Sharper Bounds for ℓ_p Sensitivity Sampling

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

Abstract

In large scale machine learning, *random sampling* is a popular way to approximate datasets by a small representative subset of examples. In particular, *sensitivity sampling* is an intensely studied technique which provides provable guarantees on the quality of approximation, while reducing the number of examples to the product of the *VC dimension* d and the *total sensitivity* $\mathfrak{S}$ in remarkably general settings. However, guarantees going beyond this general bound of $\mathfrak{S}d$ are known in perhaps only one setting, for ℓ_2 *subspace embeddings*, despite intense study of sensitivity sampling in prior work. In this work, we show the first bounds for sensitivity sampling for ℓ_p subspace embeddings for $p \neq 2$ that improve over the general $\mathfrak{S}d$ bound, achieving a bound of roughly $\mathfrak{S}^{2/p}$ for $1 \leq p < 2$ and $\mathfrak{S}^{2-2/p}$ for $2 < p < \infty$. For $1 \leq p < 2$, we show that this bound is tight, in the sense that there exist matrices for which $\mathfrak{S}^{2/p}$ samples is necessary. Furthermore, our techniques yield further new results in the study of sampling algorithms, showing that the *root leverage score sampling* algorithm achieves a bound of roughly d for $1 \leq p < 2$, and that a combination of leverage score and sensitivity sampling achieves an improved bound of roughly $d^{2/p}\mathfrak{S}^{2-4/p}$ for $2 < p < \infty$. Our sensitivity sampling results yield the best known sample complexity for a wide class of structured matrices that have small ℓ_p sensitivity.

1. Introduction

In typical large scale machine learning problems, one encounters a dataset represented by an $n \times d$ matrix $\mathbf{A}$, consisting of d features and a large number $n \gg d$ of training examples. While an extremely large n can cause various data analytic tasks to be intractable, it is often the case that not all n training examples are necessary, and *random sampling* can effectively reduce the number of training examples while approximately preserving the information necessary for downstream prediction tasks.

Uniform sampling is perhaps the simplest instance of this idea that is often used in practice. However, uniform sampling can lead to significant information loss when there are a small number of important training examples that must be kept under sampling. Thus, recent work, both in theory (Langberg & Schulman, 2010; Feldman & Langberg, 2011) and in practice (Katharopoulos & Fleuret, 2017; Johnson & Guestrin, 2018), has focused on *importance sampling* methods that sample more important examples with higher probability.

In this work, we focus on the use of importance sampling techniques for approximating objective functions for empirical risk minimization problems. Consider an objective function $f : X \to \mathbb{R}_{\geq 0}$ of the form of a sum along the coordinates, i.e.,

$$f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x}),$$

where X is some domain set and $f_i : X \to \mathbb{R}_{\geq 0}$ are non-negative loss functions for $i \in [n] = \{1, 2, \ldots, n\}$. Then, we seek algorithms that sample a subset $S \subseteq [n]$ and weights $\mathbf{w}_i$ for $i \in S$ such that, with high probability, the function

$$\tilde{f}(\mathbf{x}) \coloneqq \sum_{i \in S} \mathbf{w}_i f_i(\mathbf{x})$$

satisfies

$$\text{for every } \mathbf{x} \in X, \qquad \tilde{f}(\mathbf{x}) = (1 \pm \varepsilon) f(\mathbf{x}) \qquad (1)$$

for some accuracy parameter $0 < \varepsilon < 1$. We will also refer to the ε parameter as the *sampling error*. Note then that if $|S| \ll n$, then $\tilde{f}$ can be used as a surrogate objective function in downstream applications that can be processed much more efficiently than f itself.

