decouple the use of coresets and the bounded integral coordinates assumption, it is natural to restrict the solution space of projective clustering.

In this paper, we investigate the following question: Can the computational challenges of projective clustering be significantly reduced by restricting our attention to axis-aligned projective clustering, a special case that already encompasses a sufficiently broad range of practical applications? We give an affirmative answer from the perspective of data reduction. For every fixed $z \geq 1$, we give a direct sensitivity bound, and hence a coreset-size bound, for axis-aligned projective clustering on arbitrary real inputs, including the previously unresolved regime $j, k \geq 2$. Thus, axis alignment removes the bounded-coordinate assumption required by the known guarantees for general projective clustering.

We also study the statistical generalization of projective clustering. For unrestricted linear subspaces, Bucarelli et al. (2023) proved an excess-risk bound of $\widetilde{O}_z(\sqrt{kj^2/n})$ for the general (k, j, z) objective and improved it to $\widetilde{O}(\sqrt{kj/n})$ for $z = 2$, matching their $\Omega(\sqrt{kj/n})$ lower bound up to logarithmic factors. We show that, for axis-aligned linear subspaces on the unit sphere, the expected excess risk is $\widetilde{O}_z(\sqrt{kj/n})$ for every $z \geq 1$ and, in the parameter regime described below, is optimal up to logarithmic factors. We also establish a similar generalization bound for axis-aligned affine subspaces whose distance from the origin is at most 1.

Our Results. Our first result bounds the total sensitivity of axis-aligned projective clustering and gives a randomized coreset construction. For simplicity, we state the coreset bound in the regime $1 \leq j \leq d/2$; the formal results hold for all $1 \leq j < d$.

Theorem 1 (Informal statements of Theorems 4 and 9)**.** *Let $X \subseteq \mathbb{R}^d$ have size $n \geq 2$, let $z \geq 1$ be fixed, and let $\eta_z = \max\{z/2 - 1, 0\}$. When $k = 1$, $S(X) \lesssim_z (j+1)^{1+\eta_z}$. For $k \geq 2$ and $1 \leq r \leq k$, let*

$$\overline{\Lambda}_{r,k}(n) = \max_{1 \leq s \leq r} \left(1 + \frac{\log_+(n/k^s)}{\log(2k/s)}\right).$$

Here $\log_+(x) = \max\{\log x, 0\}$. Then

$$S(X) \leq \begin{cases} (j+1)^{z/2}(C_z k)^{j+1}\overline{\Lambda}_{j,k}(n), & j \leq k, \\ k(j+1)^{k+\eta_z}(C_z k)^k \overline{\Lambda}_{k,k}(n), & j > k, \end{cases}$$

where $C_z > 0$ depends only on z. For $1 \leq j \leq d/2$, sensitivity sampling produces, with probability at least 0.9, an ϵ-coreset whose size is the preceding sensitivity upper bound times $\widetilde{O}_z(\epsilon^{-2}kd)$.

[1] For the L_∞ objective, $\phi_F(X) = \max_{x \in X} \min_i d(x, F_i)$, and an ϵ-coreset $C \subseteq X$ satisfies $\phi_F(X) \leq (1+\epsilon)\phi_F(C)$ for every F.

Regime	Upper bound on $S(X)$	Lower bound on $S(X)$
$j > k$, constant k	$O_{k,z}(j^k \log n)$	$\Omega_k(j^k)$ (Lemma 13) $\Omega_k(j^{k-1} \log(n/j^{k-1}))$ (Lemma 10)
$j \le k$, constant j	$O_{j,z}(k^{j+1}\Lambda_{j,k}(n))$	$\Omega_{j,z}\big(k^{j+1}\Lambda_{j,k}(n)\big)$ (Lemma 12)

Table 1: Sensitivity bounds for $1 \le z \le 2$ and $j, k \ge 2$, for sufficiently large n. All displayed lower bounds hold for every $z \ge 1$, already in the minimum admissible ambient dimension $d = j + 1$. All three lower-bound constructions have aspect ratio $O(n)$; see Section 4. The lower bounds also apply to ϵ-coresets for sufficiently small constant $\epsilon > 0$.

Note that the sensitivity bound is independent of the ambient dimension. In particular, for axis-aligned projective clustering it covers the regime $j, k \ge 2$, beyond the previously known cases $j = 1$ and $k = 1$, and requires neither bounded integer coordinates nor an intermediate L_∞-coreset.

We prove two complementary lower-bound mechanisms in Section 4, both valid for every $z \ge 1$. Lemmas 10 and 12 give two multiscale constructions. The grid construction combines the polynomial dependence on k and the dependence on n on a single instance, matching the strengthened upper bound up to constants for fixed j and z. Lemma 13 uses exact point isolation and shows that total sensitivity and coreset size can be $2^{\Omega(\min\{j,k\})}$. Consequently, for fixed $j, k \ge 2$ and $z \ge 1$, the worst-case total sensitivity over n-point inputs remains $\Theta_{j,k,z}(\log n)$.

For $1 \le r \le k$, define

$$\Lambda_{r,k}(n) = 1 + \frac{\log_+(n/k^r)}{\log(2k/r)}.$$

For fixed j and $k \ge j$, $\overline{\Lambda}_{j,k}(n) = \Theta_j(\Lambda_{j,k}(n))$. Indeed, each $\Lambda_{s,k}(n)$ with $s \le j$ is at most $\Lambda_{j,k}(n) + O_j(1)$, while the reverse inequality is immediate. Together with $G_{j,k} = \Theta_j(k^{j+1})$ and $\log(a_{j,k} + 1) = \Theta_j(\log(k + 1))$, Lemma 12 gives matching bounds in the fixed-j regime.

Table 1 gives the comparison for $1 \le z \le 2$, where $\eta_z = 0$, in the two natural parameter regimes. When $j > k$ and k is constant, the dependence on j is tight up to the $\log n$ factor throughout this range: the upper and lower bounds are $O_{k,z}(j^k \log n)$ and $\Omega_k(j^k)$, respectively.

As a separate contribution, we study statistical generalization on the unit sphere. Let $\mathcal{E}_n^{\mathrm{lin}}$ and $\mathcal{E}_n^{\mathrm{aff}}$ denote the expected excess risks of empirical risk minimization over axis-aligned j-dimensional linear subspaces and axis-aligned j-dimensional affine subspaces within distance 1 of the origin, respectively. Formal definitions and proofs are given in Appendix C.

Theorem 2 (Informal versions of Theorems 20 and 21)**.** *For every constant $\gamma > 0$,*

$$\mathcal{E}_n^{\mathrm{lin}} \lesssim_{z,\gamma} \sqrt{\frac{kj}{n}} \log^{\frac{3}{2}+\gamma}\left(\frac{en}{j}\right),$$
$$\mathcal{E}_n^{\mathrm{aff}} \lesssim_{z,\gamma} \left(\sqrt{\frac{kj}{n}} + \sqrt{\frac{k}{n}} \log^2(en)\right) \log^{\frac{3}{2}+\gamma}\left(\frac{en}{j}\right).$$

We remark that the $\Omega(\sqrt{kj/n})$ lower bound of Bucarelli et al. (2023) uses $d = 2kj$ and a hard instance supported on standard basis vectors, for which the relevant distances are either zero or one and the population and empirical minimizers are coordinate subspaces. Thus, for $d \ge 2kj$ and n sufficiently large relative to kj, the same lower bound applies to every $z \ge 1$ in our axis-aligned setting, and the linear subspace bound in Theorem 2 is optimal up to logarithmic factors.

1.1 Other Related Work

A related research direction is k-clustering of axis-aligned affine subspaces, where now each data point is represented by a j-dimensional affine subspaces and and the task is to find k classical centers in $\mathbb{R}^d$. An equivalent interpretation is clustering with missing values (Wagstaff, 2004), where each data point can contain up to j missing coordinates and the distance to a center will only be evaluated in the existing coordinates. Under this model, approximation algorithms have been proposed for k-center (Gao et al., 2008; 2010; Lee & Schulman, 2013) and k-means (Eiben et al., 2021). Moreover,

coresets of size $(jk)^{O(\min(j,k))}\mathrm{poly}(d\epsilon^{-1}\log n)$ were constructed for k-means with missing values (Braverman et al., 2021).

2 PRELIMINARIES

Notation. For a metric space (X, d), let $\mathrm{diam}(X, d)$ denote its diameter and $N(X, d, \epsilon)$ its ϵ-covering number. If X is finite, its aspect ratio is defined as $\mathrm{diam}(X, d)/\min_{x\neq y}\|x - y\|_2$. For a function ψ and a finite multiset A, write $\psi(A) = \sum_{x\in A}\psi(x)$, where the sum counts multiplicity. For a map π, the image $\pi(A)$ is likewise treated as a multiset, with multiplicities retained.

The next lemma connects the sum of sensitivities of a set to that of its projections and will be used repeatedly.

Lemma 3. *Let $z \geq 1$, let C be a finite multiset in $\mathbb{R}^d$, and let $C' \subseteq C$ be a submultiset partitioned as $C' = C_1 \cup \cdots \cup C_k$. For a map $\pi : C' \to \mathbb{R}^d$, let*

$$\Delta = \sum_{x\in C'}\|x - \pi(x)\|_2^z .$$

Let $\mathcal{Q}$ be a family of nonnegative functions ψ such that $\psi^{1/z}$ is 1-Lipschitz, and suppose that some $\alpha \geq 1$ satisfies $\psi(C) \geq \Delta/\alpha$ for every $\psi \in \mathcal{Q}$. Finally, let

$$A = \sum_{s=1}^{k}\sum_{x\in C_s}\sup_{\psi\in\mathcal{Q}}\frac{\psi(\pi(x))}{\psi(\pi(C_s))},$$

where a ratio with $\psi(\pi(C_s)) = 0$ is interpreted as zero. Then

$$\sum_{x\in C'}\sup_{\psi\in\mathcal{Q}}\frac{\psi(x)}{\psi(C)} \leq 2^{2z-1}\alpha A + 2^{z-1}\alpha.$$

The proof is given in Appendix A.1.

3 UPPER BOUNDS ON SENSITIVITY AND CORESET SIZE

We now establish the sensitivity and coreset guarantees stated in the introduction. The argument first transfers the original instance to the projections induced by a constant-factor solution and then bounds each projected instance through a dimension-free reduction to set clustering. We prove the coordinate and active-coordinate reductions here; the remaining proofs appear in Appendix A.

3.1 FROM GLOBAL TO LOCAL SENSITIVITY

Fix $z \geq 1$, let $\eta_z = \max\{z/2 - 1, 0\}$, and let $X \subset \mathbb{R}^d$ have size $n \geq 2$. We first state the resulting total sensitivity bound. For $0 \leq r \leq k$, define

$$\Lambda_{0,k}(n) = 1, \qquad \Lambda_{r,k}(n) = 1 + \frac{\log_+(n/k^r)}{\log(2k/r)} \quad (1 \leq r \leq k),$$

where $\log_+(x) = \max\{\log x, 0\}$. Also define

$$\overline{\Lambda}_{r,k}(n) = \max_{0\leq s\leq r}\Lambda_{s,k}(n).$$

Theorem 4. *It holds that*

$$S(X) \leq C_{1,z}kR_{j,k}(n),$$

where

$$R_{j,k}(n) = \begin{cases}(j+1)^{1+\eta_z}, & k = 1,\\ \min\{(j+1)^{k+\eta_z}(C_{2,z}k)^r, (j+1)^{z/2}(C_{2,z}k)^j\}\overline{\Lambda}_{r,k}(n), & k \geq 2,\end{cases}$$

with $r = \min\{j, k\}$, and $C_{1,z}, C_{2,z} > 0$ are constants depending only on z.

We first reduce the global sensitivity to local sensitivities. Let $B = (B_1, \dots, B_k)$ be a constant-factor approximate solution with induced clusters $C_1, \dots, C_k$, and let α denote its approximation factor. For $x \in C_s$, let $\pi_s(x)$ be its orthogonal projection onto B_s, retaining multiplicities in $\pi_s(C_s)$. The projection-transfer lemma (Lemma 3) gives

$$S(X) \le 2^{2z-1}\alpha \sum_{s=1}^{k} \sum_{y \in \pi_s(C_s)} \sup_{F \in \mathcal{F}_{d,j,k}} \frac{\phi_F(y)}{\phi_F(\pi_s(C_s))} + 2^{z-1}\alpha. \tag{1}$$

Its proof and the details of this application are given in Appendix A.1. Equation (1) reduces the problem to local sensitivities of multisets contained in a j-dimensional axis-aligned affine subspace. We next reparametrize these local queries in $\mathbb{R}^j$, thereby removing the ambient dimension from the sensitivity analysis.

3.2 Local sensitivity after projection

Motivated by (1), let X now be a multiset contained in an axis-aligned affine j-dimensional subspace, which we identify with $\mathbb{R}^j$.

Consider a query axis-aligned affine subspace F_i. Restricted to points in X, the distance to F_i has the form

$$d(x, F_i) = \sqrt{\|(x - a_i)_{I_i}\|_2^2 + c_i^2}$$

for some index set $I_i \subseteq [j]$ (possibly empty), some vector $a_i \in \mathbb{R}^j$ with $\operatorname{supp}(a_i) \subseteq I_i$ and some constant $c_i \ge 0$.

Thus, every query $F = (F_1, \dots, F_k)$ induces a cost function of the form

$$\psi(x) = \min_{1 \le i \le k} \left(\|(x - a_i)_{I_i}\|_2^2 + c_i^2 \right)^{z/2},$$

where $I_i \subseteq [j]$, $\operatorname{supp}(a_i) \subseteq I_i$ and $c_i \ge 0$ for all $i \in [k]$. For fixed index sets $I_1, \dots, I_k \subseteq [j]$, define $\mathcal{Q}(I_1, \dots, I_k)$ to be the family of all such functions ψ. We also define the worst-case sum of sensitivities to be

$$T(j, k, n) = \sup_{\substack{\text{multiset } X \subset \mathbb{R}^j \\ |X| \le n}} \sum_{x \in X} \max_{I_1, \dots, I_k \subseteq [j]} \sup_{\psi \in \mathcal{Q}(I_1, \dots, I_k)} \frac{\psi(x)}{\psi(X)}.$$

The following local-sensitivity estimate is the technical core of the upper bound.

