Coresets for Multiple ℓ_p Regression

David P. Woodruff [1] **Taisuke Yasuda** [1]

Abstract

A *coreset* of a dataset with n examples and d features is a weighted subset of examples that is sufficient for solving downstream data analytic tasks. Nearly optimal constructions of coresets for least squares and ℓ_p linear regression with a single response are known in prior work. However, for multiple ℓ_p regression where there can be m responses, there are no known constructions with size sublinear in m. In this work, we construct coresets of size $\tilde{O}(\varepsilon^{-2}d)$ for $p < 2$ and $\tilde{O}(\varepsilon^{-p}d^{p/2})$ for $p > 2$ independently of m (i.e., dimension-free) that approximate the multiple ℓ_p regression objective at every point in the domain up to $(1 \pm \varepsilon)$ relative error. If we only need to preserve the minimizer subject to a subspace constraint, we improve these bounds by an ε factor for all $p > 1$. All of our bounds are nearly tight.

We give two application of our results. First, we settle the number of uniform samples needed to approximate ℓ_p Euclidean power means up to a $(1 + \varepsilon)$ factor, showing that $\tilde{\Theta}(\varepsilon^{-2})$ samples for $p = 1$, $\tilde{\Theta}(\varepsilon^{-1})$ samples for $1 < p < 2$, and $\tilde{\Theta}(\varepsilon^{1-p})$ samples for $p > 2$ is tight, answering a question of Cohen-Addad, Saulpic, and Schwiegelshohn. Second, we show that for $1 < p < 2$, every matrix has a subset of $\tilde{O}(\varepsilon^{-1}k)$ rows which spans a $(1+\varepsilon)$-approximately optimal k-dimensional subspace for ℓ_p subspace approximation, which is also nearly optimal.

1. Introduction

Least squares linear regression and ℓ_p linear regression are some of the most fundamental and practically valuable computational problems in statistics and optimization. In this problem, our input is an $n \times d$ matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and a response vector $\mathbf{b} \in \mathbb{R}^n$, and our goal is to output an approximate minimizer $\hat{\mathbf{x}} \in \mathbb{R}^d$ such that

$$\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_p^p \leq (1+\varepsilon)\min_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p. \tag{1}$$

Among the vast literature on ℓ_p regression, sampling algorithms and coresets, which are algorithms that select a weighted subset of the rows of $\mathbf{A}$ and $\mathbf{b}$ that suffice to solve ℓ_p regression, have played major roles in the development of efficient algorithms. That is, we seek a diagonal matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ with few non-zero entries, i.e., $\mathsf{nnz}(\mathbf{S}) \ll n$, such that the weighted subset of rows $\mathbf{SA}$ and $\mathbf{Sb}$ are sufficient to compute a solution $\hat{\mathbf{x}}$ satisfying (1). We will often refer to $\mathsf{nnz}(\mathbf{S})$ as the *sample complexity*. We focus on approaches that construct $\mathbf{S}$ by i.i.d. sampling of each of the n rows:

Definition 1.1 (ℓ_p sampling matrix)**.** Let $p \geq 1$. A random diagonal matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ is a *random ℓ_p sampling matrix with sampling probabilities* $\{q_i\}_{i=1}^n$ if for each $i \in [n]$, the ith diagonal entry is independently set to be

$$\mathbf{S}_{i,i} = \begin{cases} 1/q_i^{1/p} & \text{with probability } q_i \\ 0 & \text{otherwise} \end{cases}$$

Two well-studied guarantees for $\mathbf{S}$ are *strong coresets* and *weak coresets*. Strong coresets refer to coresets that preserve the value of the objective function at *every* point in the domain, while weak coresets only guarantee that the unconstrained minimizer is preserved. If we only care about solving the unconstrained ℓ_p regression problem, then weak coresets are sufficient to solve this problem, and it is known that weak coresets can be substantially smaller than strong coresets in certain settings (Musco et al., 2022). On the other hand, strong coresets are necessary when the objective function must be evaluated at points away from the optimum, for example for constrained optimization problems.

Definition 1.2 (Strong coreset)**.** We say that $\mathbf{S}$ is a *strong coreset* if $\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p$ simultaneously for every $\mathbf{x} \in \mathbb{R}^d$.

Definition 1.3 (Weak coreset)**.** We say that $\mathbf{S}$ is a *weak coreset* if

$$\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_p^p \leq (1+\varepsilon)\min_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p$$

[*] Equal contribution [1] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. Correspondence to: David P. Woodruff <dwoodruf@cs.cmu.edu>, Taisuke Yasuda <taisukey@cs.cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

for $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{b})\|_p^p$.

The efficient construction of coresets for ℓ_p regression has been studied in a long line of work (Clarkson, 2005; Drineas et al., 2006a;b; Dasgupta et al., 2009) culminating in the ℓ_p Lewis weight sampling algorithm (Lewis, 1978; Bourgain et al., 1989; Talagrand, 1990; Ledoux & Talagrand, 1991; Talagrand, 1995; Schechtman & Zvavitch, 2001; Cohen & Peng, 2015; Woodruff & Yasuda, 2023a), which gives an algorithm that constructs a strong coreset $\mathbf{S}$ with

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} \tilde{O}(\varepsilon^{-2}d) & p \leq 2 \\ \tilde{O}(\varepsilon^{-2}d^{p/2}) & p > 2 \end{cases}.$$

A related line of work in the *active ℓ_p regression* setting shows that weak coresets for ℓ_p regression with

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} \tilde{O}(\varepsilon^{-2}d) & p = 1 \\ \tilde{O}(\varepsilon^{-1}d) & 1 < p < 2 \\ O(\varepsilon^{-1}d) & p = 2 \\ \tilde{O}(\varepsilon^{-(p-1)}d^{p/2}) & p > 2 \end{cases}$$

can be constructed even without knowing $\mathbf{b}$ (Chen & Price, 2019; Chen & Derezinski, 2021; Parulekar et al., 2021; Musco et al., 2022; Woodruff & Yasuda, 2023b). Note that these bounds strictly improve over the strong coreset guarantees of ℓ_p Lewis weight sampling for $1 < p < 3$.

1.1. Multiple ℓ_p regression

It is often the case that we are interested in more than just one target $\mathbf{b}$ to predict, and in general, we may wish to simultaneously fit m target vectors that are given by a matrix $\mathbf{B} \in \mathbb{R}^{n\times m}$ and solve the minimization problem

$$\min_{\mathbf{X}\in\mathbb{R}^{d\times m}} \|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p = \min_{\mathbf{X}\in\mathbb{R}^{d\times m}} \sum_{j=1}^m \|\mathbf{A}\mathbf{X}\mathbf{e}_j - \mathbf{B}\mathbf{e}_j\|_p^p$$

This is known as the *multiple response ℓ_p regression* problem, or simply the *multiple ℓ_p regression* problem, and is the focus of the present work.

1.1.1. Coreset constructions for $p = 2$

For $p = 2$, the construction of strong coresets for the multiple response problem follows almost immediately from strong coresets for the single response problem due to orthogonality and the Pythagorean theorem, and we can construct $\mathbf{S}$ such that

$$\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_F^2 = (1\pm\varepsilon)\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_F^2$$

with $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(\varepsilon^{-2}d)$ samples. Indeed, assume without loss of generality that $\mathbf{A}$ has orthogonal columns, and suppose that $\mathbf{S}$ satisfies

- $\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 = (1\pm\varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2$ for every $\mathbf{x}\in\mathbb{R}^d$ (i.e., $\mathbf{S}$ is a subspace embedding)
- $\|\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\|_F^2 = (1\pm\varepsilon)\|\mathbf{A}\mathbf{X}^*-\mathbf{B}\|_F^2$ where $\mathbf{X}^*$ is the optimal minimizer
- $\|\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\|_F^2 \leq (\varepsilon^2/d)\|\mathbf{A}\|_F^2\|\mathbf{A}\mathbf{X}^*-\mathbf{B}\|_F^2 = \varepsilon^2\|\mathbf{A}\mathbf{X}^*-\mathbf{B}\|_F^2$

Then, the following argument of Section 7.5 of (Clarkson & Woodruff, 2013) shows that $\mathbf{S}$ is a strong coreset. Indeed,

$$\begin{aligned}\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_F^2 &= \|\mathbf{S}\mathbf{A}(\mathbf{X}-\mathbf{X}^*)\|_F^2 + \|\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\|_F^2 \\ &\quad + 2\operatorname{tr}\big((\mathbf{X}-\mathbf{X}^*)^\top\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\big)\end{aligned}$$

by expanding the square, and the inner product term is bounded by

$$\begin{aligned}&\left|\operatorname{tr}\big((\mathbf{X}-\mathbf{X}^*)^\top\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\big)\right| \\ &\leq \|\mathbf{X}-\mathbf{X}^*\|_F\|\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\|_F \\ &\leq \varepsilon\|\mathbf{A}(\mathbf{X}-\mathbf{X}^*)\|_F\|\mathbf{A}\mathbf{X}^*-\mathbf{B}\|_F \\ &\leq \varepsilon\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_F^2\end{aligned}$$

and $\mathbf{S}$ also preserves the quantities $\|\mathbf{S}\mathbf{A}(\mathbf{X}-\mathbf{X}^*)\|_F^2$ and $\|\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\|_F^2$ up to $(1\pm\varepsilon)$ relative error. A similar trick is available in the weak coreset setting (see, e.g., Section 3.1 of Cohen et al. (2016)), which gives a bound of $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(\varepsilon^{-1}d)$ for this guarantee. Unfortunately, almost every step in the above argument uses special properties of the ℓ_2 norm that are not available for the ℓ_p norm, and thus we will need completely different arguments to handle $p \neq 2$.

1.1.2. Challenges for $p \neq 2$

If we desire only weak coresets, then prior results on active ℓ_p regression in fact almost immediately provide a solution. These results show that a weak coreset $\mathbf{S}$ for the single response ℓ_p regression problem can be constructed independently of $\mathbf{b}$, and with the dependence of $\mathsf{nnz}(\mathbf{S})$ on the failure probability δ being polylogarithmic. Thus by setting the failure rate to $\delta = 1/10m$, we can simultaneously solve every column of $\mathbf{B}$ independently with overall probability at least $9/10$.

For strong coresets, however, such a column-wise strategy must be implemented carefully. If we consider constructing a strong coreset for a single column $j\in[m]$, then the sampling probabilities now depend on the target vector $\mathbf{B}\mathbf{e}_j$, so the sampling complexity would need to scale as m rather than $\operatorname{poly}\log(m)$ as in the previous upper bound weak coresets. On the other hand, another natural strategy is to mimic the strategy for the $p = 2$ case and take the sampling probabilities to only guarantee an ℓ_p subspace embedding for the column space of $\mathbf{A}$ and that $q_i \geq \|\mathbf{e}_i^\top\mathbf{B}^*\|_p^p/\|\mathbf{B}^*\|_{p,p}^p$ for

$\mathbf{B}^* := \mathbf{A}\mathbf{X}^* - \mathbf{B}$. This is a reasonable choice of sampling probabilities, and indeed it is not hard to see that

$$\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_{p,p}^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p$$

for any fixed $\mathbf{X}\in\mathbb{R}^{d\times m}$ with only $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(\varepsilon^{-2}d)$ samples for $p<2$ and $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(\varepsilon^{-2}d^{p/2})$ samples for $p>2$ via a Bernstein tail bound. However, it is unclear how to extend a guarantee for any single $\mathbf{X}\in\mathbb{R}^{d\times m}$ to a guarantee simultaneously for *all* $\mathbf{X}\in\mathbb{R}^{d\times m}$. Although the dependence on the failure rate δ is logarithmic, a net argument, or even more sophisticated chaining arguments, over the possible choices of $\mathbf{X}\in\mathbb{R}^{d\times m}$ seem to require a union bound over sets of size $\exp(dm)$, thus again introducing a linear dependence on m in the sample complexity $\mathsf{nnz}(\mathbf{S})$. As we show, a careful blend of these two ideas will be necessary to obtain our strong coreset result.

1.2. Strong coresets for multiple ℓ_p regression

Our first main result is the first construction of strong coresets for multiple ℓ_p regression that is independent of m.

Theorem 1.4 (Strong coresets for multiple ℓ_p regression)**.** *Let $\mathbf{A}\in\mathbb{R}^{n\times d}$, $\mathbf{B}\in\mathbb{R}^{n\times m}$, and $p\geq 1$. There is an algorithm which constructs $\mathbf{S}$ with*

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} \dfrac{O(d)}{\varepsilon^2}\left[(\log d)^2\log\dfrac{d}{\varepsilon}+\log\dfrac{1}{\delta}\right] & 1\leq p<2 \\ \dfrac{O(d^{p/2})}{\varepsilon^p}\left[(\log d)^2\log\dfrac{d}{\varepsilon}+\log\dfrac{1}{\delta}\right] & p>2 \end{cases}$$

such that with probability at least $1-\delta$,

$$\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_{p,p}^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p$$

simultaneously for every $\mathbf{X}\in\mathbb{R}^{d\times m}$. Furthermore, $\mathbf{S}$ can be constructed in $\tilde{O}(\mathsf{nnz}(\mathbf{A})+\mathsf{nnz}(\mathbf{B})+\mathrm{poly}(d))$ time.

We achieve a nearly optimal dependence on d and ε, as we show that $\Omega(d^{p/2}/\varepsilon^p)$ rows are necessary for strong coresets in Theorem 5.1 for $p>2$, while it is known that $\tilde{\Omega}(d/\varepsilon^2)$ rows are necessary even for $m=1$ for $p<2$ (Li et al., 2021a). We note that our upper bound shows that multiple ℓ_p regression is as easy as single response ℓ_p regression for $p<2$, while our lower bound demonstrates an interesting separation between the two for $p>2$.

1.2.1. Initial $\log m$ bound

Our main technique is to generalize the "partition by sensitivity" technique introduced in the active ℓ_p regression work of Musco et al. (2022) and show how this can be applied to the strong coreset setting. We describe the idea for the case of $p<2$, as the case of $p>2$ is analogous.

In the active ℓ_p regression setting, we must show that we can design sampling algorithms that preserve the objective function value, even if we do not know the target vector $\mathbf{b}$. In this setting, one of the main observations of Musco et al. (2022) is that even though we cannot preserve $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p$ itself, we can actually preserve the difference $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p - \|\mathbf{b}\|_p^p$, if $\|\mathbf{b}\|_p^p = O(\mathsf{OPT}^p)$ which is without loss of generality. To see this idea, assume (without loss of generality due to Dasgupta et al. (2009)) that we restrict our attention to $\|\mathbf{A}\mathbf{x}\|_p^p = O(\mathsf{OPT}^p)$. Then, the analysis of Musco et al. (2022) proceeds by partitioning the coordinates of $\mathbf{b}$ into two sets, those such that $|\mathbf{b}(i)|^p$ is larger than $\varepsilon^{-p}\mathbf{w}_i\mathsf{OPT}^p$ and those that are smaller than this threshold, where $\mathbf{w}_i$ is the i-th ℓ_p Lewis weight of $\mathbf{A}$. It is known that $\mathbf{w}_i$ bounds the sensitivities of $\mathbf{A}$, that is, $|[\mathbf{A}\mathbf{x}](i)|^p\leq \mathbf{w}_i\|\mathbf{A}\mathbf{x}\|_p^p$ so it follows that for any $|\mathbf{b}(i)|^p\geq\varepsilon^{-p}\mathbf{w}_i\mathsf{OPT}^p$, we have that

$$\left||[\mathbf{A}\mathbf{x}-\mathbf{b}](i)|^p - |\mathbf{b}(i)|^p\right| = O(\varepsilon)|\mathbf{b}(i)|^p$$

for any $\mathbf{x}\in\mathbb{R}^d$ with $\|\mathbf{A}\mathbf{x}\|_p^p = O(\mathsf{OPT}^p)$. On the other hand, if $|\mathbf{b}(i)|^p\leq\varepsilon^{-p}\mathbf{w}_i\mathsf{OPT}^p$, then we have by the triangle inequality that

$$\left||[\mathbf{A}\mathbf{x}-\mathbf{b}](i)|^p - |\mathbf{b}(i)|^p\right| \leq O(\varepsilon^{-p})\mathbf{w}_i\mathsf{OPT}^p.$$

Thus, up to an additive $O(\varepsilon)(\|\mathbf{S}\mathbf{b}\|_p^p + \|\mathbf{b}\|_p^p)$ error, $\left||[\mathbf{A}\mathbf{x}-\mathbf{b}](i)|^p - |\mathbf{b}(i)|^p\right|$ has sensitivities which are controlled by the ℓ_p Lewis weights of $\mathbf{A}$. This allows one to show that sampling by the ℓ_p Lewis weights of $\mathbf{A}$ preserves $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p - \|\mathbf{b}\|_p^p$ for all $\|\mathbf{A}\mathbf{x}\|_p^p = O(\mathsf{OPT}^p)$.

In order to apply this idea to the strong coreset setting, we generalize the above argument to multiple scales. That is, we replace OPT^p by an arbitrary scale $R\geq\|\mathbf{b}\|_p^p$, and show that for every $\|\mathbf{A}\mathbf{x}\|_p^p\leq O(R)$ that

$$\begin{aligned}&\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{b})\|_p^p - \|\mathbf{S}\mathbf{b}\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p - \|\mathbf{b}\|_p^p\right)\right| \\ &\leq\ \varepsilon(R+\|\mathbf{S}\mathbf{b}\|_p^p)\end{aligned}$$

Finally, we can generalize this to the following guarantee by union bounding over finitely many scales R, which holds for every $\mathbf{x}\in\mathbb{R}^d$:

$$\begin{aligned}&\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{b})\|_p^p - \|\mathbf{S}\mathbf{b}\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p - \|\mathbf{b}\|_p^p\right)\right| \\ &\leq\ \varepsilon\left(\|\mathbf{b}\|_p^p + \|\mathbf{S}\mathbf{b}\|_p^p + \|\mathbf{A}\mathbf{x}\|_p^p\right)\end{aligned} \tag{2}$$

This guarantee is in a form that can be summed over the m columns of $\mathbf{B}$. Thus, if a $\log m$ dependence is admissible, then we can apply the above result with failure probability $1/10m$, union bound over the m columns, and sum the results to obtain

$$\begin{aligned}&\left|(\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_{p,p}^p - \|\mathbf{S}\mathbf{B}\|_{p,p}^p)\right. \\ &\left.-(\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p - \|\mathbf{B}\|_{p,p}^p)\right| \\ &\leq\ \varepsilon\left(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{S}\mathbf{B}\|_{p,p}^p + \|\mathbf{A}\mathbf{X}\|_{p,p}^p\right).\end{aligned}$$

Now suppose that we additionally have

- $\|\mathbf{SB}\|_{p,p}^p = (1 \pm \varepsilon)\|\mathbf{B}\|_{p,p}^p$
- $\|\mathbf{B}\|_{p,p}^p = O(\mathsf{OPT}^p)$ (which is without loss of generality by subtracting an $O(1)$-optimal solution)

Then, we have

$$\begin{aligned}
&\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\|_{p,p}^p \\
&= \|\mathbf{AX}-\mathbf{B}\|_{p,p}^p - \|\mathbf{B}\|_{p,p}^p + \|\mathbf{SB}\|_{p,p}^p \\
&\qquad \pm O(\varepsilon)\big(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{AX}\|_{p,p}^p\big) \\
&= \|\mathbf{AX}-\mathbf{B}\|_{p,p}^p \pm \varepsilon\|\mathbf{B}\|_{p,p}^p \pm O(\varepsilon)\big(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{AX}\|_{p,p}^p\big) \\
&= \|\mathbf{AX}-\mathbf{B}\|_{p,p}^p \pm O(\varepsilon)\|\mathbf{AX}-\mathbf{B}\|_{p,p}^p
\end{aligned}$$

so we indeed have a strong coreset as desired.