Sensitivity Sampling. The *sensitivity sampling* framework, introduced by (Langberg & Schulman, 2010; Feldman & Langberg, 2011), provides one method of achieving

[1]Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, US. Correspondence to: Taisuke Yasuda <taisukey@cs.cmu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

guarantees of the form of (1). In this framework, one first computes[1] *sensitivity scores* of each coordinate $i \in [n]$:

$$\boldsymbol{\sigma}_i \coloneqq \sup_{\mathbf{x} \in X} \frac{f_i(\mathbf{x})}{\sum_{j=1}^n f_j(\mathbf{x})}.^2$$

Then, each $i \in [n]$ is independently sampled with probability p_i proportional to $\boldsymbol{\sigma}_i$, and the sampled row is assigned a weight of $1/p_i$. It is easy to see that this preserves the objective function $f(\mathbf{x})$ in expectation for every $\mathbf{x} \in X$. Furthermore, it can be shown that sampling $\tilde{O}(\varepsilon^{-2}\mathfrak{S}d)$ rows provides a $(1 \pm \varepsilon)$-factor approximation to the objective function (Braverman et al., 2016; Feldman et al., 2020), where $\mathfrak{S} = \sum_{i=1}^n \boldsymbol{\sigma}_i$ is known as the *total sensitivity* and d is the VC dimension of an associated set system.

Sampling Algorithms for ℓ_p Linear Regression. We now turn to sampling algorithms for ℓ_p *linear regression*, which is the main problem of study of this work. Consider an input matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with n rows $\mathbf{a}_i \in \mathbb{R}^d$, a label vector $\mathbf{b} \in \mathbb{R}^n$, and $1 \leq p < \infty$. We then seek to minimize

$$f(\mathbf{x}) = \sum_{i=1}^n |\langle \mathbf{a}_i, \mathbf{x} \rangle - \mathbf{b}_i|^p = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p.$$

Note that this problem is in the form of an empirical risk minimization problem as discussed previously, with $f_i(\mathbf{x}) = |\langle \mathbf{a}_i, \mathbf{x} \rangle - \mathbf{b}_i|^p$ and $X = \mathbb{R}^d$. Furthermore, in the case of ℓ_p regression, we may use the scale invariance of the ℓ_p norm to fold the weights $\mathbf{w}_i$ into the objective f_i, so we can write the sampling procedure as a diagonal map:

Definition 1.1 (ℓ_p Sampling Matrix)**.** Let $1 \leq p < \infty$. 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

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

Our goal then is to compute probabilities $\{p_i\}_{i=1}^n$ such that the associated ℓ_p random sampling matrix $\mathbf{S}$ satisfies

$$\text{for every } \mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p = (1 \pm \varepsilon)\|\mathbf{S}\mathbf{A}\mathbf{x} - \mathbf{S}\mathbf{b}\|_p^p$$

Note that we may in fact also assume that $\mathbf{b} = 0$, since we can append $\mathbf{b}$ to be one of the columns of $\mathbf{A}$. Thus, our problem is to compute probabilities $\{p_i\}_{i=1}^n$ satisfying

$$\text{for every } \mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p, \quad (2)$$

which is known as an ℓ_p *subspace embedding* of $\mathbf{A}$.

[1]While computing sensitivity scores exactly is often inefficient, efficiently computable approximations exist and suffice in many cases.

[2]We define the fraction to be 0 when the denominator is 0.

We introduce the following notation for sensitivity sampling when specifically applied to ℓ_p subspace embeddings:

Definition 1.2 (ℓ_p sensitivities)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $p \geq 1$. Then, for each $i \in [n]$, we define the *ith ℓ_p sensitivity* to be

$$\boldsymbol{\sigma}_i^p(\mathbf{A}) \coloneqq \sup_{\mathbf{x} \in \mathbb{R}^d, \mathbf{A}\mathbf{x} \neq 0} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p}.^3$$

and the *total ℓ_p sensitivity* to be $\mathfrak{S}^p(\mathbf{A}) \coloneqq \sum_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A})$.

Note that the calculation of ℓ_p sensitivities can be formulated as an ℓ_p regression problem, and can be computed efficiently using recent developments in algorithms for ℓ_p regression. Indeed, it is easy to see that

$$\frac{1}{\boldsymbol{\sigma}_i^p(\mathbf{A})} = \min_{[\mathbf{A}\mathbf{x}](i)=1} \|\mathbf{A}\mathbf{x}\|_p^p,$$

which can be efficiently approximated to high precision in nearly matrix multiplication time (Adil et al., 2019a;b; Adil & Sachdeva, 2020).

For $p = 2$, the ℓ_p sensitivities are exactly equal to the *leverage scores*.

