Near-Optimal Strong Coresets for ℓ_p Subspace Approximation

Honghao Lin* Vahab Mirrokni† David P. Woodruff‡

Abstract

We study strong coresets for ℓ_p subspace approximation. Given $A \in \mathbb{R}^{n\times d}$, the goal is to sample and rescale a small number of its rows to form SA such that

$$\|SA(I-P_F)\|_{p,2}^p = (1\pm\varepsilon)\|A(I-P_F)\|_{p,2}^p$$

simultaneously for every subspace $F \subseteq \mathbb{R}^d$ of dimension at most k, where P_F is the orthogonal projector onto F. Woodruff and Yasuda (FOCS 2025) [WY25] obtained coreset sizes $\widetilde{O}_p(k\varepsilon^{-4/p})$ for $1 \le p < 2$ and $\widetilde{O}_p(k^{p/2}\varepsilon^{-p})$ for $p > 2$. We improve these bounds to $\widetilde{O}_p(k\varepsilon^{-2})$ and $\widetilde{O}_p(k^{p/2}\varepsilon^{-2})$, respectively. For $1 \le p < 2$, our algorithm runs in $\widetilde{O}_p(\mathrm{nnz}(A) + d^\omega + k\varepsilon^{-2})$ time. The resulting coreset size matches the known sampling lower bound [LWW21] up to logarithmic factors when $k+1 \ge C\log(1/\varepsilon)$ for an absolute constant C. For $p > 2$, our algorithm runs in $\widetilde{O}_p(\mathrm{nnz}(A) + d^\omega)$ time, matching the running time of the Woodruff–Yasuda framework.

We use different techniques in the two regimes. For $1 \le p < 2$, we combine a bicriteria low-rank split with Lewis-weight sampling and empirical-process bounds independent of the output dimension. For $p > 2$, we give a sharper analysis of the Woodruff–Yasuda construction. By retaining the truncation in its sampling probabilities throughout the row-count recurrence, we show that it achieves the improved ε^{-2} dependence.

1 Introduction

Subspace approximation is a basic primitive for representing high-dimensional data by a low-dimensional linear model [DV07, DTV11, SV12, CW15]. Given a collection of points, the goal is to find a low-dimensional subspace that minimizes the aggregate distance of the points from that subspace. This model underlies classical dimensionality reduction as well as robust variants of low-rank approximation.

Let $A \in \mathbb{R}^{n\times d}$ have rows $a_1^\top, \ldots, a_n^\top$. For a subspace $F \subseteq \mathbb{R}^d$, let P_F denote the orthogonal projector onto F. For $p \ge 1$, the ℓ_p subspace approximation cost of F is

$$\mathrm{cost}_A(F) := \|A(I-P_F)\|_{p,2}^p = \sum_{i=1}^n \|a_i^\top(I-P_F)\|_2^p.$$

Writing $\mathcal{F}_k$ for the set of subspaces of dimension at most k, the optimization problem is

$$\mathsf{OPT}_A := \min_{F\in\mathcal{F}_k} \mathrm{cost}_A(F).$$

*Google Research, Carnegie Mellon University / Texas A&M University
†Google Research
‡Google Research and Carnegie Mellon University

The case $p = 2$ is principal component analysis and is solved by the singular value decomposition. The case $p = 1$ includes the median hyperplane problem, while larger values of p increasingly emphasize points with large residuals and give finite-p relaxations of the center hyperplane objective. Except for the Euclidean case $p = 2$, these problems are generally computationally hard [DTV11, GRSW12, CW15].

This hardness makes data reduction especially consequential. A common route to an approximation algorithm is to first replace the input by a much smaller instance and then run a more expensive optimization routine on that instance; the size of the reduced instance therefore directly affects both the storage cost and the eventual running time [FMSW10, CW15]. The central target is a *dimension-independent* summary: its size may depend on the intrinsic rank k, the accuracy ε, and the fixed exponent p, but not on either the number of points n or the ambient dimension d.

Coresets provide such summaries while retaining the geometric objective. A row-sampling and rescaling matrix S represents a $(1 \pm \varepsilon)$ strong row coreset if it is supported on few input rows and satisfies

$$\mathrm{cost}_{SA}(F) = (1 \pm \varepsilon)\,\mathrm{cost}_A(F) \qquad \text{simultaneously for every } F \in \mathcal{F}_k.$$

Equivalently, the coreset is a small weighted subset of the input points that preserves every rank-at-most-k subspace cost. If it contains m rows, its row span has dimension at most m, so rotational invariance also allows the retained instance to be represented in at most m dimensions. Thus a bound $m = \mathrm{poly}(k/\varepsilon)$ removes both n and d from the size of the optimization instance.

The uniformity over F is the defining strength of this guarantee. An optimizer for the sampled instance immediately yields an approximate optimizer for the original data, but the same sample can also be reused for many queries, for an adaptively chosen candidate subspace, or for a restricted family of subspaces imposed only after preprocessing. These applications are not covered by a weak coreset, which need only preserve the optimum value or an approximate minimizer [FL11, HV20]. A strong coreset is therefore a reusable surrogate for the entire objective rather than only for one optimization run.

There is also an important distinction between a coreset supported on the original input rows and a reduced instance consisting of modified points. Earlier dimension-independent constructions append an extra coordinate to each retained point [SW18, FKW21]. Such a construction still gives a compact representation, but it is not a weighted subset of the original dataset. A true row-subset coreset preserves the ambient representation and the sparsity pattern of every retained row, maintains a direct correspondence between sampled and original points, and can be passed to downstream routines without adapting them to an additional coordinate. The distinction is also theoretically meaningful: the known sampling lower bounds concern weighted subsets of the input rows [LWW21, WY23].

Obtaining dimension independence, a true row subset, nearly optimal rank dependence, and an efficient construction simultaneously was the main obstacle. Sohler and Woodruff [SW18] achieved dimension-independent size with nearly optimal dependence on k, but used modified points and an exponential-time construction. Subsequent work either retained the extra coordinate and lost additional polynomial factors [FKW21], or produced true row subsets with polynomially worse dependence on k [HV20]. The natural benchmark is $\widetilde{O}(k)$ rows for $p < 2$ and $\widetilde{O}(k^{p/2})$ rows for $p > 2$, with polynomial dependence on $1/\varepsilon$ [LWW21, WY23].

Woodruff and Yasuda [WY25] reached this benchmark in the rank parameter. They obtained the first dimension-independent strong row coresets with nearly optimal dependence on k for every $p \neq 2$, together with a $\widetilde{O}(\mathrm{nnz}(A) + d^{\omega})$-time construction. Their framework recursively samples rows according to root ridge leverage scores and returns $\widetilde{O}_p(k\varepsilon^{-4/p})$ rows for $1 \leq p < 2$. For $p > 2$,

its row count is

$$\frac{k^{p/2}}{\varepsilon^p}\left(\log\frac{k}{\varepsilon\delta}\right)^{O(p^2)}.$$

This result resolves the dimension and rank dependence, but it leaves the accuracy dependence open. For $1 \le p < 2$, the $\varepsilon^{-4/p}$ upper bound is separated from the sampling lower bound $\widetilde{\Omega}_p(k\varepsilon^{-2})$ by a polynomial factor [LWW21]. For $p > 2$, their one-round sampling theorem is applied recursively, and the analysis of the resulting row-count recurrence produces the factor ε^{-p}. It is not immediate whether this factor is required by the sampling rule itself.

> *What is the right dependence on ε for dimension-independent strong row coresets for ℓ_p subspace approximation?*

1.1 Related work

Coresets and sketches for high-dimensional geometric approximation have been studied through dimensionality reduction, sensitivity sampling, and matrix sampling techniques [FMSW10, FL11, VX12]. Earlier work developed sampling-based dimension reduction and approximation algorithms specifically for subspace objectives [DV07, DTV11, SV12]. For squared Euclidean objectives, Feldman, Schmidt, and Sohler developed dimension-independent reductions for PCA and projective clustering [FSS20]. Sohler and Woodruff subsequently obtained the first strong coresets for subspace approximation whose size is independent of both n and d for constant p, using a representative subspace and points with an appended coordinate rather than a weighted subset of the original rows [SW18]. Their construction achieves the desired dependence on k but runs in exponential time, and the appended coordinate means that downstream algorithms must be adapted to the modified instance. Feng, Kacham, and Woodruff reduced the running time to polynomial for sum-of-distances objectives, while retaining the appended coordinate and losing additional polynomial factors in the coreset size [FKW21].

A complementary line of work seeks coresets supported on the input rows. Huang and Vishnoi obtained the first dimension-independent strong coresets of this form through sensitivity sampling, but their size loses polynomial factors in the rank parameter [HV20]. Woodruff and Yasuda developed offline and online subset-selection algorithms and online strong coresets for ℓ_p subspace approximation [WY23]. Lewis-weight sampling is a central tool for ℓ_p subspace embeddings when $p < 2$ [CP15], and Munteanu and Omlor's ℓ_2-augmentation framework provides a related $\widetilde{O}_p(d\varepsilon^{-2})$ sensitivity bound for subspace embeddings [MO24].

1.2 Our results

Our results obtain ε^{-2} dependence in both regimes. For $1 \le p < 2$, the first theorem closes the gap with the sampling lower bound, up to logarithmic factors. Its pre-sparsified implementation makes these factors independent of n.

Theorem 1.1 (Strong row coresets for $1 \le p < 2$)**.** *Fix $1 \le p < 2$. Given $A \in \mathbb{R}^{n\times d}$, $1 \le k < d$, and $\varepsilon \in (0, 1/2)$, there is a randomized algorithm running in*

$$\widetilde{O}_p\Big(\operatorname{nnz}(A) + d^{\omega} + k\varepsilon^{-2}\Big)$$

time that outputs a reweighted row subset SA supported on at most

$$C_p k\varepsilon^{-2}\left(\log\frac{C_p k}{\varepsilon}\right)^{C_p}$$

rows such that, with probability at least $2/3$,

$$\mathrm{cost}_{SA}(F) = (1 \pm \varepsilon)\,\mathrm{cost}_A(F) \qquad \textit{simultaneously for every } F \in \mathcal{F}_k.$$

The sampled rows and their weights do not depend on F.

When $k + 1 \geq C \log(1/\varepsilon)$, the sampling lower bound of Li, Wang, and Woodruff [LWW21] implies $\widetilde{\Omega}_p(k\varepsilon^{-2})$ rows. Thus Theorem 1.1 is optimal in that parameter range. The upper bound itself holds without these additional restrictions.

For $p > 2$, the ε^{-p} dependence is not intrinsic to the Woodruff–Yasuda construction. We keep their one-round sampling probabilities and recursive framework, but a sharper row-count analysis yields ε^{-2} dependence while preserving the $k^{p/2}$ dependence and nearly input-sparsity running time.

Theorem 1.2 (Strong row coresets for $p > 2$)**.** *Fix* $p > 2$. *There is a constant* $C_p \geq 1$ *such that the following holds. Given* $A \in \mathbb{R}^{n\times d}$, $1 \leq k < d$, *and* $\varepsilon, \delta \in (0, 1/2)$, *there is a randomized algorithm running in*

$$\widetilde{O}_p(\mathrm{nnz}(A) + d^{\omega})$$

time that outputs a row-sampling matrix S *such that, with probability at least* $1 - \delta$,

$$\mathrm{cost}_{SA}(F) = (1 \pm \varepsilon)\,\mathrm{cost}_A(F) \qquad \textit{simultaneously for every } F \in \mathcal{F}_k,$$

and

$$\mathrm{nnz}(S) \leq \frac{C_p k^{p/2}}{\varepsilon^2} \left(\log \frac{C_p k}{\varepsilon\delta}\right)^{p+5}.$$

The logarithmic exponent $p + 5$ is a convenient explicit bound and is not optimized. Thus Theorem 1.2 improves the best known row-subset upper bound from ε^{-p} to ε^{-2} and reduces the logarithmic exponent from $O(p^2)$ to $O(p)$. Taken together, the two theorems give the dimension-independent size bound

$$\widetilde{O}_p\Big(k^{\max\{1,p/2\}}\varepsilon^{-2}\Big)$$

for every fixed $p \in [1, \infty) \setminus \{2\}$, with the regime-specific running times stated above.

1.3 Technical overview

The case $1 \leq p < 2$**.** To see the main difficulty, consider sampling indices $I_1, \ldots, I_m$ independently from a distribution q and estimating, for each query subspace F,

$$\mathrm{cost}_A(F) = \sum_i \|Q_F a_i\|_2^p \quad \text{by} \quad \widehat{C}_F := \frac{1}{m}\sum_{s=1}^{m} \frac{\|Q_F a_{I_s}\|_2^p}{q_{I_s}}, \qquad Q_F = I - P_F.$$

For any fixed F, this estimator is unbiased. If q_i is chosen according to the largest relative contribution that row i can make to a query, then no summand is too large, and Bernstein's inequality gives concentration for that particular F. However, a strong coreset requires uniform concentration over all $F \in \mathcal{F}_k$. A standard ε-net argument would cover $\mathcal{F}_k$ and take a union bound over the net, but the required covering number depends on the ambient dimension d. We instead need a dimension-independent uniform concentration bound governed only by k.

We first explain the argument assuming that we are given a subspace F_0, spanned by $\widetilde{O}(k)$ rows of A, with

$$r := \dim(F_0) = \widetilde{O}(k), \qquad \mathsf{OPT}_A := \min_{F\in\mathcal{F}_k} \mathrm{cost}_A(F), \qquad \|A(I - P_{F_0})\|_{p,2}^p \leq \Gamma_0 \mathsf{OPT}_A,$$

where Γ_0 depends only on p. Thus F_0 has low dimension and its projection residual is within a constant factor of the optimal rank-k cost. For each input row a_i, a sketched regression computes a fitted vector $b_i \in F_0$, and we set $e_i = a_i - b_i$. Let B and E be the matrices with rows b_i and e_i, respectively. The regression guarantee gives

$$A = B + E, \qquad \operatorname{rowspan}(B) = F_0, \qquad \operatorname{rank}(B) = r, \qquad R := \|E\|_{p,2}^p = \sum_i \|e_i\|_2^p \le \Gamma \mathsf{OPT}_A,$$

where Γ depends only on p. This split serves two different purposes. The rows b_i of B lie in an r-dimensional space and can therefore be controlled by Lewis weights. The residual E may have high rank, but its total mass is bounded by every query cost, since $R \le \Gamma \mathsf{OPT}_A \le \Gamma \operatorname{cost}_A(F)$.

Let w_i be the ℓ_p Lewis weight of b_i, and, when $R > 0$, let $\rho_i = \|e_i\|_2^p / R$ be the fraction of the residual mass in row i. The vector-valued Lewis-weight sensitivity bound in Lemma 2.12 shows that, for every F,

$$\|Q_F b_i\|_2^p \le w_i \sum_j \|Q_F b_j\|_2^p.$$

Moreover, the triangle inequality and the bound on R imply $\sum_j \|Q_F b_j\|_2^p = O_{p,\Gamma}(\operatorname{cost}_A(F))$. Since Q_F is a contraction, $\|Q_F e_i\|_2^p \le \|e_i\|_2^p = \rho_i R \le \Gamma \rho_i \operatorname{cost}_A(F)$. Thus w_i and ρ_i control the low-rank and residual contributions of row i, respectively.

Set $D = k + r$, define the unnormalized balanced scores

$$s_i \asymp_{p,\Gamma} w_i + D\rho_i, \qquad q_i = \frac{s_i}{\sum_j s_j}.$$

Set

$$m = \widetilde{O}_p(D\varepsilon^{-2}),$$

with a sufficiently large hidden constant, and draw $I_1, \dots, I_m$ independently with replacement from q, rescaling each sampled row a_{I_s} by $(m q_{I_s})^{-1/p}$. Since $\sum_i w_i = r$ and $\sum_i \rho_i = 1$, the total score is $O(D)$, and the resulting distribution satisfies

$$q_i \gtrsim_{p,\Gamma} \frac{w_i}{D}, \qquad q_i \gtrsim_{p,\Gamma} \rho_i.$$

In particular, every normalized row contribution is bounded by $O(D)$. This already gives the correct sample size for a fixed F. The remaining challenge is to prove concentration simultaneously over all $F \in \mathcal{F}_k$ without introducing a dependence on the ambient dimension.

For this purpose, normalize each positive-cost query and write

$$x_F(i) = \frac{Q_F b_i}{q_i^{1/p} \operatorname{cost}_A(F)^{1/p}}, \qquad y_F(i) = \frac{Q_F e_i}{q_i^{1/p} \operatorname{cost}_A(F)^{1/p}},$$

so that

$$Z_F(i) := \|x_F(i) + y_F(i)\|_2^p = \frac{\|Q_F a_i\|_2^p}{q_i \operatorname{cost}_A(F)}, \qquad \mathbb{E}_q Z_F = 1.$$

For the samples above, the relative estimator is exactly

$$\frac{\widehat{C}_F}{\operatorname{cost}_A(F)} = \frac{1}{m} \sum_{s=1}^m Z_F(I_s).$$

It is not enough to analyze the two summands separately: they need not be orthogonal, and their parallel components may reinforce or cancel. We instead describe the nonlinear loss $Z_F(i)$ by four scalar coordinates:

1. the low-rank power $\|x_F(i)\|_2^p$;
2. the power in the component of $y_F(i)$ parallel to $x_F(i)$;
3. their signed first-order interaction; and
4. the power in the component of $y_F(i)$ orthogonal to $x_F(i)$.

An exact geometric identity recovers $Z_F(i)$ from these four quantities through a Lipschitz map, including at $p = 1$, when $x_F(i) = 0$, and under exact cancellation.

The advantage of this decomposition is that each scalar class has dimension-independent complexity. The low-rank process is controlled by the Lewis geometry of the rank-r matrix B. Directional and duality-map estimates control the two parallel interaction terms, while a row-augmented subspace entropy bound controls the orthogonal residual term. The corresponding Rademacher bounds scale with r or k, rather than with the dimension of the output space $F^\perp$. Vector contraction then combines the four estimates and yields

$$\mathbb{E}\sup_F \left|\frac{1}{m}\sum_{s=1}^{m} Z_F(I_s) - 1\right| \le \widetilde{O}_p\left(\sqrt{\frac{k+\mathrm{rank}(B)}{m}} + \frac{k+\mathrm{rank}(B)}{m}\right).$$

Since $D = k + r = \widetilde{O}(k)$, the choice above uses $m = \widetilde{O}_p(k\varepsilon^{-2})$ samples. Taking the hidden constant sufficiently large, the displayed bound and Markov's inequality give a $(1 \pm \varepsilon)$ approximation simultaneously for every $F \in \mathcal{F}_k$ with constant probability. Merging repeated samples then yields a strong row coreset supported on at most m input rows.

It remains to obtain such an F_0 and implement the scores efficiently. The fast bicriteria construction of Woodruff and Yasuda [WY25, Lemma 2.3] provides, with high probability, a subspace spanned by $\widetilde{O}(k)$ input rows that satisfies the assumed dimension and residual bounds. Once F_0 is available, standard sparse embeddings, Johnson–Lindenstrauss sketches, and Lewis-weight routines give constant-factor estimates of the residual masses and Lewis weights; see Section 4.

The case $p > 2$. Woodruff and Yasuda [WY25] provide a one-round sampling primitive for sparsifying a weighted active matrix. Given a current matrix $B \in \mathbb{R}^{N\times d}$ with rows $\mathbf{b}_i^\top$, it sets

$$\lambda_B := \frac{\|B - B_k\|_F^2}{k}, \qquad \tau_i := \mathbf{b}_i^\top (B^\top B + \lambda_B I)^\dagger \mathbf{b}_i,$$

where B_k is a best rank-k approximation to B in Frobenius norm and $\dagger$ denotes the Moore–Penrose pseudoinverse. These are ridge leverage scores at the standard rank-k regularization scale [CMM17]. The primitive samples row i independently with probability

$$q_i = \min\left\{1, \frac{N^{p/2-1}\tau_i^{p/2}}{\alpha}\right\},$$

rescaling a retained row by $q_i^{-1/p}$. Their one-round theorem shows that the sampling step simultaneously preserves the cost of every rank-k subspace, up to a small multiplicative error and an additive term proportional to the optimum cost. Since every subspace cost is at least the optimum, the additive term can be absorbed into the relative-error budget after tuning its coefficient. The construction in [WY25] applies this primitive recursively. We retain the same sampling rule but sharpen the analysis of the number of rows retained in each round. The per-round errors and failure probabilities compose across the recursion.