Lemma 5. *For every $j, k \ge 1$ and $n \ge 2$, it holds that*

$$T(j, k, n) \le C_z R_{j,k}(n),$$

where $C_z > 0$ depends only on z.

To make the source of the two terms in $R_{j,k}(n)$ explicit, we record the two reductions behind Lemma 5.

Lemma 6. *For every multiset $X \subseteq \mathbb{R}^j$, every $I_1, \dots, I_k \subseteq [j]$, every $\psi \in \mathcal{Q}(I_1, \dots, I_k)$, and every $x \in X$,*

$$\frac{\psi(x)}{\psi(X)} \le (j+1)^{\eta_z} \sum_{\substack{J_1, \dots, J_k \subseteq [j] \\ |J_i| \le 1 \ \forall i}} \sup_{f \in \mathcal{Q}(J_1, \dots, J_k)} \frac{f(x)}{f(X)},$$

and

$$\frac{\psi(x)}{\psi(X)} \le (j+1)^{z/2} \max_{\substack{J_1, \dots, J_k \subseteq [j] \\ |J_i| \le 1 \ \forall i}} \sup_{f \in \mathcal{Q}(J_1, \dots, J_k)} \frac{f(x)}{f(X)}.$$

Proof. Let $T_i = I_i \cup \{0\}$, $u_{i,0}(x) = c_i$, and $u_{i,t}(x) = |x_t - (a_i)_t|$ for $t \in I_i$. Then $\psi(x) = \min_{i \in [k]} h_i(x)$, where comparison of the ℓ_2- and ℓ_z-norms gives

$$h_i(x) = \left(\sum_{t \in T_i} u_{i,t}(x)^2 \right)^{z/2} \le (j+1)^{\eta_z} \sum_{t \in T_i} u_{i,t}(x)^z.$$

Applying the elementary inequality

$$\min_{i\in[k]} \sum_{t\in T_i} b_{i,t} \le \sum_{t_1\in T_1,\dots,t_k\in T_k} \min_{i\in[k]} b_{i,t_i} \qquad (b_{i,t}\ge 0)$$

yields

$$\psi(x) \le (j+1)^{\eta_z} \sum_{t_1\in T_1,\dots,t_k\in T_k} \min_{i\in[k]} u_{i,t_i}(x)^z.$$

For each tuple $t=(t_1,\dots,t_k)\in T_1\times\cdots\times T_k$, define

$$J_i(t)=\begin{cases}\emptyset, & t_i=0,\\ \{t_i\}, & t_i\neq 0,\end{cases}$$

and

$$f_t(x) := \min_{i\in[k]} u_{i,t_i}(x)^z \in \mathcal{Q}(J_1(t),\dots,J_k(t)). \tag{2}$$

Since $f_t\le\psi$ and $t\mapsto(J_1(t),\dots,J_k(t))$ is injective,

$$\frac{\psi(x)}{\psi(X)} \le (j+1)^{\eta_z} \sum_{t\in T_1\times\cdots\times T_k} \frac{f_t(x)}{f_t(X)} \le (j+1)^{\eta_z} \sum_{\substack{J_1,\dots,J_k\subseteq[j]\\ |J_i|\le 1\ \forall i}} \sup_{f\in\mathcal{Q}(J_1,\dots,J_k)} \frac{f(x)}{f(X)}.$$

If a denominator vanishes, so does its numerator, and the same argument applies under the zero-denominator convention.

For the second inequality, choose $t_i\in\arg\max_{t\in T_i} u_{i,t}(x)$ for each $i\in[k]$. Then

$$u_{i,t_i}(x)^z \ge (j+1)^{-z/2} h_i(x), \qquad f_t(x)\ge (j+1)^{-z/2}\psi(x).$$

Together with $f_t(X)\le\psi(X)$, this gives the second inequality. □

The first inequality sums over all singleton coordinate patterns. In the second, the maximizing pattern may depend on x. We control both expressions by grouping patterns according to their active coordinate sets. For $H\subseteq[j]$, let $\mathcal{Q}_H$ be the union of $\mathcal{Q}(J_1,\dots,J_k)$ over all $J_1,\dots,J_k\subseteq[j]$ satisfying $|J_i|\le 1$ for every i and $\bigcup_{i=1}^k J_i = H$. The second ingredient is a uniform bound for each active coordinate set. Its proof uses the set-clustering sensitivity bound stated in Lemma 15 (Appendix A.2).

Lemma 7. *Let $X\subseteq\mathbb{R}^j$ be a multiset of size $1\le N\le n$, and let $H\subseteq[j]$ satisfy $|H|\le\min\{j,k\}$. Then*

$$\sum_{x\in X} \sup_{f\in\mathcal{Q}_H} \frac{f(x)}{f(X)} \le C_z (C_z k)^{|H|} \Lambda_{|H|,k}(N),$$

where $C_z>0$ depends only on z.

Proof. Write $H=\{q_1,\dots,q_r\}$, where $r=|H|$, and enumerate $X=\{x_1,\dots,x_N\}$. Set $p_0=(0,0,0)$ and note first that if $r=0$, every $f\in\mathcal{Q}_H$ is constant. The left-hand side is then at most $1=\Lambda_{0,k}(N)$. Hence assume $r\ge 1$.

$$p_s(x)=(x_{q_s},0,s)\in\mathbb{R}^2\times\{0,\dots,r\}, \qquad s\in[r].$$

For $f\in\mathcal{Q}_H$, choose $J_1,\dots,J_k\subseteq[j]$ with $|J_i|\le 1$ for every i and $\bigcup_i J_i = H$ such that $f\in\mathcal{Q}(J_1,\dots,J_k)$, and let $a_1,\dots,a_k$ and $c_1,\dots,c_k$ be its parameters. If $J_i=\{q_s\}$, define

$$b_i=((a_i)_{q_s},c_i,s)\in\mathbb{R}^2\times\{0,\dots,r\}.$$

If some J_i is empty, let $\gamma=\min_{i:J_i=\emptyset} c_i^z$ and $b_0=(\gamma^{1/z},0,0)$.

On $\mathbb{R}^2\times\{0,\dots,r\}$, define

$$d_M((u,s),(v,t))=\begin{cases}\|u-v\|_2, & s=t,\\ \|u\|_2+\|v\|_2+M, & s\neq t.\end{cases}$$

Let $U = \{p_0\}$, $P_\ell = U \cup \{p_s(x_\ell) : s \in [r]\}$, and $C_f = \{b_i : J_i \neq \emptyset\}$, including b_0 in C_f if some J_i is empty. Then $|P_\ell \setminus U| = r$ and $|C_f| \leq k$.

Choose $M > \max_{\ell \in [N]} \max_{q \in [j]} |(x_\ell)_q|$. If $J_i = \{q_s\}$, then for every $t \neq s$,

$$d_M(p_s(x_\ell), b_i) \leq |(x_\ell)_{q_s}| + \|((a_i)_{q_s}, c_i)\|_2 < M + \|((a_i)_{q_s}, c_i)\|_2 \leq d_M(p_t(x_\ell), b_i).$$

The same comparison holds with $p_t(x_\ell)$ replaced by p_0. Moreover, if b_0 is present, then $d_M(p_0, b_0) < d_M(p_s(x_\ell), b_0)$ for every s. Consequently,

$$\widetilde{d}_M(P_\ell, C_f) := \min_{p \in P_\ell,\, c \in C_f} d_M(p, c)^z = f(x_\ell)$$

for every $\ell \in [N]$ and every $f \in \mathcal{Q}_H$. Applying Lemma 15 with $K = k$ and $m = r$ gives

$$\sum_{\ell=1}^{N} \sup_{f \in \mathcal{Q}_H} \frac{f(x_\ell)}{f(X)} \leq \sum_{\ell=1}^{N} \sup_{1 \leq |C| \leq k} \frac{\widetilde{d}_M(P_\ell, C)}{\sum_{h=1}^{N} \widetilde{d}_M(P_h, C)} \leq C_{1,z}(C_{2,z}k)^r \Lambda_{r,k}(N),$$

which proves the claim. □

These reductions also explain the minimum in $R_{j,k}(n)$. Put $r_0 = \min\{j, k\}$. For $k \geq 2$, summing the first inequality of Lemma 6 over all $(j+1)^k$ patterns gives

$$T(j, k, n) \leq C_z (j+1)^{k+\eta_z} (C_z k)^{r_0} \overline{\Lambda}_{r_0,k}(n).$$

Alternatively, the second inequality and grouping by H give

$$\begin{aligned} T(j, k, n) &\leq C_z (j+1)^{z/2} \overline{\Lambda}_{r_0,k}(n) \sum_{r=0}^{r_0} \binom{j}{r} (C_z k)^r \\ &\leq C'_z (j+1)^{z/2} (C'_z k)^j \overline{\Lambda}_{r_0,k}(n). \end{aligned}$$

These are precisely the two estimates in Lemma 5; their complete derivation, including the case $k = 1$, appears in Appendix A.3. Applying the lemma to the projected clusters in (1) gives Theorem 4; see Appendix A.4 for the details.

3.3 From sensitivity to coreset size

It remains to control the complexity of the threshold ranges used by sensitivity sampling. For non-negative point weights $\lambda = (\lambda_x)_{x \in X}$, define

$$\mathcal{A}_\lambda = \{\{x \in X : \lambda_x \phi_F(x) \geq r\} : F \in \mathcal{F}_{d,j,k},\ r \geq 0\}. \tag{3}$$

Given upper bounds $s_x \geq \max\{1/|X|, \sigma_X(x)\}$ and $S = \sum_x s_x$, the sensitivity-sampling theorem of Feldman et al. (2020, Theorem 31) produces an ϵ-coreset of size

$$O\left(\frac{S+1}{\epsilon^2}\left(D \log(S+1) + \log\frac{1}{\delta}\right)\right),$$

provided that D uniformly bounds $\mathrm{VCdim}(\mathcal{A}_\lambda)$. It therefore remains only to establish the following VC-dimension estimate.

Lemma 8. *For every $\lambda = (\lambda_x)_{x \in X} \in \mathbb{R}^X_{\geq 0}$, let $\mathcal{A}_\lambda$ be as defined in* (3)*. Then*

$$\mathrm{VCdim}(\mathcal{A}_\lambda) = O\left((k(d-j)+1)\log(2k) + kj \log\frac{ed}{j}\right).$$

To see the two contributions to this bound, fix the coordinate pattern of the k subspaces. The offsets and threshold give $k(d-j)+1$ real parameters, and membership in a range is a conjunction of k quadratic inequalities, contributing $O((k(d-j)+1)\log(2k))$. Taking the union over the $\binom{d}{j}^k$ coordinate patterns adds $O(kj \log(ed/j))$. The formal argument appears in Appendix A.5. Combining the lemma with Theorem 4 and the preceding sampling theorem gives the final coreset-size bound. For brevity, let

$$D_{d,j,k} = (d-j+1)\log(2k) + j \log\frac{ed}{j},$$

$$L_{j,k}(n) = \begin{cases} \log(j+1), & k = 1, \\ \min\{j,k\}\log(j+k) + \log(2\log n), & k \geq 2. \end{cases}$$

Theorem 4 implies that $\log(S(X)+1) = O_z(L_{j,k}(n))$.

Theorem 9. *Suppose that $\epsilon, \delta \in (0, 1/2)$. There exists a randomized algorithm that, with probability at least $1 - \delta$, outputs an ϵ-coreset for axis-aligned (j, k, z)-projective clustering of size*

$$O_z\left(\frac{kR_{j,k}(n)}{\epsilon^2}\left(kD_{d,j,k}L_{j,k}(n) + \log\frac{1}{\delta}\right)\right).$$

The sampling result requires computable pointwise upper bounds, not merely the value of their sum. These are obtained by following the same reductions: start from the approximate solution B, project each induced cluster, and apply the iterative set-clustering routine to every reduced local instance. Each iteration identifies a certified subset of points and assigns pointwise upper bounds; tracing these values back through coordinate reduction produces scores $s_x \geq \max\{1/n, \sigma_X(x)\}$ satisfying

$$\sum_{x \in X} s_x \leq C_{1,z} k R_{j,k}(n).$$

Sampling proportionally to these scores then gives the stated guarantee. The construction is detailed in Appendix A.6. Notice that the total-sensitivity term is independent of d; the ambient dimension enters the coreset size only through the VC-dimension factor. We next show that the dependence on n, j, and k is unavoidable in several parameter regimes.

4 Lower Bounds on Sensitivity and Coreset Size

We complement the upper bounds with three lower bounds, all valid for every $z \geq 1$. All lower bounds already work in the minimum admissible ambient dimension $d = j + 1$. All proofs are deferred to Appendix B.

Lemma 10. *Let $j, k \geq 2$ and $z \geq 1$. For $q \in [j-1]$, let*

$$B(j, k; q) \coloneqq q\binom{j - q + k - 3}{k - 2}.$$

Suppose that $\epsilon \in (0, 1/32)$. For every $n \geq 64B(j, k; q)$, there exists a set $X \subset \mathbb{R}^{j+1}$ with $|X| = n$ such that its total sensitivity and the size of every ϵ-coreset for axis-aligned (j, k, z)-projective clustering of X are at least

$$\Omega\left(B(j, k; q) \log\frac{n}{B(j, k; q)}\right).$$

Moreover, X can be chosen such that the aspect ratio of X is at most $4n$.

Proof sketch. For notational simplicity, let $B = B(j, k; q)$ and $L = \lfloor n/B \rfloor$. Up to padding points, the instance is the union of B blocks, each containing L points at geometrically increasing scales. The blocks are indexed by a choice of one of q distinguished coordinates and a weak composition of $k - 2$ among the remaining $j - q$ coordinates, giving exactly B possibilities. The coordinates are chosen so that a query can isolate any one block. More precisely, the clipping hyperplane is placed on one of the composition coordinates, leaving the last coordinate free; we use this last coordinate to assign distinct, evenly spaced labels to all input points. This does not change any of the witness costs. Within the selected block, denote the point at scale $2^{t/z}$ by x_t and the query at scale $2^{s/z}$ by F_s. Taking the zth power of the distance gives

$$\phi_{F_s}(x_t) = \min\{2^s, 2^t\}, \qquad \phi_{F_s}(y) = 0 \quad \text{outside the selected block.}$$

Using F_t for x_t gives

$$\sigma_X(x_t) \geq \frac{1}{L - t + 2}.$$

Summing over t yields $\Omega(\log L)$ total sensitivity per block; approximating all of these truncations also forces a weighted coreset to retain $\Omega(\log L)$ points from each block. □

Remark 11. The parameter q interpolates between the dependence on j and k. Optimizing it gives the following useful specialization. One can compute that $q^\star = \lceil (j-1)/(k-1) \rceil$ is a maximizer of $B(j, k; q)$. Thus $q^\star = 1$ when $k > j$, while $q^\star = \Theta(j/k)$ when $k \leq j$. In particular, for constant k, $B(j, k; q^\star) = \Theta_k(j^{k-1})$, so Lemma 10 gives the lower bound

$$\Omega_k\left(j^{k-1} \log\frac{n}{j^{k-1}}\right)$$

whenever $n \geq 64B(j, k; q^\star)$.