Definition 1.3 (Leverage scores)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then, for each $i \in [n]$, we define the *ith leverage score* to be $\boldsymbol{\tau}_i(\mathbf{A}) \coloneqq \boldsymbol{\sigma}_i^2(\mathbf{A})$.

For ℓ_p subspace embeddings, it is known that the total sensitivity is at most d for $p \leq 2$ and at most $d^{p/2}$ for $p > 2$, where d is the dimension. Thus, sensitivity sampling applies in this setting, and has indeed been successfully applied in prior work (Braverman et al., 2020; 2021). There are also other sampling schemes, based on Lewis weights, which we discuss more below.

For $p = 2$, which corresponds to the standard least squares regression problem, it has long been known that sensitivity sampling, known as *leverage score sampling* in this case, yields ℓ_2 subspace embeddings (2) with nearly optimal sample complexity of[4] $\tilde{O}(\varepsilon^{-2}d)$ (Mahoney, 2011). However, it is also known that the total sensitivity $\mathfrak{S}$ is exactly d in this setting, and the dimension is also d. Thus, in the natural setting of ℓ_2 subspace embeddings, the bound of $\mathfrak{S}d = d^2$ for sensitivity sampling is quadratically loose in the analysis. However, to the best of our knowledge, no bound which improves over the general bound of $\mathfrak{S}d$ specifically for sensitivity sampling is known in any other setting. We thus arrive at the central question of our work:

Question 1.4.

How many samples are necessary for sensitivity sampling to output an ℓ_p subspace embedding (2)*?*

[3]We write $[\mathbf{A}\mathbf{x}](i)$ for the ith entry of the vector $\mathbf{A}\mathbf{x} \in \mathbb{R}^n$.

[4]We write $\tilde{O}(f)$ to denote $f \operatorname{poly} \log f$.

1.1. Our Contributions

In this work, we make progress towards resolving Question 1.4 by obtaining the first bounds for sensitivity sampling that go beyond the general $\mathfrak{S}d$ bound:

Theorem 1.5 (Informal Restatement of Theorem D.2 and Theorem D.4)**.** *Let $1 \le p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\alpha > 0$ and let $q_i = \min\{1, 1/n + \boldsymbol{\sigma}_i^p(\mathbf{A})/\alpha\}$ for $i \in [n]$. Then, there is an α such that the random ℓ_p sampling matrix $\mathbf{S}$ with sampling probabilities $\{q_i\}_{i=1}^n$ satisfies* (2) *and samples at most m rows with probability at least $1 - 1/\operatorname{poly}(n)$, with*

$$m = \begin{cases} \dfrac{\mathfrak{S}^p(\mathbf{A})^{2/p}}{\varepsilon^2} \operatorname{poly}\log n & 1 \le p < 2 \\ \dfrac{\mathfrak{S}^p(\mathbf{A})^{2-2/p}}{\varepsilon^2} \operatorname{poly}\log n & 2 < p < \infty \end{cases}$$

Sample Complexity Bounds for ℓ_p Sensitivity Sampling

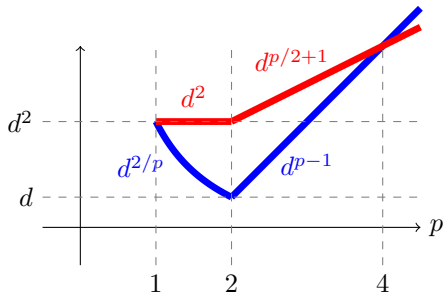

Figure 1. Dependence on d for sensitivity sampling when $\mathfrak{S}^p(\mathbf{A}) = d^{1\vee(p/2)}$ (i.e., worst-case). The y axis indicates the exponent on d in the sample complexity. The red line indicates prior bounds, blue indicates our results.

For $1 \le p < 2$, we show that our bound of Theorem 1.5 is in fact tight, in the sense that there exist matrices such that, up to logarithmic factors, $\mathfrak{S}^p(\mathbf{A})^{2/p}$ samples are necessary to obtain ℓ_p subspace embeddings. We show this by showing that random matrices have a total sensitivity of at most $\mathfrak{S}^p(\mathbf{A}) = \tilde{O}(d^{p/2})$, which implies the claim since d rows are necessary to even maintain the rank.