The parameter α is the sampling threshold in the generic-chaining argument that controls the deviation uniformly over all $F \in \mathcal{F}_k$. For a one-round target error ζ and failure probability η, it is chosen as

$$\alpha = \Theta_p\left(\frac{\zeta^2}{(\log N)^3 + \log(1/\eta)}\right).$$

Thus, up to logarithmic factors, α scales quadratically with the target accuracy. The expected number of rows that survive one round is $\sum_i q_i$, so the recursive coreset size is governed by how this thresholded sum is bounded. The available global information about the scores is the trace bound $\sum_i \tau_i \le 2k$.

For $0 < \alpha \le 1$, the earlier row-count bound in [WY25] uses the relaxation

$$\sum_i q_i \le \frac{1}{\alpha}\sum_i \min\{1, N^{p/2-1}\tau_i^{p/2}\} = O_p\left(\frac{k}{\alpha}N^{1-2/p}\right).$$

Writing N_t for the number of active rows after round t, iterating this estimate gives a row-count recurrence of the form

$$N_t \lesssim k\alpha^{-1}N_{t-1}^{1-2/p},$$

whose fixed point is $k^{p/2}\alpha^{-p/2}$. After the overall error ε is distributed across the recursive rounds, $\alpha = \widetilde{\Theta}_p(\varepsilon^2)$. The resulting fixed point therefore has an ε^{-p} dependence.

Our analysis keeps the full factor $N^{p/2-1}/\alpha$ inside the truncation. At a high level, this charges both sampling regimes to the same linear score budget. Rows whose sampling probabilities saturate at one must each carry a substantial amount of ridge leverage mass, so the trace bound limits how many such rows there can be. For the remaining rows, the superlinear contribution $\tau_i^{p/2}$ can also be dominated by a suitably scaled linear contribution in τ_i. The following pointwise inequality captures both regimes at once. For $p > 2$ and every $y \ge 0$,

$$\min\{1, y^{p/2}\} \le y.$$

Taking $y = c^{2/p}\tau_i$ yields, for every $c > 0$,

$$\min\{1, c\tau_i^{p/2}\} \le c^{2/p}\tau_i.$$

Applying this inequality directly with $c = N^{p/2-1}/\alpha$ and then using the trace bound gives

$$\sum_i q_i \le 2k\alpha^{-2/p}N^{1-2/p},$$

and the resulting recurrence is

$$N_t \lesssim k\alpha^{-2/p}N_{t-1}^{1-2/p}.$$

Its fixed point is

$$\left(k\alpha^{-2/p}\right)^{p/2} = k^{p/2}\alpha^{-1},$$

which gives ε^{-2} after substitution. Thus the gain comes from preserving the interaction between truncation and the chaining threshold before the recurrence is iterated: the threshold contributes $\alpha^{-2/p}$ rather than α^{-1} to the row-count recurrence, and the fixed-point calculation turns this local saving into the improvement from ε^{-p} to ε^{-2}.

Removing the $\log n$ dependence. Both analyses above contain logarithmic factors in the number N of rows to which the sampling procedure is applied. Running them directly on A would therefore leave a dependence on $\log n$ in the final coreset size. We first apply the Woodruff–Yasuda strong coreset construction stated in Theorem 2.4 with accuracy $\varepsilon/3$, obtaining a weighted row subset A_0 with $\mathrm{poly}_p(k/\varepsilon)$ rows. We then apply the corresponding construction described above to A_0. Since $\log(\mathrm{rows}(A_0)) = O_p(\log(k/\varepsilon))$, all logarithmic factors in the second stage depend only on k and ε.

1.4 Open problems

For $p > 2$, the known lower bound for strong coresets is

$$\widetilde{\Omega}\left(\frac{k^{p/2}}{\varepsilon} + \frac{k}{\varepsilon^2}\right),$$

obtained from lower bounds for ℓ_p subspace embeddings [LWW21, WY23]. This does not match our upper bound $\widetilde{O}_p(k^{p/2}\varepsilon^{-2})$ as a joint function of k and ε. The main remaining question is whether the $p > 2$ upper bound can be improved to match the known lower bound up to logarithmic factors, or whether a stronger lower bound captures an unavoidable interaction between the rank and accuracy parameters. In contrast, for $1 \le p < 2$ our upper bound matches the known sampling lower bound under the parameter conditions stated above.

Organization. Section 2 collects the sampling and sketching models, Lewis-weight facts, and empirical-process tools used throughout the paper. Sections 3 and 4 prove the upper bound for $1 \le p < 2$ and record the matching known lower bound. The following three sections treat $p > 2$: the one-round sampling section states the baseline primitive from [WY25], Section 6 proves the sharper row-count estimate, and Section 7 completes the recursive construction. The appendices contain the empirical-process and entropy estimates used in the $p < 2$ analysis.

2 Preliminaries

2.1 Notation and sampling models

Throughout the paper, p is fixed within the regime under consideration. Constants denoted by c_p and C_p may depend on p but on no other parameter, and their values may change from one occurrence to the next. We write $\mathrm{nnz}(A)$ for the number of nonzero entries of A and use ω for the exponent of matrix multiplication. The notation $\widetilde{O}$, $\widetilde{\Omega}$, and $\widetilde{\Theta}$ suppresses logarithmic factors in the relevant parameters. A subscript p indicates that the implicit constant may depend on p. We write

$$\|Y\|_{p,2} := \left(\sum_i \|e_i^\top Y\|_2^p\right)^{1/p}.$$

For $F \subseteq \mathbb{R}^d$, write P_F for the orthogonal projector onto F and $Q_F = I - P_F$. Let $\mathrm{Gr}(j,d)$ denote the set of j-dimensional subspaces of $\mathbb{R}^d$. The set of subspaces of $\mathbb{R}^d$ of dimension at most k is

$$\mathcal{F}_k := \bigcup_{j=0}^{k} \mathrm{Gr}(j,d).$$

We use the equivalent notation

$$\Phi_A(F) = \mathrm{cost}_A(F) = \|AQ_F\|_{p,2}^p, \qquad \mathsf{OPT}_A := \min_{F\in\mathcal{F}_k} \mathrm{cost}_A(F).$$

We use the convention that

$$X = (1 \pm a)Y \pm b$$

means $(1-a)Y - b \le X \le (1+a)Y + b$. All logarithms are natural.

Definition 2.1 (Strong row coreset)**.** A reweighted row subset is represented by a matrix $S \in \mathbb{R}^{s \times n}$ with one nonzero entry in each row. Repeated input indices may be aggregated, and the size is the number of distinct retained rows. It is a $(1 \pm \varepsilon)$ strong row coreset for rank-at-most-k ℓ_p subspace approximation if

$$\mathrm{cost}_{SA}(F) = (1 \pm \varepsilon)\,\mathrm{cost}_A(F) \qquad \text{for every } F \in \mathcal{F}_k.$$

Definition 2.2 (Bernoulli ℓ_p sampling)**.** Given inclusion probabilities $\pi_1, \dots, \pi_N \in [0,1]$, let $\xi_1, \dots, \xi_N$ be independent Bernoulli variables with $\mathbb{P}[\xi_i = 1] = \pi_i$. The associated ℓ_p sampling matrix is diagonal with

$$S_{ii} := \begin{cases} \pi_i^{-1/p}\xi_i, & \pi_i > 0, \\ 0, & \pi_i = 0. \end{cases}$$

After sampling, zero rows are deleted from the active matrix. A product of successive sampling maps can be viewed, after padding deleted rows with zeros, as a single diagonal reweighting of the original rows.

Definition 2.3 (Sampling with replacement)**.** Let q be a probability distribution on $[N]$, and draw $I_1, \dots, I_m$ independently from q. The corresponding sampling matrix $S \in \mathbb{R}^{m \times N}$ has rows

$$S_{s,*} = (m q_{I_s})^{-1/p} e_{I_s}^\top, \qquad s \in [m].$$

Thus, if M has rows $z_i^\top$, then

$$\|SM\|_{p,2}^p = \frac{1}{m}\sum_{s=1}^{m} \frac{\|z_{I_s}\|_2^p}{q_{I_s}}.$$

Repeated draws may be merged, so the support contains at most m distinct input rows.

2.2 Pre-sparsification and composition

We use the following result of Woodruff and Yasuda as a preprocessing step. The theorem is stated for weighted inputs by absorbing each positive row weight into the corresponding row scaling.

Theorem 2.4 (Woodruff–Yasuda pre-sparsification [WY25])**.** *Fix $p \in [1,\infty)\setminus\{2\}$. Given $A \in \mathbb{R}^{N \times d}$, $1 \le k < d$, and $\xi, \delta \in (0, 1/2)$, there is a randomized algorithm that returns a weighted row subset $A_0 = S_0 A$ satisfying, with probability at least $1 - \delta$,*

$$\mathrm{cost}_{A_0}(F) = (1 \pm \xi)\,\mathrm{cost}_A(F) \qquad \textit{for every } F \in \mathcal{F}_k.$$

For a constant C_p, its support size is at most

$$\begin{cases} C_p k \xi^{-4/p} \left(\log \dfrac{C_p k}{\xi\delta}\right)^{C_p}, & 1 \le p < 2, \\ C_p k^{p/2} \xi^{-p} \left(\log \dfrac{C_p k}{\xi\delta}\right)^{C_p}, & p > 2. \end{cases} \tag{2.1}$$

The running time is $\widetilde{O}_p(\mathrm{nnz}(A) + d^\omega)$.

This is the row-subset strong-coreset guarantee in [WY25, Theorems 1.3 and 1.4]. Its dependence on ξ will be larger than that of our final coreset, but its row count is independent of the input row count N.

We also use their fast bicriteria construction for the low-rank split [WY25, Lemma 2.3].

Lemma 2.5 (Woodruff–Yasuda bicriteria subspace)**.** *Fix $1 \le p < 2$. Given $A \in \mathbb{R}^{N\times d}$, $1 \le k < d$, and $\delta \in (0, 1/2)$, there is a randomized algorithm that returns a set $\mathcal{J}$ of at most*

$$C_p k \left(\log \frac{C_p N}{\delta}\right)^{C_p}$$

rows whose span F_0 satisfies

$$\mathrm{cost}_A(F_0) \le \Gamma_0 \mathsf{OPT}_A$$

with probability at least $1-\delta$, where Γ_0 depends only on p. The algorithm runs in $\widetilde{O}_p(\mathrm{nnz}(A) + |\mathcal{J}|^\omega)$ time.

Lemma 2.6 (Composition of strong row coresets)**.** *Suppose $A_0 = S_0 A$ is a $(1 \pm \xi_0)$ strong row coreset for A, and $A_1 = S_1 A_0$ is a $(1 \pm \xi_1)$ strong row coreset for A_0. Then, for every $F \in \mathcal{F}_k$,*

$$(1-\xi_0)(1-\xi_1)\,\mathrm{cost}_A(F) \le \mathrm{cost}_{A_1}(F) \le (1+\xi_0)(1+\xi_1)\,\mathrm{cost}_A(F).$$

In particular, if $0 < \varepsilon < 1/2$ and $\xi_0 = \xi_1 = \varepsilon/3$, then A_1 is a $(1 \pm \varepsilon)$ strong row coreset for A. Moreover, $S_1 S_0$ is again a row-sampling matrix for the original rows of A.

Proof. Apply the two coreset inequalities successively. For the stated choice of parameters, $(1-\varepsilon/3)^2 \ge 1-\varepsilon$ and $(1+\varepsilon/3)^2 \le 1+\varepsilon$ when $\varepsilon < 1/2$. The final assertion follows because a product of row-selection and rescaling maps still selects and rescales rows of the original matrix. □

2.3 Leverage scores and Lewis weights

Throughout this subsection, zero rows are deleted before computing Lewis weights and restored with weight zero afterward.

Definition 2.7 (Leverage scores)**.** Let $M \in \mathbb{R}^{N\times r}$ have rows $m_i^\top$. The leverage score of row i is

$$\tau_i(M) := m_i^\top (M^\top M)^\dagger m_i.$$

Thus $\sum_i \tau_i(M) = \mathrm{rank}(M)$. If M has full column rank, the pseudoinverse may be replaced by the inverse, and $\tau_i(MR) = \tau_i(M)$ for every invertible change of coordinates $R \in \mathbb{R}^{r\times r}$.

Definition 2.8 (ℓ_p Lewis weights)**.** Let $M \in \mathbb{R}^{N\times r}$ have full column rank and rows $m_i^\top$. The ℓ_p Lewis weights of M are the positive numbers $w_1, \dots, w_N$ satisfying

$$w_i = \tau_i\Big(W^{1/2-1/p} M\Big), \qquad W = \mathrm{diag}(w_1, \dots, w_N). \tag{2.2}$$

Equivalently,

$$m_i^\top \Big(M^\top W^{1-2/p} M\Big)^{-1} m_i = w_i^{2/p}. \tag{2.3}$$

We call $\widetilde{w}_i$ an α-approximation to the Lewis weights if $\alpha^{-1} w_i \le \widetilde{w}_i \le \alpha w_i$ for every positive-weight row.

Thus Lewis weights are ordinary leverage scores after a self-consistent row reweighting; when $p = 2$, they reduce to the usual leverage scores of M. For $1 \le p < 4$, Cohen and Peng prove that these weights exist and are unique [CP15, Sec. 3, in particular Cor. 3.4]. Since they are leverage scores of a reweighted matrix,

$$\sum_i w_i = \operatorname{rank}(M) = r. \tag{2.4}$$

They are invariant under invertible changes of column coordinates: M and MR have the same Lewis weights whenever R is invertible. Accordingly, for vectors lying in an r-dimensional subspace, their Lewis weights mean the weights of their coordinates in any basis of that subspace; this is independent of the chosen basis.

Theorem 2.9 (Cohen–Peng Lewis-weight sampling)**.** *Fix $1 \le p < 2$ and $0 < \varepsilon < 1$. Let $M \in \mathbb{R}^{N\times r}$ have Lewis weights w_i, and let q be a distribution satisfying $q_i \gtrsim_p w_i/r$ whenever $w_i > 0$. If S is formed from $m = \widetilde{O}_p(r\varepsilon^{-2})$ i.i.d. samples from q as in Definition 2.3, then, with constant probability,*

$$\|SMx\|_p^p = (1 \pm \varepsilon)\|Mx\|_p^p \qquad \textit{simultaneously for every } x \in \mathbb{R}^r.$$

This is the standard Lewis-weight subspace-embedding theorem of Cohen and Peng [CP15, Theorem 7.1 and Lemma 7.4]; logarithmic factors depend on the fixed p. We later prove a vector-valued refinement tailored to subspace costs.

Algorithm 1 Iterative Algorithm to Compute the ℓ_p Lewis Weights [CP15]

1: Initialize $w = \mathbf{1} \in \mathbb{R}^N$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Let $W = \operatorname{diag}(w_1, \ldots, w_N)$
4: Let $\tau \in \mathbb{R}^N$ be a constant-factor approximation of the leverage scores of $W^{1/2-1/p}M$
5: Set $w_i \leftarrow (w_i^{2/p-1}\tau_i)^{p/2}$
6: **end for**
7: **return** w

Lemma 2.10 (Cohen–Peng iteration [CP15])**.** *Fix $1 \le p < 2$, and let $M \in \mathbb{R}^{N\times r}$ have full column rank. After $T = O_p(\log\log(N+2))$ iterations of Algorithm 1, the returned vector w is a constant-factor approximation to the ℓ_p Lewis weights of M.*

Lemma 2.11 (Lewis normalization)**.** *Let $M \in \mathbb{R}^{N\times r}$ have full column rank, let $1 \le p \le 2$, and let $m_i^\top$ and w_i be its rows and Lewis weights. There exist an invertible matrix $J \in \mathbb{R}^{r\times r}$ and unit vectors $y_i \in \mathbb{R}^r$ such that*

$$m_i = w_i^{1/p} J y_i, \qquad \sum_i w_i y_i y_i^\top = I_r, \qquad \sum_i w_i = r. \tag{2.5}$$

Proof. By the convention at the beginning of this subsection, M has no zero rows and every w_i is positive. Set

$$G = M^\top W^{1-2/p} M, \qquad J = G^{1/2}, \qquad y_i = w_i^{-1/p} G^{-1/2} m_i.$$

Since M has full column rank and W is positive diagonal, G is positive definite. Thus $G^{1/2}$ and $G^{-1/2}$ are well defined. By (2.3),

$$\|y_i\|_2^2 = w_i^{-2/p} m_i^\top G^{-1} m_i = 1,$$

and the definition of y_i gives $m_i = w_i^{1/p} J y_i$. Moreover,

$$\sum_i w_i y_i y_i^\top = G^{-1/2} \left(\sum_i w_i^{1-2/p} m_i m_i^\top \right) G^{-1/2} = G^{-1/2} G G^{-1/2} = I_r.$$

Taking traces and using $\|y_i\|_2 = 1$ gives $\sum_i w_i = r$. This proves (2.5). □

The next vector-valued sensitivity estimate is the offline $p \le 2$ specialization of the online bound in [WY23, Lemma 8.15]. We include a short proof because the offline Lewis normalization makes the argument particularly direct.

Lemma 2.12 (Lewis weights bound $(p,2)$-sensitivities)**.** *Let $M \in \mathbb{R}^{N\times r}$ have full column rank, let $1 \le p \le 2$, and let $m_i^\top$ and w_i be its rows and Lewis weights. For every $s \ge 1$ and every matrix $T \in \mathbb{R}^{s\times r}$,*

$$\|T m_i\|_2^p \le w_i \sum_j \|T m_j\|_2^p. \tag{2.6}$$

Proof. Use Lemma 2.11 to write $m_j = w_j^{1/p} J y_j$ with unit y_j and $\sum_j w_j y_j y_j^\top = I_r$, and set $S = TJ$. The claim is immediate if $S = 0$. Otherwise, since $p \le 2$ and $\|S y_j\|_2 \le \|S\|_{\mathrm{op}}$,

$$\sum_j w_j \|S y_j\|_2^p \ge \|S\|_{\mathrm{op}}^{p-2} \sum_j w_j \|S y_j\|_2^2 = \|S\|_{\mathrm{op}}^{p-2} \|S\|_F^2 \ge \|S\|_{\mathrm{op}}^p.$$

The conclusion follows from $\|T m_i\|_2^p = w_i \|S y_i\|_2^p$ and $\sum_j \|T m_j\|_2^p = \sum_j w_j \|S y_j\|_2^p$. □

In particular, the usual scalar ℓ_p sensitivity of row i satisfies

$$\sigma_i^{(p)}(M) := \sup_{x : Mx \ne 0} \frac{|m_i^\top x|^p}{\|Mx\|_p^p} \le w_i. \tag{2.7}$$

2.4 Euclidean sketches

Definition 2.13 (Oblivious subspace embedding)**.** A distribution over matrices $\Pi \in \mathbb{R}^{s\times d}$ is an (r, ζ, δ) oblivious subspace embedding (OSE) if, for every fixed subspace $L \subseteq \mathbb{R}^d$ with $\dim L \le r$,

$$\mathbb{P}_\Pi \Big[(1-\zeta)\|x\|_2^2 \le \|\Pi x\|_2^2 \le (1+\zeta)\|x\|_2^2 \text{ for every } x \in L \Big] \ge 1 - \delta.$$

The subspace must be fixed independently of the draw of Π.

We use the following standard Euclidean sketching facts. For any fixed vectors $v_1, \dots, v_N$ and $0 < \zeta, \delta < 1$, a Johnson–Lindenstrauss map with $O(\zeta^{-2} \log(N/\delta))$ rows simultaneously preserves all $\|v_i\|_2^2$ within a factor $1 \pm \zeta$ with probability at least $1 - \delta$ [JL84]. For a fixed r-dimensional subspace, sparse OSEs with constant distortion have $\widetilde{O}(r)$ rows and polylogarithmic column sparsity, including at inverse-polynomial failure probability [NN13]. Consequently, multiplying a matrix by the sparse OSE takes input-sparsity time up to logarithmic factors.

2.5 Uniform convergence tools

For a fixed query, a row sample gives an unbiased estimate of its cost. A scalar concentration inequality controls its error. A strong coreset, however, requires one sample to be accurate for every query, so we must control the largest sampling error over a function class. We first record the fixed-query estimate and then collect the uniform-convergence tools used for this purpose. All function classes in our applications are countable, finite, or replaced by a countable dense subclass before these tools are applied. Thus the suprema below are measurable in every use. Symmetrization and chaining also underlie Lewis-weight and sensitivity-sampling analyses for ℓ_p objectives [CP15, MMWY22, MO24]. In the broader coreset literature, pseudodimension-based uniform convergence is a standard ingredient of general sensitivity-sampling frameworks [FL11, BFL+21].