1.2.2. Removing the m dependence

Next, we show how to completely remove the m dependence, which requires additional ideas. When applying (2) to each of the m columns, suppose that we set the failure probability to $\operatorname{poly}(\varepsilon\delta)$ instead of $O(1/m)$. Then, this guarantee will hold for a $1 - \operatorname{poly}(\varepsilon\delta)$ fraction of "good" columns, for which we can obtain $(1 \pm \varepsilon)$ approximations. On the remaining $\operatorname{poly}(\varepsilon\delta)$ fraction of "bad" columns, note that the mass of $\mathbf{B}$ on these columns is at most $\operatorname{poly}(\varepsilon\delta)\|\mathbf{B}\|_{p,p}^p$ with probability $1-\delta$ by Markov's inequality. Then on these columns, $\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\mathbf{e}_j\|_p$ is just $\|\mathbf{SAXe}_j\|_p$ up to a small total additive error of $\operatorname{poly}(\varepsilon\delta)\|\mathbf{B}\|_{p,p}^p$. In turn, we have that $\|\mathbf{SAXe}_j\|_p = (1 \pm \varepsilon)\|\mathbf{AXe}_j\|_p$ by using that $\mathbf{S}$ is an ℓ_p subspace embedding. Thus, by combining with the $(1 \pm \varepsilon)$ approximation on the rest of the "good" columns, we can still ensure that $\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\|_{p,p} = (1 \pm \varepsilon)\|\mathbf{AX}-\mathbf{B}\|_{p,p}$.

1.3. Weak coresets for multiple ℓ_p regression

In the weak coreset setting, we consider a generalized multiple ℓ_p regression problem, where we are given a design matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, an "embedding" $\mathbf{G} \in \mathbb{R}^{t\times m}$, and a target matrix $\mathbf{B} \in \mathbb{R}^{n\times m}$, and we wish to approximately minimize the objective function $\|\mathbf{AXG}-\mathbf{B}\|_{p,p}$.

As noted previously, for multiple ℓ_p regression without an embedding (i.e., $\mathbf{G} = \mathbf{I}_t$) the construction of weak coresets follows relatively straightforwardly by applying active ℓ_p regression results along each column. However, this strategy fails when we must additionally handle the embedding matrix $\mathbf{G}$, as this constraint couples the columns of $\mathbf{AX}$ together. Furthermore, we argue that handling the embedding $\mathbf{G}$ is substantially more interesting that the unconstrained case. Indeed, as we see later in Sections 1.4 and 1.5, the incorporation of the embedding $\mathbf{G}$ will allow us to handle interesting extensions of our results to settings beyond the entrywise ℓ_p norm via the use of a linear embedding into this norm. We will denote the optimal value as

$$\mathsf{OPT} := \min_{\mathbf{X}\in\mathbb{R}^{d\times t}} \|\mathbf{AXG}-\mathbf{B}\|_{p,p}$$

and let $\mathbf{X}^*$ denote the matrix achieving this optimum unless otherwise noted. We will prove the following result:

Theorem 1.5 (Weak coresets for multiple ℓ_p regression)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{G} \in \mathbb{R}^{t\times m}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, and $1 \le p < \infty$. There is an algorithm which constructs $\mathbf{S}$ independently of $\mathbf{B}$ with*

$$\mathsf{nnz}(\mathbf{S}) = \frac{O(d)}{\varepsilon^2\delta^2}\left[(\log d)^2\log\frac{d}{\varepsilon} + \log\frac{1}{\delta}\right]\left(\log\log\frac{1}{\varepsilon}\right)^2$$

for $p = 1$,

$$\mathsf{nnz}(\mathbf{S}) = \frac{O(d)}{\varepsilon\delta^2}\left[(\log d)^2\log\frac{d}{\varepsilon} + \log\frac{1}{\delta}\right]\left(\log\log\frac{1}{\varepsilon}\right)^2$$

for $1 < p < 2$, and

$$\mathsf{nnz}(\mathbf{S}) = \frac{O(d^{p/2})}{\varepsilon^{p-1}\delta^p}\left[(\log d)^2\log\frac{d}{\varepsilon} + \log\frac{1}{\delta}\right]\left(\log\log\frac{1}{\varepsilon}\right)^p$$

for $p > 2$ such that with probability at least $1-\delta$, for any $\hat{\mathbf{X}} \in \mathbb{R}^{d\times t}$ such that

$$\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B})\|_{p,p}^p \le (1+\varepsilon)\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{S}(\mathbf{AXG}-\mathbf{B})\|_{p,p}^p,$$

we have

$$\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B}\|_{p,p}^p \le (1+O(\varepsilon))\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{AXG}-\mathbf{B}\|_{p,p}^p.$$

Conditioned on the event that $\|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p = O(\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p)$ for the global optimizer $\mathbf{X}^*$, the dependence on δ can be replaced by a single $\log\frac{1}{\delta}$ factor and the $\operatorname{poly}(\log\log\frac{1}{\varepsilon})$ factor can be removed. Furthermore, $\mathbf{S}$ can be constructed in $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^\omega)$ time.

We achieve a nearly optimal dependence on d and ε, as we show that $\Omega(d^{p/2}/\varepsilon^{p-1})$ rows are necessary for weak coresets in Theorem 5.2 for $p > 2$. Our weak coreset upper bound result together with our strong coreset lower bound of Theorem 5.1 shows a tight ε factor separation between the two coreset guarantees.

Note that in the statement of Theorem 1.5, the dependence on the failure rate δ is polynomial. This is in fact necessary if we restrict our algorithm to be of the form of "sample-and-solve" algorithms whose sampling matrices $\mathbf{S}$ do not depend on $\mathbf{B}$, as demonstrated in a lower bound result of Theorem 12.8 of (Musco et al., 2022). The only reason why this dependence becomes necessary in the analysis of the upper bound is that $\|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p$ may be as large as $O\big(\frac{1}{\delta}\big)\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$ with probability at least δ, and this is the source of the hardness result of Theorem 12.8 of (Musco et al., 2022) as well. This is a mild problem and can be easily circumvented in one of two ways. The first is to simply allow the algorithm to incorporate the row norms of $\mathbf{B}$ into the sampling probabilities just as in Theorem 1.4. However, this would not give an active regression algorithm that makes only polylogarithmic in δ many queries. If we wish for such an active regression algorithm, then we can follow (Musco et al., 2022) and consider the following two-stage procedure. First, we can obtain a constant factor solution $\hat{\mathbf{X}}$ with a polylogarithmic dependence on δ by employing a "median"-like procedure (see Section 3.1 of (Musco et al., 2022)). Then, we can run $\log \frac{1}{\delta}$ copies of the algorithm, each of which succeeds with probability $1-\delta$. Then, we can sort the runs by their estimates $\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B})\|_{p,p}^p$ and discard half of the runs with the highest values of $\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B})\|_{p,p}^p$. This guarantees that the remaining runs have $\|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p = O(1)\|\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$ with probability at least $1-\delta$, which is enough for the rest of the argument to go through with only a polylogarithmic dependence on δ.

1.4. Applications: sublinear algorithms for Euclidean power means

Our first application of our results on coresets for multiple ℓ_p regression is on designing coresets for the Euclidean power means problem. In this problem, we are given as input a set of n points $\{\mathbf{b}_i\}_{i=1}^n \subseteq \mathbb{R}^t$, and we wish to find a center $\hat{\mathbf{x}} \in \mathbb{R}^t$ that minimizes the sum of the Euclidean distances to $\hat{\mathbf{x}}$, raised to the power p. That is, we seek to minimize the objective function given by

$$\sum_{i=1}^n \|\mathbf{x}-\mathbf{b}_i\|_2^p = \|\mathbf{1}\mathbf{x}^\top - \mathbf{B}\|_{p,2}^p$$

where $\mathbf{1}$ is the $n \times 1$ matrix of all ones, $\mathbf{B} \in \mathbb{R}^{n\times t}$ is the matrix with $\mathbf{b}_i$ in its n rows, and $\|\cdot\|_{p,2}$ is the $(p,2)$-norm of a matrix given by the ℓ_p norm of the Euclidean norm of the rows. This is a fundamental problem which generalizes the well-studied problems of the mean ($p=2$), geometric median ($p=1$), and minimum enclosing balls ($p=\infty$). Coresets and sampling algorithms for this problem were recently studied by Cohen-Addad et al. (2021), who showed that a uniform sample of $\tilde{O}(\varepsilon^{-(p+3)})$ points suffices to output a center $\hat{\mathbf{x}} \in \mathbb{R}^t$ such that

$$\|\mathbf{1}\hat{\mathbf{x}}^\top - \mathbf{B}\|_{p,2}^p \leq (1+\varepsilon)\min_{\mathbf{x}\in\mathbb{R}^t}\|\mathbf{1}\mathbf{x}^\top - \mathbf{B}\|_{p,2}^p = (1+\varepsilon)\mathsf{OPT}^p.$$

In comparison to the upper bounds, the lower bounds given by Cohen-Addad et al. (2021) was $\Omega(\varepsilon^{-(p-1)})$ which is off by an ε^4 factor compared to the upper bound, which was improved to $\Omega(\varepsilon^{-1})$ for $1 < p < 2$ by (Musco et al., 2022) and $\Omega(\varepsilon^{-2})$ for $p=1$ by (Chen & Derezinski, 2021; Parulekar et al., 2021).

One of the main open questions highlighted by the work of Cohen-Addad et al. (2021) is to obtain tight bounds for this problem: how many uniform samples are necessary and sufficient to output a $(1+\varepsilon)$-approximate solution to the Euclidean power means problem. Our main contribution is a nearly optimal algorithm which matches the lower bounds of Chen & Derezinski (2021); Parulekar et al. (2021); Cohen-Addad et al. (2021); Musco et al. (2022).

Theorem 1.6. *Let $\{\mathbf{b}_i\}_{i=1}^n \subseteq \mathbb{R}^d$. Then, there is a sublinear algorithm which uniformly samples at most*

$$s = \begin{cases} O(\varepsilon^{-2})\big(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\big)\log\frac{1}{\delta} & p=1 \\ O(\varepsilon^{-1})\big(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\big)\log\frac{1}{\delta} & 1<p\leq 2 \\ O(\varepsilon^{1-p})\big(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\big)\log\frac{1}{\delta} & 2<p<\infty \end{cases}$$

rows $\mathbf{b}_i$ and outputs a center $\hat{\mathbf{x}}$ such that

$$\sum_{i=1}^n \|\hat{\mathbf{x}}-\mathbf{b}_i\|_2^p \leq (1+\varepsilon)\min_{\mathbf{x}\in\mathbb{R}^d}\sum_{i=1}^n\|\mathbf{x}-\mathbf{b}_i\|_2^p$$

with probability at least $1-\delta$.

To apply the techniques developed in this work to the Euclidean power means problem, we need to embed the $(p,2)$-norm into the entrywise ℓ_p norm. To make this reduction, we use a classic result of Dvoretzky and Milman (Dvoretzky, 1961; Milman, 1971), which shows that a random subspace of a normed space is approximately Euclidean. We will need the following version of this result for ℓ_p norms:

Theorem 1.7 (Dvoretzky's theorem for ℓ_p norms (Figiel et al., 1977; Paouris et al., 2017))**.** *Let $p \geq 1$ and $0 < \varepsilon < 1/p$. Let $n \geq O(\max\{\varepsilon^{-2}k, \varepsilon^{-1}k^{p/2})$, and let $\mathbf{G} \in \mathbb{R}^{n\times k}$ be an i.i.d. random Gaussian matrix. Then, with probability at least $2/3$, $\|\mathbf{G}\mathbf{x}\|_p^p = (1\pm\varepsilon)n\|\mathbf{x}\|_2^p$ for every $\mathbf{x}\in\mathbb{R}^k$.*

Note then that if $\mathbf{G}$ is an appropriately scaled random Gaussian matrix, then we have that

$$\|\mathbf{1}\mathbf{x}^\top - \mathbf{B}\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{1}\mathbf{x}^\top\mathbf{G} - \mathbf{B}\mathbf{G}\|_{p,p}^p$$

by the above result. We may now note that the latter optimization problem is exactly of the form of an embedded ℓ_p regression problem, and thus our weak coreset results immediately apply to this problem. In fact, handling this Dvoretzky embedding is our main motivation for studying the ℓ_p regression problem with the embedding. We also note that similar reductions are possible by making use of other linear embeddings between ℓ_p norms (Wang & Woodruff, 2019; Li et al., 2021b; 2023). The full argument is given in Appendix D.1.

In addition to sharpening the bound of Cohen-Addad et al. (2021) to optimality, we note that our techniques, both algorithmically and in the analysis, are simpler than the prior work of Cohen-Addad et al. (2021). The previous algorithm required partitioning the dataset into "rings" of points with similar costs and preprocessing these rings. Furthermore, the analysis uses a specially designed chaining argument with custom net constructions that require terminal Johnson–Lindenstrauss embeddings. On the other hand, our algorithm simply runs multiple instances of a "sample-and-solve" algorithm, where the run with lowest sampled mass is kept. Furthermore, the analysis largely builds on existing net constructions for ℓ_p regression, and does not need terminal embeddings. In fact, our proof for the power means problem only need ℓ_p regression net constructions in $d = 1$ dimensions due to our use of Dvoretzky's theorem, which avoids the sophisticated constructions of Bourgain et al. (1989) for large d and only needs a standard volume argument (Remark B.15). Our partition of sensitivity can also be thought of as a coarse notion of rings, where we only consider two classes of costs, "big" and "small", whereas prior work requires finer a classification of points into rings of points whose costs are related up to a constant factor.

1.5. Applications: spanning coresets for ℓ_p subspace approximation

As a second application of our results, we give the first construction of *spanning coresets for ℓ_p subspace approximation* with nearly optimal size. The ℓ_p subspace approximation is a popular generalization of the classic Frobenius norm low rank approximation problem, where the input is a set of n points $\{\mathbf{a}_i\}_{i=1}^n$ in d dimensions, and we wish to compute a rank k subspace $F \subseteq \mathbb{R}^d$ that minimizes

$$\sum_{i=1}^n \|\mathbf{a}_i^\top (\mathbf{I}_d - \mathbf{P}_F)\|_2^p$$

where $\mathbf{P}_F$ denotes the orthogonal projection matrix onto F. Equivalently, we can write this as

$$\min_{\mathrm{rank}(F)\le k} \|\mathbf{A}(\mathbf{I}_d - \mathbf{P}_F)\|_{p,2}^p.$$

While strong and weak coresets for this problem have attracted much attention (Feldman & Langberg, 2011; Shyamalkumar & Varadarajan, 2012; Sohler & Woodruff, 2018; Huang & Vishnoi, 2020; Feng et al., 2021; Woodruff & Yasuda, 2023b), our main contribution to this line of research is on a different coreset guarantee, which we call *spanning coresets*. Spanning coresets are subsets of the points $\mathbf{a}_i$ which span a $(1+\varepsilon)$-optimal rank k subspace, and is another popular guarantee in this literature (Deshpande & Varadarajan, 2007; Shyamalkumar & Varadarajan, 2012; Clarkson & Woodruff, 2015). In addition to being an interesting object in its own right (Shyamalkumar & Varadarajan, 2012), the existence of small spanning coresets have found applications to constructions for strong and weak coresets for ℓ_p subspace approximation (Huang & Vishnoi, 2020).

Definition 1.8 (Spanning coreset)**.** Let $\{\mathbf{a}_i\}_{i=1}^n \subseteq \mathbb{R}^d$. A subset $S \subseteq [n]$ is a $(1+\varepsilon)$-*spanning coreset* if the points $\{\mathbf{a}_i\}_{i\in S}$ span a k-dimensional subspace $\hat{F}$ such that

$$\|\mathbf{A}(\mathbf{I}_d - \mathbf{P}_{\hat{F}})\|_{p,2}^p \le (1+\varepsilon) \min_{\mathrm{rank}(F)\le k} \|\mathbf{A}(\mathbf{I}_d - \mathbf{P}_F)\|_{p,2}^p.$$

Our main result is the following upper bound on the size of spanning coresets.

Theorem 1.9. *Let $\{\mathbf{a}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, $1 \le p < \infty$, $k \in \mathbb{N}$, and $0 < \varepsilon < 1$. Then, there exists a $(1+\varepsilon)$-spanning coreset S of size at most*

$$|S| = \begin{cases} O(\varepsilon^{-2}k)(\log(k/\varepsilon))^3 & p = 1 \\ O(\varepsilon^{-1}k)(\log(k/\varepsilon))^3 & 1 < p \le 2 \\ O(\varepsilon^{1-p}k^{p/2})(\log(k/\varepsilon))^3 & 2 < p < \infty \end{cases}$$

In particular, we improve the previous best result of $O(\varepsilon^{-1}k^2\log(k/\varepsilon))$ due to Theorem 3.1 of (Shyamalkumar & Varadarajan, 2012) in the k dependence for all $1 \le p < 4$. The proof of this result is given in Section D.2. Furthermore, we give the first lower bounds on the size of spanning coresets by generalizing an argument of (Deshpande & Vempala, 2006) for $p = 2$, showing that spanning coresets must have size at least $\Omega(\varepsilon^{-1}k)$ in Theorem 5.3. Together, our results settle the size of spanning coresets up to polylogarithmic factors for $1 < p < 2$. To obtain this result, we again use Dvoretzky's theorem to embed the problem to an embedded entrywise ℓ_p norm problem, and then apply our weak coreset results.