Lemma 12. *Let $j, k \geq 2$, $z \geq 1$, and $\epsilon \in (0, 1/4)$. Define $a_{j,k} = 1 + \lfloor (k-2)/(j+2) \rfloor$ and $G_{j,k} = a_{j,k}^{j+1}$. For every $n \geq \max\{2, G_{j,k}\}$, there exists a set $X \subset \mathbb{R}^{j+1}$ with $|X| = n$ such that its total sensitivity and the size of every ϵ-coreset for axis-aligned (j, k, z)-projective clustering of X are at least*

$$\Omega_{j,z}\left(G_{j,k}\left[1 + \frac{\log(n/G_{j,k})}{\log(a_{j,k}+1)}\right]\right).$$

Moreover, X can be chosen such that the aspect ratio of X is at most $O(n)$.

Proof sketch. Let $a = a_{j,k}$ and index $B = a^{j+1}$ blocks by $[a]^{j+1}$. We construct BT groups $X_{p,s}$, indexed by an address $p \in [a]^{j+1}$ and a scale $s \in [T]$. The first j coordinates encode the first j entries of p, while the last coordinate encodes p_{j+1} and s. As s increases, the groups contain fewer points but occur at larger distance scales. For every $X_{p,s}$, a collection of $(j+2)(a-1) + 2 \leq k$ axis-aligned hyperplanes defines a query $F_{p,s}$ that singles out this group. The coordinate hyperplanes give zero cost to groups whose addresses differ from p in one of its first j entries, and the remaining hyperplanes control the groups at other scales. The parameters are chosen so that $X_{p,s}$ contributes much larger to $F_{p,s}$ than all other groups together. Therefore every group contributes $\Omega(1)$ to the total sensitivity, resulting in a lower bound of $\Omega(BT)$ for total sensitivity. Choosing $T \asymp_z (\log(n/B)/\log(a+1)) + 1$ gives the claimed bound. □

For fixed j and k, Lemma 10 shows that the $\log n$ dependence in the upper bound is unavoidable. For fixed j and growing k, Lemma 12 has

$$G_{j,k} = \Theta_j(k^{j+1}), \qquad \log(a_{j,k}+1) = \Theta_j(\log(k+1)),$$

and therefore combines the polynomial block count and the dependence on n on a single instance. We next obtain complementary dependence on j and k through exact point isolation, which depends only on whether a distance is zero or positive.

Lemma 13. *Let $j, k \geq 2$, $z \geq 1$, and $\epsilon \in (0, 1)$, and let*

$$t = \left\lfloor \frac{k-1}{j+1} \right\rfloor, \qquad r = \min\left\{k, \left\lfloor \frac{j+1}{2} \right\rfloor\right\}, \qquad N_{j,k} = \max\left\{\binom{j+t+1}{j+1}, \binom{j+1}{r}\right\}.$$

For every $n \geq N_{j,k}$, there exists a set $X \subset \mathbb{R}^{j+1}$ with $|X| = n$ such that its total sensitivity and the size of every ϵ-coreset for axis-aligned (j, k, z)-projective clustering of X are at least $N_{j,k}$. Furthermore, X can be chosen such that the aspect ratio of X is $O(n)$.

Proof sketch. The common principle behind the two binomials is point isolation. Each construction produces a distinguished subset $X_0 \subseteq X$ such that, for every $x \in X_0$, there is a query $F_x \in \mathcal{F}_{j+1,j,k}$ whose union contains every point of $X \setminus \{x\}$ but not x. Consequently,

$$\phi_{F_x}(X) = \phi_{F_x}(x) > 0,$$

so $\sigma_X(x) = 1$, and every ϵ-coreset with $\epsilon < 1$ must contain x, independently of z. Two subset encodings realize this principle. The first maps each $(j+1)$-subset $\{p_1 < \cdots < p_{j+1}\} \subseteq [j+t+1]$ to $(p_1, \ldots, p_{j+1}) \in \mathbb{R}^{j+1}$. It uses $(j+1)t$ isolating hyperplanes, together with one hyperplane for padding points, and yields $\binom{j+t+1}{j+1}$ isolated points. The second maps each r-subset $C \subseteq [j+1]$ to its incidence vector and isolates it using the hyperplanes $y_c = 0$ for $c \in C$, together with one padding hyperplane when $r < k$. This yields $\binom{j+1}{r}$ isolated points. These encodings dominate in complementary regimes. For fixed j and growing k, the first term in $N_{j,k}$ is $\Theta_j(k^{j+1})$; for fixed k and growing j, the second is $\Theta_k(j^k)$. Once $k \geq \lfloor (j+1)/2 \rfloor$, the latter becomes a central binomial coefficient and is $2^{\Theta(j)}$. □

AI USE STATEMENT

The authors used ChatGPT 5.5, 5.6 Sol and 6 Pro as auxiliary tools for improving exposition, checking algebraic and logical steps in proofs, and suggesting local simplifications or corrections. In particular, GPT 6 Pro pointed out the current Lemmata 14 and 15, which improves Lemma 7 by replacing a $\log n$ factor with $\Lambda_{|H|,k}(n)$, as well as the corresponding lower bound Lemma 12. All mathematical results and final proofs were independently checked and approved by the authors, who take full responsibility for the manuscript.

REFERENCES

Charu C. Aggarwal and Philip S. Yu. Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 70–81, 2000.

Charu C. Aggarwal, Cecilia Procopiuc, Joel L. Wolf, Philip S. Yu, and Jong Soo Park. Fast algorithms for projected clustering. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, pp. 61–72, 1999.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999. ISBN 978-0-521-57353-5. doi: 10.1017/CBO9780511624216. Digitally reprinted 2009.

Vladimir Braverman, Shaofeng Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering with missing values. *Advances in Neural Information Processing Systems*, 34:17360–17372, 2021.

Maria Sofia Bucarelli, Matilde Fjeldsø Larsen, Chris Schwiegelshohn, and Mads Bech Toftrup. On generalization bounds for projective clustering. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. doi: 10.52202/075280-3140. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e30bf4765ae6b16a87fb4d7b0b3b3dec-Abstract-Conference.html.

Eduard Eiben, Fedor V Fomin, Petr A Golovach, William Lochet, Fahad Panolan, and Kirill Simonov. Eptas for k-means clustering of affine subspaces. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 2649–2659. SIAM, 2021.

Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. *SIAM Journal on Computing*, 49(3): 601–657, 2020. doi: 10.1137/18M1209854.

Dylan J. Foster and Alexander Rakhlin. ℓ_∞ vector contraction for Rademacher complexity. arXiv:1911.06468 [cs.LG], 2019. URL https://arxiv.org/abs/1911.06468.

Jie Gao, Michael Langberg, and Leonard J. Schulman. Analysis of incomplete data and an intrinsic-dimension helly theorem. *Discrete Comput. Geom.*, 40(4):537–560, October 2008. ISSN 0179-5376.

Jie Gao, Michael Langberg, and Leonard J Schulman. Clustering lines in high-dimensional space: Classification of incomplete data. *ACM Transactions on Algorithms (TALG)*, 7(1):1–26, 2010.

Steve Hanneke and Mingyue Xu. Universal rates of empirical risk minimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 116291–116331, 2024.

Sariel Har-Peled. No, coreset, no cry. In Kamal Lodaya and Meena Mahajan (eds.), *FSTTCS 2004: Foundations of Software Technology and Theoretical Computer Science, 24th International Conference, Chennai, India, December 16-18, 2004, Proceedings*, volume 3328 of *Lecture Notes in Computer Science*, pp. 324–335. Springer, 2004.

Ibrahim Jubran, Murad Tukan, Alaa Maalouf, and Dan Feldman. Sets clustering. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4994–5005. PMLR, 2020. URL https://proceedings.mlr.press/v119/jubran20a.html.

Euiwoong Lee and Leonard J Schulman. Clustering affine subspaces: hardness and algorithms. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pp. 810–827. SIAM, 2013.

Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training Gaussian mixture models at scale via coresets. *Journal of Machine Learning Research*, 18(160):1–25, 2018. URL http://jmlr.org/papers/v18/15-506.html.

Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in euclidean space. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pp. 1064–1069, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059. doi: 10.1145/3313276.3316307. URL https://doi.org/10.1145/3313276.3316307.

Cecilia M. Procopiuc. Projective clustering. In Claude Sammut and Geoffrey I. Webb (eds.), *Encyclopedia of Machine Learning*, pp. 806–811. Springer, Boston, MA, 2010. doi: 10.1007/978-0-387-30164-8_686.

Murad Tukan, Xuan Wu, Samson Zhou, Vladimir Braverman, and Dan Feldman. New coresets for projective clustering and applications. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5391–5415. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/tukan22a.html.

Kasturi R. Varadarajan and Xin Xiao. On the sensitivity of shape fitting problems. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2012)*, volume 18 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 486–497. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2012a. doi: 10.4230/LIPIcs.FSTTCS.2012.486.

Kasturi R. Varadarajan and Xin Xiao. A near-linear algorithm for projective clustering integer points. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1329–1342. SIAM, 2012b. doi: 10.1137/1.9781611973099.105.

Kiri Wagstaff. Clustering with missing values: No imputation required. In *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*, pp. 649–658. Springer, 2004.

A Proofs for the Upper Bounds

This appendix supplies the remaining proofs for Section 3. We first prove the projection-transfer lemma and state the supporting set-clustering bounds, then complete the sensitivity, VC-dimension, and coreset arguments.

A.1 Proof of Lemma 3

Proof. Let $d_x = \|x - \pi(x)\|_2$ for $x \in C'$ and let $c_z = 2^{z-1}$. If $\Delta = 0$, then $d_x = 0$ for every $x \in C'$ and hence $\pi(x) = x$. For $x \in C_r$, nonnegativity gives

$$\psi(C) \geq \psi(C_r) = \psi(\pi(C_r)).$$

Thus

$$\frac{\psi(x)}{\psi(C)} \leq \frac{\psi(\pi(x))}{\psi(\pi(C_r))}.$$

Taking the supremum over ψ and summing over $x \in C'$ gives a bound A, which is stronger than the desired result.

Assume now that $\Delta > 0$. Fix $\psi \in \mathcal{Q}$ and let $P = \psi(\pi(C'))$. Since $\psi^{1/z}$ is 1-Lipschitz,

$$\psi(x) \leq c_z(\psi(\pi(x)) + d_x^z), \qquad \psi(\pi(x)) \leq c_z(\psi(x) + d_x^z), \qquad x \in C'. \tag{4}$$

Summing the second inequality and using $\Delta \leq \alpha\psi(C)$ gives

$$P \leq c_z\big(\psi(C') + \Delta\big) \leq c_z(1+\alpha)\psi(C) \leq 2c_z\alpha\psi(C).$$

If $P > 0$, the first inequality in (4), together with the preceding bound and $\psi(C) \geq \Delta/\alpha$, yields

$$\frac{\psi(x)}{\psi(C)} \leq 2c_z^2\alpha\frac{\psi(\pi(x))}{P} + c_z\alpha\frac{d_x^z}{\Delta}. \tag{5}$$

If $P = 0$, then $\psi(\pi(x)) = 0$ for every $x \in C'$, and the same inequality, with its first term interpreted as zero, follows directly from

$$\frac{\psi(x)}{\psi(C)} \le c_z \alpha \frac{d_x^z}{\Delta}.$$

Taking the supremum over ψ and summing over $x \in C'$, we obtain

$$\sum_{x \in C'} \sup_{\psi \in \mathcal{Q}} \frac{\psi(x)}{\psi(C)} \le 2c_z^2 \alpha \sum_{s=1}^{k} \sum_{x \in C_s} \sup_{\psi \in \mathcal{Q}} \frac{\psi(\pi(x))}{\psi(\pi(C_s))} + c_z \alpha,$$

where we used $P = \psi(\pi(C')) \ge \psi(\pi(C_s))$ for $x \in C_s$. This is the claimed bound. □

A.2 Set-clustering sensitivity bounds

We first prove a deterministic batch bound that certifies all but an $m/(2K)$ fraction of the input sets at once. The formulation uses indexed candidate lists, whose entries need not be distinct; this avoids ambiguity when a candidate promoted to the common set coincides with another entry.

Lemma 14. *Let $z \ge 1$, $K, N \ge 1$, and $0 \le m \le K$. Let U be a nonempty finite subset of a metric space $(\mathcal{X}, d)$. For each $i \in [N]$, let $A_i = (a_{i,1}, \dots, a_{i,m})$ be an indexed list of candidates and let*

$$P_i = U \cup \{a_{i,1}, \dots, a_{i,m}\}.$$

For $P, C \subseteq \mathcal{X}$, write

$$\widetilde{d}(P, C) = \min_{p \in P, c \in C} d(p, c)^z.$$

Then there are a set $G \subseteq [N]$ and nonnegative reals $\{u_i\}_{i \in G}$ such that

$$|[N] \setminus G| \le \frac{m}{2K} N,$$

for every $i \in G$ and every $C \subseteq \mathcal{X}$ with $1 \le |C| \le K$,

$$\frac{\widetilde{d}(P_i, C)}{\sum_{h=1}^{N} \widetilde{d}(P_h, C)} \le u_i,$$

and

$$\sum_{i \in G} u_i \le (C_z K)^m, \tag{6}$$

where $C_z > 0$ depends only on z. Moreover, G and the scores u_i can be computed by finite enumeration.

Proof. We prove the claim by induction on m. If $m = 0$, take $G = [N]$ and $u_i = 1/N$. This is valid under the zero-denominator convention, and (6) holds with equality.