Concentration for one query. For a fixed positive-cost query, normalize each sampled contribution by the true cost. A sensitivity bound then gives a random summand X with $0 \le X \le K$ and $\mathbb{E}X = 1$. Necessarily $K \ge 1$, and $\mathrm{Var}(X) \le \mathbb{E}X^2 \le K$ and $|X - 1| \le K$. Bernstein's inequality [BLM13, Corollary 2.11] therefore gives, for independent copies $X_1, \dots, X_m$ and $0 < \eta \le 1$,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{s=1}^{m} X_s - 1\right| > \eta\right] \le 2\exp\left(-\frac{3m\eta^2}{8K}\right).$$

A zero-cost query has zero contribution from every row and is preserved exactly.

Empirical averages and symmetrization. Let $I_1, \dots, I_m$ be independent samples from a distribution q on the row indices. For a function f on the rows, write

$$P_m f = \frac{1}{m}\sum_{s=1}^{m} f(I_s), \qquad Pf = \mathbb{E}_{I\sim q}[f(I)].$$

Thus $P_m f$ is the sampled average and Pf is its expectation. For a function class $\mathcal{G}$, its absolute Rademacher complexity is

$$\mathfrak{R}_m^{\mathrm{abs}}(\mathcal{G}) = \mathbb{E}_{I,\varepsilon} \sup_{g\in\mathcal{G}} \left|\frac{1}{m}\sum_{s=1}^{m} \varepsilon_s g(I_s)\right|,$$

where the ε_s are independent uniform signs. If the sample $I_1, \dots, I_m$ is fixed, then $\mathfrak{R}_m^{\mathrm{abs}}(\mathcal{G} \mid I_1, \dots, I_m)$ denotes the same quantity with expectation only over the signs. The symmetrization lemma [vdVW96, Lemma 2.3.1] bounds the expected uniform sampling error by this signed process:

$$\mathbb{E}\sup_{g\in\mathcal{G}} |(P_m - P)g| \le 2\mathfrak{R}_m^{\mathrm{abs}}(\mathcal{G}). \tag{2.8}$$

Covering numbers and Dudley's bound. For a probability distribution ν on the row indices and real-valued functions f, g on these indices, write

$$\|f - g\|_{L_2(\nu)} = \left(\sum_i \nu_i |f(i) - g(i)|^2\right)^{1/2}.$$

Let $N(\delta, \mathcal{G}, L_2(\nu))$ be the minimum cardinality of a set $\mathcal{N} \subseteq \mathcal{G}$ such that every $g \in \mathcal{G}$ satisfies $\|g - \widetilde{g}\|_{L_2(\nu)} \le \delta$ for some $\widetilde{g} \in \mathcal{N}$. A δ-packing is a subset of $\mathcal{G}$ whose distinct elements have $L_2(\nu)$-distance greater than δ. Every maximal δ-packing is also a δ-cover. For the empirical distribution

P_m, this distance becomes

$$\|f-g\|_{L_2(P_m)} = \left(\frac{1}{m}\sum_{s=1}^{m}|f(I_s)-g(I_s)|^2\right)^{1/2}.$$

Covering numbers at all scales can be combined by Dudley's chaining argument. Applying the usual entropy integral [BLM13, Corollary 13.2] to an α-net and bounding the discarded increments by Cauchy–Schwarz gives the following truncated form; see also [BFT17, Lemma A.5] for a direct chaining proof of the normalized empirical Rademacher version. Set

$$\mathcal{G}^{\pm} = \mathcal{G}\cup(-\mathcal{G})\cup\{0\}, \qquad R_m = \sup_{g\in\mathcal{G}}\|g\|_{L_2(P_m)}.$$

If $R_m = 0$, the conditional Rademacher complexity is zero. If $R_m > 0$, then for every $0 < \alpha \le R_m$,

$$\mathfrak{R}_m^{\mathrm{abs}}(\mathcal{G}\mid I_1,\dots,I_m) \le 4\alpha + \frac{C}{\sqrt{m}}\int_{\alpha}^{R_m}\sqrt{\log N(u,\mathcal{G}^{\pm},L_2(P_m))}\,du. \tag{2.9}$$

Here symmetrizing the class converts the absolute supremum into an ordinary one, and adjoining zero makes R_m an upper bound on the relevant covering scales. The term 4α is the error from stopping the chaining at scale α.

For real-valued classes, pseudodimension is the VC dimension of the class of subgraphs [Hau92]. Equivalently, $\mathrm{Pdim}(\mathcal{G})$ is the largest v for which there are points $x_1,\dots,x_v$ and thresholds $t_1,\dots,t_v$ such that, for every $S\subseteq[v]$, some $g\in\mathcal{G}$ satisfies $g(x_j) > t_j$ for $j\in S$ and $g(x_j)\le t_j$ for $j\notin S$. The VC-subgraph covering theorem [vdVW96, Theorem 2.6.7] implies that if $\mathcal{G}$ takes values in $[0,1]$ and has pseudodimension $v\ge 1$, then, for every probability distribution ν,

$$\log N(\delta,\mathcal{G},L_2(\nu)) \le Cv\log(C/\delta), \qquad 0<\delta<1. \tag{2.10}$$

Lipschitz composition. Our loss functions are often Lipschitz functions of several simpler scalar coordinates. Fix the sample, and let $\{v_f\}$ be a class of functions from the row indices to $\mathbb{R}^a$, with coordinate functions $v_f = (v_{f,1},\dots,v_{f,a})$. Suppose that $\phi:\mathbb{R}^a\to\mathbb{R}$ satisfies $\phi(0)=0$ and $|\phi(u)-\phi(v)|\le L\|u-v\|_2$. Adjoin the zero map to the class if necessary, and let $(\varepsilon_{s,\ell})$ be independent uniform signs. Maurer's vector-contraction inequality [Mau16, Corollary 4], followed by subadditivity over the coordinates, gives

$$\begin{aligned}
&\mathbb{E}_{\varepsilon}\sup_f\left|\frac{1}{m}\sum_{s=1}^{m}\varepsilon_s\phi(v_f(I_s))\right| \\
&\quad\le 2\sqrt{2}L\,\mathbb{E}_{\varepsilon_{s,\ell}}\sup_f\frac{1}{m}\sum_{s=1}^{m}\sum_{\ell=1}^{a}\varepsilon_{s,\ell}v_{f,\ell}(I_s) \\
&\quad\le 2\sqrt{2}L\sum_{\ell=1}^{a}\mathbb{E}_{\varepsilon}\sup_f\left|\frac{1}{m}\sum_{s=1}^{m}\varepsilon_s v_{f,\ell}(I_s)\right|.
\end{aligned} \tag{2.11}$$

The factor 2 in the first inequality accounts for the absolute value; Maurer's theorem gives the factor $\sqrt{2}$ and the doubly indexed signs.

2.6 The Euclidean duality map

For x in a real Hilbert space, define

$$J_p(x) = \begin{cases}\|x\|_2^{p-2}x, & x\neq 0,\\ 0, & x=0.\end{cases}$$

Thus

$$\langle J_p(x), x\rangle = \|x\|_2^p, \qquad \|J_p(x)\|_2 = \|x\|_2^{p-1} \quad (x \neq 0).$$

At $p = 1$, this means $J_1(x) = x/\|x\|$ off zero and $J_1(0) = 0$. If g is a standard Gaussian in a finite-dimensional real Hilbert space, then, with $\gamma_p = \mathbb{E}|N(0,1)|^p$,

$$\mathbb{E}_g |\langle v, g\rangle|^p = \gamma_p \|v\|_2^p. \tag{2.12}$$

3 Coresets from a Low-Rank Split for $1 \le p < 2$

For any linear subspace $F \subseteq \mathbb{R}^d$, let P_F be the orthogonal projector onto F and let $Q_F = I - P_F$. The query family $\mathcal{F}_k$ consists of all such subspaces of dimension at most k, and the cost of $F \in \mathcal{F}_k$ is

$$\mathrm{cost}_A(F) = \|AQ_F\|_{p,2}^p = \sum_i \|Q_F a_i\|_2^p.$$

Our goal is to construct a single weighted sample of the rows of A that preserves this quantity simultaneously for every $F \in \mathcal{F}_k$.

Let $F_0 \subseteq \mathbb{R}^d$ be a bicriteria subspace satisfying

$$r := \dim(F_0) = \widetilde{O}(k), \qquad \mathrm{cost}_A(F_0) \le \Gamma_0 \mathsf{OPT}_A$$

for a constant $\Gamma_0 \ge 1$. For each row a_i, we use a vector $b_i \in F_0$ as a constant-factor approximation to its Euclidean projection onto F_0. The sketched regression procedure in Section 4 computes these vectors without forming an ambient-dimensional projection and guarantees

$$\|a_i - b_i\|_2 \le C \,\mathrm{dist}(a_i, F_0) = C\|Q_{F_0} a_i\|_2$$

for an absolute constant C. Let $e_i = a_i - b_i$, and let B and E have rows b_i and e_i, respectively. The construction also ensures that the rows of B span F_0, and hence

$$A = B + E, \qquad \mathrm{rank}(B) = r, \qquad R := \|E\|_{p,2}^p \le C^p \,\mathrm{cost}_A(F_0) \le \Gamma \mathsf{OPT}_A$$

for a constant $\Gamma \ge 1$. These are the only properties of the split used in this section.

The sampling distribution combines one score for each component of $A = B + E$. Let w_i be the ℓ_p Lewis weight of the low-rank row b_i. When $R > 0$, let $\rho_i = \|e_i\|_2^p / R$ be the fraction of the residual mass contributed by row i, and choose a distribution q satisfying

$$q_i \gtrsim_{p,\Gamma} \frac{w_i}{k + r}, \qquad q_i \gtrsim_{p,\Gamma} \rho_i.$$

Draw $m = \widetilde{O}_{p,\Gamma}((k + r)\varepsilon^{-2})$ indices $I_1, \ldots, I_m$ independently from q and retain row a_{I_s} with scaling $(mq_{I_s})^{-1/p}$. Intuitively, the Lewis-weight score controls the contribution of the low-rank component $Q_F b_i$, while the residual score controls that of $Q_F e_i$. In Section 3, we prove that this sampling rule preserves all queries simultaneously. In Section 4, we construct F_0, compute the regression fits b_i, and estimate both scores efficiently.

3.1 The low-rank-split theorem

Suppose that the decomposition $A = B + E$ satisfies

$$r = \operatorname{rank}(B), \qquad R = \|E\|_{p,2}^p \le \Gamma\, \mathsf{OPT}_A \tag{3.1}$$

for some constant $\Gamma \ge 1$. Put

$$D = k + r.$$

Let w_i denote the Lewis weight of the nonzero row b_i, with $w_i = 0$ when $b_i = 0$. When $R > 0$, put $\rho_i = \|e_i\|_2^p / R$; when $R = 0$, put $\rho_i = 0$.

Theorem 3.1 (Low-rank-split coreset)**.** *Assume $k, r \ge 1$, and let q be any probability distribution satisfying*

$$q_i \gtrsim_{p,\Gamma} \frac{w_i}{D} + \rho_i \qquad (i \in [n]). \tag{3.2}$$

Fix $0 < \varepsilon, \eta < 1/2$, and form S by drawing $m = \widetilde{O}_{p,\Gamma}(D\varepsilon^{-2}\eta^{-2})$ rows independently from q and reweighting them as in Definition 2.3. Then, with probability at least $1 - \eta$,

$$\operatorname{cost}_{SA}(F) = (1 \pm \varepsilon)\operatorname{cost}_A(F) \qquad \textit{simultaneously for every } F \in \mathcal{F}_k.$$

The suppressed factors are polylogarithmic in $n + D + \varepsilon^{-1} + \eta^{-1}$.

The balanced sensitivity lemma below supplies a distribution satisfying (3.2) and bounds each row's contribution to every query. We then normalize the query loss and decompose it into four scalar coordinate classes. Appendix A proves empirical process bounds for these classes with no dependence on d. In Section 3.1.5, we combine these estimates through symmetrization and vector contraction to prove Theorem 3.1.

If $R = 0$, then $A = B$, and the conclusion follows directly from Theorem A.9. Hence assume $R > 0$ throughout the remainder of the proof.

3.1.1 Balanced sensitivity

Lemma 3.2 (Balanced sensitivity for a low-rank split)**.** *Let $A = B + E$ satisfy (3.1) with $R > 0$, let $\rho_i = \|e_i\|^p / R$, and let w_i be the Lewis weight of b_i, with $w_i = 0$ when $b_i = 0$. With $D = k + r$, define*

$$s_i = w_i + D\rho_i. \tag{3.3}$$

Then

$$\|Q_F a_i\|_2^p \le C_{p,\Gamma} s_i \sum_j \|Q_F a_j\|_2^p \qquad (F \in \mathcal{F}_k).$$

Moreover, if $S = \sum_i s_i$ and $q_i = s_i/S$, then

$$S \le 2D, \qquad q_i \ge \frac{w_i}{2D}, \qquad q_i \ge \frac{\rho_i}{2}.$$

Proof. For a query F, write $C_F = \sum_i \|Q_F a_i\|^p$. Since $\mathsf{OPT}_A \le C_F$, $R \le \Gamma \mathsf{OPT}_A$, and Q_F is a contraction,

$$\|BQ_F\|_{p,2} \le \|AQ_F\|_{p,2} + \|EQ_F\|_{p,2} \le C_F^{1/p} + R^{1/p} \le (1 + \Gamma^{1/p}) C_F^{1/p}.$$

Lemma 2.12 and the row-wise p-triangle inequality give

$$\|Q_F a_i\|^p \le 2^{p-1}(\|Q_F b_i\|^p + \|Q_F e_i\|^p) \le 2^{p-1}((1 + \Gamma^{1/p})^p w_i + \Gamma \rho_i) C_F \le C_{p,\Gamma} s_i C_F.$$

Finally,

$$S = \sum_i w_i + D \sum_i \rho_i = r + D \le 2D.$$

The two lower bounds for q_i follow directly from (3.3). □

3.1.2 Balanced normalization

Fix a distribution q satisfying (3.2). There is a constant $c = c_{p,\Gamma} > 0$ such that

$$q_i \geq c\frac{w_i}{D}, \qquad q_i \geq c\rho_i. \tag{3.4}$$

For a query F with positive cost, abbreviate $Q = Q_F$ and write

$$C_F := \mathrm{cost}_A(F) = \sum_i \|Qa_i\|^p.$$

We normalize the low-rank and residual contributions of each row by $q_i C_F$.

The i.i.d. rescaling by $q_i^{-1/p}$ turns the Lewis scale $w_i^{1/p}$ and the residual scale $\rho_i^{1/p}$ into the row-wise factors

$$\alpha_i = (w_i/q_i)^{1/p}, \qquad \beta_i = (\rho_i/q_i)^{1/p},$$

with zero-over-zero values set to zero. The bounds in (3.4) imply

$$\alpha_i^p \leq c^{-1}D, \qquad \beta_i^p \leq c^{-1}. \tag{3.5}$$

Define the normalized residual scale

$$t_F := \left(\frac{R}{C_F}\right)^{1/p}.$$

Since $C_F \geq \mathsf{OPT}_A$ and $R \leq \Gamma\mathsf{OPT}_A$, we have $0 < t_F \leq \Gamma^{1/p}$. Let $L_\Gamma = 1 + \Gamma^{1/p}$. By the definition of C_F and the fact that orthogonal projection is nonexpansive,

$$\|AQ\|_{p,2} = C_F^{1/p}, \qquad \|EQ\|_{p,2} \leq \|E\|_{p,2} = R^{1/p} = t_F C_F^{1/p}.$$

Since $B = A - E$, it follows that

$$\|BQ\|_{p,2} \leq \|AQ\|_{p,2} + \|EQ\|_{p,2} \leq (1 + t_F)C_F^{1/p} \leq L_\Gamma C_F^{1/p}. \tag{3.6}$$

By coordinate invariance of Lewis weights and Lemma 2.11, there exist a matrix $J \in \mathbb{R}^{d\times r}$ whose columns span F_0 and vectors $y_i \in \mathbb{R}^r$ satisfying

$$b_i = w_i^{1/p} J y_i, \qquad \|y_i\|_2 \leq 1, \qquad \sum_i w_i y_i y_i^\top = I_r, \qquad \sum_i w_i = r. \tag{3.7}$$

Here $\|y_i\| = 1$ when $b_i \neq 0$, while $w_i = 0$ and $y_i = 0$ when $b_i = 0$. Likewise, write

$$e_i = R^{1/p}\rho_i^{1/p} u_i,$$

where u_i is a unit vector whenever $\rho_i > 0$ and may be chosen to be any unit vector otherwise. With these coordinates, define

$$x_F(i) = \alpha_i C_F^{-1/p} QJy_i = q_i^{-1/p} C_F^{-1/p} Qb_i, \qquad y_F(i) = t_F\beta_i Qu_i = q_i^{-1/p} C_F^{-1/p} Qe_i. \tag{3.8}$$

Consequently,

$$Z_F(i) = \|x_F(i) + y_F(i)\|^p = \frac{\|Qa_i\|^p}{q_i C_F}, \qquad \mathbb{E}_q Z_F = 1, \qquad 0 \leq Z_F(i) \leq C_{p,\Gamma,c} D. \tag{3.9}$$

The last inequality follows by combining Lemma 3.2 with (3.4). Thus the remainder of the proof uses the sampling distribution only through (3.4).

Finally, Equations (3.7) and (3.6) give

$$\sum_i w_i \|C_F^{-1/p} Q J y_i\|^p = \frac{\|BQ\|_{p,2}^p}{C_F} \le L_\Gamma^p$$

and applying Lemma 2.12 to the rows of B gives, for every i with $w_i > 0$,

$$\|C_F^{-1/p} Q J y_i\|^p = \frac{\|Q b_i\|^p}{w_i C_F} \le \frac{\|BQ\|_{p,2}^p}{C_F} \le L_\Gamma^p. \tag{3.10}$$

3.1.3 An exact four-coordinate decomposition

By (3.9), the normalized loss is the pth power of $\|x_F(i) + y_F(i)\|_2$. Fix F and i, and abbreviate $x = x_F(i)$ and $y = y_F(i)$. We first split y into its component parallel to x and its component orthogonal to x. If $x \neq 0$, set

$$\widehat{x} = \frac{x}{\|x\|_2}, \qquad c_F(i) = \langle \widehat{x}, y \rangle, \qquad y^\perp = y - c_F(i)\widehat{x}.$$

If $x = 0$, set $\widehat{x} = 0$, $c_F(i) = 0$, and $y^\perp = y$. In either case, $y^\perp$ is orthogonal to x and

$$x + y = (\|x\|_2 + c_F(i))\widehat{x} + y^\perp.$$

Consequently,

$$\|x_F(i) + y_F(i)\|_2^2 = |\|x_F(i)\|_2 + c_F(i)|^2 + \|y^\perp\|_2^2. \tag{3.11}$$

Taking the $p/2$ power in (3.11) and using (3.9) gives

$$Z_F(i) = \left(|\|x_F(i)\|_2 + c_F(i)|^2 + \|y^\perp\|_2^2 \right)^{p/2}. \tag{3.12}$$

The parallel contribution depends on the sign of $c_F(i)$. We therefore record its two magnitudes, their signed interaction, and the orthogonal residual power. Recall that $J_p(z) = \|z\|_2^{p-2} z$ for $z \neq 0$ and $J_p(0) = 0$. Define

$$\begin{aligned} T_F(i) &= \|x_F(i)\|_2^p, & V_F(i) &= |c_F(i)|^p, \\ H_F(i) &= p \, \langle J_p(x_F(i)), y_F(i) \rangle, & W_F^\perp(i) &= \operatorname{dist}(y_F(i), \operatorname{span}\{x_F(i)\})^p. \end{aligned} \tag{3.13}$$

When $x_F(i) \neq 0$,

$$H_F(i) = p \, \operatorname{sgn}(c_F(i)) \|x_F(i)\|_2^{p-1} |c_F(i)|;$$

when $x_F(i) = 0$, our convention $J_p(0) = 0$ gives $H_F(i) = 0$. Here T_F is the low-rank power, V_F is the parallel residual power, H_F records the signed interaction between the two, and $W_F^\perp$ is the orthogonal residual power.