Finally, we note that our spanning coreset lower bound implies other interesting lower bounds for coresets. First, we note that weak coresets for ℓ_p subspace approximation are automatically spanning coresets, so our lower bound for spanning coresets also gives the first nontrivial lower bound on the size of weak coresets for ℓ_p subspace approximation. Secondly, we note that our proof of Theorem 1.9 in fact shows that any upper bound on weak coresets for ℓ_p regression with an embedding implies upper bounds for spanning coresets of the same size. Thus, our spanning coreset lower bound in fact implies an $\Omega(d/\varepsilon)$ lower bound on the size of weak coresets for ℓ_p regression with an embedding, which establishes that our weak coreset upper bound for ℓ_p regression (Theorem 1.5) is also nearly optimal for $1 < p < 2$ up to polylogarithmic factors.

On the other hand, for $p > 2$, our weak coreset lower bound of Theorem 5.2 shows that our technique of reducing spanning coresets to weak coresets cannot prove a better upper bound than the result of Theorem 1.9, and thus new ideas are required to improve upon the $\tilde{O}(\varepsilon^{-1}k^2)$ spanning coreset upper bound of Theorem 3.1 of (Shyamalkumar & Varadarajan, 2012). This is an interesting open problem.

1.6. Open directions

We conclude with several potential directions for future research. One interesting question is to improve our understanding of upper bounds and lower bounds for coresets for single response ℓ_p regression.

Question 1.10. How many rows are necessary and sufficient for strong and weak coresets for single response ℓ_p regression?

For strong coresets, this questions is already nearly optimally settled for $p < 2$ with $\tilde{\Theta}(\varepsilon^{-2}d)$ rows known to be necessary and sufficient (Li et al., 2021a). For $p > 2$, however, there is still a gap in our understanding, with the best known upper bound being $\tilde{\Theta}(\varepsilon^{-2}d^{p/2})$ via ℓ_p Lewis weight sampling while the best known lower bound is only $\Omega(\varepsilon^{-1}d^{p/2} + \varepsilon^{-2}d)$. It is an interesting question to determine whether the lower bound can be improved to match the ℓ_p Lewis weight sampling upper bound or not.

For weak coresets, the deficiencies are much more glaring. There are currently no known nontrivial lower bounds for weak coresets, while the best known algorithms are the better of the two upper bounds given by active ℓ_p regression and strong coresets, both of which are substantially more restricted settings than weak coresets.

Finally, we highlight the question of obtaining a nearly optimal upper bound on spanning coresets for ℓ_p subspace approximation for $p > 2$.

Question 1.11. How many rows are necessary and sufficient for spanning coresets for ℓ_p subspace approximation?

We conjecture that our lower bound of $\Omega(k/\varepsilon)$ is tight, while the best known upper bound is the better of our Theorem 1.9 and $\tilde{O}(\varepsilon^{-1}k^2)$ (Shyamalkumar & Varadarajan, 2012).

2. Preliminaries

2.1. ℓ_p Lewis weights

Definition 2.1 (One-sided ℓ_p Lewis weights (Jambulapati et al., 2022; Woodruff & Yasuda, 2022))**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \in (0,\infty)$. Let $\gamma \in (0,1]$. Then, weights $\mathbf{w} \in \mathbb{R}^n$ are *γ-one-sided ℓ_p Lewis weights* if $\mathbf{w}_i \geq \gamma \cdot \boldsymbol{\tau}_i(\mathbf{W}^{1/2-1/p}\mathbf{A})$, where $\mathbf{W} := \operatorname{diag}(\mathbf{w})$. If $\gamma = 1$, we just say that $\mathbf{w}$ are *one-sided ℓ_p Lewis weights.*

The following theorem collects the results of (Cohen & Peng, 2015; Jambulapati et al., 2022) on the fastest known algorithms for approximating one-sided ℓ_p Lewis weights:

Theorem 2.2. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p > 0$. There is an algorithm which computes one-sided ℓ_p Lewis weights (Def. 2.1) $\mathbf{w}$ such that $d \leq \|\mathbf{w}\|_1 \leq 2d$ in $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^\omega)$ time.*

3. Strong coresets

Theorem 3.1 (Strong coresets for multiple ℓ_p regression)**.** *Let $\hat{\mathbf{X}} \in \mathbb{R}^{d\times m}$ satisfy*

$$\|\mathbf{A}\hat{\mathbf{X}} - \mathbf{B}\|_{p,p}^p \leq O(1) \min_{\mathbf{X}\in\mathbb{R}^{d\times m}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{p,p}^p$$

and let $\hat{\mathbf{B}} := \mathbf{A}\hat{\mathbf{X}} - \mathbf{B}$. Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha + \mathbf{v}_i/\beta\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$, $\mathbf{v}_i = \|\mathbf{e}_i^\top \hat{\mathbf{B}}\|_p^p / \|\hat{\mathbf{B}}\|_{p,p}^p$,

$$\alpha = \begin{cases} O(\gamma)\varepsilon^2 \left[(\log d)^2 \log n + \log \frac{1}{\delta}\right]^{-1} & p < 2 \\ \dfrac{O(\gamma^{p/2})\varepsilon^p}{\|\mathbf{w}\|_1^{p/2-1}} \left[(\log d)^2 \log n + \log \frac{1}{\delta}\right]^{-1} & p > 2 \end{cases}$$

and $\beta = O(\varepsilon^{-2}\log\frac{1}{\delta})$. Then with probability at least $1-\delta$,

$$\|\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_{p,p}^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{p,p}^p$$

simultaneously for every $\mathbf{X} \in \mathbb{R}^{d\times m}$.

Our main technical lemma is the following result which generalizes the sampling results of (Musco et al., 2022; Woodruff & Yasuda, 2023b) on preserving differences. The proof can be found in Appendix A.

Theorem 3.2. *Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and*

$$\alpha = \begin{cases} \dfrac{O(\gamma)\varepsilon^2}{\eta^{2/p}} \left[(\log d)^2 \log n + \log \frac{1}{\delta}\right]^{-1} & p < 2 \\ \dfrac{O(\gamma^{p/2})\varepsilon^p}{\eta\|\mathbf{w}\|_1^{p/2-1}} \left[(\log d)^2 \log n + \log \frac{1}{\delta}\right]^{-1} & p > 2 \end{cases}.$$

For each $\mathbf{x}^* \in \mathbb{R}^d$ and $\mathbf{b}^* = \mathbf{A}\mathbf{x}^* - \mathbf{b}$, with probability at least $1-\delta$,

$$\begin{aligned} &\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p - \|\mathbf{S}\mathbf{b}^*\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p - \|\mathbf{b}^*\|_p^p\right)\right| \\ &\leq \varepsilon\left(\|\mathbf{b}^*\|_p^p + \|\mathbf{S}\mathbf{b}^*\|_p^p + \frac{1}{\eta}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p\right) \end{aligned}$$

simultaneously for every $\mathbf{x} \in \mathbb{R}^d$.

Given Theorem 3.2, the proof of Theorem 3.1 proceeds as described in the introduction.

Proof of Theorem 3.1. By replacing $\mathbf{B}$ by $\hat{\mathbf{B}} - \mathbf{A}\hat{\mathbf{X}}$, we assume that $\|\mathbf{B}\|_p = O(\mathsf{OPT})$. We apply Theorem 3.2 with failure probability at $\varepsilon^p\delta^2$. Now let $S \subseteq [m]$ be the set of columns for which the guarantee of Theorem 3.2 fails. Note then that by Markov's inequality,

$$\sum_{j\in S} \|\mathbf{B}\mathbf{e}_j\|_p^p = O(\varepsilon^p\delta)\|\mathbf{B}\|_{p,p}^p$$

with probability at least $1-\delta$. We also have that

$$\sum_{j\in S}\|\mathbf{S}\mathbf{B}\mathbf{e}_j\|_p^p \le \frac{1}{\delta}\sum_{j\in S}\|\mathbf{B}\mathbf{e}_j\|_p^p = O(\varepsilon^p)\|\mathbf{B}\|_{p,p}^p$$

with probability at least $1-\delta$, again by Markov's inequality. Then,

$$\begin{aligned}\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\mathbf{e}_j\|_p^p &= (1\pm\varepsilon)\|\mathbf{S}\mathbf{A}\mathbf{X}\mathbf{e}_j\|_p^p \pm \frac{O(1)}{\varepsilon^{p-1}}\|\mathbf{S}\mathbf{B}\mathbf{e}_j\|_p^p \\ &= (1\pm\varepsilon)^2\|\mathbf{A}\mathbf{X}\mathbf{e}_j\|_p^p \pm \frac{O(1)}{\varepsilon^{p-1}}\|\mathbf{S}\mathbf{B}\mathbf{e}_j\|_p^p\end{aligned}$$

by using that $\mathbf{S}$ is a subspace embedding. Similarly, we have that

$$\|(\mathbf{A}\mathbf{X}-\mathbf{B})\mathbf{e}_j\|_p^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{X}\mathbf{e}_j\|_p^p \pm \frac{O(1)}{\varepsilon^{p-1}}\|\mathbf{B}\mathbf{e}_j\|_p^p.$$

Then summing over $j \in S$ gives that

$$\sum_{j\in S}\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\mathbf{e}_j\|_p^p = \sum_{j\in S}\|(\mathbf{A}\mathbf{X}-\mathbf{B})\mathbf{e}_j\|_p^p \pm O(\varepsilon)\|\mathbf{B}\|_{p,p}^p.$$

On the other hand, for $j \notin S$, Theorem 3.2 succeeds so we have

$$\begin{aligned}\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\mathbf{e}_j\|_p^p &= \|(\mathbf{A}\mathbf{X}-\mathbf{B})\mathbf{e}_j\|_p^p \\ &\quad - \|\mathbf{B}\mathbf{e}_j\|_p^p + \|\mathbf{S}\mathbf{B}\mathbf{e}_j\|_p^p \\ &\quad \pm \varepsilon\big(\|\mathbf{B}\mathbf{e}_j\|_p^p + \|\mathbf{S}\mathbf{B}\mathbf{e}_j\|_p^p + \|\mathbf{A}\mathbf{X}\mathbf{e}_j\|_p^p\big)\end{aligned}$$

Summing the guarantee over the m columns j gives

$$\begin{aligned}\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_{p,p}^p &= \|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p - \|\mathbf{B}\|_{p,p}^p + \|\mathbf{S}\mathbf{B}\|_{p,p}^p \\ &\quad \pm O(\varepsilon)\big(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{A}\mathbf{X}\|_{p,p}^p\big) \\ &= \|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p \pm \varepsilon\|\mathbf{B}\|_{p,p}^p \\ &\quad \pm O(\varepsilon)\big(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{A}\mathbf{X}\|_{p,p}^p\big) \\ &= \|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p \pm O(\varepsilon)\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p \square\end{aligned}$$

4. Weak coresets

We sketch the proof of the following result in this section. Full proofs can be found in Appendix C.

Theorem 4.1 (Weak coresets for multiple ℓ_p regression)**.** *Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 1.1) with sampling probabilities $q_i \ge \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w}\in\mathbb{R}^n$ and*

$$\alpha = O(\gamma)\varepsilon\delta^2\left[(\log d)^2\log n + \log\frac{1}{\delta}\right]^{-1}\left[\log\log\frac{1}{\varepsilon}\right]^{-2}$$

for $p < 2$ and

$$\alpha = \frac{O(\gamma^{p/2})\varepsilon^{p-1}\delta^p}{\|\mathbf{w}\|_1^{p/2-1}}\left[(\log d)^2\log n + \log\frac{1}{\delta}\right]^{-1}\left[\log\log\frac{1}{\varepsilon}\right]^{-p}$$

for $p > 2$. Then, for any $\hat{\mathbf{X}}\in\mathbb{R}^{d\times t}$ such that

$$\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B})\|_{p,p}^p \le (1+\varepsilon)\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{S}(\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B})\|_{p,p}^p,$$

we have

$$\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B}\|_{p,p}^p \le (1+O(\varepsilon))\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}^p.$$

We first establish lemmas that relate approximation quality to the closeness of solutions to the optimum in Section 4.1, and we use this in an iterative argument in Section 4.2.

4.1. Closeness of nearly optimal solutions

The following lemma uses strong convexity for $p < 2$ and a Bregman divergence bound for $p > 2$ to quantify the difference between the ℓ_p norms of two vectors.

Lemma 4.2. *For any $\mathbf{y},\mathbf{y}'\in\mathbb{R}^n$, we have*

$$\|\mathbf{y}'\|_p^2 \ge \|\mathbf{y}\|_p^2 - 2\|\mathbf{y}\|_p^{2-p}\langle\mathbf{y}^{\circ(p-1)}, \mathbf{y}-\mathbf{y}'\rangle + \frac{p-1}{2}\|\mathbf{y}-\mathbf{y}'\|_p^2$$

if $1 < p < 2$ (Lemma 8.1 of (Ben-Tal et al., 2001)) and

$$\|\mathbf{y}'\|_p^p \ge \|\mathbf{y}\|_p^p - p\langle\mathbf{y}^{\circ(p-1)}, \mathbf{y}-\mathbf{y}'\rangle + \frac{p-1}{p2^p}\|\mathbf{y}-\mathbf{y}'\|_p^p$$

if $2 \le p < \infty$ (Lemmas 3.2 and 4.6 of (Adil et al., 2019)).

We need the following elementary computation.

Lemma 4.3 (Gradients of multiple ℓ_p regression)**.** *The gradient $\nabla_{\mathbf{X}}\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}^p$ is given by the formula*

$$\sum_{i=1}^n\sum_{j=1}^m p[\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}](i,j)^{\circ(p-1)}(\mathbf{A}^\top\mathbf{e}_i)(\mathbf{e}_j^\top\mathbf{G}^\top)$$

The following lemma uses Lemmas 4.2 and 4.3 to show that if $\mathbf{X}$ achieves a nearly optimal value, then $\mathbf{X}$ must be close to the optimal solution $\mathbf{X}^*$.

Lemma 4.4 (Closeness of nearly optimal solutions)**.** *Let $p > 1$. For any $\mathbf{X}\in\mathbb{R}^{d\times t}$ such that $\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p} \le (1+\eta)\mathsf{OPT}$ with $\eta\in(0,1)$, we have that*

$$\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p} \le \begin{cases} O(\eta^{1/2})\mathsf{OPT} & p < 2 \\ O(\eta^{1/p})\mathsf{OPT} & p > 2\end{cases}$$

where $\mathbf{X}^* := \arg\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}$.

4.2. Iterative size reduction argument

We now sketch the proof of Theorem 4.1.

We will need the following initial result to seed our iterative argument. Note that the dependence on ε is suboptimal by an ε factor for every $1 < p < \infty$.

Lemma 4.5. *Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and*

$$\alpha = O(\gamma)(\varepsilon\delta)^2 \left[(\log d)^2 \log n + \log \frac{1}{\delta}\right]^{-1}$$

for $1 \leq p < 2$ and

$$\alpha = \frac{O(\gamma^{p/2})(\varepsilon\delta)^p}{\|\mathbf{w}\|_1^{p/2-1}} \left[(\log d)^2 \log n + \log \frac{1}{\delta}\right]^{-1}$$

for $2 < p < \infty$. Then, for any $\hat{\mathbf{X}} \in \mathbb{R}^{d\times t}$ such that

$$\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B})\|_{p,p}^p \leq (1+\varepsilon) \min_{\mathbf{X}\in\mathbb{R}^{d\times t}} \|\mathbf{S}(\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B})\|_{p,p}^p,$$

we have

$$\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B}\|_{p,p}^p \leq (1+O(\varepsilon)) \min_{\mathbf{X}\in\mathbb{R}^{d\times t}} \|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}^p.$$

Starting from this initial solution bound of Lemma 4.5, we can proceed via an iterative argument similar to those of (Musco et al., 2022; Woodruff & Yasuda, 2023b) which alternates between using a bound on the closeness of the solution to the optimal solution to improve the approximation (Theorem 3.2), and using a bound on the approximation to improve the closeness to the optimum (Lemma 4.4). More specifically, we can show that for $1 < p < 2$, a bound of C/ε^β on the sample complexity implies that a bound of $C/\varepsilon^{2\beta/(1+\beta)}$ is sufficient as well. Iterating this argument starting from $\beta = 2$ due to Lemma 4.5 for $O(\log\log\frac{1}{\varepsilon})$ iterations yields the desired bound of C/ε, as claimed. Similarly, for $p > 2$, a bound of C/ε^β implies a bound of $C/\varepsilon^{p\beta/(1+\beta)}$, which results in a final bound of C/ε^{p-1}, as claimed. The full details can be found in Appendix C.

5. Lower bounds

In this section, we complement our various upper bounds with matching lower bounds. In the interest of space, the proofs are given in Appendix E.

Theorem 5.1. *Let $2 < p < \infty$ be fixed. Let $\varepsilon \in (0,1)$ be less than some sufficiently small constant. Then, a strong coreset $\mathbf{S}$ for multiple ℓ_p regression requires $\mathsf{nnz}(\mathbf{S}) = \Omega(\varepsilon^{-p}d^{p/2})$ non-zero rows.*

Theorem 5.2. *Let $2 < p < \infty$ be fixed. Let $\varepsilon \in (0,1)$ be less than some sufficiently small constant. Then, a weak coreset $\mathbf{S}$ for multiple ℓ_p regression requires $\mathsf{nnz}(\mathbf{S}) = \Omega(\varepsilon^{1-p}d^{p/2})$ non-zero rows.*

Theorem 5.3. *Let $1 \leq p < \infty$ and*

$$c_p = \begin{cases} 1/6 & p \leq 2 \\ 1/(6\cdot 5^{p/2-1}) & p > 2 \end{cases}$$

Let $k \in \mathbb{N}$. Then, there is a matrix $\mathbf{B} \in \mathbb{R}^{n\times(n+1)}$ such that for every $\varepsilon \geq k/n$ and any subset of $s \leq (c_p/4)\varepsilon^{-1}k$ rows, any rank k subspace F' spanned by the s rows must have

$$\|\mathbf{B}\mathbf{P}_{F'} - \mathbf{B}\|_{p,2}^p > (1+\varepsilon) \min_{\mathrm{rank}(F)\leq k} \|\mathbf{B}\mathbf{P}_F - \mathbf{B}\|_{p,2}^p.$$

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

We thank the anonymous reviewers for useful feedback on improving the presentation of this work. David P. Woodruff and Taisuke Yasuda were supported by a Simons Investigator Award.