Suppose that $m \ge 1$ and set $\ell = \lfloor N/(2K) \rfloor$. Among all nonempty sets $B \subseteq \bigcup_i \{a_{i,1}, \dots, a_{i,m}\}$ with $|B| \le K$ and all $I \subseteq [N]$ with $|I| = N - \ell$, choose a pair minimizing

$$\Delta = \sum_{i \in I} d(A_i, B)^z,$$

where distances involving a candidate list mean the minimum over its entries. For every $i \in I$, choose an occurrence a_i in A_i and $b_i \in B$ realizing $d(A_i, B)$, and let $I_b = \{i \in I : b_i = b\}$. The nonempty sets I_b partition I.

Fix $C \subseteq \mathcal{X}$ with $1 \le |C| \le K$ and write

$$\gamma = d(U, C)^z, \qquad d_i = \widetilde{d}(P_i, C) = \min\{\gamma, d(A_i, C)^z\}, \qquad D = \sum_{i=1}^{N} d_i.$$

If $D = 0$, every relevant ratio is zero, so assume $D > 0$. There are two cases.

If more than ℓ indices satisfy $d(A_i, C)^z > \gamma$, then at least $\ell + 1$ summands of D equal γ, and hence

$$\frac{d_i}{D} \le \frac{1}{\ell + 1}$$

for every i.

Otherwise, at most ℓ indices satisfy $d(A_i, C)^z > \gamma$. Choose $J \subseteq [N]$ of size $N - \ell$ among the remaining indices, and assign every A_i, $i \in J$, to a nearest center $c \in C$. For every nonempty cluster J_c, choose $h_c \in J_c$ minimizing $d(A_{h_c}, c)$, choose b_c in A_{h_c} closest to c, and set $B' = \{b_c : J_c \neq \emptyset\}$. For $i \in J_c$, the triangle inequality gives

$$d(A_i, B')^z \le 2^{z-1}\big(d(A_i, C)^z + d(A_{h_c}, c)^z\big).$$

Since $|J_c| d(A_{h_c}, c)^z \le \sum_{i \in J_c} d(A_i, C)^z$, summing first over each cluster and then over c gives

$$\sum_{i \in J} d(A_i, B')^z \le 2^z \sum_{i \in J} d(A_i, C)^z.$$

The minimality of (B, I) and the choice of J therefore imply

$$\Delta \le 2^z D. \tag{7}$$

For each nonempty I_b and $i \in I_b$, delete the chosen occurrence a_i from the list A_i, enlarge the common set to $U_b = U \cup \{b\}$, and let P_i' be the union of U_b and the remaining $m - 1$ entries. Thus the induction hypothesis applies to $(P_i')_{i \in I_b}$. Put

$$d_i' = \widetilde{d}(P_i', C), \qquad \delta_i = d(a_i, b)^z.$$

The two sets P_i and P_i' have Hausdorff distance at most $\delta_i^{1/z}$, so

$$|d_i^{1/z} - (d_i')^{1/z}| \le \delta_i^{1/z},$$

and consequently

$$d_i \lesssim_z d_i' + \delta_i, \qquad d_i' \lesssim_z d_i + \delta_i.$$

Let $D_b' = \sum_{i \in I_b} d_i'$. Since $\sum_{i \in I} \delta_i = \Delta$, (7) yields

$$D_b' \lesssim_z D + \Delta \lesssim_z D$$

in the second case. It follows that

$$\frac{d_i}{D} \lesssim_z \frac{d_i'}{D_b'} + \frac{\delta_i}{\Delta}.$$

Here and below a zero first denominator gives a zero ratio, and δ_i / Δ is defined as zero when $\Delta = 0$. In the latter case every $\delta_i = 0$ and $P_i = P_i'$, so the displayed inequality remains valid.

Combining the two cases, there is a constant $\kappa_z \ge 1$ depending only on z such that, for every query C,

$$\frac{d_i}{D} \le \kappa_z \left(\frac{1}{\ell + 1} + \frac{d_i'}{D_b'} + \frac{\delta_i}{\Delta} \right), \qquad i \in I_b.$$

Apply the induction hypothesis independently to every nonempty I_b. It gives $G_b \subseteq I_b$ and scores u_i' satisfying

$$|I_b \setminus G_b| \le \frac{m - 1}{2K} |I_b|, \qquad \sum_{i \in G_b} u_i' \le (C_z K)^{m-1}.$$

For $i \in G_b$, define

$$u_i = \kappa_z \left(\frac{1}{\ell + 1} + u_i' + \frac{\delta_i}{\Delta} \right),$$

and let $G = \bigcup_{b \in B} G_b$. The preceding pointwise inequality shows that these scores are valid. Moreover,

$$|[N] \setminus G| \le \ell + \frac{m - 1}{2K}(N - \ell) \le \frac{m}{2K} N,$$

and

$$\sum_{i \in G} u_i \le \kappa_z \left(\frac{N}{\ell + 1} + K (C_z K)^{m-1} + 1 \right) \le (C_z K)^m$$

for a sufficiently large C_z, using $N/(\ell + 1) \le 2K$. All minimizations and recursive choices range over finite sets, so this also proves the computability assertion. □

Iterating this result gives the total-sensitivity bound used in the proof of Lemma 7.

Lemma 15. *Under the assumptions of Lemma 14, for $1 \le m \le K$ there are computable scores $v_1, \dots, v_N$ such that*

$$\frac{\widetilde{d}(P_i, C)}{\sum_{h=1}^{N} \widetilde{d}(P_h, C)} \le v_i$$

for every $i \in [N]$ and every C with $1 \le |C| \le K$, and

$$\sum_{i=1}^{N} v_i \le (C_z K)^m \left(1 + \frac{\log_+(N/K^m)}{\log(2K/m)}\right).$$

Proof. Apply Lemma 14 repeatedly to the indices that remain uncertified. At each round their number decreases by a factor at most $m/(2K)$, and the returned scores have total mass at most $(C_z K)^m$. A score valid relative to the current subinstance is also valid relative to the original instance, whose denominator is no smaller.

Stop once at most K^m indices remain and assign score 1 to each of them. The number of batch rounds is at most

$$1 + \frac{\log_+(N/K^m)}{\log(2K/m)}.$$

Summing all batch scores and the residual scores proves the stated bound after enlarging C_z. The construction is computable because every batch is computable. □

A.3 Proof of Lemma 5

Proof. If $k = 1$, Lemma 6 reduces the query family to $J = \emptyset$ and $J = \{q\}$ for $q \in [j]$. In the latter case every $f \in \mathcal{Q}(J)$ is a z-th power of a Euclidean distance to one center in $\mathbb{R}^2$, and therefore has $O_z(1)$ total sensitivity. In the former case f is constant and has total sensitivity at most 1. Summing over the $j + 1$ choices gives

$$T(j, 1, n) \le C_z (j+1)^{1+\eta_z}.$$

Suppose that $k \ge 2$. The first inequality in Lemma 6 gives

$$T(j, k, n) \le (j+1)^{\eta_z} \sum_{\substack{J_1, \dots, J_k \subseteq [j] \\ |J_i| \le 1 \ \forall i}} \sup_{\substack{\text{multiset } X \subseteq \mathbb{R}^j \\ |X| \le n}} \sum_{x \in X} \sup_{f \in \mathcal{Q}(J_1, \dots, J_k)} \frac{f(x)}{f(X)}.$$

Set $r_0 = \min\{j, k\}$. Each tuple $(J_1, \dots, J_k)$ has an active coordinate set H of size at most r_0. There are $(j+1)^k$ such tuples, so Lemma 7 implies

$$T(j, k, n) \le C_z (j+1)^{k+\eta_z} (C_z k)^{r_0} \overline{\Lambda}_{r_0, k}(n).$$

For the second estimate, the second coordinate-reduction inequality and grouping patterns by their active set give

$$\begin{aligned} T(j, k, n) &\le (j+1)^{z/2} \sup_{\substack{\text{multiset } X \subseteq \mathbb{R}^j \\ |X| \le n}} \sum_{x \in X} \sum_{\substack{H \subseteq [j] \\ |H| \le \min\{j,k\}}} \sup_{f \in \mathcal{Q}_H} \frac{f(x)}{f(X)} \\ &\le C_{1,z} (j+1)^{z/2} \overline{\Lambda}_{r_0, k}(n) \sum_{r=0}^{r_0} \binom{j}{r} (C_{2,z} k)^r \\ &\le C'_{1,z} (j+1)^{z/2} (C'_{2,z} k)^j \overline{\Lambda}_{r_0, k}(n). \end{aligned}$$

Taking the smaller of the two estimates proves the lemma. □

A.4 Proof of Theorem 4

Proof. Let $B = (B_1, \dots, B_k)$ be a constant-factor approximate solution with approximation factor α, and let $C_1, \dots, C_k$ be its induced clusters. For $x \in C_s$, write $\pi_s(x)$ for its orthogonal projection onto B_s. Then

$$\Delta := \sum_{s=1}^{k} \sum_{x \in C_s} \|x - \pi_s(x)\|_2^z = \phi_B(X) \le \alpha \,\mathrm{OPT} \le \alpha \phi_F(X) \qquad (F \in \mathcal{F}_{d,j,k}).$$

Moreover, $x \mapsto \phi_F(x)^{1/z}$ is 1-Lipschitz. Apply Lemma 3 with $C = C' = X$, $\pi(x) = \pi_s(x)$ for $x \in C_s$, and $\mathcal{Q} = \{\phi_F : F \in \mathcal{F}_{d,j,k}\}$. This gives (1).

Each multiset $\pi_s(C_s)$ lies in an axis-aligned affine j-dimensional subspace and has at most n points. Hence Lemma 5 bounds its local sum of sensitivities by $C_z R_{j,k}(n)$. Summing over the k clusters in (1) and absorbing the constants depending only on z proves the theorem. □

A.5 Proof of Lemma 8

To prove Lemma 8, we need a result on the VC dimension of Boolean combinations of polynomial inequalities. The form we use here is an immediate corollary of (Anthony & Bartlett, 1999, Theorem 8.3).

Theorem 16 (Corollary of (Anthony & Bartlett, 1999, Theorem 8.3))**.** *Let $\mathcal{P} \subseteq \mathbb{R}^D$ and $\mathcal{H} = \{H_\theta : \theta \in \mathcal{P}\}$ be a family of subsets of a domain Q. Let maps $p_1, \dots, p_s : \mathbb{R}^D \times Q \to \mathbb{R}$ be such that, for every $x \in Q$, each $p_t(\cdot, x)$ is a polynomial of degree at most ℓ. Suppose there exists a Boolean function $B : \{0,1\}^s \to \{0,1\}$ such that $x \in H_\theta$ if and only if*

$$B\big(\mathbf{1}_{\{p_1(\theta,x)\ge 0\}}, \dots, \mathbf{1}_{\{p_s(\theta,x)\ge 0\}}\big) = 1.$$

Then

$$\mathrm{VCdim}(\mathcal{H}) \le 2D \log_2(12 s \ell).$$

We are now ready to prove Lemma 8.

Proof of Lemma 8. Recall that $\phi_F(x) = \min_{1\le i\le k} d(x, F_i)^z$, where each F_i is an axis-aligned j-dimensional affine subspace. Thus, for some $I_i \subseteq [d]$ with $|I_i| = d - j$ and $a_i \in \mathbb{R}^{I_i}$, $d(x, F_i)^2 = \|x_{I_i} - a_i\|_2^2$.

Fix $\lambda = (\lambda_x)_{x\in X} \in \mathbb{R}^X_{\ge 0}$ and $\mathcal{I} = (I_1, \dots, I_k)$. Define

$$\mathcal{A}_{\lambda,\mathcal{I}} = \{\{x \in X : \lambda_x \phi_F(x) \ge r\} : F \in \mathcal{F}_{d,j,k} \text{ and is consistent with } \mathcal{I},\ r \ge 0\}.$$

The affine subspaces and the threshold are together parameterized by

$$\mathcal{P}_{\mathcal{I}} = \prod_{i=1}^{k} \mathbb{R}^{I_i} \times \mathbb{R}_{\ge 0} \subseteq \mathbb{R}^D, \qquad \text{where} \quad D = k(d-j) + 1.$$

Reparameterize the threshold by $\rho = r^{1/z}$. For each $\theta = (a_1, \dots, a_k, \rho) \in \mathcal{P}_{\mathcal{I}}$, define

$$p_i(\theta, x) = \lambda_x^{2/z} \|x_{I_i} - a_i\|_2^2 - \rho^2, \quad i \in [k].$$

Then $\lambda_x \phi_F(x) \ge r$ if and only if $p_t(\theta, x) \ge 0$ for every $t \in [k]$. For each fixed x, $p_i(\theta, x)$ is a polynomial of degree 2. We apply Theorem 16 with $Q = X$, $\mathcal{P} = \mathcal{P}_{\mathcal{I}}$, $s = k$, $\ell = 2$ and B equal to the conjunctive function, and we obtain that

$$\mathrm{VCdim}(\mathcal{A}_{\lambda,\mathcal{I}}) = O(D \log(2k)).$$

There are $M = \binom{d}{j}^k$ possible patterns $\mathcal{I}$ and $\mathcal{A}_\lambda = \bigcup_{\mathcal{I}} \mathcal{A}_{\lambda,\mathcal{I}}$. Hence, by the standard finite union bound (see, e.g., (Hanneke & Xu, 2024, Lemma 28)),

$$\mathrm{VCdim}(\mathcal{A}_\lambda) = O\Big(\max_{\mathcal{I}} \mathrm{VCdim}(\mathcal{A}_{\lambda,\mathcal{I}}) + \log M\Big) = O\left(D \log(2k) + kj \log \frac{ed}{j}\right). \qquad \square$$

A.6 Proof of Theorem 9

Proof. The stated size bound follows by substituting Theorem 4 and Lemma 8 into the sensitivity-sampling bound in Section 3. It remains to verify that the required pointwise sensitivity upper bounds can be computed.