Apply Lemma A.6 with $a = \|x_F(i)\|_2$, $b = |c_F(i)|$, and $\sigma = \operatorname{sgn}(c_F(i))$, taking $\sigma = 1$ when $c_F(i) = 0$. It shows that the parallel power $|\|x_F(i)\|_2 + c_F(i)|^p$ is a C_p-Lipschitz function of $(T_F(i), V_F(i), H_F(i))$ on the set of feasible coordinate triples. The outer $\ell_{2/p}$ norm is one-Lipschitz with respect to the ℓ_1 distance of the parallel and orthogonal powers. Consequently,

$$|Z_F(i) - Z_G(i)| \le C_p \big(|T_F(i) - T_G(i)| + |V_F(i) - V_G(i)| + |H_F(i) - H_G(i)| + |W_F^\perp(i) - W_G^\perp(i)| \big). \tag{3.14}$$

We next bound the empirical complexity of each coordinate class separately.

3.1.4 Control of the coordinate processes

Low-rank power. For $\|BQ_F\|_{p,2} > 0$, set

$$h_F(i) = \frac{\|Q_F b_i\|^p}{q_i \|BQ_F\|_{p,2}^p}, \qquad \lambda_F = \frac{\|BQ_F\|_{p,2}^p}{C_F} \le L_\Gamma^p.$$

Then $T_F = \lambda_F h_F$. By (3.4),

$$K := \max_{w_i>0} \frac{w_i}{q_i} \le c^{-1} D.$$

Since $0 \le \lambda_F \le L_\Gamma^p$,

$$\mathfrak{R}_m^{\mathrm{abs}}(\{T_F\}) \le L_\Gamma^p \mathfrak{R}_m^{\mathrm{abs}}(\{h_F\}).$$

Applying Theorem A.9 therefore gives

$$\mathfrak{R}_m^{\mathrm{abs}}(\{T_F\}) \le \widetilde{O}_{p,\Gamma,c}\left(\sqrt{D/m} + D/m\right). \tag{3.15}$$

If $\|BQ_F\|_{p,2} = 0$, then T_F is identically zero.

Parallel residual power. Equation (3.8) gives

$$x_F(i) = q_i^{-1/p} C_F^{-1/p} Q_F b_i, \qquad y_F(i) = t_F \beta_i Q_F u_i.$$

The scalar multiplying $Q_F b_i$ in the first identity is positive. Hence, when $Q_F b_i \neq 0$,

$$\frac{x_F(i)}{\|x_F(i)\|_2} = \frac{Q_F b_i}{\|Q_F b_i\|_2} \in F^\perp.$$

Recalling the definition of $c_F(i)$ and using $\langle v, Q_F u_i\rangle = \langle v, u_i\rangle$ for $v \in F^\perp$, we obtain

$$\begin{aligned} c_F(i) &= \langle x_F(i)/\|x_F(i)\|_2, y_F(i)\rangle \\ &= t_F\beta_i \langle Q_F b_i/\|Q_F b_i\|_2, Q_F u_i\rangle \\ &= t_F\beta_i \langle Q_F b_i/\|Q_F b_i\|_2, u_i\rangle . \end{aligned}$$

This formula also holds when $Q_F b_i = 0$ if a normalized zero vector is interpreted as zero. Since $b_i = w_i^{1/p} J y_i$, define

$$g_F(i) := \langle Q_F J y_i/\|Q_F J y_i\|_2, \beta_i u_i\rangle = \langle J_1(Q_F J y_i), \beta_i u_i\rangle .$$

Then $c_F = t_F g_F$ and $V_F = t_F^p |g_F|^p$. The vectors $\beta_i u_i$ have uniformly bounded norm by (3.5). Since $t_F^p \le \Gamma$,

$$\mathfrak{R}_m^{\mathrm{abs}}(\{V_F\}) \le \Gamma \mathfrak{R}_m^{\mathrm{abs}}(\{|g_F|^p\}).$$

Now Lemma A.1, the truncated Dudley bound (2.9), and the Lipschitz map $g \mapsto |g|^p$ give

$$\mathfrak{R}_m^{\mathrm{abs}}(\{V_F\}) \le \widetilde{O}_{p,\Gamma,c}(\sqrt{r/m}). \tag{3.16}$$

Signed interaction. Using (3.8) and the same projection identity as in the parallel residual calculation gives

$$H_F(i) = p\, t_F \left\langle J_p(\alpha_i C_F^{-1/p} Q_F J y_i), \beta_i u_i \right\rangle. \tag{3.17}$$

Lemma A.2 controls classes of the form

$$i \longmapsto p \left\langle J_p(\alpha_i A y_i), z_i \right\rangle, \qquad A : \mathbb{R}^r \to \mathcal{H}.$$

In (3.17), the corresponding choices are

$$A = C_F^{-1/p} Q_F J, \qquad z_i = \beta_i u_i.$$

The bounds used in the preceding paragraph give $\|z_i\|_2 \le c^{-1/p}$ and $t_F \le \Gamma^{1/p}$. Since t_F is independent of the row index, it contributes only a constant factor to the absolute Rademacher supremum.

Let $\nu = m^{-1} \sum_{s=1}^m \delta_{I_s}$ be the empirical law of the sampled indices. Recall that $P_m f = m^{-1} \sum_{s=1}^m f(I_s)$, and hence $\sum_i \nu_i f(i) = P_m f$. Under the choices above, the empirical p-energy in Lemma A.2 is

$$E_\nu = \sup_F \sum_i \nu_i \|\alpha_i C_F^{-1/p} Q_F J y_i\|^p = \sup_F P_m T_F. \tag{3.18}$$

Equations (3.5) and (3.10) imply, pointwise in i and F,

$$\|\alpha_i C_F^{-1/p} Q_F J y_i\|^p \le C_{p,\Gamma,c} D.$$

Thus $E_\nu \le C_{p,\Gamma,c} D$ deterministically. We first bound every logarithmic factor involving E_ν by a polylogarithmic function of $D+2$ and absorb it into the $\widetilde{O}$ notation. Applying Corollary A.3 now gives, for $p > 1$,

$$\mathfrak{R}_m^{\mathrm{abs}}(\{H_F\} \mid I_1, \dots, I_m) \le \widetilde{O}_{p,\Gamma,c}\left(\left(E_\nu^{1-1/p} + E_\nu^{(1-1/p)/2} \right) \sqrt{r/m} \right) \tag{3.19}$$

At $p = 1$, the endpoint statement of Corollary A.3 directly gives $\widetilde{O}_{1,\Gamma,c}(\sqrt{r/m})$.

For $p > 1$, it remains to control the random empirical p-energy E_ν in expectation. For every query with $\|BQ_F\|_{p,2} > 0$, the normalized low-rank contribution satisfies $\mathbb{E}_q h_F = 1$ and $T_F = \lambda_F h_F$, where $\lambda_F \le L_\Gamma^p$. If $\|BQ_F\|_{p,2} = 0$, then $T_F = 0$. Consequently,

$$E_\nu \le L_\Gamma^p \left(1 + \sup_{\|BQ_F\|_{p,2} > 0} |P_m h_F - 1| \right).$$

Symmetrization and Theorem A.9 imply that, for $m \ge \widetilde{\Omega}_{p,\Gamma,c}(D)$,

$$\mathbb{E} \sup_{\|BQ_F\|_{p,2} > 0} |P_m h_F - 1| \le 2\mathfrak{R}_m^{\mathrm{abs}}(\{h_F\}) \le \widetilde{O}_{p,c}\left(\sqrt{D/m} + D/m \right) \le C_{p,\Gamma,c}.$$

It follows that

$$\mathbb{E} E_\nu \le C_{p,\Gamma,c}. \tag{3.20}$$

Both exponents of E_ν in (3.19) lie in $(0,1)$. After the deterministic treatment of the logarithmic factors above, Jensen's inequality applies to these pure powers. Together with (3.20), it gives

$$\mathbb{E}\mathfrak{R}_m^{\mathrm{abs}}(\{H_F\} \mid I_1, \dots, I_m) \le \widetilde{O}_{p,\Gamma,c}(\sqrt{r/m}). \tag{3.21}$$

Together with the endpoint estimate above, this proves (3.21) for the full range $1 \le p < 2$.

Orthogonal residual power. By definition, $W_F^{\perp}(i) = \operatorname{dist}(y_F(i), \operatorname{span}\{x_F(i)\})^p$. Equation (3.8) shows that $\operatorname{span}\{x_F(i)\} = \operatorname{span}\{Q_F b_i\}$ and $y_F(i) = t_F \beta_i Q_F u_i$. Therefore

$$W_F^{\perp}(i) = t_F^p \beta_i^p \operatorname{dist}(Q_F u_i, \operatorname{span}\{Q_F b_i\})^p. \tag{3.22}$$

The distance in (3.22) can be written without projected vectors:

$$\begin{aligned}\operatorname{dist}(Q_F u_i, \operatorname{span}\{Q_F b_i\}) &= \inf_{a\in\mathbb{R}} \|Q_F(u_i - a b_i)\|_2 \\ &= \inf_{a\in\mathbb{R},\, f\in F} \|u_i - a b_i - f\|_2 \\ &= \operatorname{dist}(u_i, F + \operatorname{span}\{b_i\}).\end{aligned}$$

Consequently,

$$W_F^{\perp}(i) = t_F^p \beta_i^p \operatorname{dist}(u_i, F + \operatorname{span}\{b_i\})^p.$$

After rescaling this class to have envelope at most one, Corollary A.5 and the truncated Dudley bound (2.9) give

$$\mathfrak{R}_m^{\mathrm{abs}}(\{W_F^{\perp}\}) \le \widetilde{O}_{p,\Gamma,c}(\sqrt{k/m}). \tag{3.23}$$

The $O_{p,\Gamma,c}(m^{-1})$ truncation terms for V_F and $W_F^{\perp}$ are absorbed into their square-root bounds. The D/m term in (3.15) remains because the low-rank class has envelope of order D.

3.1.5 Proof of Theorem 3.1

For $m \ge \widetilde{\Omega}_{p,\Gamma,c}(D)$, the symmetrization bound (2.8), with the measurability convention from Section 2, gives

$$\mathbb{E} \sup_{\substack{F\in\mathcal{F}_k \\ C_F>0}} |P_m Z_F - 1| \le 2\mathbb{E}_{I,\varepsilon} \sup_{\substack{F\in\mathcal{F}_k \\ C_F>0}} \left| \frac{1}{m} \sum_s \varepsilon_s Z_F(I_s) \right|.$$

For each F and i, collect the four coordinates into

$$v_F(i) = \big(T_F(i), V_F(i), H_F(i), W_F^{\perp}(i)\big) \in \mathbb{R}^4.$$

On the set of feasible coordinate vectors, let ϕ be the decoder satisfying $\phi(v_F(i)) = Z_F(i)$. It maps the origin to zero, and (3.14) gives

$$|\phi(u) - \phi(v)| \le C_p \|u - v\|_1 \le 2C_p \|u - v\|_2.$$

Extend ϕ to all of $\mathbb{R}^4$ if necessary without increasing its Lipschitz constant. Applying Maurer's vector-contraction inequality (2.11) and then separating the four coordinates gives

$$\begin{aligned}&\mathbb{E}_{I,\varepsilon} \sup_{\substack{F\in\mathcal{F}_k \\ C_F>0}} \left| \frac{1}{m} \sum_s \varepsilon_s Z_F(I_s) \right| \\ &\quad \le C_p \Big(\mathfrak{R}_m^{\mathrm{abs}}(\{T_F\}) + \mathfrak{R}_m^{\mathrm{abs}}(\{V_F\}) + \mathbb{E}_I \mathfrak{R}_m^{\mathrm{abs}}(\{H_F\} \mid I_1, \dots, I_m) + \mathfrak{R}_m^{\mathrm{abs}}(\{W_F^{\perp}\}) \Big).\end{aligned}$$

Substituting (3.15), (3.16), (3.21), and (3.23), and recalling that $D = k + r$, yields

$$\mathbb{E} \sup_{\substack{F\in\mathcal{F}_k \\ C_F>0}} |P_m Z_F - 1| \le \widetilde{O}_{p,\Gamma,c} \left(\sqrt{\frac{k+r}{m}} + \frac{k+r}{m} \right). \tag{3.24}$$

For every fixed sample, the normalized loss $F \mapsto P_m Z_F$ is continuous on each positive-cost stratum, so the countable-dense-subclass convention indeed gives the displayed supremum. Zero-cost queries satisfy $AQ_F = SAQ_F = 0$ and are handled separately.

Taking $m = \widetilde{O}_{p,\Gamma}(D\varepsilon^{-2}\eta^{-2})$ makes the right-hand side of (3.24) at most $\eta\varepsilon$ after adjusting the hidden constant. Markov's inequality therefore gives, with probability at least $1-\eta$,

$$1 - \varepsilon \le P_m Z_F \le 1 + \varepsilon \qquad (F \in \mathcal{F}_k). \tag{3.25}$$

By Definition 2.3, the sampled matrix SA has at most m nonzero rows, and

$$P_m Z_F = \frac{\|SAQ_F\|_{p,2}^p}{\|AQ_F\|_{p,2}^p}$$

for every positive-cost query, proving the required $1 \pm \varepsilon$ cost guarantee. □

4 Input-Sparsity-Time Construction for $1 \le p < 2$

In this section, we give an efficient implementation of the sampling procedure in Theorem 3.1. In the proof of Theorem 1.1, we apply it to the Woodruff–Yasuda coreset A_0 from Theorem 2.4. Since A_0 has $\mathrm{poly}_p(k/\varepsilon)$ rows, the resulting logarithmic factors depend only on k and ε, rather than on the original row count n.

Proposition 4.1 (Implementing the low-rank-split sampling procedure)**.** *Fix $1 \le p < 2$. Given $A \in \mathbb{R}^{N\times d}$, $1 \le k < d$, and $0 < \varepsilon < 1/2$, there is a randomized algorithm running in*

$$\widetilde{O}_p\Big(\mathrm{nnz}(A) + k^\omega + k\varepsilon^{-2}\Big)$$

time that returns a $(1\pm\varepsilon)$ strong row coreset with probability at least $5/6$. Its support size is at most

$$C_p k\varepsilon^{-2}\left(\log\frac{C_p(N+k)}{\varepsilon}\right)^{C_p}. \tag{4.1}$$

The suppressed factors in the running time are polylogarithmic in $N + d + \varepsilon^{-1}$.

Proof. We describe the construction and indicate where each logarithmic factor enters. The claim is immediate when $A = 0$, so assume otherwise. Apply Lemma 2.5 with failure probability $1/24$. It returns a set $\mathcal{J}$ of

$$m_0 \le C_p k \log^{C_p}(N+2)$$

rows whose span F_0 has dimension $r_0 \le m_0$ and satisfies

$$\mathrm{cost}_A(F_0) \le \Gamma_0 \mathsf{OPT}_A \tag{4.2}$$

for a constant Γ_0 depending only on p. The construction takes $\widetilde{O}_p(\mathrm{nnz}(A) + m_0^\omega)$ time.

We next construct the split used in Theorem 3.1. Draw a sparse right OSE $\Pi \in \mathbb{R}^{s\times d}$ with $s = \widetilde{O}(m_0)$ and inverse-polynomial failure probability. Each of the N subspaces

$$F_0 + \mathrm{span}\{a_i\}, \qquad i \in [N],$$

has dimension at most $r_0 + 1$, where $a_i^\top$ denotes the ith row of A. A union bound therefore shows that, with high probability, Π has constant distortion on all of them simultaneously. Condition on

this event, and let the columns of $V \in \mathbb{R}^{s\times r_0}$ form an orthonormal basis for ΠF_0. Define b_i by the sketched least-squares regression

$$b_i = \operatorname*{argmin}_{b\in F_0} \|\Pi(a_i - b)\|_2, \qquad e_i = a_i - b_i.$$

The OSE is injective on F_0, so the minimizer is unique. Since $VV^\top$ is the orthogonal projector onto ΠF_0, the regression condition is equivalently

$$\Pi b_i = VV^\top \Pi a_i. \tag{4.3}$$

If $b_i^* = P_{F_0} a_i$, sketched optimality and the OSE inequalities give

$$\|e_i\|_2 \le C\|a_i - b_i^*\|_2.$$

Consequently, for the matrices B, E with rows b_i, e_i,

$$A = B + E, \qquad \operatorname{rank}(B) = r_0, \qquad \|E\|_{p,2}^p \le C_p \operatorname{cost}_A(F_0) \le \Gamma_1 \mathsf{OPT}_A.$$

The equality $\operatorname{rank}(B) = r_0$ follows because the selected rows spanning F_0 are fitted exactly.

It remains to estimate the two parts of the balanced score. Set $H = A\Pi^\top$. By (4.3), the ith row of $H(I - VV^\top)$ is

$$a_i^\top \Pi^\top (I - VV^\top) = \left((I - VV^\top)\Pi a_i\right)^\top = (\Pi e_i)^\top.$$

Since $e_i \in F_0 + \operatorname{span}\{a_i\}$, the OSE event gives $c\|e_i\|_2 \le \|\Pi e_i\|_2 \le C\|e_i\|_2$ for every i, where $c, C > 0$ are absolute constants. Applying a Johnson–Lindenstrauss map on the right preserves the norms of these N rows simultaneously up to a constant factor. Raising the resulting row-norm estimates to the pth power gives values $\widetilde{r}_i$ satisfying

$$c_p\|e_i\|_2^p \le \widetilde{r}_i \le C_p\|e_i\|_2^p \qquad \text{for every } i \in [N]. \tag{4.4}$$

We next compute the Lewis weights of B in r_0 dimensions. Consider the coordinate map

$$T : F_0 \longrightarrow \mathbb{R}^{r_0}, \qquad T(x) = V^\top \Pi x.$$

This map is invertible because Π is injective on F_0 and the columns of V form a basis for ΠF_0. Moreover,

$$T(b_i) = V^\top \Pi b_i = V^\top \Pi a_i = (HV)_{i,*}^\top.$$

Thus the matrix

$$C = HV \in \mathbb{R}^{N\times r_0}$$

has $T(b_i)^\top$ as its ith row. Since the rows b_i span F_0 and T is invertible, C has full column rank. By the coordinate invariance of Lewis weights in Definition 2.8, C and B have the same Lewis weights.

Run the Cohen–Peng iteration from Lemma 2.10 on C. At iteration t, let W_t be the current diagonal row-weight matrix and apply an input-sparsity-time leverage-score routine to $W_t C$ [CW13]. Using the factorization $C = HV = A\Pi^\top V$, the routine can be implemented without forming the dense matrix C. Together with Lemma 2.10, this yields simultaneous constant-factor estimates $\widetilde{w}_i$ of the Lewis weights. Fresh sketches and a union bound cover the $O_p(\log\log(N+2))$ iteration calls.

Using the residual estimates from (4.4), set

$$\widetilde{\rho}_i = \frac{\widetilde{r}_i}{\sum_j \widetilde{r}_j}, \qquad \widetilde{q}_i = \frac{\widetilde{w}_i + (k + r_0)\widetilde{\rho}_i}{\sum_j (\widetilde{w}_j + (k + r_0)\widetilde{\rho}_j)}$$

when the residual mass is positive. If the residual mass is zero, put $\widetilde{\rho}_i = 0$ and normalize only the $\widetilde{w}_i$ in the definition of $\widetilde{q}_i$. Constant-factor accuracy and $\sum_i w_i = r_0$ imply

$$\widetilde{q}_i \gtrsim_p \frac{w_i}{k + r_0}, \qquad \widetilde{q}_i \gtrsim_p \rho_i.$$

Thus Theorem 3.1 applies with $\eta = 1/12$.

Finally, $r_0 \le C_p k \log^{C_p}(N+2)$. Expanding the logarithmic factors in Theorem 3.1 and increasing C_p gives the support bound (4.1). Assigning total failure probability $1/24$ to the sketching routines, the bicriteria, sketching, and sampling events fail with total probability at most $1/24+1/24+1/12 = 1/6$.