References

Adil, D., Kyng, R., Peng, R., and Sachdeva, S. Iterative refinement for ℓ_p-norm regression. In Chan, T. M. (ed.), *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pp. 1405–1424. SIAM, 2019.

Ben-Tal, A., Margalit, T., and Nemirovski, A. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM J. Optim.*, 12(1):79–108, 2001. ISSN 1052-6234. doi: 10.1137/S1052623499354564.

Bourgain, J., Lindenstrauss, J., and Milman, V. Approximation of zonoids by zonotopes. *Acta Math.*, 162(1-2):73–141, 1989. ISSN 0001-5962. doi: 10.1007/BF02392835.

Chen, X. and Derezinski, M. Query complexity of least absolute deviation regression via robust uniform convergence. In Belkin, M. and Kpotufe, S. (eds.), *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1144–1179. PMLR, 2021.

Chen, X. and Price, E. Active regression via linear-sample sparsification. In Beygelzimer, A. and Hsu, D. (eds.), *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pp. 663–695. PMLR, 2019.

Clarkson, K. L. Subgradient and sampling algorithms for ℓ_1 regression. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, pp. 257–266, USA, 2005. Society for Industrial and Applied Mathematics.

Clarkson, K. L. and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In Boneh, D., Roughgarden, T., and Feigenbaum, J. (eds.), *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pp. 81–90. ACM, 2013. doi: 10.1145/2488608.2488620.

Clarkson, K. L. and Woodruff, D. P. Input sparsity and hardness for robust subspace approximation. In Guruswami, V. (ed.), *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pp. 310–329. IEEE Computer Society, 2015. doi: 10.1109/FOCS.2015.27.

Cohen, M. B. and Peng, R. L_p row sampling by lewis weights. In Servedio, R. A. and Rubinfeld, R. (eds.), *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pp. 183–192. ACM, 2015. doi: 10.1145/2746539.2746567.

Cohen, M. B., Nelson, J., and Woodruff, D. P. Optimal approximate matrix product in terms of stable rank. In Chatzigiannakis, I., Mitzenmacher, M., Rabani, Y., and Sangiorgi, D. (eds.), *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, volume 55 of *LIPIcs*, pp. 11:1–11:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016. doi: 10.4230/LIPICS.ICALP.2016.11.

Cohen-Addad, V., Saulpic, D., and Schwiegelshohn, C. Improved coresets and sublinear algorithms for power means in euclidean spaces. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 21085–21098, Virtual, 2021.

Dasgupta, A., Drineas, P., Harb, B., Kumar, R., and Mahoney, M. W. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009.

Deshpande, A. and Varadarajan, K. R. Sampling-based dimension reduction for subspace approximation. In Johnson, D. S. and Feige, U. (eds.), *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pp. 641–650. ACM, 2007. doi: 10.1145/1250790.1250884.

Deshpande, A. and Vempala, S. S. Adaptive sampling and fast low-rank matrix approximation. In Díaz, J., Jansen, K., Rolim, J. D. P., and Zwick, U. (eds.), *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 9th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, APPROX 2006 and 10th International Workshop on Randomization and Computation, RANDOM 2006, Barcelona, Spain, August 28-30 2006, Proceedings*, volume 4110 of *Lecture Notes in Computer Science*, pp. 292–303. Springer, 2006. doi: 10.1007/11830924_28.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 2006*, pp. 1127–1136. ACM Press, 2006a.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In Azar, Y. and Erlebach, T. (eds.), *Algorithms - ESA 2006, 14th Annual European Symposium, Zurich, Switzerland, September 11-13, 2006, Proceedings*, volume 4168 of *Lecture Notes in Computer Science*, pp. 304–314. Springer, 2006b. doi: 10.1007/11841036_29.

Dvoretzky, A. Some results on convex bodies and Banach spaces. In *Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960)*, pp. 123–160. Jerusalem Academic Press, Jerusalem; Pergamon, Oxford, 1961.

Feldman, D. and Langberg, M. A unified framework for approximating and clustering data. In Fortnow, L. and Vadhan, S. P. (eds.), *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pp. 569–578. ACM, 2011. doi: 10.1145/1993636.1993712.

Feng, Z., Kacham, P., and Woodruff, D. P. Dimensionality reduction for the sum-of-distances metric. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3220–3229. PMLR, 2021.

Figiel, T., Lindenstrauss, J., and Milman, V. D. The dimension of almost spherical sections of convex bodies. *Acta Math.*, 139(1-2):53–94, 1977. ISSN 0001-5962. doi: 10.1007/BF02392234.

Huang, L. and Vishnoi, N. K. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In Makarychev, K., Makarychev, Y., Tulsiani, M., Kamath, G., and Chuzhoy, J. (eds.), *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pp. 1416–1429. ACM, 2020. doi: 10.1145/3357713.3384296.

Jambulapati, A., Liu, Y. P., and Sidford, A. Improved iteration complexities for overconstrained *p*-norm regression.

In Leonardi, S. and Gupta, A. (eds.), *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pp. 529–542. ACM, 2022.

Ledoux, M. and Talagrand, M. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 1991. ISBN 978-3-642-20211-7. Isoperimetry and processes, Reprint of the 1991 edition.

Lewis, D. R. Finite dimensional subspaces of L_p. *Studia Mathematica*, 63(2):207–212, 1978.

Li, Y., Wang, R., and Woodruff, D. P. Tight bounds for the subspace sketch problem with applications. *SIAM J. Comput.*, 50(4):1287–1335, 2021a. doi: 10.1137/20M1311831.

Li, Y., Woodruff, D. P., and Yasuda, T. Exponentially improved dimensionality reduction for ℓ_1: Subspace embeddings and independence testing. In Belkin, M. and Kpotufe, S. (eds.), *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3111–3195. PMLR, 2021b.

Li, Y., Lin, H., and Woodruff, D. P. ℓ_p-regression in the arbitrary partition model of communication. In Neu, G. and Rosasco, L. (eds.), *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4902–4928. PMLR, 2023.

Makarychev, K., Makarychev, Y., and Razenshteyn, I. P. Performance of johnson-lindenstrauss transform for *k*-means and *k*-medians clustering. In Charikar, M. and Cohen, E. (eds.), *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pp. 1027–1038. ACM, 2019. doi: 10.1145/3313276.3316350.

Milman, V. D. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Priložen.*, 5(4):28–37, 1971.

Musco, C., Musco, C., Woodruff, D. P., and Yasuda, T. Active linear regression for ℓ_p norms and beyond. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 - November 3, 2022*, pp. 744–753. IEEE, 2022. doi: 10.1109/FOCS54457.2022.00076.

Paouris, G., Valettas, P., and Zinn, J. Random version of Dvoretzky's theorem in ℓ_p^n. *Stochastic Process. Appl.*, 127(10):3187–3227, 2017. ISSN 0304-4149. doi: 10.1016/j.spa.2017.02.007.

Parampalli, U., Tang, X., and Boztas, S. On the construction of binary sequence families with low correlation and large sizes. *IEEE Trans. Inf. Theory*, 59(2):1082–1089, 2013.

Parulekar, A., Parulekar, A., and Price, E. L1 regression with Lewis weights subsampling. In Wootters, M. and Sanità, L. (eds.), *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2021, August 16-18, 2021, University of Washington, Seattle, Washington, USA (Virtual Conference)*, volume 207 of *LIPIcs*, pp. 49:1–49:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

Schechtman, G. and Zvavitch, A. Embedding subspaces of l_p into l_p^n, $0 < p < 1$. *Mathematische Nachrichten*, 227 (1):133–142, 2001.

Shyamalkumar, N. D. and Varadarajan, K. R. Efficient subspace approximation algorithms. *Discret. Comput. Geom.*, 47(1):44–63, 2012. doi: 10.1007/s00454-011-9384-2.

Sohler, C. and Woodruff, D. P. Strong coresets for k-median and subspace approximation: Goodbye dimension. In Thorup, M. (ed.), *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pp. 802–813. IEEE Computer Society, 2018. doi: 10.1109/FOCS.2018.00081.

Talagrand, M. Embedding subspaces of L_1 into l_1^N. *Proc. Amer. Math. Soc.*, 108(2):363–369, 1990.

Talagrand, M. Embedding subspaces of L_p in l_p^N. In *Geometric aspects of functional analysis (Israel, 1992–1994)*, volume 77 of *Oper. Theory Adv. Appl.*, pp. 311–325. Birkhäuser, Basel, 1995.

Vershynin, R. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.

Wang, R. and Woodruff, D. P. Tight bounds for ℓ_p oblivious subspace embeddings. In Chan, T. M. (ed.), *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pp. 1825–1843. SIAM, 2019. doi: 10.1137/1.9781611975482.110.

Woodruff, D. P. and Yasuda, T. High-dimensional geometric streaming in polynomial space. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 - November 3, 2022*, pp. 732–743. IEEE, 2022.

Woodruff, D. P. and Yasuda, T. Online Lewis weight sampling. In Bansal, N. and Nagarajan, V. (eds.), *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023*, pp. 4622–4666. SIAM, 2023a. doi: 10.1137/1.9781611977554.ch175.

Woodruff, D. P. and Yasuda, T. New subset selection algorithms for low rank approximation: Offline and online. In Saha, B. and Servedio, R. A. (eds.), *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*, pp. 1802–1813. ACM, 2023b. doi: 10.1145/3564246.3585100.

Woodruff, D. P. and Yasuda, T. Sharper bounds for ℓ_p sensitivity sampling. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37238–37272. PMLR, 2023c.

A. ℓ_p Lewis weight sampling for differences

Throughout this section, we fix the following notation:

Definition A.1.

- Let $1 \leq p < \infty$.
- Let $\varepsilon \in (0,1)$ be an accuracy parameter and let $\delta \in (0,1)$ be a failure probability parameter.
- Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$.
- Let $\mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights for $\mathbf{A}$ such that $\max_{i=1}^n \mathbf{w}_i \leq w$.
- Let $\mathbf{x}^* \in \mathbb{R}^d$ any center, let $\eta \in (0,1)$ be a proximity parameter, and let $R \geq \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ be a scale parameter.
- For each $i \in [n]$ and $\mathbf{x} \in \mathbb{R}^d$, let

$$\Delta_i(\mathbf{x}) \coloneqq |[\mathbf{A}\mathbf{x} - \mathbf{b}](i)|^p - |[\mathbf{A}\mathbf{x}^* - \mathbf{b}](i)|^p$$

Our main result of the section is the following:

Theorem A.2. *Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and*

$$\alpha = \begin{cases} O(\gamma)\dfrac{\varepsilon^2}{\eta^{2/p}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p < 2 \\ O(\gamma^{p/2})\dfrac{\varepsilon^p}{\eta\|\mathbf{w}\|_1^{p/2-1}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p > 2 \end{cases}.$$

Then for each $\mathbf{x}^* \in \mathbb{R}^d$ and $R \geq \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$, with probability at least $1-\delta$,

$$\sup_{\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R} \left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p - \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p\right)\right| \leq \varepsilon(R + \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p)$$

We will prove Theorem A.2 throughout this section. Before doing so, we state the following more convenient form of the result:

Theorem 3.2. *Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and*

$$\alpha = \begin{cases} \dfrac{O(\gamma)\varepsilon^2}{\eta^{2/p}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p < 2 \\ \dfrac{O(\gamma^{p/2})\varepsilon^p}{\eta\|\mathbf{w}\|_1^{p/2-1}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p > 2 \end{cases}.$$

For each $\mathbf{x}^* \in \mathbb{R}^d$ and $\mathbf{b}^* = \mathbf{A}\mathbf{x}^* - \mathbf{b}$, with probability at least $1-\delta$,

$$\begin{aligned} &\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p - \|\mathbf{S}\mathbf{b}^*\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p - \|\mathbf{b}^*\|_p^p\right)\right| \\ \leq\ &\varepsilon\left(\|\mathbf{b}^*\|_p^p + \|\mathbf{S}\mathbf{b}^*\|_p^p + \frac{1}{\eta}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p\right) \end{aligned}$$

simultaneously for every $\mathbf{x} \in \mathbb{R}^d$.

Proof. We apply Theorem A.2 with δ set to δ/L for $L = O(\log(1/\delta\varepsilon))$ and R set to $2^l\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ for $l \in [L]$. By a union bound, the conclusion holds simultaneously for every $l \in [L]$ with probability at least $1-\delta$. Furthermore, by Markov's inequality, $\|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p = O(1/\delta)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ with probability at least $1-\delta$.

If $\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \leq 2^L\|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p = \mathrm{poly}(1/\delta\varepsilon)\|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p$, then the result follows immediately from applying the conclusion of Theorem A.2 at the appropriate scale $l \in [L]$. Otherwise, we have that $\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \geq \mathrm{poly}(1/\delta\varepsilon)\|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p$, in which case

$$\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*)\|_p^p \geq \Omega(1)\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \geq \mathrm{poly}(1/\delta\varepsilon)\|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p$$

so

$$\begin{aligned}\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{b})\|_p^p - \|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p &= (1\pm\varepsilon)\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*)\|_p^p \pm \frac{(1+\varepsilon)^{p-1}}{\varepsilon^{p-1}}\|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p \\ &= (1\pm\varepsilon)\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*)\|_p^p \pm \frac{(1+\varepsilon)^{p-1}}{\delta\varepsilon^{p-1}}\|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p \\ &= (1\pm O(\varepsilon))\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*)\|_p^p\end{aligned}$$

and similarly,

$$\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p = (1\pm O(\varepsilon))\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p.$$

Thus it suffices to have that

$$\left|\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*)\|_p^p - \|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p\right| \leq \frac{\varepsilon}{\eta}\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p.$$

In fact, standard ℓ_p Lewis weight sampling guarantees give

$$\left|\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*)\|_p^p - \|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p\right| \leq \begin{cases}\dfrac{\varepsilon}{\eta^{1/p}}\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p & p<2 \\ \dfrac{\varepsilon^{p/2}}{\eta^{1/2}}\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p & p>2\end{cases}$$

which is stronger. □

Throughout our proof of Theorem A.2, we will assume without loss of generality that $\mathbf{S}_{i,i}^p > 1$, that is we only consider rows that are sampled with probability $q_i < 1$, since rows that are kept with probability $q_i = 1$ do not contribute towards the sampling error. Note first that we can write

$$\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{b})\|_p^p - \|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p\right)\right| = \left|\sum_{i=1}^n (\mathbf{S}_{i,i}^p - 1)\Delta_i(\mathbf{x})\right|.$$

The supremum of this quantity, normalized by $(R + \|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p)^l$, over $\left\{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R\right\}$ is a random variable. We will bound the l-th moment of this random variable for $l = O(\log\frac{1}{\delta} + \log n)$.

We start with a standard symmetrization procedure (see, e.g., (Cohen & Peng, 2015; Chen & Derezinski, 2021)).

Lemma A.3 (Symmetrization)**.**

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{S}}\left[\frac{1}{(R+\|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p)^l}\sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p\leq\eta R}\left|\sum_{i=1}^n(\mathbf{S}_{i,i}^p-1)\Delta_i(\mathbf{x})\right|^l\right] \\ &\leq\ 2^l \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n,\mathbf{S}}\left[\frac{1}{(R+\|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p)^l}\sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p\leq\eta R}\left|\sum_{i=1}^n\boldsymbol{\varepsilon}_i\mathbf{S}_{i,i}^p\Delta_i(\mathbf{x})\right|^l\right]\end{aligned}$$

Next, we replace the Rademacher process on the right hand side of Lemma A.3 by one which "removes" $\mathbf{S}_{i,i}^p$, that is, one of the form

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\left[\sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p\leq\eta R}\left|\sum_{i=1}^n\boldsymbol{\varepsilon}_i\Delta_i(\mathbf{x})\right|^l\right]. \tag{3}$$

This is roughly done by noting that if we take $\mathbf{SA}$ to be a "part of" $\mathbf{A}$, then the domain $\left\{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R\right\}$ only dilates by a constant factor as $\mathbf{S}$ preserves ℓ_p norms in the column space of $\mathbf{A}$. More formally, we have the following lemma:

Lemma A.4. *Let $\mathbf{B} \in \mathbb{R}^{m\times d}$ satisfy $\|\mathbf{B}\mathbf{x}\|_p^p \le C\|\mathbf{A}\mathbf{x}\|_p^p$ for every $\mathbf{x} \in \mathbb{R}^d$. For every fixing of $\mathbf{S}$, let*

$$\mathbf{B}_{\mathbf{S}} \coloneqq \begin{pmatrix} \mathbf{S}\mathbf{A} \\ \mathbf{B} \end{pmatrix}$$

be the concatenation of $\mathbf{S}\mathbf{A}$ and $\mathbf{B}$, and let

$$F_{\mathbf{S}} = \sup_{\|\mathbf{A}\mathbf{x}\|_p^p \le 1} \left| \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p - \|\mathbf{A}\mathbf{x}\|_p^p \right|.$$

Suppose that for every fixing of $\mathbf{S}$ and $R' \ge R + \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p$, we have that

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{B}_{\mathbf{S}}\mathbf{x} - \mathbf{B}_{\mathbf{S}}\mathbf{x}^*\|_p^p \le \eta R'} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i \mathbf{S}_{i,i}^p \Delta_i(\mathbf{x}) \right| \le \varepsilon^l \delta R'^l$$

Then,

$$\mathop{\mathbf{E}}_{\mathbf{S}} \frac{1}{(R + \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p)^l} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i \mathbf{S}_{i,i}^p \Delta_i(\mathbf{x}) \right|^l \le (2\varepsilon)^l \delta\Big((1+C)^l + \mathop{\mathbf{E}}_{\mathbf{S}}[F_{\mathbf{S}}^l] \Big)$$

Proof. Note that

$$\|\mathbf{B}_{\mathbf{S}}(\mathbf{x} - \mathbf{x}^*)\|_p^p = \|\mathbf{S}\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_p^p + \|\mathbf{B}(\mathbf{x} - \mathbf{x}^*)\|_p^p \le (1 + F_{\mathbf{S}} + C)\|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_p^p$$

so

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i \mathbf{S}_{i,i}^p \Delta_i(\mathbf{x}) \right|^l &\le \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{B}_{\mathbf{S}}\mathbf{x} - \mathbf{B}_{\mathbf{S}}\mathbf{x}^*\|_p^p \le (1+F_{\mathbf{S}}+C)\eta R} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i \mathbf{S}_{i,i}^p \Delta_i(\mathbf{x}) \right|^l \\
&\le \varepsilon^l \delta (1 + F_{\mathbf{S}} + C)^l (R + \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p)^l \\
&\le \varepsilon^l \delta 2^{l-1} ((1+C)^l + F_{\mathbf{S}}^l)(R + \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p)^l \qquad \text{Fact B.1}
\end{aligned}$$

Taking expectations on both sides proves the lemma. □

Note that if $\mathbf{S}$ is the ℓ_p Lewis weight sampling matrix, then $\mathbf{E}[|F_{\mathbf{S}}|^l]$ in Lemma A.4 is known to be bounded by $O(1)^l$ (that is, $\mathbf{S}$ is an $O(1)$-approximate ℓ_p subspace embedding) by standard results on ℓ_p Lewis weight sampling (Cohen & Peng, 2015; Woodruff & Yasuda, 2023a).