Choose a minimum-cost k-tuple B among the subspaces $\{u \in \mathbb{R}^d : u_I = y_I\}$, where $|I| = d - j$ and $y \in X$. Enumeration takes $O(nkd(n\binom{d}{j})^k)$ arithmetic operations. For each nonempty optimal cluster C with assigned subspace $F = \{u \in \mathbb{R}^d : u_I = a\}$, choose $y \in C$ minimizing $d(y, F)$. The candidate $\widehat{F} = \{u \in \mathbb{R}^d : u_I = y_I\}$ satisfies $d(x, \widehat{F}) \le d(x, F) + d(y, F)$ for every $x \in C$. Hence

$$\sum_{x \in C} d(x, \widehat{F})^z \le 2^{z-1} \left(\sum_{x \in C} d(x, F)^z + |C| d(y, F)^z \right) \le 2^z \sum_{x \in C} d(x, F)^z.$$

Summing over the optimal clusters shows that the minimum-cost candidate tuple B has approximation factor $\alpha = 2^z$. Let $C_1, \dots, C_k$ be its induced clusters and $\Delta = \phi_B(X)$. It remains to compute the local bounds used in Lemma 5.

For $k = 1$, deterministic one-center sensitivity upper bounds as in the proof of Lemma 5 give local scores whose total mass is at most $C_z R_{j,1}(N)$. Suppose now that $k \ge 2$. Fix a projected local instance $Y = \{y_1, \dots, y_N\} \subseteq \mathbb{R}^j$ and set $r_0 = \min\{j, k\}$. For every $H \subseteq [j]$ with $1 \le |H| \le r_0$, use the metric construction in the proof of Lemma 7 and the computable scores from Lemma 15. This gives scores $v_{\ell,H}$ such that

$$v_{\ell,H} \ge \sup_{f \in \mathcal{Q}_H} \frac{f(y_\ell)}{f(Y)}, \qquad \sum_{\ell=1}^{N} v_{\ell,H} \le C_z (C_z k)^{|H|} \Lambda_{|H|,k}(N).$$

For $H = \emptyset$, take $v_{\ell,\emptyset} = 1/N$.

Let $\mathcal{J}$ be the collection of all tuples $\mathbf{J} = (J_1, \dots, J_k)$ with $J_i \subseteq [j]$ and $|J_i| \le 1$, and write $H(\mathbf{J}) = \bigcup_i J_i$. Define

$$u_\ell^{(1)} = (j+1)^{\eta_z} \sum_{\mathbf{J} \in \mathcal{J}} v_{\ell, H(\mathbf{J})},$$
$$u_\ell^{(2)} = (j+1)^{z/2} \sum_{\substack{H \subseteq [j] \\ |H| \le r_0}} v_{\ell,H}, \qquad u_\ell = \min\{u_\ell^{(1)}, u_\ell^{(2)}\}.$$

The two pointwise inequalities in Lemma 6 show that u_ℓ upper-bounds the local sensitivity of y_ℓ. Summing the preceding score bounds as in the proof of Lemma 5 gives

$$\sum_{\ell=1}^{N} u_\ell \le C_z R_{j,k}(N) \le C_z R_{j,k}(n).$$

Reuse each active-set computation across all coordinate patterns.

In either case, assign the local scores to the corresponding occurrences $\pi_s(x) \in \pi_s(C_s)$. For $x \in C_s$ with local score u_x, set

$$s_x = \max\left\{ \frac{1}{n}, 2^{2z-1} \alpha u_x + 2^{z-1} \alpha \frac{\|x - \pi_s(x)\|_2^z}{\Delta} \right\},$$

using $s_x = \max\{1/n, u_x\}$ when $\Delta = 0$. The pointwise proof of Lemma 3 shows deterministically that these scores satisfy $s_x \ge \max\{1/n, \sigma_X(x)\}$ and

$$\sum_{x \in X} s_x \le C_{1,z} k R_{j,k}(n).$$

Apply the sensitivity-sampling theorem with failure probability δ, sampling independently with $p_x = s_x / \sum_y s_y$. For a sample of size m, give each occurrence of x weight $1/(mp_x)$ and combine repetitions. □

B Proofs for the Lower Bounds

We prove Lemmas 10, 12, and 13 in turn. The first two give the composition-block and grid multiscale constructions; the third embeds a set-clustering construction in two ways, corresponding to the two terms in $N_{j,k}$.

B.1 Proof of Lemma 10

Proof. The construction consists of $B(j,k;q)$ blocks. Within each block, geometrically increasing scales force an $\Omega(\log(n/B(j,k;q)))$ contribution, while the remaining coordinates allow a query to isolate that block.

Choose $\mathcal{H} \subseteq [j]$ of size q, and let $\mathcal{I} = [j] \setminus \mathcal{H}$. Define

$$A = \left\{ (h,p) : h \in \mathcal{H},\ p = (p_i)_{i\in\mathcal{I}} \in \mathbb{Z}_{\geq 0}^{\mathcal{I}},\ \sum_{i\in\mathcal{I}} p_i = k-2 \right\}.$$

Then

$$|A| = q\binom{j-q+k-3}{k-2} = B(j,k;q).$$

Let

$$L = \left\lfloor \frac{n}{B(j,k;q)} \right\rfloor$$

and $D > 2^{L/z}$. For every $(h,p) \in A$ and $s = 1,\dots,L$, let $x_{h,p,s} \in \mathbb{R}^{j+1}$ have coordinates

$$(x_{h,p,s})_i \coloneqq \begin{cases} 2^{s/z}, & i = h, \\ Dp_i, & i \in \mathcal{I}, \\ 0, & i \in \mathcal{H}\setminus\{h\} \text{ or } i = j+1. \end{cases}$$

Let

$$X_0 = \{x_{h,p,s} : (h,p) \in A,\ s = 1,\dots,L\}.$$

Then $|X_0| = B(j,k;q)L \leq n$. If $|X_0| < n$, add $n - |X_0|$ distinct points whose first j coordinates are zero. Let X denote the resulting set. Enumerate its points by $\nu \in [n]$ and replace their last coordinates by $kD\nu/n$. Distinct points are now at least kD/n apart.

We next associate with each block and scale a query that vanishes outside that block. Fix $(h,p) \in A$ and $s \in \{1,\dots,L\}$. Define

$$F_0 = \{u \in \mathbb{R}^{j+1} : u_h = 0\}, \qquad f_0(x) = d(x,F_0) = |x_h|,$$

and for each $i \in \mathcal{I}$ and each integer $\ell = 0,\dots,p_i - 1$, define

$$F_{i,\ell} = \{u \in \mathbb{R}^{j+1} : u_i = D\ell\}, \qquad f_{i,\ell}(x) = d(x,F_{i,\ell}) = |x_i - D\ell|.$$

Finally, fix $i_0 \in \mathcal{I}$ and let

$$F_\gamma = \{u \in \mathbb{R}^{j+1} : u_{i_0} = Dp_{i_0} + 2^{s/z}\}.$$

These are exactly

$$1 + \sum_{i\in\mathcal{I}} p_i + 1 = k$$

axis-aligned j-dimensional affine subspaces. Let $\phi_{h,p,s}$ denote their induced cost. Thus, for every $x \in X$,

$$\phi_{h,p,s}(x) = \min\left\{ d(x,F_\gamma)^z,\ f_0(x)^z,\ \min_{\substack{i\in\mathcal{I},\\ 0\leq\ell\leq p_i-1}} f_{i,\ell}(x)^z \right\}.$$

Here and below, the inner minimum is interpreted as $+\infty$ when its index set is empty.

For points in the same block $X_{h,p}$, we have

$$\phi_{h,p,s}(x_{h,p,t}) = \min\{2^s, 2^t\}.$$

Indeed,

$$d(x_{h,p,t}, F_\gamma) = 2^{s/z}, \qquad f_0(x_{h,p,t}) = 2^{t/z},$$

while for every $i \in \mathcal{I}$ and $0 \le \ell \le p_i - 1$,

$$f_{i,\ell}(x_{h,p,t}) = |Dp_i - D\ell| \ge D > 2^{L/z} \ge \max\{2^{s/z}, 2^{t/z}\}.$$

On the other hand, suppose that $(h', r) \in A$ and $(h', r) \ne (h, p)$. If $h' \ne h$, then

$$f_0(x_{h',r,t}) = 0.$$

If $h' = h$, then $r \ne p$, and since

$$\sum_{i\in\mathcal{I}} r_i = \sum_{i\in\mathcal{I}} p_i = k - 2,$$

there exists some coordinate $i \in \mathcal{I}$ such that $r_i < p_i$. Taking $\ell = r_i$, we have $0 \le \ell \le p_i - 1$, and therefore

$$f_{i,\ell}(x_{h,r,t}) = |Dr_i - D\ell| = 0.$$

Hence

$$\phi_{h,p,s}(x_{h',r,t}) = 0 \qquad \forall (h', r) \ne (h, p).$$

The same holds for all added points x, since $f_0(x) = 0$.

We first prove the lower bound on the total sensitivity. For each $x_{h,p,t}$, use the query $\phi_{h,p,t}$. This query vanishes outside $X_{h,p}$, and

$$\phi_{h,p,t}(x_{h,p,t}) = 2^t, \qquad \phi_{h,p,t}(X) = \sum_{u=1}^{L} \min\{2^t, 2^u\} = 2^t(L - t + 2) - 2.$$

Consequently,

$$\sigma_X(x_{h,p,t}) \ge \frac{2^t}{2^t(L-t+2)-2} \ge \frac{1}{L-t+2}.$$

Summing over $t = 1, \dots, L$ gives $\Omega(\log L)$ total sensitivity in each block. Summing over all $B(j, k; q)$ blocks shows that

$$S(X) = \Omega\left(B(j,k;q) \log \frac{n}{B(j,k;q)}\right).$$

We now use the same block-isolating queries to prove the coreset lower bound. Let $C \subseteq X$ be an ϵ-coreset. We prove that, for each fixed $(h, p) \in A$, the coreset must contain $\Omega(\log L)$ points from the block $X_{h,p}$.

For this fixed (h, p), denote the coreset weight assigned to $x_{h,p,t}$ by $w_t \ge 0$, with $w_t = 0$ if $x_{h,p,t} \notin C$. Since $\phi_{h,p,s}$ vanishes outside $X_{h,p}$, the coreset condition for $\phi_{h,p,s}$ becomes

$$(1 - \epsilon)Q_s \le \widetilde{Q}_s \le (1 + \epsilon)Q_s,$$

where

$$Q_s = \sum_{t=1}^{L} \min\{2^s, 2^t\} \qquad \text{and} \qquad \widetilde{Q}_s = \sum_{t=1}^{L} w_t \min\{2^s, 2^t\}.$$

Define

$$h_s \coloneqq \frac{Q_s}{2^s}, \qquad g_s \coloneqq \frac{\widetilde{Q}_s}{2^s}.$$

Then

$$h_s = \sum_{t\le s} 2^{t-s} + \sum_{t>s} 1 = 2 - 2^{1-s} + L - s$$

and

$$g_s = \sum_{t\le s} w_t 2^{t-s} + \sum_{t>s} w_t.$$

For every $R = 16 \cdot 2^r$ satisfying $R \le L/4$, consider the interval

$$I_R = \{L - 2R + 1, \dots, L - R\}.$$

We claim that I_R contains an index t with $w_t > 0$. Otherwise, throughout the interval from $s_0 = L - 2R$ to $s_2 = L - R$, there exist $U, V \ge 0$ such that

$$g_s = U2^{-s} + V.$$

Let $s_1 = L - 3R/2$. Since $R \le L/4$, we have $s_1 \ge 5L/8$ and $s_2 \ge 3L/4$. Then

$$g_{s_1} - g_{s_2} = 2^{-R/2}(g_{s_0} - g_{s_1}).$$

Using $h_s = L - s + 2 - 2^{1-s}$, we have

$$\begin{aligned}
h_{s_1} - h_{s_2} &= \frac{R}{2} - 2^{1-s_1}\left(1 - 2^{-R/2}\right) \ge \frac{R}{2} - 1, \\
h_{s_1} + h_{s_2} &= \frac{5R}{2} + 4 - 2^{1-s_1} - 2^{1-s_2} \le \frac{5R}{2} + 4, \\
h_{s_0} - h_{s_1} &= \frac{R}{2} - 2^{1-s_0}\left(1 - 2^{-R/2}\right) \le \frac{R}{2}, \\
h_{s_0} + h_{s_1} &= \frac{7R}{2} + 4 - 2^{1-s_0} - 2^{1-s_1} \le \frac{7R}{2} + 4.
\end{aligned}$$

Since $|g_s - h_s| \le \epsilon h_s$, it follows that

$$g_{s_1} - g_{s_2} \ge h_{s_1} - h_{s_2} - \epsilon(h_{s_1} + h_{s_2}) \ge \frac{R}{2} - 1 - \epsilon\left(\frac{5R}{2} + 4\right) \ge \frac{R}{3},$$

and

$$g_{s_0} - g_{s_1} \le h_{s_0} - h_{s_1} + \epsilon(h_{s_0} + h_{s_1}) \le \frac{R}{2} + \epsilon\left(\frac{7R}{2} + 4\right) \le R.$$

Together with the preceding identity, this gives

$$\frac{R}{3} \le g_{s_1} - g_{s_2} \le 2^{-R/2}R < \frac{R}{3},$$

a contradiction. Therefore I_R contains an index t with $w_t > 0$.

The intervals I_R corresponding to

$$R = 16 \cdot 2^r, \qquad 0 \le r \le \left\lfloor \log_2 \frac{L}{64} \right\rfloor,$$

are pairwise disjoint. Each contains an index with positive weight, so the coreset contains $\Omega(\log L)$ points from $X_{h,p}$. Since the queries $\phi_{h,p,s}$ vanish on all other blocks, summing over the $|A| = B(j,k;q)$ blocks gives

$$|C| = \Omega(B(j,k;q)\log L) = \Omega\left(B(j,k;q)\log\frac{n}{B(j,k;q)}\right).$$

At last, we compute the aspect ratio. For a block point,

$$\|x_{h,p,s}\|_2 \le 2^{s/z} + D\|p\|_1 + kD \le D + (k-2)D + kD < 2kD.$$

Recall that distinct points are at least kD/n apart,

$$\frac{\operatorname{diam}(X)}{\min_{x \ne y}\|x - y\|_2} \le \frac{4kD}{kD/n} = 4n.$$

□

Optimizing the parameter. For $1 \le q \le j - 2$,

$$\frac{B(j,k;q+1)}{B(j,k;q)} = \frac{(q+1)(j-q-1)}{q(j-q+k-3)},$$

which is at least 1 exactly when $q \le (j-1)/(k-1)$. Hence $q^\star = \lceil (j-1)/(k-1) \rceil$ is a maximizer, as asserted in the main text.