We next analyze the running time of the remaining steps. Applying the sparse right OSE to A takes $\widetilde{O}(\mathrm{nnz}(A))$ time, and computing the basis V takes $\widetilde{O}(m_0^\omega)$ time because $s = \widetilde{O}(m_0)$. Computing the residual estimates in (4.4) with the Johnson–Lindenstrauss sketch takes $\widetilde{O}(\mathrm{nnz}(A) + m_0^\omega)$ time. Each Cohen–Peng iteration takes $\widetilde{O}_p(\mathrm{nnz}(A) + m_0^\omega)$ time. The $O_p(\log\log(N+2))$ iterations are absorbed by the logarithmic factors in $\widetilde{O}_p$. Finally, forming the sampling distribution and drawing the $\widetilde{O}_p(k\varepsilon^{-2})$ weighted rows take $\widetilde{O}_p(\mathrm{nnz}(A) + k\varepsilon^{-2})$ time. Since $m_0 \le C_p k \log^{C_p}(N+2)$, the total running time is

$$\widetilde{O}_p\Big(\mathrm{nnz}(A) + k^\omega + k\varepsilon^{-2}\Big),$$

as claimed. □

Proof of Theorem 1.1. For the row bound independent of n, first apply Theorem 2.4 to A with accuracy $\varepsilon/3$ and failure probability $1/12$. Denote its output by $A_0 = S_0 A$ and its number of rows by N_0. Equation (2.1) gives

$$N_0 \le C_p k \varepsilon^{-4/p} \left(\log \frac{C_p k}{\varepsilon}\right)^{C_p}. \tag{4.5}$$

Condition on this event and apply Proposition 4.1 to A_0 with accuracy $\varepsilon/3$. Since

$$\log \frac{C_p(N_0 + k)}{\varepsilon} \le C_p \log \frac{C_p k}{\varepsilon},$$

the resulting matrix $A_1 = S_1 A_0$ has at most

$$C_p k \varepsilon^{-2} \left(\log \frac{C_p k}{\varepsilon}\right)^{C_p}$$

rows after adjusting C_p. By Lemma 2.6, $A_1 = S_1 S_0 A$ is a $(1 \pm \varepsilon)$ strong row coreset for A. The two stages fail with total probability at most $1/12 + 1/6 = 1/4$, which is below $1/3$.

Because A_0 is a weighted row subset of A, $\mathrm{nnz}(A_0) \le \mathrm{nnz}(A)$, and $k^\omega \le d^\omega$. The total running time is therefore

$$\widetilde{O}_p\Big(\mathrm{nnz}(A) + d^\omega + k\varepsilon^{-2}\Big).$$

□

Lower bound. For every fixed $1 \le p < 2$, the upper bound in Theorem 1.1 is tight up to logarithmic factors whenever $0 < \varepsilon < 1/2$ and $k + 1 \ge C \log(1/\varepsilon)$. Woodruff and Yasuda observed the following reduction to sampling-based subspace embeddings [WY23, Sec. 1.3]. Set $d_0 = k + 1$, and for a unit vector $x \in \mathbb{R}^{d_0}$ let $F_x = x^\perp$. Then

$$\|B(I - P_{F_x})\|_{p,2}^p = \|Bx\|_p^p.$$

Thus a strong row coreset for residual queries over all subspaces of rank at most k gives a for-all sampling-based ℓ_p embedding. The lower bound of Li, Wang, and Woodruff [LWW21, Theorem 6.1], applied in dimension $d_0 = k + 1$, therefore requires $\widetilde{\Omega}_p(k\varepsilon^{-2})$ rows under the stated conditions. Consequently, the worst-case strong row coreset size is $\widetilde{\Theta}_p(k\varepsilon^{-2})$ in this parameter range.

5 One-Round Sampling for $p > 2$

5.1 Ridge leverage scores

Let $A \in \mathbb{R}^{N\times d}$, and let A_k be a best rank-k approximation to A in Frobenius norm. Set

$$\lambda_A := \frac{\|A - A_k\|_F^2}{k}.$$

The ridge leverage score of row i is

$$\tau_i^{\lambda_A}(A) := a_i^\top (A^\top A + \lambda_A I)^\dagger a_i,$$

where $\dagger$ denotes the Moore–Penrose pseudoinverse. When the matrix is clear, we abbreviate this score to τ_i. Following Woodruff and Yasuda [WY25], we call sampling with weights $\left(\tau_i^{\lambda_A}(A)\right)^{p/2}$ *root ridge leverage score sampling*.

Lemma 5.1 (Ridge leverage score trace bound [CMM17])**.** *For $\lambda_A = \|A - A_k\|_F^2/k$,*

$$\sum_{i=1}^{N} \tau_i^{\lambda_A}(A) \le 2k.$$

Proof. If $\lambda_A = 0$, then $\|A - A_k\|_F = 0$ and hence $\operatorname{rank}(A) \le k$. In this case,

$$\sum_{i=1}^{N} \tau_i^{0}(A) = \operatorname{tr}\Big(A^\top A (A^\top A)^\dagger\Big) = \operatorname{rank}(A) \le k.$$

Assume now that $\lambda_A > 0$. Let $\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$ be the singular values of A. Then

$$\begin{aligned}\sum_{i=1}^{N} \tau_i^{\lambda_A}(A) &= \operatorname{tr}\Big(A^\top A (A^\top A + \lambda_A I)^\dagger\Big) \\ &= \sum_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda_A}.\end{aligned}$$

The first k summands contribute at most k. For the remaining summands,

$$\sum_{j>k} \frac{\sigma_j^2}{\sigma_j^2 + \lambda_A} \le \frac{1}{\lambda_A} \sum_{j>k} \sigma_j^2 = k.$$

□

5.2 The Woodruff–Yasuda one-round theorem

We use the following normalization of the Woodruff–Yasuda one-round guarantee for $p > 2$ [WY25, Theorem 3.3].

Theorem 5.2 (Woodruff–Yasuda one-round theorem)**.** *Fix $p > 2$. There are constants $c_p > 0$ and $C_p \ge 1$ such that the following holds. Let $A \in \mathbb{R}^{N\times d}$ with $N \ge 3$, let $\zeta, \eta \in (0, 1/2)$, and suppose*

$$N \ge \frac{C_p k^{p/2}}{\zeta}.$$

Set

$$\alpha := \frac{c_p \zeta^2}{(\log N)^3 + \log(1/\eta)}.$$

Let S be an ℓ_p sampling matrix whose probabilities satisfy

$$q_i \geq \min\left\{1, \frac{N^{p/2-1}(\tau_i^{\lambda_A}(A))^{p/2}}{\alpha}\right\}.$$

Then, with probability at least $1-\eta$, simultaneously for every $F \in \mathcal{F}_k$,

$$\mathrm{cost}_{SA}(F) = (1 \pm \zeta)\,\mathrm{cost}_A(F) \pm C_p \zeta \left(\log \frac{C_p N}{\zeta \eta}\right)^{p/2} \mathsf{OPT}_A.$$

5.3 Uniform parameter tuning

The next lemma tunes the internal parameter ζ so that both the multiplicative and additive errors are at most a prescribed value θ. The upper bound $M \geq N$ allows the same tuning to be used throughout the recursion.

Lemma 5.3 (Uniform tuning)**.** *Fix $M \geq N \geq 3$ and $\theta, \eta \in (0, 1/2)$. Let*

$$L := \log\left(\frac{C_p M}{\theta \eta}\right), \qquad \zeta := \frac{c_p \theta}{L^{p/2}},$$

where $c_p > 0$ is sufficiently small. Suppose

$$N \geq C_p k^{p/2} \theta^{-1} L^{p/2}.$$

Let S be obtained by applying Theorem 5.2 with internal accuracy ζ and failure probability η. Then, with probability at least $1-\eta$, simultaneously for every $F \in \mathcal{F}_k$,

$$\mathrm{cost}_{SA}(F) = (1 \pm \theta)\,\mathrm{cost}_A(F) \pm \theta \mathsf{OPT}_A.$$

Moreover, the sampling threshold obeys

$$\alpha \geq \frac{c_p \theta^2}{L^p\big((\log M)^3 + \log(1/\eta)\big)}.$$

Proof. The assumed lower bound on N implies the size condition in Theorem 5.2, after changing C_p. Also $\zeta \leq \theta$. For the additive term, observe that

$$\begin{aligned}\log \frac{C_p N}{\zeta \eta} &\leq \log \frac{C_p M}{\theta \eta} + \log \frac{\theta}{\zeta} \\ &\leq L + \log(c_p^{-1} L^{p/2}) \leq C_p L.\end{aligned}$$

Thus, by taking c_p sufficiently small,

$$C_p \zeta \left(\log \frac{C_p N}{\zeta \eta}\right)^{p/2} \leq \theta.$$

Together with $\zeta \leq \theta$, this turns the one-round guarantee in Theorem 5.2 into

$$\mathrm{cost}_{SA}(F) = (1 \pm \theta)\,\mathrm{cost}_A(F) \pm \theta \mathsf{OPT}_A$$

simultaneously for every $F \in \mathcal{F}_k$. Finally,

$$\alpha = \frac{c_p \zeta^2}{(\log N)^3 + \log(1/\eta)} \geq \frac{c_p \theta^2}{L^p\big((\log M)^3 + \log(1/\eta)\big)},$$

after another adjustment of c_p. □

6 Threshold-aware row-count analysis

The retained row count is governed by a sum of truncated powers of ridge leverage scores. The crucial point is that keeping the entire sampling scale inside the truncation allows this sum to be bounded by the linear ridge leverage score mass in Lemma 5.1.

Lemma 6.1 (Threshold summation). *Let $v > 1$, $c > 0$, and $\tau_1, \dots, \tau_N \geq 0$. Then*

$$\sum_{i=1}^{N} \min\{1, c\tau_i^v\} \leq c^{1/v} \sum_{i=1}^{N} \tau_i.$$

In particular, if $p > 2$, $v = p/2$, and $\sum_i \tau_i \leq 2k$,

$$\sum_{i=1}^{N} \min\{1, c\tau_i^{p/2}\} \leq 2kc^{2/p}.$$

Proof. Set $y_i = c^{1/v}\tau_i$. For every $y \geq 0$,

$$\min\{1, y^v\} \leq y.$$

Indeed, if $0 \leq y \leq 1$, then $y^v \leq y$, while if $y > 1$, then $\min\{1, y^v\} = 1 < y$. Therefore

$$\min\{1, c\tau_i^v\} = \min\{1, y_i^v\} \leq y_i = c^{1/v}\tau_i.$$

Summing over i proves the claim. Equality at $\tau_i = c^{-1/v}$ shows that the factor $c^{1/v}$ is best possible. □

Corollary 6.2 (Expected one-round row count). *Fix $p > 2$ and $1 \leq k < d$. Let $A \in \mathbb{R}^{N\times d}$ and $\alpha > 0$, and set $\lambda_A = \|A - A_k\|_F^2/k$. Let S be the ℓ_p sampling matrix associated with*

$$q_i = \min\left\{1, \frac{N^{p/2-1}\tau_i^{p/2}}{\alpha}\right\},$$

where $\tau_i = \tau_i^{\lambda_A}(A)$. Then

$$\mathbb{E}[\mathrm{nnz}(S)] \leq 2k\alpha^{-2/p}N^{1-2/p}.$$

Proof. Apply Lemma 6.1 with $c = N^{p/2-1}/\alpha$ and use Lemma 5.1:

$$\begin{aligned}\mathbb{E}[\mathrm{nnz}(S)] &= \sum_{i=1}^{N} q_i \\ &\leq 2k\left(\frac{N^{p/2-1}}{\alpha}\right)^{2/p} = 2k\alpha^{-2/p}N^{1-2/p}.\end{aligned}$$

□

To use the improved expectation bound in the recursive construction, we also need a high-probability bound on the retained row count.

Lemma 6.3 (Bernoulli row-count concentration). *Let Z be a sum of independent Bernoulli random variables with mean μ. For every $\eta \in (0, 1/2)$,*

$$\mathbb{P}\big[Z > 3(\mu + \log(1/\eta))\big] \leq \eta.$$

Proof. Set $L := \log(1/\eta)$ and

$$t := 2\sqrt{\mu L} + 2L.$$

Bernstein's inequality gives

$$\mathbb{P}[Z - \mu \geq t] \leq \exp\left(-\frac{t^2}{2(\mu + t/3)}\right) \leq e^{-L} = \eta.$$

Here the second inequality follows from $t^2 \geq 2L(\mu + t/3)$. Moreover, $2\sqrt{\mu L} \leq \mu + L$, and hence

$$\mu + t \leq 2\mu + 3L \leq 3(\mu + L).$$

The claimed bound follows. □

Combining Corollary 6.2 and Lemma 6.3, one sampling round on an N-row matrix produces at most

$$O\Big(k\alpha^{-2/p}N^{1-2/p} + \log(1/\eta)\Big)$$

rows with probability at least $1 - \eta$.

7 Recursive Construction for $p > 2$

We obtain the final coreset by applying the one-round sampler recursively to the current weighted matrix. The recursion stops when the number of active rows falls below a prescribed threshold or after a bounded number of rounds. The next subsection analyzes the resulting row-count recurrence and shows that the latter number of rounds suffices to reach its fixed-point bound. We then apply the Woodruff–Yasuda pre-sparsification from Theorem 2.4. The reduced input size ensures that a single recursive reduction achieves the desired row bound.

7.1 Inner recursive sampling

Fix an active matrix $A^{(0)} \in \mathbb{R}^{N\times d}$ with $N \geq 1$, a target relative error $\rho \in (0, 1/2)$, and a failure probability $\eta \in (0, 1/2)$. Set

$$\gamma := 1 - \frac{2}{p} \in (0, 1)$$

and

$$r := \max\left\{1, \left\lceil \frac{\log\log(\max\{e^e, N\})}{-\log\gamma} \right\rceil\right\}.$$

Choose

$$\theta := \frac{c\rho}{r}, \qquad \eta_0 := \frac{\eta}{4r},$$

where $c > 0$ is a sufficiently small absolute constant, and define

$$L := \log\left(\frac{C_p N}{\theta\eta_0}\right), \qquad \zeta := \frac{c_p\theta}{L^{p/2}}.$$

Finally, set the stopping threshold

$$N_{\min} := \frac{C_p k^{p/2}}{\zeta} = \frac{C_p}{c_p} k^{p/2}\theta^{-1}L^{p/2}.$$

The parameter choices imply, after increasing C_p if necessary,

$$r \le C_p L, \qquad \log N \le L, \qquad \log(1/\eta_0) \le L, \qquad L \le C_p \log\left(\frac{C_p N}{\rho\eta}\right). \tag{7.1}$$

These relations follow from $\theta\eta_0 = c\rho\eta/(4r^2)$ and $r = O_p(\log\log(\max\{e^e, N\}))$.

In round t, let $A^{(t-1)}$ be the current weighted active matrix and let N_{t-1} be its number of rows. If $N_{t-1} \le N_{\min}$, stop and output the current matrix. Otherwise compute the ridge leverage scores of $A^{(t-1)}$ at

$$\lambda_{t-1} := \frac{\|A^{(t-1)} - (A^{(t-1)})_k\|_F^2}{k},$$

set

$$\alpha_t := \frac{c_p \zeta^2}{(\log N_{t-1})^3 + \log(1/\eta_0)},$$

and sample independently with

$$q_i^{(t)} := \min\left\{1, \frac{N_{t-1}^{p/2-1}(\tau_i^{\lambda_{t-1}}(A^{(t-1)}))^{p/2}}{\alpha_t}\right\}.$$

Let $A^{(t)} = S_t A^{(t-1)}$ after deleting zero rows. The sampler only deletes rows, so $N_t \le N_{t-1} \le N$ in every round. The procedure executes at most r rounds.

7.2 Row-count contraction and its fixed point

We first isolate the deterministic fixed-point calculation. We then combine it with the high-probability row bound from Section 6.

Lemma 7.1 (Fixed-point recurrence; cf. [MMWY22])**.** *Let $\gamma \in (0,1)$ and let $N_0 \ge 3$. Suppose*

$$N_t \le K N_{t-1}^{\gamma} \qquad (t = 1, \dots, r),$$

where $K \ge 1$ and

$$r \ge \frac{\log\log N_0}{-\log\gamma}.$$

Then

$$N_r \le e K^{1/(1-\gamma)}.$$

In particular, when $\gamma = 1 - 2/p$,

$$N_r \le e K^{p/2}.$$

Proof. If some $N_t = 0$, the conclusion is immediate. Otherwise set $a_t := \log N_t$. Then

$$a_t \le \gamma a_{t-1} + \log K.$$

Unrolling gives

$$a_r \le \gamma^r \log N_0 + (\log K)\sum_{j=0}^{r-1}\gamma^j \le 1 + \frac{\log K}{1-\gamma}.$$

Exponentiating proves the claim. □

Proposition 7.2 (High-probability row bound)**.** *Fix $p > 2$, let $A^{(0)} \in \mathbb{R}^{N\times d}$ with $N \geq 1$ and $1 \leq k < d$, and let $\rho, \eta \in (0, 1/2)$. Let the recursive procedure and the parameter L be as defined in Section 7.1. With probability at least $1 - \eta/2$, the output has at most*

$$C_p \frac{k^{p/2}}{\rho^2} L^{p+5}$$

rows.

Proof. The definition of L is the common tuning scale in Lemma 5.3 with $M = N$, target error θ, and failure probability η_0. Since $N_{t-1} \leq N$, the thresholds satisfy

$$\alpha_t \geq \alpha_* := \frac{c_p \theta^2}{L^p((\log N)^3 + \log(1/\eta_0))}.$$

Conditioned on all previous rounds, Corollary 6.2 gives

$$\mu_t := \mathbb{E}[N_t \mid A^{(t-1)}] \leq 2k\alpha_t^{-2/p} N_{t-1}^{1-2/p}.$$

By Lemma 6.3, except with conditional probability at most η_0,

$$\begin{aligned} N_t &\leq C_p k \alpha_t^{-2/p} N_{t-1}^{1-2/p} + C_p \log(1/\eta_0) \\ &\leq K N_{t-1}^{1-2/p}, \end{aligned}$$

where

$$K := C_p \left(k\alpha_*^{-2/p} + \log(1/\eta_0) \right).$$

After increasing C_p, we have $K \geq 1$. The last inequality uses $N_{t-1}^{1-2/p} \geq 1$. A union bound shows that this recurrence holds in every executed round with probability at least $1 - r\eta_0 \geq 1 - \eta/4$.

If the algorithm stops early, the output has at most

$$N_{\min} \leq C_p k^{p/2} \theta^{-1} L^{p/2} \leq C_p \frac{k^{p/2}}{\rho^2} L^{p+5}$$

rows, using Equation (7.1) and $\rho < 1$. Otherwise, the algorithm executes all r rounds. In this case $N > N_{\min} \geq 3$, and the definition of r gives

$$r \geq \frac{\log\log N}{-\log(1 - 2/p)}.$$

Thus all hypotheses of Lemma 7.1 hold, and it yields

$$\begin{aligned} N_r &\leq eK^{p/2} \\ &\leq C_p \left(k^{p/2} \alpha_*^{-1} + (\log(1/\eta_0))^{p/2} \right). \end{aligned}$$

Now

$$\begin{aligned} \alpha_*^{-1} &\leq C_p \theta^{-2} L^p ((\log N)^3 + \log(1/\eta_0)) \\ &\leq C_p \rho^{-2} r^2 L^{p+3} \\ &\leq C_p \rho^{-2} L^{p+5}. \end{aligned}$$

The last two inequalities use $\theta = c\rho/r$ and Equation (7.1). Also $(\log(1/\eta_0))^{p/2} \leq L^{p/2}$, which is absorbed by the same bound because $k \geq 1$ and $\rho < 1$. This proves the claim after adjusting constants. □

7.3 Accumulation of the approximation error

The row-count analysis does not by itself control the distortion accumulated across the recursive rounds. We next show that the mixed additive-multiplicative guarantees from the executed rounds compose to a multiplicative guarantee.

For $t \geq 0$, write

$$C_t(F) := \mathrm{cost}_{A^{(t)}}(F), \qquad \mathsf{OPT}_t := \min_{F \in \mathcal{F}_k} C_t(F).$$

Lemma 7.3 (Composition of additive-multiplicative rounds)**.** *Let $\theta \in (0, 1/2)$ and let $T \geq 0$ be an integer. Suppose that for every $t = 1, \ldots, T$ and every $F \in \mathcal{F}_k$,*

$$C_t(F) = (1 \pm \theta) C_{t-1}(F) \pm \theta \mathsf{OPT}_{t-1}.$$

If $T\theta \leq c_0$ for a sufficiently small absolute constant c_0, then

$$C_T(F) = (1 \pm CT\theta) C_0(F) \qquad \text{for every } F \in \mathcal{F}_k,$$

where C is an absolute constant.