Furthermore, we can design $\mathbf{B}$ such that the ℓ_p Lewis weights of $\mathbf{B}_{\mathbf{S}}$ are uniformly bounded by α, where α is the oversampling parameter such that $\mathbf{S}$ samples the ith row with probability $\min\{1, \mathbf{w}_i/\alpha\}$. For $p < 2$, this simply follows by taking $\mathbf{B}$ to be a flattening of $\mathbf{A}$ where every row is duplicated $1/\alpha$ times due to the monotonicity of ℓ_p Lewis weights (Cohen & Peng, 2015). For $p > 2$, monotonicity of ℓ_p Lewis weights does not hold, but Theorem 5.2 of (Woodruff & Yasuda, 2023a) nonetheless shows that γ-one-sided ℓ_p Lewis weights can be constructed for $\mathbf{B}_{\mathbf{S}}$ with $\gamma = \Omega(1)$ that makes a similar argument go through.

Finally, it remains to bound the Rademacher process of the form of (3), where $\mathbf{A}$ has γ-one-sided ℓ_p Lewis weights uniformly bounded by $w = \alpha$. We will prove the following in Section B. Assuming this theorem, Theorem A.2 follows by setting $w = \alpha$ as stated.

Theorem A.5. *For all $l \in \mathbb{N}$, we have*

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x}) \right|^l \le (\varepsilon R)^l \tag{4}$$

where

$$\varepsilon = \begin{cases} O(w\eta^{2/p})^{1/2}\gamma^{-1/2}\Big[\big((\log d)^2 \log n\big)^{1+1/l} + l\Big]^{1/2} & p < 2 \\ O(w\eta\|\mathbf{w}\|_1^{p/2-1})^{1/p}\gamma^{-1/2}\Big[\big((\log d)^2 \log n\big)^{1+1/l} + l\Big]^{1/p} & p > 2 \end{cases}.$$

B. Rademacher process bounds

We continue to fix our notation from Definition A.1. We will prove Theorem A.5 in this section.

We split the sum in (4) into two parts: the part that is bounded by the γ-one-sided Lewis weights of $\mathbf{A}$, and the part that is not. To this end, define a threshold

$$\tau := \begin{cases} \dfrac{\eta}{\gamma^{p/2}\varepsilon^p} & p < 2 \\ \dfrac{\eta\|\mathbf{w}\|_1^{p/2-1}}{\gamma^{p/2}\varepsilon^p} & p > 2 \end{cases}$$

where ε will be determined later, and define the set of "good" entries $G \subseteq [n]$ as

$$G := \{i \in [n] : |[\mathbf{A}\mathbf{x}^* - \mathbf{b}](i)| \leq \tau \mathbf{w}_i R\} \tag{5}$$

We then bound

$$\begin{aligned} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R} \left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x})\right|^l &\leq 2^{l-1} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R} \left|\sum_{i\in G} \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x})\right|^l \\ &+ 2^{l-1} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R} \left|\sum_{i\in [n]\setminus G} \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x})\right|^l \end{aligned}$$

using the Fact B.1, and separately estimate each term. We can think of the first term as the "sensitivity" term, where each term in the sum is bounded by the Lewis weights of $\mathbf{A}$, and the latter term as the "outlier" term, where each term in the sum is much larger than the corresponding Lewis weights.

B.1. Preliminaries

We repeatedly use the following inequalities.

Fact B.1. *For any $p \geq 1$ and any $a, b \in \mathbb{R}$, $|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p) = O(|a|^p + |b|^p)$.*

Fact B.2 (Corollary A.2, (Makarychev et al., 2019))**.** *For any $p \geq 1$, $\varepsilon > 0$, and any $a, b \in \mathbb{R}$, $|a+b|^p \leq (1+\varepsilon)|a|^p + \frac{(1+\varepsilon)^{p-1}}{\varepsilon^{p-1}}|b|^p$.*

Fact B.3. *For any $p \geq 1$ and any $a, b \in \mathbb{R}$, $|a|^p - |b|^p \leq p|a-b|(|a|^{p-1} + |b|^{p-1})$.*

We will need the notion of weighted ℓ_p norms $\|\cdot\|_{\mathbf{w},p}$:

Definition B.4. Let $\mathbf{w} \in \mathbb{R}^n$ be non-negative weights. Then for $\mathbf{y} \in \mathbb{R}^n$, we define

$$\|\mathbf{y}\|_{\mathbf{w},p} := \left(\sum_{i=1}^n \mathbf{w}_i |\mathbf{y}(i)|^p\right)^{1/p}.$$

B.1.1. ℓ_p Lewis weights

Lemma B.5 (One-sided Lewis weights bound sensitivities)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p < \infty$. Let $\mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Then,*

$$\sup_{\mathbf{x}\in\operatorname{rowspan}(\mathbf{A})\setminus\{0\}} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \leq \begin{cases} \gamma^{-p/2}\|\mathbf{w}\|_1^{p/2-1}\cdot\mathbf{w}_i & p > 2 \\ \gamma^{-1}\cdot\mathbf{w}_i & p < 2 \end{cases}$$

Lemma B.6. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $\mathbf{w}$ be γ-one-sided ℓ_p Lewis weights for $\mathbf{A}$. Then,*

$$\left\|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{x}\right\|_2 \leq \begin{cases} \|\mathbf{w}\|_1^{1/2-1/p}\|\mathbf{A}\mathbf{x}\|_p & p > 2 \\ \gamma^{1/2-1/p}\|\mathbf{A}\mathbf{x}\|_p & p < 2 \end{cases}$$

Lemma B.7. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $0 < p < \infty$. The following hold: Let $\mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights, and let $\mathbf{R}$ be a change of basis matrix $\mathbf{R}$ such that $\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}$ is an orthonormal matrix. Then, for each $i \in [n]$,*

$$\mathbf{w}_i \geq \gamma^{p/2}\cdot\left\|\mathbf{e}_i^\top\mathbf{A}\mathbf{R}\right\|_2^p.$$

B.1.2. Gaussian processes

Theorem B.8 (Gaussian comparison, Equation 4.8, (Ledoux & Talagrand, 1991))**.** *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be convex and let $\{\mathbf{x}_i\}_{i=1}^n$ be a finite sequence in a Banach space. Then,*

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} F\left(\left\|\sum_{i=1}^n \boldsymbol{\varepsilon}_i\mathbf{x}_i\right\|\right) \leq \mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)} F\left(\left(\frac{\pi}{2}\right)^{1/2}\left\|\sum_{i=1}^n \mathbf{g}_i\mathbf{x}_i\right\|\right).$$

Theorem B.9 (Dudley's entropy integral, Theorem 8.1.6, (Vershynin, 2018))**.** *Let $(X_t)_{t\in T}$ be a Gaussian process with pseudo-metric $d_X(s,t) \coloneqq \|X_s - X_t\|_2$. Let $E(T, d_X, u)$ denote the minimal number of d_X-balls of radius u required to cover T. Then, for every $z \geq 0$, we have that*

$$\mathbf{Pr}\left\{\sup_{s,t\in T}|X_s - X_t| \geq C\left[\int_0^\infty \sqrt{\log E(T,d_X,u)}\,du + z\cdot\operatorname{diam}(T)\right]\right\} \leq 2\exp(-z^2)$$

Integrating the tail bound gives moment bounds. The following is taken from Lemma 6.8 of (Woodruff & Yasuda, 2023c).

Lemma B.10 (Moment bounds)**.** *Let Λ be a Gaussian process with domain T and distance d_X. Let $\mathcal{E} \coloneqq \int_0^\infty \sqrt{\log E(T,d_X,u)}\,du$ and $\mathcal{D} = \operatorname{diam}(T)$. Then, for $l \in \mathbb{N}$,*

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)}[|\Lambda|^l] \leq (2\mathcal{E})^l(\mathcal{E}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l$$

B.2. Estimates on the outlier term

We first bound the outlier terms ($i \notin G$), which is much easier.

Lemma B.11. *With probability* 1*, we have that*

$$\sup_{\|\mathbf{Ax}-\mathbf{Ax}^*\|_p^p\leq\eta R}\sum_{i\in[n]\setminus G}|\Delta_i(\mathbf{x})| \leq O(\varepsilon)R.$$

Proof. For each $i \in [n]\setminus G$, we have that

$$\begin{aligned}
|[\mathbf{Ax}-\mathbf{b}](i)| &\in |[\mathbf{Ax}^*-\mathbf{b}](i)| \pm |[\mathbf{Ax}^*-\mathbf{Ax}](i)| \\
&\in |[\mathbf{Ax}^*-\mathbf{b}](i)| \pm \gamma^{-1/2}\|\mathbf{w}\|_1^{1/2-1/p}\mathbf{w}_i^{1/p}\|\mathbf{Ax}^*-\mathbf{Ax}\|_p && \text{Lemma B.5} \\
&\in |[\mathbf{Ax}^*-\mathbf{b}](i)| \pm \gamma^{-1/2}\eta^{1/p}\|\mathbf{w}\|_1^{1/2-1/p}\mathbf{w}_i^{1/p}R^{1/p} \\
&\in |[\mathbf{Ax}^*-\mathbf{b}](i)| \pm \varepsilon|[\mathbf{Ax}^*-\mathbf{b}](i)| && i\in[n]\setminus G
\end{aligned}$$

Thus,

$$|\Delta_i(\mathbf{x})| \leq O(\varepsilon)|[\mathbf{Ax}^*-\mathbf{b}](i)|^p$$

so

$$\sum_{i\in[n]\setminus G}|\Delta_i(\mathbf{x})| \leq \sum_{i=1}^n O(\varepsilon)|[\mathbf{Ax}^*-\mathbf{b}](i)|^p = O(\varepsilon)\|\mathbf{Ax}^*-\mathbf{b}\|_p^p \leq O(\varepsilon)R. \qquad \square$$

B.3. Estimates on the sensitivity term

Next, we estimate the sensitivity term ($i \in G$),

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\sup_{\|\mathbf{Ax}-\mathbf{Ax}^*\|_p^p\leq\eta R}\left|\sum_{i\in G}\boldsymbol{\varepsilon}_i\Delta_i(\mathbf{x})\right|^l.$$

To estimate this moment, we obtain a subgaussian tail bound via the tail form of Dudley's entropy integral, and then integrate it. We will crucially use that $|\Delta_i(\mathbf{x})|$ for $i \in G$ is bounded over all $\|\mathbf{Ax}-\mathbf{Ax}^*\|_p^p \leq \eta R$, which gives the following sensitivity bound:

Lemma B.12. *For all $i \in G$, and $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R$, we have $|[\mathbf{A}\mathbf{x} - \mathbf{b}](i)|^p \leq O(\tau \mathbf{w}_i R)$ and $|\Delta_i(\mathbf{x})| \leq O(\tau \mathbf{w}_i R)$.*

Proof. We have

$$\begin{aligned}
|[\mathbf{A}\mathbf{x} - \mathbf{b}](i)|^p &\leq 2^{p-1}(|[\mathbf{A}\mathbf{x}^* - \mathbf{b}](i)|^p + |[\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*](i)|^p) && \text{Fact B.1} \\
&\leq 2^{p-1}\tau \mathbf{w}_i R + 2^{p-1}\gamma^{-p/2}\eta \|\mathbf{w}\|_1^{0\vee(p/2-1)}\mathbf{w}_i R && i \in G \text{ (see (5)) and Lemma B.5} \\
&\leq O(\tau \mathbf{w}_i R)
\end{aligned}$$

The bound on $\Delta_i(\mathbf{x})$ follows easily from the above calculation. □

B.3.1. Bounding low-sensitivity entries

We now separately handle entries $i \in G$ with small Lewis weight. To do this end, define

$$J := \left\{ i \in G : \mathbf{w}_i \geq \frac{\varepsilon}{\tau n} \right\}.$$

We then bound the mass on the complement of J:

Lemma B.13. *For all $\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R$, we have that*

$$\sum_{i \in [n] \setminus J} |\Delta_i(\mathbf{x})| \leq O(\varepsilon R)$$

Proof. We have that for each $i \in [n] \setminus J$, $\mathbf{w}_i \leq \varepsilon / \tau n$ so by Lemma B.12,

$$\sum_{i \in [n] \setminus J} |\Delta_i(\mathbf{x})| \leq \sum_{i \in [n] \setminus J} O(\tau \mathbf{w}_i R) \leq \sum_{i \in [n] \setminus J} \frac{O(\varepsilon)}{n} R \leq O(\varepsilon R)$$

□

B.3.2. Bounding high-sensitivity entries: Gaussian processes

Finally, it remains to bound the Rademacher process only on the entries indexed by $i \in J$. By a Gaussian comparison theorem (Theorem B.8), we may bound the Rademacher process above by a Gaussian process instead, that is,

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon} \sim \{\pm 1\}^n} \sup_{\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R} \left| \sum_{i \in J} \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x}) \right|^l \leq \left(\frac{\pi}{2}\right)^{l/2} \mathop{\mathbf{E}}_{\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)} \sup_{\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R} \left| \sum_{i \in J} \mathbf{g}_i \Delta_i(\mathbf{x}) \right|^l. \tag{6}$$

We can now appeal to the theory of Gaussian processes to bound this quantity. Define a Gaussian process by

$$X_{\mathbf{x}} := \sum_{i \in J} \mathbf{g}_i \Delta_i(\mathbf{x})$$

with pseudo-metric

$$d_X(\mathbf{x}, \mathbf{x}') := \left(\mathop{\mathbf{E}}_{\mathbf{g}} |X_{\mathbf{x}} - X'_{\mathbf{x}}|^2 \right)^{1/2} = \left(\sum_{i \in J} (\Delta_i(\mathbf{x}) - \Delta_i(\mathbf{x}'))^2 \right)^{1/2}$$

We will use Dudley's entropy integral (Theorem B.9) to bound the tail of this quantity, and then integrate to obtain moment bounds.

Using the sensitivity bound of Lemma B.12, we obtain a bound on the pseudo-metric d_X.

Lemma B.14. *Let $q = O(\log(\tau n/\varepsilon))$. For $\mathbf{x}, \mathbf{x}' \in T$ for $T = \{\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R\}$, we have that*

$$d_X(\mathbf{x}, \mathbf{x}') \leq \begin{cases} O(w^{1/2})\eta^{1/p-1/2}\|\mathbf{W}^{-1/p}\mathbf{A}(\mathbf{x}-\mathbf{x}')\|_{\mathbf{w},q}^{p/2}R^{1/2} & p < 2 \\ O(w^{1/2})\tau^{1/2-1/p}\|\mathbf{W}^{-1/p}\mathbf{A}(\mathbf{x}-\mathbf{x}')\|_{\mathbf{w},q}R^{1-1/p} & p > 2 \end{cases}$$

and

$$\operatorname{diam}(T) = \sup_{\mathbf{x},\mathbf{x}' \in T} d_X(\mathbf{x}, \mathbf{x}') \leq \begin{cases} O(w^{1/2}\eta^{1/p}\gamma^{-1/2}R) & p < 2 \\ O(\varepsilon w^{1/2}\tau^{1/2}R) & p > 2 \end{cases}$$

Proof. Let $\mathbf{y} = \mathbf{A}\mathbf{x} - \mathbf{b}$ and $\mathbf{y}' = \mathbf{A}\mathbf{x}' - \mathbf{b}$. Note then that

$$\begin{aligned} d_X(\mathbf{x}, \mathbf{x}')^2 &= \sum_{i \in J}(\Delta_i(\mathbf{x}) - \Delta_i(\mathbf{x}'))^2 = \sum_{i \in J}(|\mathbf{y}(i)|^p - |\mathbf{y}'(i)|^p)^2 \\ &\leq p^2 \sum_{i \in J}|\mathbf{y}(i) - \mathbf{y}'(i)|^2(|\mathbf{y}(i)|^{p-1} + |\mathbf{y}'(i)|^{p-1})^2 && \text{Fact B.3} \end{aligned}$$

For $p < 2$, we have that

$$\begin{aligned} d_X(\mathbf{x}, \mathbf{x}')^2 &\leq p^2\|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^p \sum_{i \in J}(|\mathbf{y}(i) - \mathbf{y}'(i)|)^{2-p}(|\mathbf{y}(i)|^{p-1} + |\mathbf{y}'(i)|^{p-1})^2 \\ &\leq 2p^2\|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^p \sum_{i \in J}(|\mathbf{y}(i) - \mathbf{y}'(i)|)^{2-p}(|\mathbf{y}(i)|^{2p-2} + |\mathbf{y}'(i)|^{2p-2}) \\ &\leq 2p^2\|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^p\|\mathbf{y} - \mathbf{y}'\|_p^{2-p}(\|\mathbf{y}\|_p^{2p-2} + \|\mathbf{y}'\|_p^{2p-2}) && \text{Hölder's inequality} \\ &\leq O(\eta^{2/p-1})\|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^p R. \end{aligned}$$

where Hölder's inequality is applied with exponents $\frac{p}{2-p} > 1$ and $\frac{p}{2p-2} > 1$. For $p > 2$, we have that

$$\begin{aligned} d_X(\mathbf{x}, \mathbf{x}')^2 &\leq 2p^2\|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^2 \sum_{i=1}^n |\mathbf{y}(i)|^{2p-2} + |\mathbf{y}'(i)|^{2p-2} \\ &\leq 2p^2 \max\{\|\mathbf{y}|_J\|_\infty, \|\mathbf{y}'|_J\|_\infty\}^{p-2}\|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^2 \sum_{i=1}^n |\mathbf{y}(i)|^p + |\mathbf{y}'(i)|^p \\ &\leq O(1)(\tau w R)^{1-2/p}\|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^2 R && \text{Lemma B.12} \end{aligned}$$

Furthermore, we have that

$$\begin{aligned} \|(\mathbf{y} - \mathbf{y}')|_J\|_\infty &= \|(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}')|_J\|_\infty \\ &= \|\mathbf{W}^{1/p}(\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}')|_J\|_\infty \\ &\leq w^{1/p}\|(\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}')|_J\|_\infty \\ &\leq 2w^{1/p}\|\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}'\|_{\mathbf{w},q} \end{aligned}$$

where the last step follows from the fact that $\mathbf{w}_i \geq \varepsilon/\tau n$ for $i \in J$ and $q = O(\log(\tau n/\varepsilon))$. Combining these bounds gives the claimed bound on $d_X(\mathbf{x}, \mathbf{x}')$.