B.2 Proof of Lemma 12

Proof. Let $a = a_{j,k}$ and $B = a^{j+1}$. Set

$$\eta = (100a)^{-1/z}, \qquad A = 1 + \lceil 100a(2^z + 1) \rceil, \qquad \gamma = 1 + \frac{2}{z}. \tag{8}$$

Then $A^{2/z} \geq 4$ and $\log A = \Theta_z(\log(a+1))$. For a positive integer T and $s \in [T]$, define

$$m_s = A^{T-s}, \qquad r_s = A^{2(s-1)/z}.$$

Choose T maximal subject to

$$BA^{\gamma(T-1)} \leq n. \tag{9}$$

Such a $T \geq 1$ exists because $n \geq B$. The number of non-padding points constructed below is at most n. When $T = 1$, this number is $B \leq n$. When $T \geq 2$, we have $A > 2$ and $A^{2(T-1)/z} \geq 4$, and therefore

$$B \sum_{s=1}^{T} m_s = B \frac{A^T - 1}{A - 1} < 2BA^{T-1} \leq \frac{1}{2} BA^{\gamma(T-1)} \leq n. \tag{10}$$

Index the B blocks by $p = (p_1, \ldots, p_{j+1}) \in [a]^{j+1}$ and set

$$\delta = \frac{\eta}{A^{T-1} + 1}, \qquad D = 8r_T.$$

For $p \in [a]^{j+1}$, $s \in [T]$, and $h \in [m_s]$, let

$$\theta_{s,h} = \frac{\eta r_s h}{m_s + 1}$$

and define $x_{p,s,h} \in \mathbb{R}^{j+1}$ by

$$(x_{p,s,h})_i = Dp_i \quad (i \in [j]), \qquad (x_{p,s,h})_{j+1} = Dp_{j+1} + r_s + \theta_{s,h}.$$

For each (p, s), let

$$X_{p,s} = \{x_{p,s,h} : h \in [m_s]\}, \qquad X_p = \bigcup_{s=1}^{T} X_{p,s}.$$

Thus the block X_p consists of T groups. The m_s points in $X_{p,s}$ share their first j coordinates and are equally spaced in the last coordinate between $Dp_{j+1} + r_s$ and $Dp_{j+1} + (1+\eta)r_s$. Let $X_0 = \bigcup_{p \in [a]^{j+1}} X_p$. If $|X_0| < n$, add the distinct padding points

$$y_h = (D + h\delta, D, \ldots, D), \qquad h = 1, \ldots, n - |X_0|,$$

and let X denote the resulting point set, then $|X| = n$.

For every block-level pair (p, s), let $F_{p,s}$ consist of the axis-aligned hyperplanes below, where $u_i = c$ abbreviates $\{u \in \mathbb{R}^{j+1} : u_i = c\}$:

$$\begin{array}{ll} u_i = Dv & (i \in [j],\ v \in [a] \setminus \{p_i\}), \\ u_{j+1} = Dv & (v \in [a]), \\ u_{j+1} = Dv + r_s & (v \in [a] \setminus \{p_{j+1}\}), \\ u_1 = Dp_1 + r_s. & \end{array}$$

The number of hyperplanes is

$$j(a-1) + a + (a-1) + 1 = (j+2)(a-1) + 2 \leq k.$$

Repeating an existing hyperplane if necessary gives exactly k hyperplanes. Let $\phi_{p,s}$ denote the resulting clustering cost. The first family gives zero cost to blocks whose first j address entries differ from p. The base hyperplanes in the last coordinate control lower levels and include $u_{j+1} = D$, on which all padding points lie. The shifted last-coordinate hyperplanes control the other blocks at level s, while $u_1 = Dp_1 + r_s$ fixes the target cost and controls higher levels.

Let

$$M_s = m_s r_s^z = A^{T+s-2}.$$

For $x_{p,s,h} \in X_{p,s}$, the hyperplane $u_1 = Dp_1 + r_s$ is at distance r_s, whereas the base hyperplane $u_{j+1} = Dp_{j+1}$ is at distance $r_s + \theta_{s,h}$. Every other hyperplane is farther than r_s by the choice $D = 8r_T$. Hence $\phi_{p,s}(x) = r_s^z$ for every $x \in X_{p,s}$.

If $q \in [a]^{j+1}$ differs from p in one of its first j coordinates, every point in X_q lies on one of the hyperplanes $u_i = Dq_i$ and has zero cost. Thus only the a addresses $q = (p_1, \ldots, p_j, q_{j+1})$ remain. For $x_{q,t,h} \notin X_{p,s}$, the following three cases are exhaustive:

$$\phi_{p,s}(x_{q,t,h}) \le \begin{cases} (1+\eta)^z r_t^z, & t < s, \\ \eta^z r_s^z, & t = s \text{ and } q_{j+1} \neq p_{j+1}, \\ r_s^z, & t > s. \end{cases} \tag{11}$$

The three bounds use, respectively, the base hyperplane $u_{j+1} = Dq_{j+1}$, the shifted hyperplane $u_{j+1} = Dq_{j+1} + r_s$, and the special hyperplane $u_1 = Dp_1 + r_s$.

Summing the pointwise bounds in (11) over all non-target points gives

$$\phi_{p,s}(X \setminus X_{p,s}) \le L_{p,s} := a(1+\eta)^z \sum_{t<s} m_t r_t^z + a \sum_{t>s} m_t r_s^z + (a-1) m_s \eta^z r_s^z.$$

Here the factor a counts the surviving choices of q_{j+1}, and the padding points contribute zero because they lie on $u_{j+1} = D$. Since

$$\frac{m_t r_t^z}{M_s} = A^{t-s} \quad (t < s), \qquad \frac{m_t r_s^z}{M_s} = A^{s-t} \quad (t > s),$$

dividing the preceding bound by M_s gives

$$\frac{L_{p,s}}{M_s} \le \frac{a((1+\eta)^z + 1)}{A - 1} + a\eta^z \le \frac{1}{100} + \frac{1}{100} = \frac{1}{50}.$$

Indeed, $\eta \le 1$ and (8) imply $(1+\eta)^z \le 2^z$, $A - 1 \ge 100a(2^z + 1)$, and $a\eta^z = 1/100$. Consequently,

$$M_s \le \phi_{p,s}(X) \le \frac{51}{50} M_s. \tag{12}$$

For each $x \in X_{p,s}$, using $\phi_{p,s}$ in the definition of sensitivity gives

$$\sigma_X(x) \ge \frac{r_s^z}{(51/50) m_s r_s^z} = \frac{50}{51 m_s}.$$

Thus every block-level class contributes at least $50/51$ to total sensitivity, and

$$S(X) \ge \frac{50}{51} BT. \tag{13}$$

We next prove the coreset lower bound. Let (C, w) be an ϵ-coreset and write

$$W_{p,s} = \sum_{x \in C \cap X_{p,s}} w(x).$$

The upper coreset inequality for $\phi_{p,s}$ and (12) imply

$$W_{p,s} r_s^z \le \phi_{p,s}(C, w) \le (1+\epsilon) \frac{51}{50} m_s r_s^z,$$

so $W_{p,s} \le c_0 m_s$, where $c_0 = (1+\epsilon)51/50 < 3/2$. Since (p, s) was arbitrary, the same bound $W_{q,t} \le c_0 m_t$ holds for every group $X_{q,t}$. Now fix (p, s) and suppose that $C \cap X_{p,s}$ is empty. Padding points have zero cost, and the pointwise bound in (11) is constant within each remaining group. Applying the weight bound to each such group therefore gives

$$\phi_{p,s}(C, w) \le c_0 L_{p,s} \le \frac{3}{100} M_s.$$

On the other hand, the lower coreset inequality and (12) give

$$\phi_{p,s}(C, w) \ge (1-\epsilon)\phi_{p,s}(X) > \frac{3}{4} M_s,$$

a contradiction. Hence C contains a point from every $X_{p,s}$ and

$$|C| \geq BT. \tag{14}$$

It remains to estimate T and the aspect ratio. By maximality in (9),

$$T = 1 + \left\lfloor \frac{\log(n/B)}{\gamma \log A} \right\rfloor, \qquad T = \Theta_z\left(1 + \frac{\log(n/B)}{\log(a+1)}\right),$$

Combining this with (13) and (14) proves the claimed lower bounds.

Finally, we bound the aspect ratio of X. For any distinct $h, h' \in [m_s]$,

$$\|x_{p,s,h} - x_{p,s,h'}\|_2 = \frac{\eta r_s |h - h'|}{m_s + 1} \geq \frac{\eta r_s}{m_s + 1} \geq \frac{\eta}{A^{T-1} + 1} = \delta.$$

For points in the same block X_p, say $x \in X_{p,s}$, and $y \in X_{p,s'}$ ($s' > s$),

$$|x_{j+1} - y_{j+1}| \geq r_{s+1} - (1+\eta) r_s = (A^{2/z} - 1 - \eta) r_s \geq 2.$$

If two blocks differ in one of their first j address entries, they are separated by at least D in that coordinate. If they differ only in their last address entry, their last coordinates differ by at least $D - 2r_T = 3D/4$. Finally, every non-padding point has last coordinate at least $D + 1$, whereas every padding point has last coordinate D. Distinct padding points are separated by δ. Since $\delta \leq 1$, these observations imply

$$\min_{x \neq y \in X} \|x - y\|_2 \geq \delta.$$

The scale budget gives

$$\frac{r_T}{\delta} = \eta^{-1} A^{2(T-1)/z}(A^{T-1} + 1) \leq 2\eta^{-1} A^{\gamma(T-1)} \leq \frac{2\eta^{-1} n}{B}.$$

Let $b = (D, \ldots, D)$. Every hard point is within $aD\sqrt{j+1}$ of b, and every padding point is within $n\delta$ of b. If $a \geq 2$, then, since $j \geq 2$ and $z \geq 1$,

$$\frac{a\sqrt{j+1}\,\eta^{-1}}{B} = 100^{1/z}\sqrt{j+1}\,a^{1/z - j} \leq 100\sqrt{j+1}\,a^{1-j} \leq 100.$$

Consequently,

$$\frac{\operatorname{diam}(X)}{\delta} \leq 16a\sqrt{j+1}\,\frac{r_T}{\delta} + 2n \leq 3202n.$$

If $a = 1$, all points agree in coordinates $2, \ldots, j$, so $\operatorname{diam}(X) \leq n\delta + 2r_T$. Since then $B = 1$ and $\eta^{-1} = 100^{1/z} \leq 100$,

$$\frac{\operatorname{diam}(X)}{\delta} \leq n + 4\eta^{-1} n \leq 401n.$$

Thus in all cases the aspect ratio is at most $3202n$. □

B.3 Proof of Lemma 13

Proof. We use two embeddings of the set-clustering construction from (Jubran et al., 2020, Theorem 4.3), one for each term in $N_{j,k}$. That construction considers every m-subset $P \subseteq [m+s]$ and its complement $C_P = [m+s] \setminus P$; the query associated with C_P has positive cost on P and zero cost on every other m-subset.

Suppose first that

$$N_{j,k} = \binom{j+t+1}{j+1}.$$

Then $t \geq 1$. Let $m = j + 1$ and $s = t$. For each

$$P = \{p_1 < \cdots < p_m\} \subseteq [m+s],$$

let $x_P = (p_1, \ldots, p_m) \in \mathbb{R}^m$, and define

$$X_0 = \{x_P : P \subseteq [m+s],\ |P| = m\}.$$

Let

$$F_0 = \{y \in \mathbb{R}^m : y_m = 0\}.$$

Let $X = X_0$. If $|X_0| < n$, for each $a = 1, \dots, n - |X_0|$, add

$$y_a = (1 + a, 2, 3, \dots, m - 1, 0) \in F_0$$

to X.

For each $i \in [m]$ and $c \in C_P$, let

$$F_{i,c} = \{y \in \mathbb{R}^m : y_i = c\}.$$

For each P, the collection

$$\mathcal{F}_P = \{F_0\} \cup \{F_{i,c} : i \in [m],\ c \in C_P\}$$

has $1+ms \le k$ axis-aligned $(m-1)$-dimensional affine subspaces, contains every point of $X \setminus \{x_P\}$, and does not contain x_P.

Suppose next that

$$N_{j,k} = \binom{j+1}{r}.$$

Let $m = j + 1 - r$ and $s = r$, so that $m + s = j + 1$. For every m-subset $P \subseteq [j + 1]$, let

$$x_P = \mathbf{1}_{C_P} \in \mathbb{R}^{j+1},$$

and define

$$X_0 = \{x_P : P \subseteq [j + 1],\ |P| = m\}.$$

If $r < k$, let

$$F_0 = \{y \in \mathbb{R}^{j+1} : y_1 = 2\}.$$

For $a = 1, \dots, n - |X_0|$, add

$$y_a = 2e_1 + ae_2$$

to X_0, and denote the resulting set by X. If $r = k$, for $a = 1, \dots, n - |X_0|$, add instead

$$y_a = (a + 2)e_1$$

to X_0, and again denote the resulting set by X.

For each $c \in C_P$, let

$$F_c = \{y \in \mathbb{R}^{j+1} : y_c = 0\}.$$

If $r < k$, define

$$\mathcal{F}_P = \{F_0\} \cup \{F_c : c \in C_P\},$$

and if $r = k$, define

$$\mathcal{F}_P = \{F_c : c \in C_P\}.$$

In the former case, the added points lie on F_0. In the latter case, $r = k \ge 2$, so every added point lies on F_c for some $c \in C_P$. Moreover, if $Q \ne P$, then there is a $c \in Q \cap C_P$, and hence $(x_Q)_c = 0$. Thus, in both cases, $\mathcal{F}_P$ has at most k axis-aligned j-dimensional affine subspaces, contains every point of $X \setminus \{x_P\}$, and does not contain x_P.