Theorem 1.6. *Let $1 \le p < 2$. Let $n = d^{O(p)}$ be large enough, and let $\mathbf{A}$ be a random $n \times d$ standard Gaussian matrix. Then, with probability at least $2/3$, $\mathfrak{S}^p(\mathbf{A}) \le O(d \log d)^{p/2}$.*

Furthermore, we show the first bounds showing that the total ℓ_p sensitivity cannot be any smaller than the construction in Theorem 1.6, up to logarithmic factors:

Theorem 1.7. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $1 \le p < \infty$. Then,*

$$\mathfrak{S}^p(\mathbf{A}) \ge \begin{cases} d/2 & p > 2 \\ d^{p/2}/2 & p < 2 \end{cases}$$

Comparison to ℓ_p Lewis Weight Sampling. We now compare our results to the well-known ℓ_p *Lewis weight sampling* technique, which yields nearly optimal bounds for ℓ_p subspace embeddings for matrices with worst-case total ℓ_p sensitivity (Cohen & Peng, 2015; Woodruff & Yasuda, 2023). More specifically, it is known that the total ℓ_p sensitivity is at most d for $p < 2$ and at most $d^{p/2}$ for $p > 2$ for all matrices $\mathbf{A} \in \mathbb{R}^{n\times d}$, and ℓ_p Lewis weight sampling provides a way to reduce the number of rows to at most roughly $d^{1\vee(p/2)}$ rows[5] for all matrices $\mathbf{A}$:

Theorem 1.8 (ℓ_p Lewis weight sampling)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p > 0$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix with sampling probabilities $\{q_i\}_{i=1}^n$ proportional to the ℓ_p Lewis weights. Then, $\mathbf{S}$ samples $O(\varepsilon^{-2} d^{1\vee(p/2)} (\log d)^2 \log(d/\varepsilon))$ rows and satisfies* (2)*.*

In particular, ℓ_p Lewis weight sampling yields better sampling complexity bounds than our Theorem 1.5 when $p < 2$ or when the ℓ_p total sensitivity $\mathfrak{S}^p(\mathbf{A})$ is close to $d^{p/2}$. However, our results on ℓ_p sensitivity sampling have a number of advantages over the existing ℓ_p Lewis weight sampling results. First, Lewis weight sampling, to the best of our knowledge, is not known to admit bounds better than $d^{p/2}$ for matrices with very small total sensitivity $\mathfrak{S}^p(\mathbf{A}) \ll d^{p/2}$ for $p > 2$. Thus, for matrices with substantially smaller total sensitivity, ℓ_p sensitivity sampling yields far improved bounds. We illustrate a number of such applications below. Second, sensitivity sampling generalizes to sampling problems beyond ℓ_p subspace embeddings and has been applied to subspace embeddings for general M-estimators and Orlicz norms (Clarkson & Woodruff, 2015a;b; Tukan et al., 2020; Musco et al., 2022), logistic regression (Munteanu et al., 2018) and other generalized linear models (Munteanu et al., 2022), as well as general shape fitting problems including clustering and subspace approximation (Huang & Vishnoi, 2020) and projective clustering (Varadarajan & Xiao, 2012). Thus, our techniques may be useful for improving the analyses of a broad range of sampling problems.

Other Sampling Algorithms. In addition to ℓ_p sensitivity sampling, our techniques yield other new results on sampling algorithms for ℓ_p subspace embeddings.

Our next result is a new analysis of *root leverage score sampling*, which is a popular method for efficiently computing upper bounds to sensitivity scores for loss functions of at most quadratic growth, including ℓ_p losses for $1 \le p < 2$, the Huber loss, and the logistic loss (Clarkson & Woodruff, 2015b; Munteanu et al., 2018; Ghadiri et al., 2021). In this technique, the p_i are set proportionally to the *square root* of the ℓ_2 leverage scores (Definition 1.3). While the sum of the root leverage scores can be as large as $\sqrt{nd}$, this sampling procedure can be recursively applied for $O(\log\log n)$

[5] We write $a \vee b$ to denote the maximum of a and b.

iterations to reduce the sample complexity to $\mathrm{poly}(d)$.