Finally, we have by Lemma B.5 that

$$\|\mathbf{W}^{-1/p}\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_\infty = \max_{i=1}^{n} \frac{|[\mathbf{A}(\mathbf{x} - \mathbf{x}^*)](i)|}{\mathbf{w}_i} \leq \begin{cases} \gamma^{-1/p}\|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_p & p < 2 \\ \gamma^{-1/2}\|\mathbf{w}\|_1^{1/2-1/p}\|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_p & p > 2 \end{cases}$$

so we have the claimed diameter bound for the set $\{\|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|_p^p \leq \eta R\}$. □

The following entropy bounds are obtained from (Woodruff & Yasuda, 2023c), which in turn largely follow (Bourgain et al., 1989).

Remark B.15. The following entropy bounds are net necessary if we only need this result for $d = 1$, for example for applications to Euclidean power means. In this case, standard volume arguments suffice (see, e.g., Lemma 2.4 of Bourgain et al. (1989)).

Lemma B.16. *Let $1 \geq \mathbf{w} \in \mathbb{R}^n$ be non-negative weights. Let $2 \leq q < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be such that $\mathbf{W}^{1/2}\mathbf{A}$ is orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Let $B^p_{\mathbf{w}}(\mathbf{A}) \coloneqq \{\mathbf{x} : \|\mathbf{A}\mathbf{x}\|_{\mathbf{w},p} \leq 1\}$. Then,*

$$\log E(B^2_{\mathbf{w}}(\mathbf{A}), B^q_{\mathbf{w}}(\mathbf{A}), t) \leq O(1)\frac{n^{2/q}q\cdot\tau}{t^2}$$

and

$$\log E(B^p_{\mathbf{w}}(\mathbf{A}), B^q_{\mathbf{w}}(\mathbf{A}), t) \leq O(1)\frac{1}{t^p}\left(\frac{\log d}{2-p} + \log n + n^{2/q}q\right)\tau.$$

for $p < 2$.

We may now evaluate Dudley's entropy integral.

Lemma B.17 (Entropy integral bound for $p < 2$)**.** *We have that*

$$\int_0^\infty \sqrt{\log E(B^p(\mathbf{A}), d_X, t)}\,dt \leq O(w^{1/2}\gamma^{-1/2}\eta^{1/2}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log d$$

Proof. Note that it suffices to integrate the entropy integral to $\operatorname{diam}(T)$, which is bounded in Lemma B.14. Note also that T is just a translation of $(\eta R)^{1/p}\cdot B^p(\mathbf{A})$, so we have

$$\begin{aligned}
\log E(T, d_X, t) &= \log E((\eta R)^{1/p}\cdot B^p(\mathbf{A}), d_X, t) \\
&= \log E((\eta R)^{1/p}\cdot B^p(\mathbf{A}), K\|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}^{p/2}, t) && \text{Lemma B.14} \\
&= \log E(B^p_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), B^q_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), t^{2/p}/K^{2/p}(\eta R)^{1/p})
\end{aligned}$$

where $K = O(w^{1/2}\eta^{1/p-1/2}R^{1/2})$.

For small radii less than λ for a parameter λ to be chosen, we use a standard volume argument, which shows that

$$\log E(B^p_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), B^q_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), t) \leq O(d)\log\frac{n}{t}$$

so

$$\begin{aligned}
\int_0^\lambda \sqrt{\log E(T, d_X, t)}\,dt &\leq \int_0^\lambda \sqrt{d\log\frac{nK^{2/p}(\eta R)^{1/p}}{t^{2/p}}}\,dt \\
&\leq \lambda\sqrt{d\log(n(\eta^{2/p}w)^{1/p})} + \sqrt{d}\int_0^\lambda\sqrt{\log\frac{R^{2/p}}{t^{2/p}}}\,dt \\
&\leq \lambda\sqrt{d\log(n(\eta^{2/p}w)^{1/p})} + \sqrt{d}\cdot O(\lambda)\sqrt{\log\frac{R}{\lambda}} \\
&\leq O(\lambda)\sqrt{d\log\frac{n(\eta^{2/p}w)^{1/p}R}{\lambda}}
\end{aligned}$$

On the other hand, for large radii larger than λ, we use the bounds of Lemma B.16. Note that the entropy bounds do not change if we replace $\mathbf{A}$ by $\mathbf{AR}$, where $\mathbf{R}$ is the change of basis matrix such that $\mathbf{W}^{1/2-1/p}\mathbf{AR}$ is orthonormal. Then by the properties of γ-one-sided ℓ_p Lewis weights (Lemma B.7), we have

$$\|\mathbf{e}_i^\top\mathbf{W}^{-1/p}\mathbf{AR}\|_2^2 = \mathbf{w}_i^{-2/p}\|\mathbf{e}_i^\top\mathbf{AR}\|_2^2 \leq \gamma^{-1}.$$

Then, Lemma B.16 gives

$$\log E(B^p_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), B^q_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), t^{2/p}/K^{2/p}(\eta R)^{1/p}) = \frac{O(w\eta^{2/p}R^2)}{\gamma t^2}\log\frac{\tau n}{\varepsilon}$$

so the entropy integral gives a bound of

$$\frac{O(w^{1/2}\eta^{1/p}R)}{\gamma^{1/2}}\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\int_{\lambda}^{\mathrm{diam}(T)}\frac{1}{t}\,dt = \frac{O(w^{1/2}\eta^{1/p}R)}{\gamma^{1/2}}\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log\frac{\mathrm{diam}(T)}{\lambda}.$$

We choose $\lambda = \mathrm{diam}(T)/\sqrt{d}$, which yields the claimed conclusion. □

An analogous result and proof holds for $p > 2$.

Lemma B.18 (Entropy integral bound for $p > 2$)**.** *Let $2 < p < \infty$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $0 \le \mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Let $w = \max_{i\in[n]}\mathbf{w}_i$. Then,*

$$\int_0^\infty \sqrt{\log E(B^p(\mathbf{A}), d_X, t)}\,dt \le O(w^{1/2}\varepsilon\tau^{1/2}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log d$$

Proof. Note that it suffices to integrate the entropy integral to $\mathrm{diam}(T)$, which is bounded in Lemma B.14. Note also that T is just a translation of $(\eta R)^{1/p}\cdot B^p(\mathbf{A})$, so we have

$$\begin{aligned}
\log E(T, d_X, t) &= \log E((\eta R)^{1/p}\cdot B^p(\mathbf{A}), d_X, t) \\
&= \log E((\eta R)^{1/p}\cdot B^p(\mathbf{A}), K\|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}, t) && \text{Lemma B.14} \\
&= \log E(B^p_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), B^q_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), t/K(\eta R)^{1/p})
\end{aligned}$$

where $K = O(w^{1/2}\tau^{1/2-1/p}R^{1-1/p})$.

For small radii less than λ for a parameter λ to be chosen, we use a standard volume argument, which shows that

$$\log E(B^p_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), B^q_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), t) \le O(d)\log\frac{n}{t}$$

so

$$\begin{aligned}
\int_0^\lambda \sqrt{\log E(T, d_X, t)}\,dt &\le \int_0^\lambda \sqrt{d\log\frac{nK(\eta R)^{1/p}}{t}}\,dt \\
&\le \lambda\sqrt{d\log(nw^{1/2}\eta^{1/p}\tau^{1/2-1/p})} + \sqrt{d}\int_0^\lambda\sqrt{\log\frac{R}{t}}\,dt \\
&\le \lambda\sqrt{d\log(nw^{1/2}\eta^{1/p}\tau^{1/2-1/p})} + \sqrt{d}\cdot O(\lambda)\sqrt{\log\frac{R}{\lambda}} \\
&\le O(\lambda)\sqrt{d\log\frac{nw^{1/2}\eta^{1/p}\tau^{1/2-1/p}R}{\lambda}}
\end{aligned}$$

On the other hand, for large radii larger than λ, we use the bounds of Lemma B.16. Note that the entropy bounds do not change if we replace $\mathbf{A}$ by $\mathbf{AR}$, where $\mathbf{R}$ is the change of basis matrix such that $\mathbf{W}^{1/2-1/p}\mathbf{AR}$ is orthonormal. Then by the properties of γ-one-sided ℓ_p Lewis weights (Lemma B.7), we have

$$\|\mathbf{e}_i^\top\mathbf{W}^{-1/p}\mathbf{AR}\|_2^2 = \mathbf{w}_i^{-2/p}\|\mathbf{e}_i^\top\mathbf{AR}\|_2^2 \le \gamma^{-1}.$$

Then, Lemma B.6 and Lemma B.16 give

$$\begin{aligned}
&\log E(B^p_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), B^q_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), t/K(\eta R)^{1/p}) \\
\le\ &\log E(B^2_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), B^q_{\mathbf{w}}(\mathbf{W}^{-1/p}\mathbf{A}), t/K(\eta R)^{1/p}\|\mathbf{w}\|_1^{1/2-1/p}) \\
\le\ &\frac{K^2(\eta R)^{2/p}\|\mathbf{w}\|_1^{1-2/p}}{\gamma t^2}\log\frac{\tau n}{\varepsilon} \\
\le\ &\frac{O(w)\varepsilon^2\tau R^2}{t^2}\log\frac{\tau n}{\varepsilon}
\end{aligned}$$

so the entropy integral gives a bound of

$$O(w^{1/2}\varepsilon\tau^{1/2}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\int_{\lambda}^{\operatorname{diam}(T)}\frac{1}{t}\,dt = O(w^{1/2}\varepsilon\tau^{1/2}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log\frac{\operatorname{diam}(T)}{\lambda}.$$

We choose $\lambda = \operatorname{diam}(T)/\sqrt{d}$, which yields the claimed conclusion. □

We are now ready to prove Theorem A.5.

Proof of Theorem A.5. We have by Lemma B.10 that the Gaussian process of (6) is bounded by

$$(2\mathcal{E})^l(\mathcal{E}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l$$

where

$$\mathcal{E} \leq \begin{cases} O(w^{1/2}\gamma^{-1/2}\eta^{1/p}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log d & p < 2 \\ O(\varepsilon w^{1/2}\tau^{1/2}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log d & p > 2 \end{cases}$$

by Lemmas B.17 and B.18 and

$$\mathcal{D} \leq \begin{cases} O(w^{1/2}\eta^{1/p}\gamma^{-1/2}R) & p < 2 \\ O(\varepsilon w^{1/2}\tau^{1/2}R) & p > 2 \end{cases}$$

by Lemma B.14. This gives a bound of $(\alpha R)^l$ on (6), where

$$\alpha = \begin{cases} O(w^{1/2}\eta^{1/p}\gamma^{-1/2})\left[\left(\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log d\right)^{1+1/l} + \sqrt{l}\right] & p < 2 \\ O(\varepsilon w^{1/2}\tau^{1/2}R)\left[\left(\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log d\right)^{1+1/l} + \sqrt{l}\right] & p > 2 \end{cases}$$

We now set $\alpha = \varepsilon$ and solve for the ε that we can obtain. From this, we see that we can set

$$\varepsilon = \begin{cases} O(w\eta^{2/p})^{1/2}\gamma^{-1/2}\Big[\big((\log d)^2\log n\big)^{1+1/l} + l\Big]^{1/2} & p < 2 \\ O(w\eta\|\mathbf{w}\|_1^{p/2-1})^{1/p}\gamma^{-1/2}\Big[\big((\log d)^2\log n\big)^{1+1/l} + l\Big]^{1/p} & p > 2 \end{cases}.$$

□

C. Missing proofs for weak coresets

C.1. Proof of the closeness lemma

Proof of Lemma 4.4. First note that

$$\begin{aligned}\Big\langle(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})^{\circ(p-1)}, \mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{A}\mathbf{X}\mathbf{G}\Big\rangle &= \sum_{i=1}^n\sum_{j=1}^m[\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}](i,j)^{\circ(p-1)}[\mathbf{A}(\mathbf{X}^*-\mathbf{X})\mathbf{G}](i,j) \\ &= \sum_{i=1}^n\sum_{j=1}^m[\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}](i,j)^{\circ(p-1)}\Big\langle(\mathbf{A}^\top\mathbf{e}_i)(\mathbf{e}_j^\top\mathbf{G}^\top), \mathbf{X}^*-\mathbf{X}\Big\rangle \\ &= \Bigg\langle\sum_{i=1}^n\sum_{j=1}^m[\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}](i,j)^{\circ(p-1)}(\mathbf{A}^\top\mathbf{e}_i)(\mathbf{e}_j^\top\mathbf{G}^\top), \mathbf{X}^*-\mathbf{X}\Bigg\rangle.\end{aligned}$$

The left term in the product is the gradient of the objective at the optimum by Lemma 4.3, so this is just 0 for any $\mathbf{X}$. Then for $p < 2$, we have by Lemma 4.2 that

$$\|\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}\|_{p,p}^2 + \frac{p-1}{2}\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^2 \leq \|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}^2 \leq (1+\eta)^2\|\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}\|_{p,p}^2$$

which rearranges to

$$\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p} \leq O(\eta^{1/2})\mathsf{OPT}.$$

and for $p > 2$, we have by Lemma 4.2 that

$$\|\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p + \frac{p-1}{p2^p}\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p \leq \|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}^p \leq (1+\eta)^p\|\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$$

which rearranges to

$$\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p} \leq O(\eta^{1/p})\mathsf{OPT}.$$

□

C.2. Proof of the initial weak coreset bound

Proof of Lemma 4.5. We first show that

$$\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p \leq O\left(\frac{1}{\delta}\right)\mathsf{OPT}^p$$

with probability at least $1-\delta$. By using the fact that $\mathbf{S}$ is an $O(1)$-approximate ℓ_p subspace embedding, we have that

$$\begin{aligned}
\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p &\leq \|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G})\|_{p,p}^p \\
&\leq 2^{p-1}\Big(\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B})\|_{p,p}^p + \|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p\Big) && \text{Fact B.1} \\
&\leq 2^{p+1}\|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p && \text{Approximate optimality of } \hat{\mathbf{X}}
\end{aligned}$$

The latter quantity is at most $O(\frac{1}{\delta})\mathsf{OPT}^p$ with probability at least $1-\delta$ by Markov's inequality. Thus, we may replace the optimization of $\hat{\mathbf{X}}$ over all $\mathbf{X}\in\mathbb{R}^{d\times t}$ with optimization over the ball $\{\mathbf{X} : \|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p = O(\frac{1}{\delta})\mathsf{OPT}^p\}$.

The rest of the proof now mimics the proof of Theorem 3.1. We apply Theorem 3.2 with accuracy parameter ε set to $\varepsilon\delta$, failure parameter set to $(\varepsilon\delta)^p\delta^2$, and proximity parameter η set to 1. Let $S\subseteq[m]$ be the set of columns for which Theorem 3.2 fails. Then by applying Markov's inequality twice as in the proof of Theorem 3.1, we have that

$$\sum_{j\in S}\|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})\mathbf{e}_j\|_p^p = O((\varepsilon\delta)^p)\mathsf{OPT}^p$$

and

$$\sum_{j\in S}\|(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})\mathbf{e}_j\|_p^p = O((\varepsilon\delta)^p)\mathsf{OPT}^p$$

and thus it follows that

$$\sum_{j\in S}\|\mathbf{S}(\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B})\mathbf{e}_j\|_p^p = \sum_{j\in S}\|(\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B})\mathbf{e}_j\|_p^p \pm O(\varepsilon\delta)\big(\|\mathbf{A}(\mathbf{X}-\mathbf{X}^*)\mathbf{G}\|_p^p + \mathsf{OPT}^p\big).$$

Summing this result with the rest of the columns $j\notin S$ gives that

$$\begin{aligned}
&\Big|\big(\|\mathbf{S}(\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B})\|_{p,p}^p - \|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p\big) - \big(\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}^p - \|\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p\big)\Big| \\
\leq\ &\varepsilon\delta\big(\|\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p + \|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p + \|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p\big) \leq O(\varepsilon)\mathsf{OPT}^p
\end{aligned}$$

Thus, in the ball $\{\mathbf{X} : \|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p = O(\frac{1}{\delta})\mathsf{OPT}^p\}$, we have that

$$\|\mathbf{S}(\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B})\|_{p,p}^p = \|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}^p + (\|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p - \|\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p) \pm O(\varepsilon)\mathsf{OPT}^p.$$

It follows that $\hat{\mathbf{X}}$ must minimize $\|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}^p$ up to an additive $O(\varepsilon)\mathsf{OPT}^p$. □

C.3. Proof of the weak coreset construction

Proof of Theorem 4.1. Let

$$C = \begin{cases} O(\gamma^{-1})\delta^{-2}\|\mathbf{w}\|_1\left[(\log d)^2 \log n + \log\frac{1}{\delta}\right] & p < 2 \\ O(\gamma^{-p/2})\delta^{-p}\|\mathbf{w}\|_1^{p/2}\left[(\log d)^2 \log n + \log\frac{1}{\delta}\right] & p > 2 \end{cases}$$

We will make use of the fact that $\|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B})\|_{p,p}^p = O(\frac{1}{\delta})\|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B})\|_{p,p}^p$ with probability at least $1-\delta$ by Markov's inequality.