In either construction, $|X_0| = N_{j,k}$. Add, if necessary, axis-aligned j-dimensional affine subspaces that avoid X so that $|\mathcal{F}_P| = k$. We have

$$\phi_{\mathcal{F}_P}(x_P) > 0, \qquad \phi_{\mathcal{F}_P}(X) = \phi_{\mathcal{F}_P}(x_P).$$

It follows that $\sigma_X(x_P) = 1$. Summing over P gives

$$S(X) \ge |X_0| = N_{j,k}.$$

Finally, let C be an ϵ-coreset. If some $x_P \in X_0$ does not belong to C, then $\phi_{\mathcal{F}_P}(C) = 0$, while $\phi_{\mathcal{F}_P}(X) > 0$, contradicting the coreset guarantee because $\epsilon < 1$. Therefore $X_0 \subseteq C$, and

$$|C| \ge |X_0| = N_{j,k}. \qquad \square$$

It remains to compute the aspect ratio. In the first construction, distinct points of X_0 are distinct integer vectors and hence are at distance at least 1. The padding points are mutually at distance at least 1, and every padding point has last coordinate 0, whereas for every $x_P \in X_0$,

$$(x_P)_m = p_m \geq m,$$

so its distance from every padding point is also at least 1. Moreover, if $P = \{p_1 < \cdots < p_m\} \subseteq [m+s]$, then

$$i \leq p_i \leq i + s.$$

Since

$$|X_0| = \binom{m+s}{m} \geq 1 + ms,$$

we have

$$s\sqrt{m} \leq ms \leq |X_0| \leq n, \qquad m + s \leq |X_0| \leq n.$$

For two points of X_0, their distance is at most $s\sqrt{m} \leq n$, while two padding points are at distance at most n. For $x_P \in X_0$ and a padding point y_a, the contribution from the first coordinate is at most n, that from coordinates $2, \ldots, m-1$ is at most $s\sqrt{m-2} \leq n$, and that from the last coordinate is at most $m + s \leq n$. Hence

$$\|x_P - y_a\|_2 \leq \sqrt{3}n.$$

Thus the first construction has minimum pairwise distance at least 1 and diameter at most $\sqrt{3}n$.

In the second construction, the points of X_0 are distinct binary vectors of equal Hamming weight and therefore have pairwise distance at least $\sqrt{2}$. The explicit padding points above have mutual distance at least 1 and are at distance at least 1 from X_0. Since

$$j + 1 \leq \binom{j+1}{r} = |X_0| \leq n,$$

their union also has diameter at most $\sqrt{2}n$.

Therefore, in either construction,

$$\min_{x \neq y} \|x - y\|_2 \geq 1 \qquad \text{and} \qquad \operatorname{diam}(X) \leq 2n.$$

Hence the aspect ratio of X is at most $2n$.

C Generalization Bounds

Throughout this section, we suppress dependence on d, j, k, z unless these parameters vary. Define the two model classes

$$\mathfrak{F}_{\text{lin}} = \{F = (F_1, \ldots, F_k) \in \mathcal{F}_{d,j,k} : 0 \in F_i \text{ for every } i \in [k]\}$$

and

$$\mathfrak{F}_{\text{aff}} = \{F = (F_1, \ldots, F_k) \in \mathcal{F}_{d,j,k} : d(0, F_i) \leq 1 \text{ for every } i \in [k]\}.$$

For $\star \in \{\text{lin}, \text{aff}\}$ and $z \geq 1$, let

$$\phi_F(x) = \min_{1 \leq r \leq k} d(x, F_r)^z, \qquad \mathcal{L}^\star_{k,z} = \{\phi_F : F \in \mathfrak{F}_\star\}.$$

When k and z are understood, we abbreviate $\mathcal{L}^\star_{k,z}$ as $\mathcal{L}^\star$. Let $\mathcal{D}$ be a distribution on $\mathbb{S}^{d-1}$, and let $X = (x_1, \ldots, x_n) \sim \mathcal{D}^n$. The empirical and population risks of $F \in \mathfrak{F}_\star$ are

$$\widehat{\mathcal{R}}_X(F) = \frac{1}{n} \sum_{i=1}^n \phi_F(x_i), \qquad \mathcal{R}_{\mathcal{D}}(F) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \phi_F(x).$$

For $\star \in \{\text{lin}, \text{aff}\}$, let

$$\widehat{F}^\star_X \in \operatorname*{arg\,min}_{F \in \mathfrak{F}_\star} \widehat{\mathcal{R}}_X(F).$$

Its expected excess risk is

$$\mathcal{E}^\star_n = \mathop{\mathbb{E}}_{X \sim \mathcal{D}^n} \mathcal{R}_{\mathcal{D}}(\widehat{F}^\star_X) - \inf_{F \in \mathfrak{F}_\star} \mathcal{R}_{\mathcal{D}}(F).$$

For a class $\mathcal{L}$ of loss functions, let its empirical and worst-case Rademacher complexities be

$$\mathfrak{R}_X(\mathcal{L}) = \mathop{\mathbb{E}}_{\xi} \sup_{\ell \in \mathcal{L}} \frac{1}{n} \sum_{i=1}^{n} \xi_i \ell(x_i), \qquad \mathfrak{R}_n(\mathcal{L}) = \sup_{X \in (\mathbb{S}^{d-1})^n} \mathfrak{R}_X(\mathcal{L}),$$

where $\xi_1, \ldots, \xi_n$ are independent Rademacher random variables. The expected excess risk of empirical risk minimization is controlled by this quantity. Indeed, standard symmetrization gives

$$\mathcal{E}_n^\star \le 2 \mathop{\mathbb{E}}_{X \sim \mathcal{D}^n} \mathfrak{R}_X(\mathcal{L}^\star) \le 2\mathfrak{R}_n(\mathcal{L}^\star).$$

C.1 Preliminaries

We present a few useful lemmata for the proofs of the generalization bounds.

Lemma 17 (Truncated Dudley's integral)**.** *Let $x_1, \ldots, x_n$ be fixed points and $\mathcal{F}$ be a class of real-valued functions on these points. For $f, g \in \mathcal{F}$, define the sample metric*

$$d(f,g) = \left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \right)^{1/2}.$$

Assume that $(\mathcal{F}, d)$ is totally bounded and $D = \operatorname{diam}(\mathcal{F}, d)$. Let $g_1, \ldots, g_n \sim N(0,1)$ be independent standard Gaussians. Then for every $0 < \tau \le D$,

$$\mathop{\mathbb{E}}_{g} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i f(x_i) \lesssim \int_{\tau}^{D} \sqrt{\log N(\mathcal{F}, d, \epsilon)} d\epsilon + \tau\sqrt{n},$$

where the implicit constant is absolute.

Lemma 18 (Terminal embedding (Narayanan & Nelson, 2019))**.** *Let $T \subseteq \mathbb{R}^d$ be a nonempty finite set and $\epsilon \in (0,1)$. There exists a map*

$$\Phi : \mathbb{R}^d \to \mathbb{R}^m, \qquad m \lesssim \frac{\log(e|T|)}{\epsilon^2},$$

such that for every terminal $t \in T$ and every point $x \in \mathbb{R}^d$,

$$(1-\epsilon) \left\| x - t \right\|_2 \le \left\| \Phi(x) - \Phi(t) \right\|_2 \le (1+\epsilon) \left\| x - t \right\|_2.$$

Lemma 19 ((Foster & Rakhlin, 2019))**.** *Suppose that $\mathcal{F}$ is a family of functions $f : X \to \mathbb{R}^K$ and $\phi_1, \ldots, \phi_n : \mathbb{R}^K \to \mathbb{R}$ are L-Lipschitz functions in the ℓ_∞ norm. For a real-valued function class $\mathcal{G}$ and independent Rademacher random variables $\xi_1, \ldots, \xi_n$, define its worst-case Rademacher sum complexity by*

$$\mathfrak{R}_n^\Sigma(\mathcal{G}) = \sup_{x_1, \ldots, x_n \in X} \mathop{\mathbb{E}}_{\xi} \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \xi_i g(x_i).$$

For each $\gamma > 0$, there exists a constant $C > 0$ such that if $|\phi_i(f(x))| \le \beta$ and $\|f(x)\|_\infty \le \beta$ for all $i \in [n]$, $x \in X$, and $f \in \mathcal{F}$, then

$$\mathop{\mathbb{E}}_{\xi} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \xi_i \phi_i(f(x_i)) \le CL\sqrt{K} R_{\max} \log^{\frac{3}{2}+\gamma} \left(\frac{e\beta n}{R_{\max}} \right), \qquad R_{\max} = \max_{r \in [K]} \mathfrak{R}_n^\Sigma(\mathcal{F}|_r).$$

C.2 Upper Bounds

Theorem 20. *Suppose that $n \ge j$ and $z \ge 1$. For every $\gamma > 0$, it holds that*

$$\mathcal{E}_n^{\mathrm{lin}} \le 2\mathfrak{R}_n(\mathcal{L}^{\mathrm{lin}}) \lesssim_{z,\gamma} \sqrt{\frac{kj}{n}} \log^{\frac{3}{2}+\gamma} \left(\frac{en}{j} \right).$$

Proof. Fix $X = (x_1, \ldots, x_n) \subset \mathbb{S}^{d-1}$. For notational brevity, let $\mathcal{H}_{d,j}$ denote the set of axis-aligned j-dimensional linear subspaces of $\mathbb{R}^d$.

We first prove the result for $z = 1$. We begin with $k = 1$ and then use Lemma 19 for general k.

When $k = 1$,

$$\mathfrak{R}_X(\mathcal{L}_{1,1}^{\text{lin}}) = \mathop{\mathbb{E}}_{\xi} \sup_{F \in \mathcal{H}_{d,j}} \frac{1}{n} \sum_i \xi_i d(x_i, F).$$

Each $F \in \mathcal{H}_{d,j}$ can be identified with an index set $S \subseteq [d]$ with $|S| = j$ such that $d(x, F) = \|x_{S^c}\|_2$. Therefore,

$$\mathfrak{R}_X(\mathcal{L}_{1,1}^{\text{lin}}) \lesssim \frac{1}{\sqrt{n}} \mathop{\mathbb{E}}_{g} \sup_{S \in \binom{[d]}{j}} Z_S, \quad \text{where} \quad Z_S = \frac{1}{\sqrt{n}} \sum_i g_i \|(x_i)_{S^c}\|_2.$$

Here we used the standard comparison of Rademacher complexity by Gaussian complexity. The canonical distance of this Gaussian process $\{Z_S\}$ is

$$d(S,T) = \left(\frac{1}{n} \sum_{i=1}^n (\|(x_i)_{S^c}\|_2 - \|(x_i)_{T^c}\|_2)^2 \right)^{\frac{1}{2}} \leq \left(\frac{1}{n} \sum_{i=1}^n \|(x_i)_{S \triangle T}\|_2^2 \right)^{\frac{1}{2}}.$$

Define

$$\sigma_s^2 = \frac{1}{n} \sum_{i=1}^n |(x_i)_s|^2, \quad s \in [d]$$

and hence

$$d(S,T) \leq \left(\sum_{s \in S \triangle T} \sigma_s^2 \right)^{\frac{1}{2}}.$$

Let $b_s \sim N(0, \sigma_s^2)$ and $Y_S = \sum_{s \in S} b_s$. Then $\mathbb{E}(Y_S - Y_T)^2 = \sum_{s \in S \triangle T} \sigma_s^2$. Hence, by the classical Sudakov-Fernique comparison inequality,

$$\mathbb{E} \sup_{|S|=j} Z_S \leq \mathbb{E} \sup_{|S|=j} Y_S.$$

Pointwise,

$$\sup_{|S|=j} Y_S \leq \sqrt{j} \left(\sum_{s=1}^d b_s^2 \right)^{1/2}.$$

Since $\sum_s \sigma_s^2 = n^{-1} \sum_i \|x_i\|_2^2 = 1$, Jensen's inequality gives $\mathbb{E} \sup_{|S|=j} Y_S \leq \sqrt{j}$. Consequently,

$$\mathop{\mathbb{E}}_{g} \sup_{S \in \binom{[d]}{j}} \frac{1}{n} \sum_{i=1}^n g_i \|(x_i)_{S^c}\|_2 \leq \sqrt{\frac{j}{n}}, \tag{15}$$

and hence $\mathfrak{R}_X(\mathcal{L}_{1,1}^{\text{lin}}) \lesssim \sqrt{j/n}$. Since this estimate is uniform over all sequences $x_1, \ldots, x_n \in \mathbb{S}^{d-1}$, the worst-case coordinate Rademacher sum complexity $\mathfrak{R}_n^{\Sigma}$ in Lemma 19 is $O(\sqrt{nj})$. Applying that lemma to the 1-Lipschitz map $(u_1, \ldots, u_k) \mapsto \min_r u_r$ yields

$$\mathfrak{R}_X(\mathcal{L}_{k,1}^{\text{lin}}) \lesssim_\gamma \sqrt{\frac{kj}{n}} \log^{\frac{3}{2}+\gamma} \left(\frac{en}{j} \right).$$

Every loss in $\mathcal{L}_{k,1}^{\text{lin}}$ takes values in $[0,1]$, and $\mathcal{L}^{\text{lin}} = \{\ell^z : \ell \in \mathcal{L}_{k,1}^{\text{lin}}\}$. Since $t \mapsto t^z$ is z-Lipschitz on $[0,1]$, the scalar contraction inequality gives

$$\mathfrak{R}_X(\mathcal{L}^{\text{lin}}) \lesssim z\, \mathfrak{R}_X(\mathcal{L}_{k,1}^{\text{lin}}).$$

Taking the supremum over X and using the symmetrization bound above proves the theorem. □

For $d \geq 2kj$ and n sufficiently large relative to kj, the dependence on k, j and n is optimal up to logarithmic factors. Indeed, the lower bound construction of Bucarelli et al. (2023) uses $d = 2kj$ and is supported on the standard basis vectors. Its population and empirical minimizers are spanned by subsets of these vectors and are therefore axis-aligned. Since the distance from a standard basis

vector to any such coordinate subspace is either zero or one, the construction and its $\Omega(\sqrt{kj/n})$ excess-risk lower bound apply to $\mathfrak{F}_{\mathrm{lin}}$ for every $z \geq 1$.