As with sensitivity sampling, the only previously known analyses of root leverage score sampling proceed by a naïve union bound, which can only reduce the number of samples to d^2, which is loose by a d factor. Our techniques for analyzing sensitivity sampling can be modified to show that root leverage score sampling in fact leads to a *nearly optimal* number of samples:

Theorem 1.9 (Informal Restatement of Theorem D.6 and Theorem D.8)**.** *Let $1 \leq p < 2$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\alpha > 0$ and let $p_i = \min\{1, \boldsymbol{\tau}_i(\mathbf{A})^{p/2}/\alpha\}$ for $i \in [n]$. Then, there is an α such that the random ℓ_p sampling matrix $\mathbf{S}$ with sampling probabilities $\{p_i\}_{i=1}^n$ satisfies* (2) *and samples at most m rows with probability at least $1 - 1/\mathrm{poly}(n)$, with*

$$m = \frac{n^{1-p/2}d^{p/2}}{\varepsilon^2}\,\mathrm{poly}\log n.$$

Recursively applying this result gives a matrix $\mathbf{S}$ satisfying (2) *with at most m rows, for*

$$m = \frac{d}{\varepsilon^{4/p}}\,\mathrm{poly}\log n.$$

Our analysis of root leverage score sampling provides a promising direction towards resolving the problem of designing nearly optimal sampling algorithms for preserving subspaces under the *Huber loss*, which has been raised as an important question on sampling algorithms in a number of works (Adil & Sachdeva, 2020; Ghadiri et al., 2021; Musco et al., 2022) for its applications in Huber regression and fast algorithms for ℓ_p regression. Here, the best known upper bound is a sampling algorithm reducing the number of rows to roughly[6] $d^{4-2\sqrt{2}} \approx d^{1.172}$, whereas d is conjectured to be possible (Musco et al., 2022). Root leverage scores are known to upper bound the sensitivities for the Huber loss (Clarkson & Woodruff, 2015a;b; Ghadiri et al., 2021), so our Theorem 1.9 suggests that root leverage score sampling may yield a sampling algorithm reducing the number of samples to d for the Huber loss as well.

We additionally show that by incorporating ℓ_2 leverage scores into ℓ_p sensitivity sampling, we can obtain sampling guarantees that further improve over the guarantee of Theorem 1.5 for $p > 2$. We note that our proof of this result uses a recursive "flattening and sampling" scheme for this result, rather than a direct sampling result as in our earlier results.

Theorem 1.10 (Informal Restatement of Theorem D.10)**.** *Let $2 < p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Then, there is an efficient algorithm which computes a matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ satisfying* (2)*, for*

$$m = \frac{d^{2/p}\mathfrak{S}^p(\mathbf{A})^{1-2/p}}{\varepsilon^2}\,\mathrm{poly}\log n.$$

[6] A polylogarithmic dependence on n is necessary and sufficient here, but we omit this from our discussion for simplicity.

Although Theorem 1.10 does not specifically use sensitivity sampling, to the best of our knowledge, it is the best known sampling result for ℓ_p subspace embeddings with small ℓ_p total sensitivity $\mathfrak{S}^p(\mathbf{A})$.

Applications. We now show several examples in structured regression problems in which our new sensitivity sampling results give the best known sample complexity results for ℓ_p subspace embeddings for $p > 2$. We start by presenting a couple of lemmas which show that certain natural classes of matrices have total ℓ_p sensitivity $\ll d^{p/2}$.

The first result is a lemma extracted from a result of (Meyer et al., 2022) bounding the total ℓ_p sensitivity for a sparse perturbation of low rank matrices:

Lemma 1.11 (Sensitivity Bounds for Low Rank + Sparse Matrices (Meyer et al., 2022))**.** *Let $\mathbf{A} = \mathbf{K} + \mathbf{S} \in \mathbb{R}^{n\times d}$ for a rank k matrix $\mathbf{K}$ and an $\mathbf{S}$ with at most s nonzero entries per row. Let $1 \leq p < \infty$. Then, $\mathfrak{S}^p(\mathbf{A}) \leq d^s(k+s)^p$.*

We provide a self-contained proof in Appendix A.3.

In a second example, we show that "concatenated Vandermonde" matrices, which were studied in, e.g., (Avron et al., 2013), also have small total ℓ_p sensitivity. These matrices naturally arise as the result of applying a polynomial feature map to a matrix.