We will first give the argument for $p < 2$. Suppose that C/ε^β rows are needed for a $(1+\varepsilon)$-approximate weak coreset. Now choose a such that $a - 2 = -a\beta$, that is, $a = 2/(1+\beta)$. Then for $\eta^{2/p} = \varepsilon^a$, $C\eta^{2/p}/(\varepsilon\delta)^2 = C/\eta^{(2/p)\beta}$ rows yields a $(1+\eta^{2/p})$-approximate weak coreset. Then, a $(1+\eta^{2/p})$-approximate minimizer $\mathbf{X}$ satisfies

$$\|\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p \leq O(\eta)\|\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B}\|_{p,p}^p$$

by Lemma 4.4. For all such $\mathbf{X}$, an argument as done in Theorem 3.1 and Lemma 4.5 shows that $\|\mathbf{S}(\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{B})\|_{p,p}^p - \|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B})\|_{p,p}^p$ and $\|\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{B}\|_{p,p}^p - \|\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B}\|_{p,p}^p$ are close up to an additive error of

$$\varepsilon\delta\left(\|\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B}\|_{p,p}^p + \|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B})\|_{p,p}^p + \frac{1}{\eta}\|\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p\right) = O(\varepsilon)\|\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B}\|_{p,p}^p$$

Thus, $C/\eta^{(2/p)\beta}$ rows in fact gives a $(1+O(\varepsilon))$-approximate minimizer. That is, if C/ε^β rows is sufficient for $(1+\varepsilon)$-approximation, then $C/\eta^{(2/p)\beta} = C/\varepsilon^{a\beta} = C/\varepsilon^{2\beta/(1+\beta)}$ rows is sufficient for $(1+\varepsilon)$-approximation as well. We may now iterate this argument. Consider the sequence β_i given by

$$\beta_0 = 2, \qquad \beta_{i+1} = \frac{2\beta_i}{1+\beta_i}.$$

The solution to this recurrence is given by the following lemma, with $p = 2$:

Lemma C.1. *Let $p > 1$ and let $\{\beta_i\}_{i=0}^\infty$ be defined by the recurrence relation $\beta_0 = p$ and $\beta_{i+1} = p\beta_i/(1+\beta_i)$. Then,*

$$\beta_i = \frac{1}{p^{-i}(p^{-1} - (p-1)^{-1}) + (p-1)^{-1}}$$

Proof. Note that $\frac{1}{\beta_{i+1}} = \frac{1}{p}\frac{1}{\beta_i} + \frac{1}{p}$ so the sequence $\{a_i\}_{i=0}^\infty$ given by $a_i = 1/\beta_i$ satisfies the linear recurrence $a_{i+1} = \frac{1}{p}a_i + \frac{1}{p}$. Note that this recurrence has the fixed point $a = 1/(p-1)$, so the sequence $a_i' = a_i - a$ satisfies $a_{i+1}' = \frac{1}{p}a_i'$, which gives, $a_i' = p^{-i}a_0'$. Thus, $a_i - a = p^{-i}(a_0 - a)$ so

$$\begin{aligned} \beta_i = \frac{1}{a_i} &= \frac{1}{p^{-i}(a_0 - a) + a} \\ &= \frac{1}{p^{-i}(p^{-1} - (p-1)^{-1}) + (p-1)^{-1}}. \end{aligned}$$

□

Thus, applying this argument $O(\log\log\frac{1}{\varepsilon})$ times yields that $\beta_i \leq 1 + O(1/\log(\frac{1}{\varepsilon}))$ which means that reading only $O(1)C/\varepsilon$ entries suffices. Union bounding over the success of the $O(\log\log\frac{1}{\varepsilon})$ rounds completes the argument.

Next, let $p > 2$. Suppose that C/ε^β rows are needed for a $(1+\varepsilon)$-approximate weak coreset. Now choose a such that $a - p = -a\beta$, that is, $a = p/(1+\beta)$. Then for $\eta = \varepsilon^a$, $C\eta/\varepsilon^p = C/\eta^\beta$ rows yields a $(1+\eta)$-approximate weak coreset. Then, a $(1+\eta)$-approximate minimizer $\mathbf{X}$ satisfies

$$\|\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p \leq O(\eta)\|\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B}\|_{p,p}^p$$

by Lemma 4.4. For all such $\mathbf{X}$, an argument as done in Theorem 3.1 and Lemma 4.5 shows that $\|\mathbf{S}(\mathbf{AXG}-\mathbf{B})\|_{p,p}^p - \|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p$ and $\|\mathbf{AXG}-\mathbf{B}\|_{p,p}^p - \|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$ are close up to an additive error of

$$\varepsilon\left(\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p + \frac{1}{\eta}\|\mathbf{AXG}-\mathbf{AX}^*\mathbf{G}\|_{p,p}^p\right) = O(\varepsilon)\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$$

Thus, C/η^β rows in fact gives a $(1+O(\varepsilon))$-approximate minimizer. That is, if C/ε^β rows is sufficient for $(1+\varepsilon)$-approximation, then $C/\eta^\beta = C/\varepsilon^{a\beta} = C/\varepsilon^{p\beta/(1+\beta)}$ rows is sufficient for $(1+\varepsilon)$-approximation as well. We may now iterate this argument. Consider the sequence β_i given by

$$\beta_1 = p, \qquad \beta_{i+1} = \frac{p\beta_i}{1+\beta_i}.$$

Then by Lemma C.1, applying this argument $O(\log\log\frac{1}{\varepsilon})$ times yields that $\beta_i \leq (p-1) + O(1/\log(\frac{1}{\varepsilon}))$ which means that reading only $O(1)C/\varepsilon^{p-1}$ entries suffices. Union bounding over the success of the $O(\log\log\frac{1}{\varepsilon})$ rounds completes the argument. □

D. Missing proofs for applications

D.1. Sublinear algorithm for Euclidean power means

Theorem 1.6. *Let $\{\mathbf{b}_i\}_{i=1}^n \subseteq \mathbb{R}^d$. Then, there is a sublinear algorithm which uniformly samples at most*

$$s = \begin{cases} O(\varepsilon^{-2})\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\right)\log\frac{1}{\delta} & p=1 \\ O(\varepsilon^{-1})\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\right)\log\frac{1}{\delta} & 1<p\leq 2 \\ O(\varepsilon^{1-p})\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\right)\log\frac{1}{\delta} & 2<p<\infty \end{cases}$$

rows $\mathbf{b}_i$ and outputs a center $\hat{\mathbf{x}}$ such that

$$\sum_{i=1}^n \|\hat{\mathbf{x}}-\mathbf{b}_i\|_2^p \leq (1+\varepsilon)\min_{\mathbf{x}\in\mathbb{R}^d}\sum_{i=1}^n\|\mathbf{x}-\mathbf{b}_i\|_2^p$$

with probability at least $1-\delta$.

Proof. We will assume without loss of generality that by reading $O(\log\frac{1}{\delta})$ rows of $\mathbf{B}$, we can identify an $O(1)$-approximate solution $\hat{\mathbf{x}}$ (see, e.g., Section 3.1 of (Musco et al., 2022)). Thus by subtracting off this solution, we may assume that $\|\mathbf{B}\|_{p,2}^p = O(\mathsf{OPT}^p)$.

We then use Dvoretzky's thoerem to embed this problem into the entrywise ℓ_p norm, so that

$$\|\mathbf{1x}^\top - \mathbf{B}\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{1x}^\top\mathbf{G}-\mathbf{BG}\|_{p,p}^p$$

for every center $\mathbf{x}\in\mathbb{R}^d$. This is now in a form where we may apply our weak coreset results for multiple ℓ_p regression of Theorem 1.5. Note that in this particular setting, the $\mathbf{A}$ matrix corresponds to the $n\times d$ all ones matrix with $d=1$, and the ℓ_p Lewis weights can be taken to be uniform.

Now consider running $L = O(\log\frac{1}{\delta})$ independent instances of the weak coreset algorithm, each which has the property that the algorithm makes at most

$$O(\varepsilon^{-\rho})\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\right) \tag{7}$$

queries for $\rho=2$ for $p=1$, $\rho=1$ for $1<p<2$, and $\rho=p-1$ for $2<p<\infty$, and that if $\|\mathbf{S}(\mathbf{1}(\mathbf{x}^*)^\top\mathbf{G}-\mathbf{BG})\|_{p,p}^p = O(\|\mathbf{1}(\mathbf{x}^*)^\top\mathbf{G}-\mathbf{BG}\|_{p,p}^p)$ for the optimal solution $\mathbf{x}^*$, then it succeeds with probability at least $1-\delta/L$. By a union bound, this holds for all L instances.

By Markov's inequality, each instance satisfies $\|\mathbf{SBG}\|_{p,p}^p = O(\|\mathbf{BG}\|_{p,p}^p)$ with probability at least $9/10$, so at least $2/3$ of the L instances must satisfy this bound with probability at least $1-\delta$. By Dvoretzky's theorem, this means that $\|\mathbf{SB}\|_{p,2}^p = O(\|\mathbf{B}\|_{p,2}^p)$. Then, if we restrict our attention to the $(2/3)L$ instances with the smallest values of $\|\mathbf{SB}\|_{p,2}^p$, then all of these instances must output a correct $(1+\varepsilon)$-approximately optimal solution, simultaneously with probability $1-\delta$. This gives a query bound of L times (7). □

D.2. Spanning coresets for ℓ_p subspace approximation

We show that weak coreset construction imply spanning sets for ℓ_p subspace approximation.

Theorem 1.9. *Let* $\{\mathbf{a}_i\}_{i=1}^n \subseteq \mathbb{R}^d$*,* $1 \le p < \infty$*,* $k \in \mathbb{N}$*, and* $0 < \varepsilon < 1$*. Then, there exists a* $(1+\varepsilon)$*-spanning coreset* S *of size at most*

$$|S| = \begin{cases} O(\varepsilon^{-2}k)(\log(k/\varepsilon))^3 & p = 1 \\ O(\varepsilon^{-1}k)(\log(k/\varepsilon))^3 & 1 < p \le 2 \\ O(\varepsilon^{1-p}k^{p/2})(\log(k/\varepsilon))^3 & 2 < p < \infty \end{cases}$$

Proof. By first computing a strong coreset of size $\mathrm{poly}(k/\varepsilon)$ (Huang & Vishnoi, 2020), we can assume that $n, d = \mathrm{poly}(k/\varepsilon)$.

Let $\mathbf{P} = \mathbf{V}\mathbf{V}^\top$ be the rank k projection that minimizes $\|\mathbf{A}\mathbf{P} - \mathbf{A}\|_{p,2}^p$. Note then that

$$\min_{\mathbf{X} \in \mathbb{R}^{k \times d}} \|\mathbf{A}\mathbf{V}\mathbf{X} - \mathbf{A}\|_{p,2}^p = \|\mathbf{A}\mathbf{P} - \mathbf{A}\|_{p,2}^p.$$

We then use Dvoretzky's theorem to embed this problem into the entrywise ℓ_p norm, so that

$$\|\mathbf{A}\mathbf{V}\mathbf{X} - \mathbf{A}\|_{p,2}^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{V}\mathbf{X}\mathbf{G} - \mathbf{A}\mathbf{G}\|_{p,p}^p$$

for every $\mathbf{X} \in \mathbb{R}^{k \times d}$, for some fixed $\mathbf{G} \in \mathbb{R}^{d \times m}$ with $m = \mathrm{poly}(d/\varepsilon)$. Then by our weak coreset result for multiple ℓ_p regression (Theorem 4.1), there is a diagonal matrix $\mathbf{S}$ with

$$\mathsf{nnz}(\mathbf{S}) \le \begin{cases} O(\varepsilon^{-2}k)(\log(k/\varepsilon))^3 & p = 1 \\ O(\varepsilon^{-1}k)(\log(k/\varepsilon))^3 & 1 < p \le 2 \\ O(\varepsilon^{1-p}k^{p/2})(\log(k/\varepsilon))^3 & 2 < p < \infty \end{cases}$$

such that any $(1+\varepsilon)$-approximate minimizer $\hat{\mathbf{X}}$ of $\|\mathbf{S}(\mathbf{A}\mathbf{V}\mathbf{X}\mathbf{G} - \mathbf{A}\mathbf{G})\|_{p,p}^p$ satisfies

$$\|\mathbf{A}\mathbf{V}\hat{\mathbf{X}}\mathbf{G} - \mathbf{A}\mathbf{G}\|_{p,p}^p \le (1+\varepsilon) \min_{\mathbf{X} \in \mathbb{R}^{k \times d}} \|\mathbf{A}\mathbf{V}\mathbf{X}\mathbf{G} - \mathbf{A}\mathbf{G}\|_{p,p}^p.$$

We will take $\hat{\mathbf{X}}$ to be

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X} \in \mathbb{R}^{k \times d}} \|\mathbf{S}(\mathbf{A}\mathbf{V}\mathbf{X} - \mathbf{A})\|_{p,2}^p$$

which is indeed a $(1+\varepsilon)$-approximate minimizer of $\|\mathbf{S}(\mathbf{A}\mathbf{V}\mathbf{X}\mathbf{G} - \mathbf{A}\mathbf{G})\|_{p,p}^p$ by Dvoretzky's theorem. Then, again by Dvoretzky's theorem, we then have for this $\hat{\mathbf{X}}$ that

$$\begin{aligned} \|\mathbf{A}\mathbf{V}\hat{\mathbf{X}} - \mathbf{A}\|_{p,2}^p &\le (1 + O(\varepsilon)) \min_{\mathbf{X} \in \mathbb{R}^{k \times d}} \|\mathbf{A}\mathbf{V}\mathbf{X} - \mathbf{A}\|_{p,2}^p \\ &= (1 + O(\varepsilon))\|\mathbf{A}\mathbf{P} - \mathbf{A}\|_{p,2}^p. \end{aligned}$$

Finally, note that $\hat{\mathbf{X}}$ has row span contained in the row span of $\mathbf{S}\mathbf{A}$, since otherwise $\|\mathbf{S}(\mathbf{A}\mathbf{V}\mathbf{X} - \mathbf{A})\|_{p,2}^p$ can be reduced by projecting the rows of $\mathbf{X}$ onto $\mathrm{rowspan}(\mathbf{S}\mathbf{A})$. Then, if $\mathbf{P}_F$ is the projection matrix onto $F = \mathrm{rowspan}(\hat{\mathbf{X}})$, then for each row $i \in [n]$ of $\mathbf{A}$,

$$\|\mathbf{P}_F\mathbf{a}_i - \mathbf{a}_i\|_2 = \min_{\mathbf{x} \in F} \|\mathbf{x} - \mathbf{a}_i\|_2 \le \|\hat{\mathbf{X}}^\top\mathbf{V}^\top\mathbf{a}_i - \mathbf{a}_i\|_2$$

so

$$\|\mathbf{A}\mathbf{P}_F - \mathbf{A}\|_{p,2}^p \le \|\mathbf{A}\mathbf{V}\hat{\mathbf{X}} - \mathbf{A}\|_{p,2}^p.$$

We thus conclude that there is a rank k subspace in the row span of $\mathbf{S}\mathbf{A}$ that is $(1+\varepsilon)$-approximately optimal. □

E. Missing proofs for coreset lower bounds

We provide missing proofs from Section 5.

We will use the following lemma from coding theory.

Theorem E.1 ((Parampalli et al., 2013)). *For any $p \geq 1$ and $d = 2^k - 1$ for some integer k, there exists a set $S \subseteq \{-1, 1\}^d$ and a constant C_p depending only on p which satisfy*

- $|S| = d^p$
- *For any $s, t \in S$ such that $s \neq t$, $|\langle s, t \rangle| \leq C_p \sqrt{d}$*

E.1. Strong coresets

Proof of Theorem 5.1. Let $s = d^{p/2}$ and let $S \subseteq \{\pm 1\}^d$ be a set of $|S| = s$ points given by Theorem E.1 such that $\langle \mathbf{a}, \mathbf{a}' \rangle \leq C_{p/2}\sqrt{d} = O(\sqrt{d})$ for some $C_{p/2}^p \geq 1$, for every distinct $\mathbf{a}, \mathbf{a}' \in S$. Let $m = s\varepsilon^{-p}$, let $\mathbf{A} \in \{\pm 1\}^{m \times d}$ be the matrix with ε^{-p} copies of $\mathbf{a}$ in its rows for each $\mathbf{a} \in S$, and let $\mathbf{B} = d \cdot \mathbf{I}_m$ be the $m \times m$ identity matrix scaled by d. For each row $i \in [m]$, we say that $i' \in [s]$ is its *group number* if $\mathbf{e}_i^\top \mathbf{A}$ is the i'-th point in S.

Suppose for contradiction that $\mathbf{S}$ is a strong coreset with $\mathsf{nnz}(\mathbf{S}) \leq m/16$ such that

$$\|\mathbf{S}(\mathbf{AX} - \mathbf{B})\|_{p,p}^p = \left(1 \pm \frac{\varepsilon}{12 C_{p/2}^p}\right) \|\mathbf{AX} - \mathbf{B}\|_{p,p}^p$$

for every $\mathbf{X} \in \mathbb{R}^{d \times m}$. Then, there is a subset $T \subseteq [m]$ with $|T| = m/16$ such that $\mathbf{S}$ is supported on T. For each $i' \in [s]$, let $T_{i'} \subseteq T$ denote the rows of T whose rows in $\mathbf{A}$ with group number $i' \in [s]$, so $\sum_{i'=1}^s |T_{i'}| = |T|$. Then by averaging, there are at least $(3/4)s$ groups $i' \in [s]$ such that $|T_{i'}| \leq \varepsilon^{-p}/2$. Thus, we may assume without loss of generality that $|T_{i'}| = \varepsilon^{-p}$ for the first $(1/4)s$ groups, $|T_{i'}| = \varepsilon^{-p}/2$ for the last $(3/4)s$ groups, and $|T| = (5/8)m$.

Let $W := \sum_{i=1}^m |\mathbf{S}_{i,i}|^p$ denote the total weight mass of $\mathbf{S}$. Note then that by querying $\mathbf{X} = 0$, we must have that

$$\|\mathbf{SB}\|_{p,p}^p = W = (1 \pm \varepsilon)\|\mathbf{B}\|_{p,p}^p = \left(1 \pm \frac{\varepsilon}{12 C_{p/2}^p}\right) m.$$

Let W_1 denote the sum of $|\mathbf{S}_{i,i}|^p$ on the first $(1/4)s$ groups, and let W_2 denote the sum of $|\mathbf{S}_{i,i}|^p$ on the last $(3/4)s$ groups. We will assume that $W_1 \leq m/4$, since the case of $W_1 \geq m/4$ is symmetric.