Proof. Let F^* minimize C_0. Since $\mathsf{OPT}_{t-1} \leq C_{t-1}(F^*)$,

$$C_t(F^*) \leq (1+\theta) C_{t-1}(F^*) + \theta \mathsf{OPT}_{t-1} \leq (1+2\theta) C_{t-1}(F^*).$$

Hence

$$\mathsf{OPT}_t \leq C_t(F^*) \leq (1+2\theta)^t \mathsf{OPT}_0 \leq 2\mathsf{OPT}_0$$

when $T\theta$ is sufficiently small. For arbitrary F, unrolling the upper recurrence gives

$$\begin{aligned} C_T(F) &\leq (1+\theta)^T C_0(F) + \theta \sum_{s=0}^{T-1} (1+\theta)^{T-1-s} \mathsf{OPT}_s \\ &\leq e^{T\theta} C_0(F) + 2T\theta e^{T\theta} \mathsf{OPT}_0 \\ &\leq (1 + CT\theta) C_0(F), \end{aligned}$$

because $\mathsf{OPT}_0 \leq C_0(F)$. Similarly,

$$\begin{aligned} C_T(F) &\geq (1-\theta)^T C_0(F) - \theta \sum_{s=0}^{T-1} (1-\theta)^{T-1-s} \mathsf{OPT}_s \\ &\geq (1 - T\theta) C_0(F) - 2T\theta \mathsf{OPT}_0 \\ &\geq (1 - 3T\theta) C_0(F). \end{aligned}$$

This proves the claim. □

7.4 The recursive reduction

The row-count and approximation analyses above combine into the following reduction. We will apply it once to the pre-sparsified matrix.

Theorem 7.4 (Recursive reduction)**.** *Fix $p > 2$. Given $A \in \mathbb{R}^{N \times d}$ with $N \geq 1$, $k \geq 1$, and $\rho, \eta \in (0, 1/2)$, the recursive procedure above outputs a weighted row subset A' such that, with probability at least $1 - \eta$,*

$$\mathrm{cost}_{A'}(F) = (1 \pm \rho)\, \mathrm{cost}_A(F) \qquad \text{for every } F \in \mathcal{F}_k,$$

and A' has at most

$$C_p \frac{k^{p/2}}{\rho^2} \left(\log \frac{C_p N}{\rho \eta} \right)^{p+5} \qquad \text{rows.}$$

Proof. The row bound follows from Proposition 7.2 and Equation (7.1), since

$$L \le C_p \log\left(\frac{C_p N}{\rho\eta}\right).$$

For the approximation guarantee, condition on the current active matrix in each round. Because the algorithm continues only when $N_{t-1} > N_{\min}$, Lemma 5.3 applies with $M = N$, target θ, and failure probability η_0. Thus, except with conditional probability at most η_0, the round satisfies

$$C_t(F) = (1 \pm \theta) C_{t-1}(F) \pm \theta \mathsf{OPT}_{t-1}$$

simultaneously for every $F \in \mathcal{F}_k$. Let $T \le r$ be the number of rounds actually executed. A union bound over at most r rounds gives all of these events with probability at least $1 - r\eta_0 \ge 1 - \eta/4$. Since $T\theta \le r\theta = c\rho$, Lemma 7.3 gives a $(1 \pm \rho)$ multiplicative guarantee when c is sufficiently small. Combining this event with Proposition 7.2 gives success probability at least $1 - 3\eta/4 \ge 1 - \eta$. □

Remark 7.5 (Implementation and running time)**.** The implementation does not require the exact value of λ_A or the exact ridge leverage scores. Using the routines in [CMM17, WY25], one may compute a constant-factor estimate of λ_A together with high-probability score estimates that, after adjusting constants, satisfy

$$\tau_i^{\lambda_A}(A) \le \widehat{\tau}_i \le C\tau_i^{\lambda_A}(A), \qquad \sum_i \widehat{\tau}_i = O(k).$$

Using $\widehat{\tau}_i$ in the sampling probabilities only increases the probabilities required by Theorem 5.2, while Lemma 6.1 applies directly to the linear mass $\sum_i \widehat{\tau}_i$. Thus every approximation and row-count bound above remains valid up to constants. The score estimation and sampling primitives run in $\widetilde{O}_p(\mathrm{nnz}(A) + d^\omega)$ time over the recursive reduction as shown in [WY25]. Allocating the unused constant fraction of the failure budget to the estimation events preserves the success probability in Theorem 7.4.

7.5 Completion of the $p > 2$ construction

Proof of Theorem 1.2. Apply Theorem 2.4 to A with accuracy $\varepsilon/3$ and failure probability $\delta/2$. On its success event, the resulting weighted row subset $A_0 = S_0 A$ satisfies

$$N_0 \le C_p k^{p/2} \varepsilon^{-p} \left(\log \frac{C_p k}{\varepsilon\delta}\right)^{C_p} \tag{7.2}$$

and preserves every cost within a factor $1 \pm \varepsilon/3$.

Condition on this event and apply Theorem 7.4 once to A_0, with $\rho = \varepsilon/3$ and $\eta = \delta/2$. The output $A_1 = S_1 A_0$ has at most

$$\frac{C_p k^{p/2}}{\varepsilon^2} \left(\log \frac{C_p N_0}{\varepsilon\delta}\right)^{p+5}$$

rows. By Equation (7.2), and because p is fixed,

$$\log \frac{C_p N_0}{\varepsilon\delta} \le C_p \log \frac{C_p k}{\varepsilon\delta}.$$

After adjusting C_p, this is the row bound in Theorem 1.2.

The two stages fail with total probability at most δ. On their common success event, Lemma 2.6 shows that A_1 preserves all rank-at-most-k subspace costs within a factor $1 \pm \varepsilon$. Moreover, $A_1 = (S_1 S_0)A$ remains a weighted row subset of the original input.

The first stage takes $\widetilde{O}_p(\mathrm{nnz}(A) + d^\omega)$ time by Theorem 2.4. By Remark 7.5, the second stage takes $\widetilde{O}_p(\mathrm{nnz}(A_0) + d^\omega)$ time. Since $\mathrm{nnz}(A_0) \le \mathrm{nnz}(A)$, the total running time is $\widetilde{O}_p(\mathrm{nnz}(A) + d^\omega)$. □

Acknowledgements

The proof for $p > 2$ was first obtained using a fully automated Gemini-based agentic system developed internally at Google. The proof for $1 \leq p < 2$ was obtained with contributions from the authors, the aforementioned system, and a harness built on top of GPT-5.6 Sol. The authors have verified the proofs and edited them for clarity of presentation, and take responsibility for the final version. The authors would like to thank Taisuke Yasuda for helpful discussions.

References

[BFL+21] Vladimir Braverman, Dan Feldman, Harry Lang, Adiel Statman, and Samson Zhou. Efficient coreset constructions via sensitivity sampling. In *Proceedings of the 13th Asian Conference on Machine Learning (ACML)*, volume 157 of *Proceedings of Machine Learning Research*, pages 948–963. PMLR, 2021.

[BFT17] Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc., 2017.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[CMM17] Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1758–1777. SIAM, 2017.

[CP15] Michael B. Cohen and Richard Peng. ℓ_p row sampling by lewis weights. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 183–192. ACM, 2015. Full version: arXiv:1412.0588.

[CW13] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 81–90. ACM, 2013.

[CW15] Kenneth L. Clarkson and David P. Woodruff. Input sparsity and hardness for robust subspace approximation. In *Proceedings of the 56th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 310–329. IEEE, 2015.

[CW25] Moses Charikar and Erik Waingarten. The Johnson–Lindenstrauss lemma for clustering and subspace approximation: From coresets to dimension reduction. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3172–3209. SIAM, 2025. Full version: arXiv:2205.00371.

[DTV11] Amit Deshpande, Madhur Tulsiani, and Nisheeth K. Vishnoi. Algorithms and hardness for subspace approximation. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 482–496. SIAM, 2011.

[DV07] Amit Deshpande and Kasturi R. Varadarajan. Sampling-based dimension reduction for subspace approximation. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, pages 641–650. ACM, 2007.

[FKW21] Zhili Feng, Praneeth Kacham, and David P. Woodruff. Dimensionality reduction for the sum-of-distances metric. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 3220–3229. PMLR, 2021.

[FL11] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 569–578. ACM, 2011.

[FMSW10] Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P. Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 630–649. SIAM, 2010.

[FSS20] 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. Preliminary version in SODA 2013.

[GJ95] Paul W. Goldberg and Mark R. Jerrum. Bounding the Vapnik–Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 1995.

[GRSW12] Venkatesan Guruswami, Prasad Raghavendra, Rishi Saket, and Yi Wu. Bypassing UGC from some optimal geometric inapproximability results. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 699–717. SIAM, 2012.

[Hau92] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[HV20] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in Euclidean spaces: Importance sampling is nearly optimal. In *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing (STOC)*, pages 1416–1429. ACM, 2020.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

[KR15] Michael Kerber and Sharath Raghvendra. Approximation and streaming algorithms for projective clustering via random projections. In *Proceedings of the 27th Canadian Conference on Computational Geometry (CCCG)*, pages 179–185, 2015.

[LWW21] Yi Li, Ruosong Wang, and David P. Woodruff. Tight bounds for the subspace sketch problem with applications. *SIAM Journal on Computing*, 50(4):1287–1335, 2021.

[Mag07] Avner Magen. Dimensionality reductions in ℓ_2 that preserve volumes and distance to affine spaces. *Discrete & Computational Geometry*, 38(1):139–153, 2007.

[Mau16] Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In *Algorithmic Learning Theory (ALT)*, volume 9925 of *Lecture Notes in Computer Science*, pages 3–17. Springer, 2016.

[MMWY22] Cameron Musco, Christopher Musco, David P. Woodruff, and Taisuke Yasuda. Active linear regression for ℓ_p norms and beyond. In *Proceedings of the 63rd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 744–753. IEEE, 2022.

[MO24] Alexander Munteanu and Simon Omlor. Optimal bounds for ℓ_p sensitivity sampling via ℓ_2 augmentation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 36769–36796. PMLR, 2024.

[NN13] Jelani Nelson and Huy L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 117–126. IEEE, 2013.

[SV12] Nariankadu D. Shyamalkumar and Kasturi R. Varadarajan. Efficient subspace approximation algorithms. *Discrete & Computational Geometry*, 47(1):44–63, 2012.

[SW18] Christian Sohler and David P. Woodruff. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *Proceedings of the 59th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 802–813. IEEE, 2018.

[Tal90] Michel Talagrand. Embedding subspaces of L_1 into ℓ_1^N. *Proceedings of the American Mathematical Society*, 108(2):363–369, 1990.

[Tal95] Michel Talagrand. Embedding subspaces of L_p in ℓ_p^N. In Joram Lindenstrauss and Vitali D. Milman, editors, *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1992–94*, volume 77 of *Operator Theory: Advances and Applications*, pages 311–326. Birkhäuser Basel, 1995.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996.

[VX12] Kasturi R. Varadarajan and Xin Xiao. On the sensitivity of shape fitting problems. In Deepak D'Souza, Jaikumar Radhakrishnan, and Kavitha Telikepalli, editors, *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2012)*, volume 18 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 486–497, Dagstuhl, Germany, 2012. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[WY23] David P. Woodruff and Taisuke Yasuda. New subset selection algorithms for low rank approximation: Offline and online. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1802–1813. ACM, 2023. Full version: arXiv:2304.09217.

[WY25] David P. Woodruff and Taisuke Yasuda. Root ridge leverage score sampling for ℓ_p subspace approximation. In *Proceedings of the 66th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 2266–2291. IEEE, 2025. Full version: arXiv:2407.03262.

A Auxiliary Process Estimates for $1 \le p < 2$

The proof of Theorem 3.1 decomposes each normalized query loss into the four coordinates in (3.13). The directional and duality-map estimates below control V_F and H_F, respectively. The row-augmented subspace estimate controls $W_F^{\perp}$, and the final Lewis-weight estimate controls T_F. The scalar composition lemma then recovers the full loss from these four quantities. We formulate the estimates independently of the coreset notation so that their hypotheses are transparent when they are invoked in the main proof.

A.1 Directional and duality-map classes

Both classes in this subsection are indexed by linear maps out of an r-dimensional space, but their values may lie in an arbitrarily large Hilbert space. The estimates therefore retain the parameter r while avoiding any dependence on the ambient dimension. The first class records the pairing of the unit direction $J_1(Ay_i)$ with z_i. The second replaces J_1 by J_p, which also records the magnitude of Ay_i. In both proofs, one Gaussian sketch is shared by the entire function class. After sketching, the functions are controlled by at most Lr real parameters, independently of the dimension of the Hilbert space. For the applications in this paper, the reader may simply take $\mathcal{H} = \mathbb{R}^D$ for an arbitrary D; the Hilbert-space formulation emphasizes that the bounds do not depend on D.

Lemma A.1 (Directional entropy at $p = 1$)**.** *Let $r, M \ge 1$ and $B \ge 0$, let $\mathcal{H}$ be a real Hilbert space, and let ν be a probability law supported on at most M indices. Let $y_i \in \mathbb{R}^r$ and $z_i \in \mathcal{H}$ satisfy $\|z_i\| \le B$. For arbitrary linear maps $A : \mathbb{R}^r \to \mathcal{H}$, define*

$$g_A(i) = \langle J_1(Ay_i), z_i \rangle .$$

Then, for every $\delta > 0$,

$$\log N(\delta, \{g_A\}, L_2(\nu)) \le CB^2 r \delta^{-2} \log\big(C(M + r + 2)\big). \tag{A.1}$$

The convention $J_1(0) = 0$ is included.

Proof. Since the zero map $A = 0$ gives the zero function and $|g_A(i)| \le B$, a single $L_2(\nu)$-ball centered at zero covers the class when $\delta \ge B$. Hence assume $\delta < B$, so $L = \lceil CB^2\delta^{-2} \rceil \ge 2$ after fixing the absolute constant. Fix an arbitrary finite δ-packing $\mathcal{P}$. Only the finitely many functions in this packing matter, so choose one representing map A for each of them and use $\mathcal{P}$ also for this set of representatives. Only the finitely many vectors Ay_i with $A \in \mathcal{P}$ and $i \in \operatorname{supp}(\nu)$, together with the vectors z_i for those same indices, matter. Let $g_1, \dots, g_L$ be independent standard Gaussian vectors in their finite-dimensional span and put

$$\widehat{g}_A(i) = \frac{1}{\gamma_1 L} \sum_{\ell=1}^{L} \operatorname{sgn}\big(\langle g_\ell, Ay_i \rangle\big) \langle g_\ell, z_i \rangle ,$$

Here $\operatorname{sgn}(0) = 0$, and γ_1 is the Gaussian moment defined in (2.12). For $x \ne 0$, rotational invariance gives $\mathbb{E}[\operatorname{sgn}(\langle g, x \rangle) g] = \gamma_1 J_1(x)$; both sides are zero when $x = 0$. Taking the inner product with z_i gives $\mathbb{E}\widehat{g}_A(i) = g_A(i)$ for $A \in \mathcal{P}$, and $\mathbb{E}|\widehat{g}_A(i) - g_A(i)|^2 \le CB^2/L$. Averaging over $\mathcal{P}$ and applying Markov's inequality shows that the expected fraction of maps for which the $L_2(\nu)$ error exceeds $\delta/8$ is at most $1/8$, after increasing the constant in L. Fix a realization for which at least $7/8$ of the packing is approximated within $\delta/8$.

For fixed $g_1, \dots, g_L$, write $X = GA \in \mathbb{R}^{L \times r}$. The coefficients $\langle g_\ell, z_i \rangle$ are now fixed, so the sketch is determined by the ternary signs of $(Xy_i)_\ell$. As the row $X_{\ell,*}$ varies in $\mathbb{R}^r$, these are the signs

of $M_0 := |\operatorname{supp}(\nu)| \le M$ linear polynomials in r variables. Discard the identically zero forms, whose signs are fixed, and let $M_1 \le M_0$ be the number that remain. If $M_1 \ge r$, the zero-inclusive sign-assignment bound of Goldberg and Jerrum [GJ95, Corollary 2.1] gives at most $(8eM_1/r)^r$ patterns; if $M_1 < r$, the trivial bound is 3^{M_1}. In either case the number is at most $[C(M+r+2)]^r$. Multiplying over the L rows gives at most $[C(M+r+2)]^{Lr}$ possible sketches. The retained packing therefore satisfies the logarithmic bound in (A.1), and restoring the discarded elements changes its size by at most a factor of $8/7$. The same estimate holds for every finite packing, so a maximal packing yields the required cover. The sign count includes zero signs, and hence no separate continuity argument is needed. □

Lemma A.2 (Duality-map entropy)**.** *Let $r, M \ge 1$ and $B \ge 0$, let $\mathcal{H}$ be a real Hilbert space, and let ν be supported on at most M indices. Let $y_i \in \mathbb{R}^r$, $\alpha_i \ge 0$, and $z_i \in \mathcal{H}$ satisfy $\|z_i\| \le B$. Let $\mathcal{A}$ be a nonempty family of linear maps $A : \mathbb{R}^r \to \mathcal{H}$. Fix $1 \le p \le 2$ and define*

$$h_A(i) = p\,\langle J_p(\alpha_i A y_i), z_i\rangle\,.$$

If $1 < p \le 2$, define the empirical p-energy of the class by

$$E_\nu = \sup_{A\in\mathcal{A}} \sum_i \nu_i \|\alpha_i A y_i\|^p,$$

and assume $E_\nu < \infty$. Then, for every $\delta > 0$,

$$\log N(\delta, \{h_A\}, L_2(\nu)) \le \widetilde{O}_{p,B}\Big(r E_\nu^{2-2/p}\delta^{-2}\Big)\,. \tag{A.2}$$

The logarithms depend only on $M + r + 2$, $1 + 1/\delta$, and $1 + E_\nu$. In this range,

$$\sup_A \|h_A\|_{L_2(\nu)} \le C_{p,B} E_\nu^{1-1/p}. \tag{A.3}$$

At $p = 1$, the entropy is $\widetilde{O}_B(r\delta^{-2})$ for every $\delta > 0$, and the class has radius $O(B)$.

Proof. Assume first $1 < p \le 2$, $B > 0$, and $E_\nu > 0$. Put $\psi_p(0) = 0$ and $\psi_p(t) = |t|^{p-2}t$ for $t \ne 0$. Fix an arbitrary finite δ-packing $\mathcal{P}$, choose one representing map A for each function, and use $\mathcal{P}$ also for this set of representatives. Let $\mathcal{H}_0$ be the span of the finitely many vectors Ay_i with $A \in \mathcal{P}$ and $i \in \operatorname{supp}(\nu)$, together with the vectors z_i for those same indices. Let $g_1, \dots, g_L$ be independent standard Gaussians in $\mathcal{H}_0$ and define

$$\widehat{h}_A(i) = \frac{p}{\gamma_p L} \sum_{\ell=1}^{L} \psi_p(\langle g_\ell, \alpha_i A y_i\rangle)\,\langle g_\ell, z_i\rangle\,. \tag{A.4}$$

Differentiating $\mathbb{E}|\,\langle g, x\rangle\,|^p = \gamma_p \|x\|^p$ gives

$$\mathbb{E}[\psi_p(\langle g, x\rangle) g] = \gamma_p J_p(x),$$

and hence $\mathbb{E}\widehat{h}_A = h_A$. To bound the variance despite the correlation between the two Gaussian factors, suppose first that $x \ne 0$. Write $\widehat{x} = x/\|x\|$ and $z = a\widehat{x} + z_\perp$, where $z_\perp \perp x$. For a standard scalar Gaussian ξ, independence of the parallel and orthogonal Gaussian components gives

$$\begin{aligned}\mathbb{E}\,|\psi_p(\langle g, x\rangle)\,\langle g, z\rangle|^2 &= \|x\|^{2p-2}\left(a^2\mathbb{E}|\xi|^{2p} + \|z_\perp\|^2\mathbb{E}|\xi|^{2p-2}\right)\\ &\le C_p\|z\|^2\|x\|^{2p-2}.\end{aligned}$$

The same bound is immediate when $x = 0$. Since ν is a probability law and $2p-2 \le p$, monotonicity of moments gives, for every nonnegative sequence (a_i),

$$\sum_i \nu_i a_i^{2p-2} \le \left(\sum_i \nu_i a_i^p \right)^{(2p-2)/p}.$$

Applying this inequality to $a_i = \|\alpha_i A y_i\|$ yields

$$\sup_A \mathbb{E}\|\widehat{h}_A - h_A\|_{L_2(\nu)}^2 \le \frac{C_{p,B} E_\nu^{2-2/p}}{L}. \tag{A.5}$$

The pointwise bound $|h_A(i)| \le pB\|\alpha_i A y_i\|^{p-1}$ and the same moment inequality give

$$\|h_A\|_{L_2(\nu)} \le pB \left(\sum_i \nu_i \|\alpha_i A y_i\|^{2p-2} \right)^{1/2} \le pB E_\nu^{1-1/p}. \tag{A.6}$$

Thus the class has diameter at most $2pBE_\nu^{1-1/p}$. When $\delta \ge 2pBE_\nu^{1-1/p}$, a single ball centered at any member of the class gives a cover. In the remaining case, choose

$$L = \left\lceil C_{p,B} E_\nu^{2-2/p} \delta^{-2} \right\rceil. \tag{A.7}$$

The nontrivial-scale assumption makes the quantity inside the ceiling bounded below by a positive constant depending only on p, B.