Definition 1.12 (Vandermonde matrix)**.** Given a vector $\mathbf{a} \in \mathbb{R}^n$, the degree q Vandermonde matrix $V^q(\mathbf{a}) \in \mathbb{R}^{n\times(q+1)}$ is defined entrywise as $V^q(\mathbf{a})_{i,j} = \mathbf{a}_i^j$ for $j = 0, 1, \ldots, q$.

Definition 1.13 (Polynomial feature map)**.** Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times k}$ and an integer q, we define the matrix $V^q(\mathbf{A}) \in \mathbb{R}^{n\times k(q+1)}$ to be the horizontal concatenation of the Vandermonde matrices $V^q(\mathbf{A}\mathbf{e}_1), V^q(\mathbf{A}\mathbf{e}_2), \ldots, V^q(\mathbf{A}\mathbf{e}_k)$.

We show the following result, proven in Appendix A.3.

Lemma 1.14 (Sensitivity Bounds for Matrices Under Polynomial Feature Maps)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times k}$ and let q be an integer. Let $1 \leq p < \infty$. Then, $\mathfrak{S}^p(V^q(\mathbf{A})) \leq (pq+1)^k$.*

This generalizes a result of (Meyer et al., 2022), which bounds the ℓ_p sensitivities of a single Vandermonde matrix.

In the low-sensitivity matrices of Lemma 1.11 and Lemma 1.14, it is in fact possible to apply Lewis weight sampling to obtain sampling bounds that match these sensitivity bounds, by using the *tensoring trick* (Meyer et al., 2022). However, when a tiny amount of noise is added to these matrices, then algebraic tricks such as tensoring break down, and the sensitivity bounds derived from Lewis weights increase substantially to $d^{p/2}$ for $p > 2$. On the other hand, sensitivity sampling itself is robust with respect to the addition of noise, as it depends only on norms rather than brittle quantities such as rank. Indeed, we have the following fact, which we prove in Appendix A.4:

Lemma 1.15. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a rank d matrix with minimum singular value $\sigma_{\min}$. Let $\mathbf{E} \in \mathbb{R}^{n\times d}$ be an arbitrary perturbation matrix with*

$$\|\mathbf{E}\|_2 \leq \frac{\sigma_{\min}}{2n^{1+1/p}}.$$

Then, $\mathfrak{S}^p(\mathbf{A}+\mathbf{E}) \leq 2^p(\mathfrak{S}^p(\mathbf{A})+1)$.

Thus, for small perturbations of structured matrices with small ℓ_p sensitivity as specified by Lemma 1.15, Theorem 1.5 and Theorem 1.10 give the tightest known bounds on the sample complexity for ℓ_p subspace embeddings. Such perturbations may arise due to roundoff error or finite precision on a computer, and no prior bounds beating Lewis weight sampling or the naïve $\mathfrak{S}d$ bound for sensitivity sampling were known for the applications above.

1.2. Other Related Work

The problem of designing sampling algorithms for ℓ_p subspace embeddings has a long and rich history, dating back to works in the functional analysis literature and culminating in the Lewis weight sampling result (Lewis, 1978; Schechtman, 1987; Bourgain et al., 1989; Talagrand, 1990; Ledoux & Talagrand, 1991; Talagrand, 1995; Schechtman & Zvavitch, 2001). More recently, the theoretical computer science community has studied this problem for its applications to ℓ_p linear regression and other empirical risk minimization problems. The early works of (Clarkson, 2005; Dasgupta et al., 2009) obtained sampling algorithms for ℓ_p regression based on sensitivity score upper bounds given by various constructions of ℓ_p *well-conditioned bases* for $\mathbf{A}$. The theory of Lewis weight sampling was brought to the theoretical computer science literature by (Cohen & Peng, 2015), and has been improved both in the computation of the weights (Lee, 2016; Fazel et al., 2022; Jambulapati et al., 2022) and sampling guarantees (Woodruff & Yasuda, 2022; 2023).

2. Sampling Error Bounds

We now give a more detailed discussion of our techniques and results. While our discussion in this section will present most of the major ideas necessary to prove our new results, all of the full proofs will be deferred to the appendix due to space considerations.