We now turn to $\mathfrak{F}_{\mathrm{aff}}$. The restriction $d(0, F_r) \leq 1$ ensures that the losses are uniformly bounded; without a bound on the translations, the Rademacher complexity of the unrestricted loss class is infinite. Each affine loss has the form

$$\phi_F(x) = \min_{1 \leq r \leq k} \left\| x_{S_r^c} - a_r \right\|_2^z, \qquad S_r \in \binom{[d]}{j}, \quad a_r \in B_2^{S_r^c}.$$

Theorem 21. *Suppose that $n \geq j$ and $z \geq 1$. For every $\gamma > 0$, it holds that*

$$\mathcal{E}_n^{\mathrm{aff}} \leq 2\mathfrak{R}_n(\mathcal{L}^{\mathrm{aff}}) \lesssim_{z,\gamma} \left(\sqrt{\frac{kj}{n}} + \sqrt{\frac{k}{n}} \log^2(en) \right) \log^{\frac{3}{2}+\gamma} \left(\frac{en}{j} \right).$$

We need the following auxiliary lemmata on covering numbers. Let $\mathcal{Q}$ denote all functions of the form

$$h_{S,a,\alpha}(x) = \sqrt{\|x_{S^c} - a\|_2^2 + \alpha^2} - \|x_{S^c}\|_2, \quad a \in \mathbb{R}^{S^c}, \quad \alpha \geq 0, \quad \|a\|_2^2 + \alpha^2 \leq 1,$$

endowed with the distance

$$d_X(h_{S,a,\alpha}, h_{T,b,\beta}) = \left(\frac{1}{n} \sum_{i=1}^n (h_{S,a,\alpha}(x_i) - h_{T,b,\beta}(x_i))^2 \right)^{1/2}, \tag{16}$$

and let

$$\sigma_s^2 = \frac{1}{n} \sum_{i=1}^n |(x_i)_s|^2, \qquad \rho(S,T) = \left(\sum_{s \in S \triangle T} \sigma_s^2 \right)^{1/2}.$$

Lemma 22. *Let $h_{S,a,\alpha} \in \mathcal{Q}$ and $T \in \binom{[d]}{j}$. There exists $h_{T,b,\beta} \in \mathcal{Q}$ such that*

$$d_X(h_{S,a,\alpha}, h_{T,b,\beta}) \leq 2\rho(S,T).$$

Proof. Let

$$b = a_{S^c \cap T^c} \oplus 0_{S \setminus T} \in \mathbb{R}^{T^c}, \qquad \beta = \sqrt{\left\| a_{T \setminus S} \right\|_2^2 + \alpha^2}.$$

Then $\|b\|_2^2 + \beta^2 = \|a\|_2^2 + \alpha^2 \leq 1$. To compare the two distance terms, regard them as the norms of

$$u = \left(x_{S^c \cap T^c} - a_{S^c \cap T^c},\ x_{T \setminus S} - a_{T \setminus S},\ 0_{S \setminus T},\ \alpha \right),$$
$$v = \left(x_{S^c \cap T^c} - a_{S^c \cap T^c},\ -a_{T \setminus S},\ x_{S \setminus T},\ \alpha \right).$$

Their norms are respectively $\sqrt{\|x_{S^c} - a\|_2^2 + \alpha^2}$ and $\sqrt{\|x_{T^c} - b\|_2^2 + \beta^2}$. Hence the reverse triangle inequality gives

$$\left| \sqrt{\|x_{S^c} - a\|_2^2 + \alpha^2} - \sqrt{\|x_{T^c} - b\|_2^2 + \beta^2} \right| \leq \|x_{S \triangle T}\|_2.$$

Likewise,

$$\left| \|x_{S^c}\|_2 - \|x_{T^c}\|_2 \right| \leq \|x_{S \triangle T}\|_2.$$

It follows pointwise that

$$|h_{S,a,\alpha}(x) - h_{T,b,\beta}(x)| \leq 2 \|x_{S \triangle T}\|_2.$$

Taking the empirical L_2-norm proves

$$d_X(h_{S,a,\alpha}, h_{T,b,\beta}) \leq 2\rho(S,T). \qquad \square$$

Lemma 23. *For every $\epsilon \in (0,1)$, it holds that*

$$N\left(\binom{[d]}{j}, \rho, \epsilon \right) \leq \left(\frac{2e}{\epsilon^2} \right)^j.$$

Proof. Let $\tau = \epsilon^2/(2j)$, let $H = \{s \in [d] : \sigma_s^2 \geq \tau\}$, and let $M = \lfloor 2j/\epsilon^2 \rfloor$. Thus $|H| \leq M$ and $M \geq 2j$. If $S, T \in \binom{[d]}{j}$ agree on H,

$$\rho(S,T) \leq (2j\tau)^{1/2} = \epsilon.$$

For every $U \subseteq H$ that occurs as $S \cap H$ for some $S \in \binom{[d]}{j}$, choose one such S as a representative. These representatives form an ϵ-net. Therefore,

$$N\left(\binom{[d]}{j}, \rho, \epsilon\right) \leq \sum_{\ell=0}^{j} \binom{M}{\ell} \leq \left(\frac{eM}{j}\right)^j \leq \left(\frac{2e}{\epsilon^2}\right)^j. \qquad \square$$

Lemma 24. *For each $S \in \binom{[d]}{j}$, let $\mathcal{Q}_S$ be the subset of $\mathcal{Q}$ consisting of functions with index set S. For every $\epsilon \in (0,1)$, it holds that*

$$\log N(\mathcal{Q}_S, d_X, \epsilon) \lesssim \frac{\log(en)}{\epsilon^2} \log \frac{C}{\epsilon},$$

where $C > 1$ is an absolute constant.

Proof. Let $h_{S,a,\alpha} \in \mathcal{Q}_S$, let $y_0 = 0$, and let $y_i = (x_i)_{S^c}$ for $i \in [n]$. Then

$$h_{S,a,\alpha}(x_i) = \|(y_i, 0) - (a, \alpha)\|_2 - \|(y_i, 0)\|_2 .$$

Applying the terminal embedding (Lemma 18) to the terminals

$$\{(y_i, 0) : 0 \leq i \leq n\},$$

we obtain a map $\Phi : \mathbb{R}^{S^c} \oplus \mathbb{R} \to \mathbb{R}^m$ with $m \lesssim (\log(en))/\eta^2$ $(0 < \eta < 1)$ such that the terminal distances are preserved within a factor $1 \pm \eta$. After translating Φ so that $\Phi(0) = 0$, let

$$u_i = \Phi((y_i, 0)), \qquad v_{a,\alpha} = \Phi((a, \alpha)).$$

For $v \in \mathbb{R}^m$, define its evaluation vector on X by

$$G_v(x_i) = \|u_i - v\|_2 - \|u_i\|_2 .$$

The terminal-embedding guarantee gives

$$\left(\frac{1}{n}\sum_{i=1}^{n} \left(G_{v_{a,\alpha}}(x_i) - h_{S,a,\alpha}(x_i)\right)^2\right)^{1/2} \leq 3\eta.$$

Moreover,

$$\|v_{a,\alpha}\|_2 \leq (1+\eta)\,\|(a,\alpha)\|_2 \leq 2.$$

Choose an internal $(\epsilon/3)$-net $\mathcal{V}$ of the admissible vectors $v_{a,\alpha}$. Since they lie in $2B_2^m$,

$$|\mathcal{V}| \leq \left(\frac{C}{\epsilon}\right)^m.$$

For every $v_{a,\alpha}$, choose $v_{b,\beta} \in \mathcal{V}$ within distance $\epsilon/3$. The reverse triangle inequality gives

$$d_X(h_{S,a,\alpha}, h_{S,b,\beta}) \leq 3\eta + \|v_{a,\alpha} - v_{b,\beta}\|_2 + 3\eta < \epsilon$$

when $\eta = \epsilon/12$. Hence the corresponding functions $h_{S,b,\beta}$ form an ϵ-net of $\mathcal{Q}_S$, and

$$\log N(\mathcal{Q}_S, d_X, \epsilon) \lesssim \frac{\log(en)}{\epsilon^2} \log \frac{C}{\epsilon}. \qquad \square$$

Now we can obtain the following entropy bound for $\mathcal{Q}$.

Lemma 25. *For every $\epsilon \in (0, 2]$, it holds that*

$$\log N(\mathcal{Q}, d_X, \epsilon) \lesssim \left(j + \frac{\log(en)}{\epsilon^2}\right) \log \frac{C_1}{\epsilon}$$

for some absolute constant $C_1 > 2$.

Proof. Use Lemma 23 to construct an $(\epsilon/3)$-net $\mathcal{S}$ on $\binom{[d]}{j}$ under distance $\rho(\cdot,\cdot)$. For each $S \in \mathcal{S}$, use Lemma 24 to construct an $(\epsilon/3)$-net $\mathcal{N}_S$ on $\mathcal{Q}_S$ under distance $d_X(\cdot,\cdot)$. Let $\mathcal{N} = \bigcup_{S\in\mathcal{S}} \mathcal{N}_S$.

Indeed, let $h_{T,b,\beta} \in \mathcal{Q}$. There exists $S \in \mathcal{S}$ such that $\rho(S,T) < \epsilon/3$. By Lemma 22, there exist a and α such that $d_X(h_{S,a,\alpha}, h_{T,b,\beta}) \le 2\rho(S,T) < 2\epsilon/3$. Then there exists $h_{S,a',\alpha'} \in \mathcal{N}_S$ such that $d_X(h_{S,a,\alpha}, h_{S,a',\alpha'}) < \epsilon/3$. Hence, $d_X(h_{S,a',\alpha'}, h_{T,b,\beta}) < \epsilon$, so $\mathcal{N}$ is an ϵ-net of $\mathcal{Q}$.

Therefore,

$$\begin{aligned}\log N(\mathcal{Q}, d_X, \epsilon) \le \log|\mathcal{N}| &\le \log N\left(\binom{[d]}{j}, \rho, \epsilon/3\right) + \max_S \log N(\mathcal{Q}_S, d_X, \epsilon/3) \\ &\lesssim \left(j + \frac{\log(en)}{\epsilon^2}\right)\log\frac{C_1}{\epsilon}.\end{aligned}$$

□

Given the preceding entropy bound of $\mathcal{Q}$, we are now ready to prove Theorem 21.

Proof of Theorem 21. Fix $X = (x_1, \dots, x_n) \subset \mathbb{S}^{d-1}$. We first prove the result for $z = 1$ and again consider $k = 1$ first. By Gaussian–Rademacher comparison,

$$\mathfrak{R}_X(\mathcal{L}^{\text{aff}}_{1,1}) \lesssim \mathop{\mathbb{E}}_{g} \sup_{S\in\binom{[d]}{j}, a\in B_2^{S^c}} \frac{1}{n}\sum_{i=1}^n g_i \left\|(x_i)_{S^c} - a\right\|_2.$$

In view of (15) and subadditivity of the supremum, let

$$\Gamma := \mathbb{E} \sup_{S\in\binom{[d]}{j}, a\in B_2^{S^c}} \sum_{i=1}^n \frac{1}{\sqrt{n}} g_i h_{S,a,0}(x_i). \tag{17}$$

Indeed, (15) gives

$$\mathfrak{R}_X(\mathcal{L}^{\text{aff}}_{1,1}) \lesssim \sqrt{\frac{j}{n}} + \frac{\Gamma}{\sqrt{n}}.$$

The canonical distance of the stochastic process in (17) is exactly the distance $d_X(\cdot,\cdot)$ defined in (16).

Applying Lemma 25 and the truncated form of Dudley's integral yields

$$\begin{aligned}\Gamma &\lesssim \epsilon_0\sqrt{n} + \int_{\epsilon_0/2}^2 \sqrt{\log N(\mathcal{Q}, d_X, \epsilon)}\,d\epsilon, \quad \epsilon_0 \in (0,2) \\ &\lesssim \epsilon_0\sqrt{n} + \sqrt{j} + \sqrt{\log(en)}\int_{\epsilon_0/2}^2 \frac{\sqrt{\log(C_1/\epsilon)}}{\epsilon}d\epsilon \\ &\lesssim \epsilon_0\sqrt{n} + \sqrt{j} + \sqrt{\log(en)}\cdot\left(\log\frac{C_2}{\epsilon_0}\right)^{\frac{3}{2}}.\end{aligned}$$

Letting $\epsilon_0 \asymp 1/\sqrt{n}$ gives $\Gamma \lesssim \sqrt{j} + \log^2(en)$, and hence

$$\mathfrak{R}_X(\mathcal{L}^{\text{aff}}_{1,1}) \lesssim \sqrt{\frac{j}{n}} + \frac{\log^2(en)}{\sqrt{n}}.$$

This bound is uniform over all samples on the unit sphere. Thus the worst-case coordinate Rademacher sum complexity satisfies

$$R_{\max} \lesssim \sqrt{nj} + \sqrt{n}\log^2(en).$$

Since $\|x_{S^c} - a\|_2 \le 2$, we may take $\beta = 2$ in Lemma 19. Let $R_0 = C\sqrt{n}(\sqrt{j} + \log^2(en))$ be the preceding upper bound for $R_{\max}$, with C sufficiently large. Using also the trivial bound $R_{\max} \le 2n$, the elementary bound

$$u\log^{\frac{3}{2}+\gamma}(e\beta n/u) \lesssim_\gamma R_0\log^{\frac{3}{2}+\gamma}(en/j), \qquad 0 < u \le \min\{R_0, 2n\},$$

follows from $n \geq j$. Applying Lemma 19 to the coordinatewise minimum gives

$$\mathfrak{R}_X(\mathcal{L}_{k,1}^{\mathrm{aff}}) \lesssim_\gamma \sqrt{k}\left(\sqrt{\frac{j}{n}} + \frac{\log^2(en)}{\sqrt{n}}\right)\log^{\frac{3}{2}+\gamma}\left(\frac{en}{j}\right).$$

Every loss in $\mathcal{L}_{k,1}^{\mathrm{aff}}$ takes values in $[0,2]$, and $\mathcal{L}^{\mathrm{aff}} = \{\ell^z : \ell \in \mathcal{L}_{k,1}^{\mathrm{aff}}\}$. Since $t \mapsto t^z$ is $(z2^{z-1})$-Lipschitz on $[0,2]$, the scalar contraction inequality gives

$$\mathfrak{R}_X(\mathcal{L}^{\mathrm{aff}}) \lesssim z2^{z-1}\mathfrak{R}_X(\mathcal{L}_{k,1}^{\mathrm{aff}}).$$

Taking the supremum over X and using the symmetrization bound above proves the theorem. □