Write $Gx = (\langle g_\ell, x \rangle)_{\ell=1}^L$, $X_A = GA$, and equip $\mathbb{R}^{L \times r}$ with the quotient seminorm

$$\|X\|_{\nu,\alpha,p} = \left(\sum_i \nu_i \alpha_i^p \|X y_i\|_2^p \right)^{1/p}.$$

This seminorm can vanish on a nonzero matrix, because only its action on the vectors $\alpha_i y_i$ in the support of ν matters. We therefore identify two matrices when their difference has zero seminorm. Let $\mathcal{N} = \{X : \|X\|_{\nu,\alpha,p} = 0\}$. The quotient $\mathcal{V} = \mathbb{R}^{L \times r}/\mathcal{N}$ is a normed space of dimension at most Lr; the sketched function depends only on the class of X in this quotient. Gaussian moments give

$$\mathbb{E}\|X_A\|_{\nu,\alpha,p}^p \le C_p L^{p/2} E_\nu.$$

We next choose one sketch that works for a large subset of the packing. Increase the constant in (A.7) so that, for every $A \in \mathcal{P}$,

$$\mathbb{P}\{\|\widehat{h}_A - h_A\|_{L_2(\nu)} > \delta/8\} \le 1/64, \qquad \mathbb{P}\{\|X_A\|_{\nu,\alpha,p} > C_p \sqrt{L} E_\nu^{1/p}\} \le 1/64.$$

The first inequality follows from (A.5) and Markov's inequality, and the second follows from the preceding pth-moment estimate. The expected fraction of packing elements failing each inequality is at most $1/64$; a second Markov inequality makes each fraction at most $1/8$ except on an event of probability at most $1/8$.

For every $i \in \mathrm{supp}(\nu)$, Gaussian concentration for the norm of a Gaussian vector gives

$$\mathbb{P}\left\{ \|Gz_i\| > CB\sqrt{L + \log(8M+8)} \right\} \le \frac{1}{8M+8}.$$

A union bound makes this simultaneous envelope event have probability at least $7/8$. The two fraction bounds and the envelope event therefore hold simultaneously with positive probability. Fix

such a realization of G and remove the two exceptional subsets. At least $3|\mathcal{P}|/4$ elements remain; each retained function is approximated within $\delta/8$, and each retained X_A lies in the quotient ball

$$\|X_A\|_{\nu,\alpha,p} \le C_p\sqrt{L}\,E_\nu^{1/p}. \tag{A.8}$$

Moreover, the retained sketched functions form a $3\delta/4$-packing by the triangle inequality.

For a matrix X, let $\widehat{h}_X$ denote the function obtained from (A.4) by replacing X_A with X. The Hölder estimate $|\psi_p(a)-\psi_p(b)| \le C_p|a-b|^{p-1}$, followed by Cauchy–Schwarz, and $\|v\|_{2p-2}^{p-1} \le L^{(2-p)/2}\|v\|_2^{p-1}$ imply, row by row and then in $L_2(\nu)$,

$$\begin{aligned}\|\widehat{h}_X - \widehat{h}_{X'}\|_{L_2(\nu)} &\le C_pB\kappa L^{(1-p)/2}\left(\sum_i \nu_i\|\alpha_i(X-X')y_i\|^{2p-2}\right)^{1/2}\\ &\le C_pB\kappa L^{(1-p)/2}\|X-X'\|_{\nu,\alpha,p}^{p-1}, \qquad \kappa = \left(1+\frac{\log(8M+8)}{L}\right)^{1/2}.\end{aligned}$$

The last step again uses monotonicity of moments under the probability law ν. Consequently, two quotient points within distance

$$\eta = c_{p,B}\sqrt{L}\,(\delta/(B\kappa))^{1/(p-1)}$$

produce sketched functions within $\delta/8$, after reducing $c_{p,B}$ if necessary. If η exceeds the radius in (A.8), one net point suffices. Otherwise, take a maximal η-separated subset of the ball. The radius-$\eta/2$ balls around its points are disjoint and lie in the ball with radius enlarged by $\eta/2$. Comparing their volumes in the at most Lr dimensional space $\mathcal{V}$ gives at most

$$\left(1+C_{p,B}E_\nu^{1/p}(B\kappa/\delta)^{1/(p-1)}\right)^{Lr} \tag{A.9}$$

mesh points. No two retained $3\delta/4$-separated sketches can be assigned to the same mesh point. Hence $|\mathcal{P}|$ is at most twice (A.9). Substituting (A.7) and absorbing the logarithm of the parenthesis into $\mathrm{polylog}(M+r+2, 1+\delta^{-1}, 1+E_\nu)$ proves (A.2). Since the argument applies to every finite packing, a maximal packing gives a cover after changing δ by an absolute constant.

The asserted radius bound (A.3) follows from (A.6). If $E_\nu = 0$, every $\alpha_i Ay_i$ vanishes on the support of ν, so the class is the zero function. If $B = 0$ the same conclusion is immediate.

At $p = 1$, positive α_i disappear inside J_1, while zero α_i give a fixed zero coordinate. Apply the packing argument of Lemma A.1 to the vectors $\alpha_i y_i$ (or, equivalently, delete the zero coordinates). Its use of the same Gaussian sketch for every function, together with the zero-inclusive sign count, gives $\widetilde{O}_B(r\delta^{-2})$, including the strata $Ay_i = 0$, and $|h_A(i)| \le B$ gives the asserted radius. This also covers $E_\nu = 0$ at the endpoint. □

Corollary A.3 (Conditional duality-map complexity)**.** *For an empirical law $\nu = m^{-1}\sum_{s=1}^m \delta_{I_s}$, the class in Lemma A.2 satisfies, conditionally on the sampled indices,*

$$\mathfrak{R}_m^{\mathrm{abs}}(\{h_A\} \mid I_1,\dots,I_m) \le \widetilde{O}_{p,B}\left(\left(E_\nu^{1-1/p} + E_\nu^{(1-1/p)/2}\right)\sqrt{\frac{r}{m}}\right)$$

for $p > 1$, and $\widetilde{O}_B(\sqrt{r/m})$ for $p = 1$.

Proof. Suppose first that $p > 1$. If $E_\nu = 0$, the class is zero by Lemma A.2. Otherwise apply the truncated Dudley bound (2.9) to the class with zero and negatives adjoined, using (A.2) and (A.3). Adjoining these functions increases each covering number by at most a constant factor. Truncate

the entropy integral at the $L_2(\nu)$ radius divided by m and write $a = 1 - 1/p$. Up to the stated logarithmic factors, (A.2) gives

$$\sqrt{\log N(u, \{h_A\}, L_2(\nu))} \le \widetilde{O}_{p,B}\left(\frac{\sqrt{r}\, E_\nu^a}{u}\right),$$

while (A.3) bounds the integration range by $O_{p,B}(E_\nu^a)$. Substitution into the Dudley integral gives the $E_\nu^a\sqrt{r/m}$ term. For $E_\nu \le 1$, the additional lower-scale logarithm is absorbed using $x^a(1+\log(1/x))^c \le C_{a,c}x^{a/2}$; for $E_\nu > 1$, it is already among the logarithms in $1 + E_\nu$. At $p = 1$, the same truncated Dudley bound applied to the endpoint entropy and radius in Lemma A.2 gives the stated estimate directly. □

A.2 Entropy of row-augmented subspace distances

The next estimate is used for the component of a residual vector orthogonal to both the query subspace and its associated low-rank row. It bounds the metric entropy of this class in terms of the query dimension k, with no dependence on the ambient dimension.

Related uses of random projections for affine-space distances, projective clustering, and subspace approximation appear in [Mag07, KR15, CW25]. Here we need a different guarantee: an $L_2(q)$ packing bound for the row-dependent augmented spaces $F + \operatorname{span}\{b_i\}$.

Theorem A.4 (Row-augmented subspace entropy)**.** *Let $\mathcal{H}$ be a finite-dimensional real Hilbert space and let q be a finitely supported probability law. Let $u_i \in \mathcal{H}$ be unit vectors and let $b_i \in \mathcal{H}$ be arbitrary fixed vectors, including zero. For every k-dimensional subspace $F \subseteq \mathcal{H}$, define*

$$f_F(i) = \operatorname{dist}(u_i, F + \operatorname{span}\{b_i\})^p, \qquad 1 \le p \le 2.$$

For $0 < \delta \le 1$ and $k \ge 1$,

$$\log N(\delta, \{f_F\}, L_2(q)) \le C_p k \delta^{-2} \operatorname{polylog}(k + 2, 1/\delta), \tag{A.10}$$

independently of the number of rows and the ambient dimension.

Proof. Fix a finite δ-packing $\mathcal{P}$. We first compress it to $k + O_p(\delta^{-2})$ dimensions and retain a constant fraction whose pairwise distances survive the compression. We then bound the pseudodimension of the compressed class using Grassmann charts.

Let

$$s = \lceil C_p \delta^{-2} \rceil, \qquad L = k + 1 + s,$$

and take a Gaussian map $R : \mathcal{H} \to \mathbb{R}^L$ represented in orthonormal bases by a matrix with independent $N(0, 1/L)$ entries. Put $G_i(F) = F + \operatorname{span}\{b_i\}$, $m_i(F) = \dim G_i(F) \in \{k, k+1\}$, and $c_m = (L/(L-m))^{p/2}$. Define

$$h_{F,R}(i) = c_{m_i(F)} \operatorname{dist}(Ru_i, RG_i(F))^p.$$

The normalization c_m removes the deterministic shrinkage caused by the dimension of the orthogonal complement. Without it, the projected distance contains a factor $(1 - m/L)^{p/2}$, whose squared deviation from one need not be $O(1/s)$ when $k \gg s$.

To identify the distribution of the projected distance, decompose u_i into its projection onto $G_i(F)$ and an orthogonal residual $v_{i,F}$. Put $\rho_{i,F} = \|v_{i,F}\| = \operatorname{dist}(u_i, G_i(F))$ and $\ell_{i,F} = L - m_i(F)$.

Conditional on $R|_{G_i(F)}$, the vector $Rv_{i,F}$ is independent of $RG_i(F)$ and has covariance $\rho_{i,F}^2 I_L/L$. Its projection onto $(RG_i(F))^\perp$ therefore has $\ell_{i,F}$ Gaussian degrees of freedom, and

$$h_{F,R}(i) = \rho_{i,F}^p \left(\frac{X_{\ell_{i,F}}}{\ell_{i,F}} \right)^{p/2}, \qquad X_{\ell_{i,F}} \sim \chi^2_{\ell_{i,F}}, \quad \ell_{i,F} \geq s. \tag{A.11}$$

almost surely. Since $\rho_{i,F} \leq 1$, $\mathbb{E}(X_\ell/\ell - 1)^2 = 2/\ell$, and $|x^{p/2} - 1| \leq |x - 1|$ for $0 < p/2 \leq 1$, we obtain

$$\mathbb{E}_R \|h_{F,R} - f_F\|^2_{L_2(q)} \leq 2/s.$$

Increase the constant in s until the right-hand side is at most $\delta^2/2^{13}$. For each $F \in \mathcal{P}$, Markov gives

$$\mathbb{P}\{\|h_{F,R} - f_F\|_{L_2(q)} > \delta/8\} \leq 1/128.$$

Thus the expected fraction of elements of $\mathcal{P}$ that fail this bound is at most $1/128$, and another application of Markov's inequality makes that fraction at most $1/8$ except on an event of probability at most $1/16$.

Since u_i is a unit vector and Gaussian moments are bounded for fixed $p \leq 2$, $\mathbb{E}_R \sum_i q_i \|Ru_i\|^{2p} \leq C_p$. Markov therefore shows that the event

$$\sum_i q_i \|Ru_i\|^{2p} \leq C_p. \tag{A.12}$$

has probability at least $7/8$, after enlarging C_p. A Gaussian map is also injective on each of the finitely many relevant spaces $G_i(F)$ with probability one, since each has dimension at most $k+1 \leq L$. We may therefore fix a realization R for which the packing-retention and envelope events hold and all relevant restrictions are injective. In particular, $RG_i(F)$ has dimension $k+1$ whenever $b_i \notin F$. Discard the packing elements that fail the approximation bound. The retained set $\mathcal{P}_R$ has size at least $7|\mathcal{P}|/8$, and its sketched functions are a $3\delta/4$-packing by the triangle inequality.

Fix this R and put $x_i = Ru_i$ and $a_i = Rb_i$. Injectivity on $G_i(F)$ implies $m_i(RF) = m_i(F)$, so the retained sketches belong to the class obtained by allowing every $E \in \mathrm{Gr}(k, L)$. Define

$$h_E(i) = c_{m_i(E)} \operatorname{dist}(x_i, E + \operatorname{span}\{a_i\})^p, \quad m_i(E) = \begin{cases} k, & a_i \in E, \\ k+1, & a_i \notin E. \end{cases}$$

With $c_* = c_{k+1}$ and $R_i^{\mathrm{env}} = c_* \|x_i\|^p$, we have $0 \leq h_E(i) \leq R_i^{\mathrm{env}}$ and $\mathcal{R}^2 := \sum_i q_i (R_i^{\mathrm{env}})^2 \leq C_p c_*^2$ by (A.12). Set $\eta = \delta/16$. If $\mathcal{R} \leq \eta$, one ball centered at zero covers the class, so the retained packing has at most one element. Otherwise, discard the zero-envelope coordinates and set

$$g_E(i) = h_E(i)/R_i^{\mathrm{env}}, \qquad \widehat{q}_i = q_i (R_i^{\mathrm{env}})^2/\mathcal{R}^2.$$

Then $0 \leq g_E \leq 1$ and

$$\|h_E - h_{E'}\|_{L_2(q)} = \mathcal{R} \|g_E - g_{E'}\|_{L_2(\widehat{q})}. \tag{A.13}$$

We next bound the pseudodimension of the compressed class, treating the degenerate branch $a_i \in E$ separately. Write $\ell = L - k = s + 1$. On a standard Grassmann chart parameterize $E^\perp$ by the columns of

$$Z_B = \begin{pmatrix} B \\ I_\ell \end{pmatrix}, \qquad B \in \mathbb{R}^{k \times \ell}, \quad Q_B = Z_B (Z_B^\top Z_B)^{-1} Z_B^\top.$$

Here Q_B is the orthogonal projector onto $E^\perp$. There are $\binom{L}{\ell}$ such charts and $d_{\mathrm{ch}} = k\ell$ parameters per chart. Let $\Delta = \det(I_\ell + B^\top B) > 0$. For fixed $x = x_i$, $a = a_i$, write

$$x^\top Q_B x = N_X/\Delta, \quad a^\top Q_B a = N_A/\Delta, \quad x^\top Q_B a = N_C/\Delta,$$

where the numerators have degree $O(\ell)$ and $N_A \geq 0$. If $N_A > 0$,

$$\operatorname{dist}(x, E + \operatorname{span}\{a\})^2 = \frac{N_X N_A - N_C^2}{\Delta N_A}; \tag{A.14}$$

if $N_A = 0$, equivalently $a \in E$, the squared distance is N_X/Δ. For a pseudodimension threshold τ_i at sample point i, the inequality $g_E(i) > \tau_i$ is identically true when $\tau_i < 0$ and identically false when $\tau_i \geq 1$. For $0 \leq \tau_i < 1$, set, separately for $m = k, k+1$,

$$\theta_{i,m} = \left(\frac{\tau_i R_i^{\mathrm{env}}}{c_m}\right)^{2/p}.$$

After raising the threshold inequality to the power $2/p$, $g_E(i) > \tau_i$ is equivalent to

$$(N_A = 0 \text{ and } N_X - \theta_{i,k}\Delta > 0) \quad \text{or} \quad (N_A > 0 \text{ and } N_X N_A - N_C^2 - \theta_{i,k+1}\Delta N_A > 0).$$

Thus the zero-denominator branch is included as a separate semialgebraic stratum. The thresholds may vary with i, as required in the definition of pseudodimension. If $v \leq d_{\mathrm{ch}}$, the desired pseudodimension bound is immediate, so assume $v > d_{\mathrm{ch}}$. Across v thresholded rows, the predicates on one chart involve $O(v)$ polynomials of degree $O(\ell)$ in d_{ch} real parameters. The zero-inclusive sign-assignment bound of Goldberg and Jerrum [GJ95, Corollary 2.1] gives at most

$$\left(\frac{C\ell v}{d_{\mathrm{ch}}}\right)^{d_{\mathrm{ch}}} \leq (C\ell v)^{d_{\mathrm{ch}}}$$

labelings on one chart. Summing over the charts gives at most

$$\binom{L}{\ell}(C\ell v)^{d_{\mathrm{ch}}}$$

labelings. If v points were shattered, then

$$v \leq \log_2 \binom{L}{\ell} + d_{\mathrm{ch}} \log_2(C\ell v).$$

Since $\log \binom{L}{\ell} \leq \ell \log(eL/\ell) \leq C d_{\mathrm{ch}} \log(L+2)$ and $d_{\mathrm{ch}} = k\ell \leq L^2$, solving this inequality for v gives

$$\operatorname{Pdim}(\{g_E\}) \leq C d_{\mathrm{ch}} \log(L+2).$$

The pseudodimension entropy bound (2.10) and (A.13), with $c_* = [1 + (k+1)/s]^{p/2}$, give

$$\log N(\eta, \{h_E\}, L_2(q)) \leq Ck(s+1)\log(L+2)\log(C\mathcal{R}/\eta).$$

Indeed $\mathcal{R} \leq C_p c_*$ by (A.12), and

$$\log(C\mathcal{R}/\eta) \leq C_p(\log(k+2) + \log(1/\delta)).$$

Since $s = O_p(\delta^{-2})$, this is the right-hand side of (A.10). Each $L_2(q)$-ball of radius η contains at most one member of the retained $3\delta/4$-packing. Hence the cover bounds $|\mathcal{P}_R|$, and $|\mathcal{P}| \leq (8/7)|\mathcal{P}_R|$ gives the same logarithmic estimate for the original packing, up to an additive constant. The estimate holds for every finite δ-packing, and a maximal packing gives the required δ-cover. □

Corollary A.5 (Rank-at-most row-augmented entropy)**.** *Let $\mathcal{H}$ be a finite-dimensional real Hilbert space, let q be a finitely supported probability law, let $u_i \in \mathcal{H}$ be unit vectors, and let $b_i \in \mathcal{H}$ be fixed vectors. Fix $k \geq 1$ and $B, T_0 > 0$. Suppose that $0 \leq \beta_i^p \leq B$, and associate with every subspace $F \subseteq \mathcal{H}$ of dimension at most k a scalar $t_F \in [0, T_0]$. Then the class*

$$i \longmapsto t_F^p \beta_i^p \operatorname{dist}(u_i, F + \operatorname{span}\{b_i\})^p \qquad (F \subseteq \mathcal{H},\ \dim F \leq k)$$

has

$$\log N(\delta, \mathcal{G}, L_2(q)) \leq \widetilde{O}_{p,B,T_0}(k\delta^{-2}) \qquad (0 < \delta \leq 1),$$

where $\mathcal{G}$ denotes the displayed class.