We now construct a query $\mathbf{X} \in \mathbb{R}^{d \times m}$ with the j-th column given by

$$\mathbf{X}\mathbf{e}_j = \begin{cases} \varepsilon \cdot \mathbf{e}_j^\top \mathbf{A} & j \in T \\ 0 & j \notin T \end{cases}$$

Note then that for each $i, j \in [m]$,

$$\mathbf{e}_i^\top \mathbf{A}\mathbf{X}\mathbf{e}_j = \begin{cases} \varepsilon d & \mathbf{e}_i^\top \mathbf{A} = \mathbf{e}_j^\top \mathbf{A}, j \in T \\ \varepsilon C_{p/2}\sqrt{d} & \mathbf{e}_i^\top \mathbf{A} \neq \mathbf{e}_j^\top \mathbf{A}, j \in T \\ 0 & j \notin T \end{cases}$$

Let $i \in [m]$ and let $i' \in [s]$ be its group number. Then the cost of row i if $i \in T$ is

$$\begin{aligned} \|\mathbf{e}_i^\top \mathbf{A}\mathbf{X} - \mathbf{e}_i^\top \mathbf{B}\|_p^p = \sum_{j=1}^m \left|\mathbf{e}_i^\top \mathbf{A}\mathbf{X}\mathbf{e}_j - \mathbf{B}(i,j)\right|^p &= \underbrace{(1-\varepsilon)^p d^p}_{i=j} + (|T_{i'}| - 1) \cdot \underbrace{\varepsilon^p d^p}_{\mathbf{e}_i^\top \mathbf{A} = \mathbf{e}_j^\top \mathbf{A}} + (|T| - |T_{i'}|) \cdot \underbrace{\varepsilon^p C_{p/2}^p d^{p/2}}_{\mathbf{e}_i^\top \mathbf{A} \neq \mathbf{e}_j^\top \mathbf{A}} \\ &= (1 - p\varepsilon + |T_{i'}|\varepsilon^p + (5/8)C_{p/2}^p + o(\varepsilon))d^p \end{aligned}$$

while the cost of row $i \in [m]$ if $i \notin T$ is

$$\begin{aligned}\|\mathbf{e}_i^\top \mathbf{A}\mathbf{X} - \mathbf{e}_i^\top \mathbf{B}\|_p^p = \sum_{j=1}^m \left|\mathbf{e}_i^\top \mathbf{A}\mathbf{X}\mathbf{e}_j - \mathbf{B}(i,j)\right|^p &= \underbrace{d^p}_{i=j} + |T_{i'}| \cdot \underbrace{\varepsilon^p d^p}_{\mathbf{e}_i^\top \mathbf{A} = \mathbf{e}_j^\top \mathbf{A}} + (|T| - |T_{i'}|) \cdot \underbrace{\varepsilon^p C_{p/2}^p d^{p/2}}_{\mathbf{e}_i^\top \mathbf{A} \neq \mathbf{e}_j^\top \mathbf{A}} \\ &= (1 + |T_i'|\varepsilon^p + (5/8)C_{p/2}^p + o(\varepsilon))d^p.\end{aligned}$$

Let

$$\begin{aligned}c_1 &= (1 - p\varepsilon + 1 + (5/8)C_{p/2}^p + o(\varepsilon))d^p \\ c_2 &= (1 - p\varepsilon + (1/2) + (5/8)C_{p/2}^p + o(\varepsilon))d^p \\ c_3 &= (1 + (1/2) + (5/8)C_{p/2}^p + o(\varepsilon))d^p\end{aligned}$$

Then, the total true cost is at least

$$\begin{aligned}\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{p,p}^p &= \frac{m}{4}c_1 + \frac{3m}{8}c_2 + \frac{3m}{8}c_3 \\ &= \frac{m}{4}c_1 + \frac{3m}{4}c_2 + \frac{3m}{8}(c_3 - c_2) \\ &\geq \frac{m}{4}c_1 + \frac{3m}{4}c_2 + \frac{3m}{4} \cdot (\varepsilon - o(\varepsilon))d^p\end{aligned}$$

while the strong coreset estimate is at most

$$\begin{aligned}\|\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_{p,p}^p &= W_1 c_1 + W_2 c_2 \\ &= W_1(c_1 - c_2) + (W_1 + W_2)c_2 \\ &\leq \frac{m}{4}(c_1 - c_2) + \left(1 + \frac{\varepsilon}{12C_{p/2}^p}\right)mc_2 \\ &\leq \frac{m}{4}c_1 + \frac{3m}{4}c_2 + \frac{\varepsilon}{4}md^p.\end{aligned}$$

Furthermore,

$$\frac{\varepsilon}{12C_{p/2}^p}\left(\frac{m}{4}c_1 + \frac{3m}{4}c_2 + \frac{\varepsilon}{4}md^p\right) \leq \frac{\varepsilon}{4}md^p$$

so $(1 + \frac{\varepsilon}{12C_{p/2}^p})\|\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_{p,p}^p < \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{p,p}^p$ and thus $\mathbf{S}$ fails to be a strong coreset. Rescaling ε by constant factors gives the desired result. □

E.2. Weak coresets

Proof of Theorem 5.2. Our hard instance is identical to the one of Theorem 5.1, except that each group has $\varepsilon^{1-p}/2C_{p/2}^p$ copies rather than ε^{-p} copies.

Note that if $\mathbf{S}$ does not sample some row $i \in [m]$, then the i-th column of $\mathbf{S}\mathbf{B}$ is all zeros, so the solution obtained by the weak coreset is $\mathbf{X}\mathbf{e}_i = 0$, which has objective function value $\|\mathbf{B}\mathbf{e}_i\|_p^p = d^p$. On the other hand, the optimal value is at most $(1 - \varepsilon)^p d^p$ since we can set $\mathbf{X}\mathbf{e}_i = \varepsilon\mathbf{A}^\top \mathbf{e}_i$ so that

$$\begin{aligned}\|(\mathbf{A}\mathbf{X} - \mathbf{B})\mathbf{e}_i\|_p^p &\leq (1 - \varepsilon)^p d^p + \frac{\varepsilon^{1-p}}{2C_{p/2}^p} \cdot \varepsilon^p d^p + d^{p/2}\frac{\varepsilon^{1-p}}{2C_{p/2}^p} \cdot C_{p/2}^p \varepsilon^p d^{p/2} \\ &\leq (1 - \varepsilon)^p d^p + \frac{\varepsilon}{2} \cdot d^p + \frac{\varepsilon}{2} \cdot d^p \\ &\leq ((1 - \varepsilon)^p + \varepsilon)d^p\end{aligned}$$

which is a $(1 + \varepsilon)$ factor smaller for all ε sufficiently small. Thus, if $\mathsf{nnz}(\mathbf{S}) \leq m/2$, then the solution $\mathbf{X}$ that minimizes $\|\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_{p,p}^p$ must be at least an additive $\varepsilon d^p \cdot m/2$ more expensive than the optimal solution, and thus it fails to be a $(1 + \varepsilon/2)$-optimal solution. □

E.3. Spanning coresets

We generalize an argument of Section 4 of (Deshpande & Vempala, 2006).

Lemma E.2. *Let $1 \leq p < \infty$ and*

$$c_p = \begin{cases} 1/6 & p \leq 2 \\ 1/(6 \cdot 5^{p/2-1}) & p > 2 \end{cases}$$

Then, there is a matrix $\mathbf{A} \in \mathbb{R}^{n\times(n+1)}$ such that for every $\varepsilon \geq 1/n$ and any subset of $s \leq c_p\varepsilon^{-1}$ rows, any rank 1 subspace F' spanned by the s rows must have

$$\|\mathbf{A}\mathbf{P}_{F'} - \mathbf{A}\|_{p,2}^p > (1+\varepsilon) \min_{\operatorname{rank}(F)\leq 1} \|\mathbf{A}\mathbf{P}_F - \mathbf{A}\|_{p,2}^p.$$

Proof. Let $n \leq \varepsilon^{-1}$ and let $\mathbf{A}$ be the $n \times (n+1)$ matrix given by $[R \cdot \mathbf{1}_n, \mathbf{I}_n]$ for some large enough $R > 0$. That is, $\mathbf{A}$ is R along the first column and the $n \times n$ identity for the last n columns. Note that the optimal value is upper bounded by

$$n((1-\varepsilon)^2 + \varepsilon^2 \cdot (n-1))^{p/2} = n(1 - 2\varepsilon + \varepsilon^2 n)^{p/2} = n(1-\varepsilon)^{p/2}.$$

Let $\mathbf{x} \in \mathbb{R}^s$ be the coefficients of a linear combination of s rows of $\mathbf{A}$. We may assume the coefficients are non-negative, since making the coefficients negative can only increase the cost. Note first that $1/2 \leq \|\mathbf{x}\|_1 \leq 3/2$ since otherwise

$$n \cdot |R - R\|\mathbf{x}\|_1|^p \geq n \cdot R/2$$

which cannot be $(1+\varepsilon)$-approximately optimal for $R \geq 2$.

The cost of the i-th row is $\left((1-\mathbf{x}_i)^2 + \|\mathbf{x}\|_2^2 - \mathbf{x}_i^2\right)^{p/2} = \left(1 - 2\mathbf{x}_i + \|\mathbf{x}\|_2^2\right)^{p/2}$. If $\|\mathbf{x}\|_2 \geq 2$, then

$$\left(1 - 2\mathbf{x}_i + \|\mathbf{x}\|_2^2\right)^{p/2} \geq (1 - 2\|\mathbf{x}\|_2 + \|\mathbf{x}\|_2^2)^{p/2} = (\|\mathbf{x}\|_2 - 1)^p \geq 1$$

so this cannot produce a $(1+\varepsilon)$-approximately optimal solution. Thus, assume $\|\mathbf{x}\|_2 \leq 2$. Then,

$$\left(1 - 2\mathbf{x}_i + \|\mathbf{x}\|_2^2\right)^{p/2} = \left(1 + \|\mathbf{x}\|_2^2\right)^{p/2}\left(1 - \frac{2}{1+\|\mathbf{x}\|_2^2}\mathbf{x}_i\right)^{p/2} \geq \left(1 + \|\mathbf{x}\|_2^2\right)^{p/2}\left(1 - \frac{p}{1+\|\mathbf{x}\|_2^2}\mathbf{x}_i\right)$$

so summing over the rows gives a cost of

$$\begin{aligned}
\left(1 + \|\mathbf{x}\|_2^2\right)^{p/2}\left(n - \frac{p}{1+\|\mathbf{x}\|_2^2}\|\mathbf{x}\|_1\right) &= \left(1 + \|\mathbf{x}\|_2^2\right)^{p/2} n - p(1+\|\mathbf{x}\|_2^2)^{p/2-1}\|\mathbf{x}\|_1 \\
&\geq \left(1 + \|\mathbf{x}\|_1^2/s\right)^{p/2} n - p(1+\|\mathbf{x}\|_2^2)^{p/2-1}\|\mathbf{x}\|_1 \qquad \text{since } 1/2 \leq \|\mathbf{x}\|_1 \leq 3/2 \\
&\geq (1 + 1/2s)^{p/2} n - (3/2)p(1+\|\mathbf{x}\|_2^2)^{p/2-1} \\
&\geq (1 + p/4s)n - (3/2)p(1+\|\mathbf{x}\|_2^2)^{p/2-1} \\
&\geq \begin{cases} (1+p/4s)n - (3/2)p & p \leq 2 \\ (1+p/4s)n - (3/2)p \cdot 5^{p/2-1} & p > 2 \end{cases}
\end{aligned}$$

Thus, this fails to be a $(1+\varepsilon)$-approximately optimal solution for

$$(p/4s)n \geq \begin{cases} (3/2)p & p \leq 2 \\ (3/2)p \cdot 5^{p/2-1} & p > 2 \end{cases}$$

that is,

$$s \leq \begin{cases} n/6 & p \leq 2 \\ n/(6 \cdot 5^{p/2-1}) & p > 2 \end{cases}.$$

□

We now extend Lemma E.2 to a general rank k lower bound.

Proof of Theorem 5.3. Let $n = \varepsilon^{-1}$ and let $\mathbf{B}$ be a $kn \times k(n+1)$ block diagonal matrix with the $n \times (n+1)$ matrix construction $\mathbf{A} \in \mathbb{R}^{n\times(n+1)}$ of Lemma E.2 on the block diagonal. Consider any set S of s rows of $\mathbf{B}$, and let S_i denote the set of $|S_i| = s_i$ rows supported on the i-th block for each $i \in [k]$. Let F_i denote the optimal subspace spanned by the rows S_i on the ith block.

Let $T \subseteq [k]$ denote the set of $i \in [k]$ such that $s_i \leq c_p n$. If $i \in T$, then we by Lemma E.2 that

$$\|\mathbf{A}\mathbf{P}_{F_i} - \mathbf{A}\|_{p,2}^p > \left(1 + \frac{c_p}{s_i}\right) \min_{\operatorname{rank}(F)\leq k} \|\mathbf{A}\mathbf{P}_F - \mathbf{A}\|_{p,2}^p$$

Then, the additive error from these rows is bounded below by

$$\begin{aligned}
\sum_{i\in T} \frac{c_p}{s_i} \min_{\operatorname{rank}(F)\leq k} \|\mathbf{A}\mathbf{P}_F - \mathbf{A}\|_{p,2}^p &\geq |T| \cdot \frac{c_p|T|}{\sum_{i\in[k]:s_i\leq c_p n} s_i} \min_{\operatorname{rank}(F)\leq k} \|\mathbf{A}\mathbf{P}_F - \mathbf{A}\|_{p,2}^p && \text{AM-HM} \\
&\geq |T| \cdot \frac{c_p|T|}{s} \min_{\operatorname{rank}(F)\leq k} \|\mathbf{A}\mathbf{P}_F - \mathbf{A}\|_{p,2}^p \\
&\geq \frac{c_p|T|^2}{ks} \min_{\operatorname{rank}(F)\leq k} \|\mathbf{B}\mathbf{P}_F - \mathbf{B}\|_{p,2}^p
\end{aligned}$$

Note that $|T| \geq k/2$ by averaging, so

$$\frac{c_p|T|^2}{ks} \geq \frac{c_p k}{4s} \geq \varepsilon$$

which proves the theorem. □

F. Experimental evaluation

We show that empirically, we indeed see that the trade-off between the number of uniform samples and the approximation quality is independent of the dimension m in the setting of Euclidean power means. We do this by plotting the sample size against the resulting relative error for $m \in \{100, 500\}$, where an m-dimensional dataset is constructed by sampling m random features from the MNIST dataset. The results are shown in Figure 1 and the experiment code is provided in Section F.1.

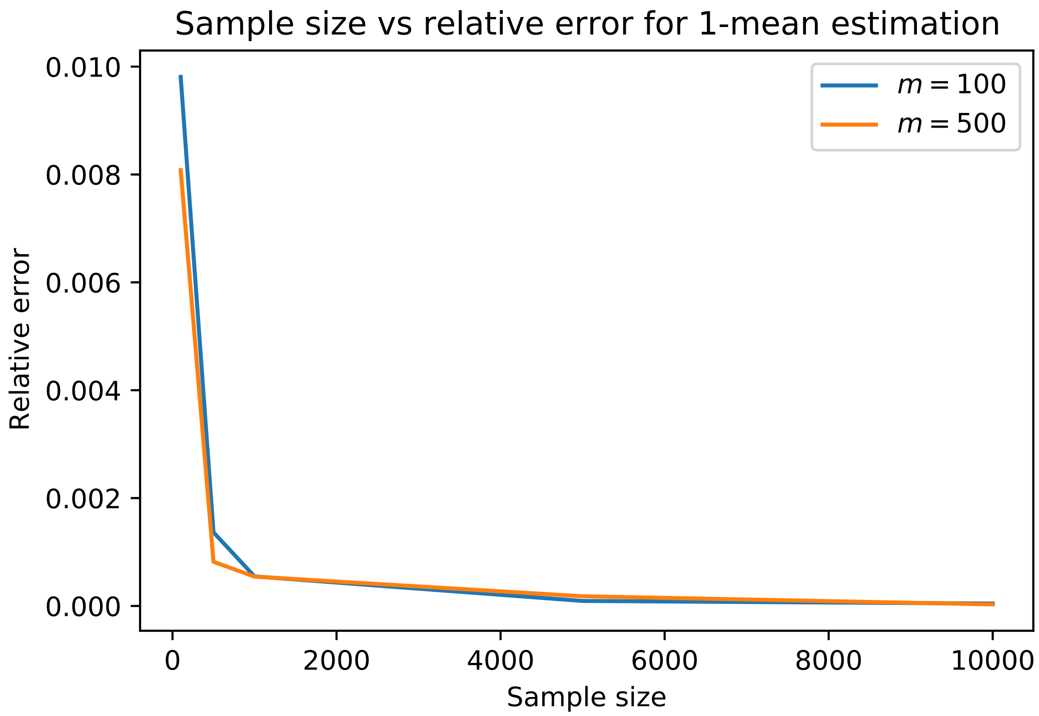

Figure 1. Sample size vs relative error for 1-mean estimation

F.1. Experiment code

We provide the code snippet for the experimental evaluation below.

```
from keras.datasets import mnist
import numpy as np
import matplotlib.pyplot as plt
import tensorflow as tf

np.random.seed(2024)

(train_X, train_y), (test_X, test_y) = mnist.load_data()
train_X = train_X.reshape(len(train_X), -1)
train_X = train_X / np.max(train_X)
n, d = train_X.shape

def power_mean_loss(train_ds, x, p=1):
    x = np.expand_dims(x, axis=0)
    x = np.repeat(x, repeats=n, axis=0)
    e = train_ds - x
    e = np.linalg.norm(e, axis=-1)
    e = np.power(e, p)
    return np.sum(e) / n

def run(train_ds, max_iter=200, p=1):
    n, d = train_ds.shape
    x0 = np.zeros(d)
    x = tf.Variable(initial_value=x0)
    opt = tf.keras.optimizers.Adam(learning_rate=0.5)
    x.assign(x0)
    def power_mean_loss_tf():
        e = train_ds - x
        e = tf.norm(e, axis=-1)
        e = tf.math.pow(e, p)
        return tf.reduce_sum(e) / n
    losses = []
    while opt.iterations < max_iter:
        opt.minimize(power_mean_loss_tf, var_list=[x])
        loss = power_mean_loss_tf().numpy()
        if np.isnan(loss):
            print(x.numpy())
        losses.append(loss)
    return x.numpy(), losses

n, d = train_X.shape
sample_sizes = [100, 500, 1000, 5000, 10000]
for m in [100, 500]:
    cols = np.random.choice(d, m)
    train_m = train_X[:, cols]
    x, losses = run(train_m)
    OPT = losses[-1]
    estimates = []
    for sample_size in sample_sizes:
        train_sample = np.random.choice(n, sample_size)
        train_sample = train_m[train_sample, :]
        x, losses = run(train_sample)
        estimates.append(power_mean_loss(train_m, x))
    relative_errors = [(e / OPT) - 1 for e in estimates]
    print(’relative errors’, relative_errors)
```