Proof. For each $1 \leq j \leq k$, apply Theorem A.4 to the stratum $\dim F = j$. Indeed, for $a, a' \in [0, T_0^p]$ and two such subspaces F, F', the triangle inequality gives

$$\|a\beta^p f_F - a'\beta^p f_{F'}\|_{L_2(q)} \leq BT_0^p \|f_F - f_{F'}\|_{L_2(q)} + B|a - a'|.$$

Here multiplication by β^p is row-wise. Thus a suitably rescaled cover from Theorem A.4, together with a grid of mesh $c_{B,T_0}\delta$ for $a = t_F^p$, covers this stratum. The $j = 0$ stratum contains only $F = \{0\}$. Taking the union of the $k + 1$ covers adds only $\log(k + 1)$ to the entropy. □

A.3 A scalar composition lemma

The following deterministic lemma shows that the parallel power is a Lipschitz function of the first three scalar coordinates introduced in Section 3.1.3. Its bound includes zero amplitudes, both signs of the parallel residual, and exact cancellation.

Lemma A.6 (Lipschitz scalar decoder)**.** *Fix $1 \leq p \leq 2$. For $a, b \geq 0$ and $\sigma \in \{-1, 1\}$ define*

$$T = a^p, \qquad V = b^p, \qquad H = p\sigma a^{p-1} b, \qquad U = |a + \sigma b|^p.$$

At $p = 1$ set $H = 0$ when $a = 0$. For another $a', b' \geq 0$ and $\sigma' \in \{-1, 1\}$, define T', V', H', U' analogously. Then

$$|U - U'| \leq C_p(|T - T'| + |V - V'| + |H - H'|), \tag{A.15}$$

where C_p depends only on p. This includes sign changes, exact cancellation, and zero amplitudes.

Proof. Assume first $1 < p \leq 2$. Put $S = T + V$. When $S > 0$, let

$$t = T/S, \qquad \alpha(t) = t^{1/p}, \qquad \beta(t) = (1 - t)^{1/p},$$

$$h(t) = p\, t^{(p-1)/p}(1 - t)^{1/p}, \qquad u_\sigma(t) = |\alpha(t) + \sigma\beta(t)|^p.$$

Then $(T, V, H, U) = S(t, 1 - t, \sigma h(t), u_\sigma(t))$. We claim

$$|u_\sigma(t) - u_\sigma(s)| \leq C_p(|t - s| + |h(t) - h(s)|). \tag{A.16}$$

Write $\rho = (p - 1)/p$. In $(0, 1)$,

$$u'_+(t) = (\alpha + \beta)^{p-1}[t^{-\rho} - (1 - t)^{-\rho}],$$

$$|u'_-(t)| = |\alpha - \beta|^{p-1}[t^{-\rho} + (1 - t)^{-\rho}], \qquad \frac{h'(t)}{h(t)} = \frac{\rho - t}{t(1 - t)}.$$

For fixed p, these formulas give

$$|u'_\sigma(t)| \le C_p(1 + |h'(t)|)$$

on $(0, \rho/2]$ and $[3/4, 1)$: near zero, $t^{-\rho}$ is controlled by $|h'(t)|$, and near one the same is true of $(1-t)^{-\rho}$. Both u_σ and h are C_p-Lipschitz on the middle interval $[\rho/2, 3/4]$, while h is monotone on each endpoint interval. Thus, integrating on any one of the three intervals gives (A.16). If $[s, t]$ crosses one boundary, apply these bounds on the two pieces and use the Lipschitz bound for h on the middle piece. If it crosses both boundaries, then $|t - s| \ge 3/4 - \rho/2 \ge 1/2$, and the claim follows directly from the boundedness of u_σ. This proves (A.16) on $[0, 1]$, including the cancellation point of u_-, by continuity.

For two tuples of the same sign, say $S \ge S' > 0$,

$$|Su_\sigma(t) - S'u_\sigma(t')| \le C_p|S - S'| + C_pS'(|t - t'| + |h(t) - h(t')|).$$

Moreover,

$$|S - S'| \le |T - T'| + |V - V'|, \qquad S'|t - t'| \le |T - T'| + |S - S'|,$$
$$S'|h(t) - h(t')| \le |H - H'| + 2|S - S'|.$$

This proves (A.15) for equal signs, including $S' = 0$ by a direct limit.

Now suppose that $\sigma' = -\sigma$. Keep the sign σ while changing the amplitudes, and define

$$\widetilde{U}' = |a' + \sigma b'|^p, \qquad \widetilde{H}' = p\sigma(a')^{p-1}b' = -H'.$$

The equal-sign case gives

$$|U - \widetilde{U}'| \le C_p(|T - T'| + |V - V'| + |H - \widetilde{H}'|).$$

Since

$$|H - \widetilde{H}'| = p|a^{p-1}b - (a')^{p-1}b'| \le |H - H'|,$$

it remains only to flip the sign of the second tuple. For $c, d \ge 0$,

$$(c + d)^p - |c - d|^p \le C_p c^{p-1}d.$$

Indeed, this follows from the mean-value theorem when $d \le c$. When $d \ge c$, the mean-value theorem gives the bound $C_p d^{p-1}c$, and $d^{p-1}c \le c^{p-1}d$ because $p \le 2$. Hence in all cases the sign-flip cost satisfies

$$|\widetilde{U}' - U'| \le C_p(a')^{p-1}b' \le C_p|H - H'|.$$

The triangle inequality proves (A.15) for opposite signs.

At $p = 1$, if $a, a' > 0$, the reverse triangle inequality gives

$$||a + \sigma b| - |a' + \sigma' b'|| \le |a - a'| + |\sigma b - \sigma' b'| = |T - T'| + |H - H'|.$$

If $a = 0$, then $H = 0$ by convention and $U = b = V$, independently of σ. Thus, when $a = 0 < a'$,

$$|U - U'| \le |b - b'| + |b' - |a' + \sigma' b'|| \le |V - V'| + |T - T'|.$$

If $a = a' = 0$, then $|U - U'| = |V - V'|$. The remaining asymmetric case follows by interchanging the two tuples. □

A.4 Lewis-weight control of the low-rank term

This section bounds the empirical complexity of normalized low-rank costs. We first use the arbitrary-sampling formulation of Cohen and Peng [CP15, Theorem 7.1] and their moment reduction [CP15, Lemma 7.4] to extract a scalar first-moment estimate from the Lewis-weight concentration bounds of Talagrand [Tal90, Tal95]. We then derive a scalar Rademacher bound and lift it to Hilbert-valued linear maps by a Gaussian-mixture representation. The resulting estimate has no dependence on the dimension of the output space.

Corollary A.7 (First-moment Cohen–Peng bound)**.** *Let $1 \le p < 2$ and $r \ge 1$, and suppose $(w_i, y_i)_{i\in[n]}$ with $w_i \ge 0$ satisfy $\sum_i w_i y_i y_i^\top = I_r$ with unit y_i on the positive support. Let q be a probability distribution with $q_i > 0$ whenever $w_i > 0$. Put*

$$\mathcal{L}(x) = \sum_i w_i |\langle y_i, x\rangle|^p, \qquad \ell_x(i) = \frac{w_i}{q_i} |\langle y_i, x\rangle|^p, \qquad K = \max_{w_i>0} \frac{w_i}{q_i},$$

with $\ell_x(i) = 0$ when $w_i = q_i = 0$. For i.i.d. indices $I_1, \dots, I_j \sim q$, put

$$P_j f = \frac{1}{j}\sum_{s=1}^{j} f(I_s), \qquad \Delta_j = \mathbb{E}_I \sup_{\mathcal{L}(x)=1} |P_j \ell_x - 1|.$$

There are constants C_p, c_p such that, for $0 < \eta \le 1/2$ and $N_0 = n + r + j + K + \eta^{-1} + 2$,

$$\frac{j}{K} \ge C_p \eta^{-2} \log^{c_p} N_0 \quad \Longrightarrow \quad \Delta_j \le C_p \eta. \tag{A.17}$$

Proof. Let $\mathcal{A}$ have rows $w_i^{1/p} y_i^\top$. After deleting its zero rows, the Gram matrix in the defining Lewis-weight equation is

$$\sum_{i:w_i>0} w_i^{2/p} w_i^{1-2/p} y_i y_i^\top = \sum_i w_i y_i y_i^\top = I_r,$$

and the corresponding quadratic form for row i is $w_i^{2/p}\|y_i\|_2^2 = w_i^{2/p}$. Thus the ℓ_p Lewis weights of $\mathcal{A}$ are w_i. Moreover, $\mathcal{L}(x) = \|\mathcal{A}x\|_p^p$.

In the sampling notation of Cohen and Peng [CP15, Theorem 7.1], set $\pi_i = j q_i$. Since $\sum_i \pi_i = j$, their procedure draws j rows independently, chooses row i with probability q_i, and scales it by $\pi_i^{-1/p}$. Hence the sampled pth-power norm is

$$\|S\mathcal{A}x\|_p^p = \frac{1}{j}\sum_{s=1}^{j} \frac{w_{I_s}}{q_{I_s}} |\langle y_{I_s}, x\rangle|^p = P_j \ell_x, \qquad \min_{w_i>0} \frac{\pi_i}{w_i} = \frac{j}{K}.$$

Apply the moment reduction [CP15, Lemma 7.4] with $\ell = 1$ and a fixed constant value of its parameter δ. The required first-moment Rademacher bounds are supplied by [CP15, Lemma 7.2] at $p = 1$ and [CP15, Lemma 7.3] for $1 < p < 2$. These bounds require a target Lewis-weight scale $\bar{u} = c_p \eta^2 / \log^{c_p} N_0$.

The statement of [CP15, Lemma 7.4] is presented under the temporary restriction $\bar{u} = \Omega(1/r)$. Its proof uses the auxiliary matrix from [CP15, Lemma B.1], which has $O(r^2)$ rows with Lewis weights $O(1/r)$. When $\bar{u} < c/r$, split each auxiliary row into $t = \lceil C/(r\bar{u})\rceil$ collinear copies, each scaled by $t^{-1/p}$. This operation preserves its contribution to the pth-power norm and divides its

Lewis weight by t, so every auxiliary copy has weight at most $\bar{u}$. It also leaves the Lewis Gram matrix unchanged: if the original row is a with Lewis weight w, then the t copies contribute

$$t\left(\frac{w}{t}\right)^{1-2/p}(t^{-1/p}a)(t^{-1/p}a)^\top = w^{1-2/p}aa^\top.$$

Appending the sampled rows cannot increase these weights [CP15, Lemma 5.5].

Each occurrence of sampled row i carries comparison weight w_i/π_i. The hypothesis in (A.17) gives

$$\frac{w_i}{\pi_i} \le \frac{K}{j} \le \bar{u}$$

after adjusting the constants and logarithmic power. This verifies the required sampling condition in [CP15, Lemma 7.4]. The resulting augmented matrix has

$$j + O(r^2 t) = O\left(j + r^2 + \frac{r}{\bar{u}}\right) \le N_0^{C_p}.$$

rows, so its logarithm remains $O_p(\log N_0)$. The auxiliary rows are used only in the comparison argument and do not change the sampling law q. Consequently, [CP15, Lemmas 7.2–7.4] give $\Delta_j \le C_p\eta$ under the condition in (A.17). □

Lemma A.8 (Scalar Lewis-weight empirical bound)**.** *Let $1 \le p < 2$ and $r \ge 1$, let $(w_i, y_i)_{i\in[n]}$ with $w_i \ge 0$ satisfy $\sum_i w_i y_i y_i^\top = I_r$ with unit y_i on the positive support, and let q be a probability distribution with $q_i > 0$ whenever $w_i > 0$. Put*

$$\mathcal{L}(x) = \sum_i w_i |\langle y_i, x\rangle|^p, \qquad \ell_x(i) = \frac{w_i}{q_i}|\langle y_i, x\rangle|^p \quad (\mathcal{L}(x) = 1).$$

Again $\ell_x(i) = 0$ when $w_i = q_i = 0$. If $K = \max_{w_i>0} w_i/q_i$, then

$$\mathbb{E}_{I,\varepsilon} \sup_{\mathcal{L}(x)=1} \left|\frac{1}{m}\sum_{s=1}^m \varepsilon_s \ell_x(I_s)\right| \le C_p\Lambda_p(n + r + m + K + 2)\left(\sqrt{\frac{K}{m}} + \frac{K}{m}\right), \tag{A.18}$$

where I_s are i.i.d. with law q and $\Lambda_p(N) = \log^{c_p} N$ for a constant c_p depending only on p. The absolute value is inside the supremum, and the sampling is with replacement.

Proof. Let $\lambda_p(N) = \log^{d_p} N$, where d_p is large enough to absorb the logarithmic powers in Corollary A.7, and put

$$\Delta_j = \mathbb{E}_I \sup_{\mathcal{L}(x)=1} |P_j\ell_x - 1|.$$

We first record a pointwise envelope. For $x \ne 0$, the bound $|\langle y_i, x\rangle| \le \|x\|$ and the inequality $p - 2 \le 0$ give

$$\mathcal{L}(x) = \sum_i w_i |\langle y_i, x\rangle|^p \ge \|x\|^{p-2}\sum_i w_i |\langle y_i, x\rangle|^2 = \|x\|^p.$$

Since $\|y_i\| = 1$ on the positive support, $0 \le \ell_x(i) \le K$. Also,

$$r = \sum_{i:w_i>0} q_i\frac{w_i}{q_i} \le K\sum_{i:w_i>0} q_i \le K,$$

so $K \ge r \ge 1$. The first-moment estimate in Corollary A.7 implies that

$$\Delta_j \le C_p\lambda_p(n + r + j + K + 2)\sqrt{K/j} \tag{A.19}$$

whenever $j \geq C_p K \lambda_p(n+r+j+K+2)^2$. Indeed, set $\eta = C_p \lambda_p(n+r+j+K+2)\sqrt{K/j}$. A sufficiently large constant in the lower bound on j ensures that $\eta \leq 1/2$, and the term η^{-1} in (A.17) changes only the logarithmic factor.

We now pass from the unsigned empirical error to the signed process. Let $M_+ = |\{s : \varepsilon_s = 1\}|$ and $M_- = m - M_+$. Conditional on the signs, the two groups consist of independent samples from q. Write $P^+_{M_+}$ and $P^-_{M_-}$ for their empirical averages. When both groups are nonempty,

$$\frac{1}{m}\sum_{s=1}^{m} \varepsilon_s \ell_x(I_s) = \frac{M_+}{m}(P^+_{M_+}\ell_x - 1) - \frac{M_-}{m}(P^-_{M_-}\ell_x - 1) + \frac{M_+ - M_-}{m}.$$

Put

$$L = \lambda_p(n+r+m+K+2), \qquad j_0 = \left\lceil C_p K L^2 \right\rceil,$$

with a sufficiently large constant C_p. If $m \geq 4j_0$, then on the event $\{M_+, M_- \geq m/4\}$, (A.19) applies to both groups. The complementary event has probability at most $2e^{-m/16}$, the supremum is at most K, and its contribution is absorbed because $Ke^{-m/16} \leq C_p\sqrt{K/m}$ in this range. Moreover, $\mathbb{E}|M_+ - M_-|/m \leq m^{-1/2} \leq \sqrt{K/m}$. Taking expectations in the preceding decomposition therefore gives

$$\mathbb{E}_{I,\varepsilon} \sup_{\mathcal{L}(x)=1} \left| \frac{1}{m}\sum_{s=1}^{m} \varepsilon_s \ell_x(I_s) \right| \leq C_p L \sqrt{\frac{K}{m}}.$$

It remains to treat $m < 4j_0$. Write

$$a_j = \mathbb{E} \sup_{\mathcal{L}(x)=1} \left| \sum_{s=1}^{j} \varepsilon_s \ell_x(I_s) \right|.$$

The map $f \mapsto \sup_{\mathcal{L}(x)=1} |f(x)|$ is convex, and the next signed summand has conditional mean zero. Conditional Jensen therefore gives $a_j \leq a_{j+1}$. Set $m_0 = 4j_0$. Since m_0 is at most a fixed polylogarithmic multiple of $n+r+m+K+2$,

$$\lambda_p(n+r+m_0+K+2) \leq C_p L.$$

Choosing the constant in j_0 sufficiently large makes m_0 satisfy the large-sample condition with its own logarithmic factor. The bound just proved then gives

$$a_{m_0} \leq C_p \lambda_p(n+r+m_0+K+2)\sqrt{K m_0} \leq C_p K L^2.$$

By monotonicity,

$$\frac{a_m}{m} \leq \frac{a_{m_0}}{m} \leq C_p L^2 \frac{K}{m}.$$

Choose the exponent c_p in the statement so that $\Lambda_p(N) = \log^{c_p} N$ dominates $\lambda_p(N)^2$. The large- and small-sample bounds then give (A.18). □

Theorem A.9 (Lewis-weight bound for low-rank terms)**.** *Let $1 \leq p < 2$ and $r \geq 1$. For $i \in [n]$, let $w_i \geq 0$ and $y_i \in \mathbb{R}^r$ satisfy*

$$\sum_{i=1}^{n} w_i y_i y_i^\top = I_r, \qquad \|y_i\|_2 = 1 \quad \text{whenever } w_i > 0,$$

and set $y_i = 0$ whenever $w_i = 0$. Let q be any probability distribution with $q_i > 0$ whenever $w_i > 0$, and put $K = \max_{w_i>0} w_i/q_i$. For a linear map $T : \mathbb{R}^r \to \mathcal{H}$ with $S_T = \sum_j w_j \|Ty_j\|^p > 0$, define

$$h_T(i) = \begin{cases} \dfrac{w_i \|Ty_i\|^p}{q_i S_T}, & w_i > 0, \\ 0, & w_i = 0. \end{cases}$$

Then

$$\mathfrak{R}_m^{\mathrm{abs}}(\{h_T\}) \le C_p \,\mathrm{polylog}(n + r + m + K + 2) \left(\sqrt{\frac{K}{m}} + \frac{K}{m} \right). \tag{A.20}$$

Zero-cost maps are assigned the zero function.

Proof. We use a Gaussian scalarization of Euclidean norms. A closely related reduction from matrix-valued $(p, 2)$-norm sampling to scalar ℓ_p sampling appears in [WY25, Lemma 3.2]. Here the same idea represents each normalized Hilbert-valued function h_T as a mixture of scalar functions to which Lemma A.8 applies. Put

$$\mathcal{L}(x) = \sum_i w_i \, |\langle y_i, x\rangle|^p .$$

For $\mathcal{L}(x) = 1$ put

$$\ell_x(i) = \begin{cases} \dfrac{w_i}{q_i} \, |\langle y_i, x\rangle|^p , & w_i > 0, \\ 0, & w_i = 0. \end{cases}$$

Let γ_T be standard Gaussian probability measure on $\mathrm{span}\{Ty_i\}$, let $g \sim \gamma_T$, and set $x_g = T^* g$. By (2.12), $\mathbb{E}_g \mathcal{L}(x_g) = \gamma_p S_T$. Define the probability measure

$$d\mu_T(g) = \frac{\mathcal{L}(x_g)}{\gamma_p S_T} \, d\gamma_T(g), \qquad \widehat{x}_g = \frac{x_g}{\mathcal{L}(x_g)^{1/p}} \quad \text{when } \mathcal{L}(x_g) > 0.$$

Define $\widehat{x}_g$ arbitrarily on the remaining set, which has zero μ_T-measure. For $w_i > 0$,

$$\begin{aligned} \int \ell_{\widehat{x}_g}(i) \, d\mu_T(g) &= \frac{w_i}{q_i \gamma_p S_T} \mathbb{E}_g |\, \langle Ty_i, g\rangle \,|^p \\ &= \frac{w_i \|Ty_i\|^p}{q_i S_T} = h_T(i). \end{aligned} \tag{A.21}$$

When $w_i = 0$, both sides of (A.21) are zero. Equation (A.21) places every h_T in the closed convex hull of $\{\ell_x : \mathcal{L}(x) = 1\}$. For each fixed sample and sign vector, the map $f \mapsto |m^{-1} \sum_s \varepsilon_s f(I_s)|$ is convex, so its supremum over this convex hull is no larger than its supremum over the scalar class. Therefore,

$$\sup_T \left| \frac{1}{m} \sum_s \varepsilon_s h_T(I_s) \right| \le \sup_{\mathcal{L}(x)=1} \left| \frac{1}{m} \sum_s \varepsilon_s \ell_x(I_s) \right|,$$

and Lemma A.8 proves (A.20). □

Remark A.10 (Necessity of the K/m term)**.** The linear term in (A.20) cannot in general be omitted. Take $n = r$, $y_i = e_i$, $w_i = 1$, and $q_i = 1/r$ for $i \in [r]$. Then $K = r$ and the class contains $r\mathbf{1}_{\{i=j\}}$ for every $j \in [r]$, so its one-sample absolute complexity is r.