2.1. Prior Approaches

We start by describing the standard proof of the $\tilde{O}(\varepsilon^{-2}\mathfrak{S}d)$ sample complexity bound for sensitivity sampling (Schechtman, 1987). Using Bernstein bounds, it can be shown that sampling $O(\varepsilon^{-2}\mathfrak{S}\log\frac{1}{\delta})$ rows preserves the ℓ_p norm of a fixed vector $\mathbf{Ax}$ up to a $(1\pm\varepsilon)$ factor with probability at least $1-\delta$. One can then consider an ε-*net* N, which is a set of size roughly $1/\varepsilon^d$ such that any ℓ_p unit vector $\mathbf{Ax}$ is ε-close to some $\mathbf{Ax}' \in N$. By a union bound, the norm preservation guarantee holds simultaneously for every $\mathbf{Ax}' \in N$ with constant probability, if we set $\delta = \varepsilon^d$. Now for an arbitrary ℓ_p unit vector $\mathbf{Ax}$, the norm preservation guarantee holds for an ε-close point $\mathbf{Ax}' \in N$, which implies the norm preservation guarantee for $\mathbf{Ax}$ itself by a standard argument. Finally, scale invariance ensures that the same conclusion holds for all vectors $\mathbf{Ax}$, rather than just unit vectors.

To improve over this argument, the ℓ_p Lewis weight sampling technique was developed in a line of work from the functional analysis literature (Lewis, 1978; Bourgain et al., 1989; Talagrand, 1990; Ledoux & Talagrand, 1991; Talagrand, 1995), which incorporates *chaining arguments*. Chaining arguments are a way of improving ε-net arguments by using a *sequence* of ε-nets at different "scales", rather than using a single scale of ε, and using tighter bounds for net constructions (i.e., smaller cardinality nets) at larger scales (see, e.g., (Nelson, 2016) for a survey of chaining applications in computer science).

We now delve into a discussion of the overall strategy towards bounding the *sampling error* of our ℓ_p sampling results, which is a random variable Λ depending on $\mathbf{S}$ given by

$$\Lambda \coloneqq \sup_{\|\mathbf{Ax}\|_p=1} \left| \|\mathbf{SAx}\|_p^p - 1 \right|. \tag{3}$$

Note that if $\Lambda \leq \varepsilon$, then $\|\mathbf{SAx}\|_p^p = (1\pm\varepsilon)\|\mathbf{Ax}\|_p^p$ for every $\mathbf{x} \in \mathbb{R}^d$. In our discussions in this section, we will focus on bounding Λ in expectation, although our full proofs in the appendix will bound higher moments of Λ to obtain high probability bounds.

2.2. Generalized Chaining Bounds for ℓ_p Subspace Embeddings

Our main technical lemma towards bounding (3) is a generalization of the chaining argument framework for Lewis weight sampling, which bounds the sampling error of ℓ_p sampling algorithms by the leverage scores and ℓ_p sensitivities of the sampled matrix $\mathbf{SA}$, rather than by the Lewis weights of $\mathbf{SA}$.

First, we introduce our sampling bounds obtained by generalizing the chaining arguments of (Bourgain et al., 1989; Ledoux & Talagrand, 1991). In this result, we obtain the following bound on a certain Rademacher process, which can be interpreted as the sampling error of a uniform sampling process, as we describe in Section 2.4.

Lemma 2.1 (Rademacher Process Bound, Simplified)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $1 \leq p < \infty$. Let $\tau \geq \boldsymbol{\tau}_i(\mathbf{A})$ and $\sigma \geq$*

$\boldsymbol{\sigma}_i^p(\mathbf{A})$ *for every* $i \in [n]$. *Let*

$$E \coloneqq \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x}\|_p=1} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i |[\mathbf{A}\mathbf{x}](i)|^p \right|.$$

Then,

$$E \leq \begin{cases} O(\tau^{1/2})(\log n)^{3/2} & p < 2 \\ O(\tau^{1/2})(\sigma n)^{1/2-1/p}(\log n)^{3/2} & p > 2 \end{cases}$$

This result follows from a Gaussianization argument (Lemma D.1), followed by an application of Dudley's entropy integral theorem (Theorem C.2), and then bounding the entropy integral in Lemma C.6 and Lemma C.7.

As we discuss in Section 2.4, Lemma 2.1 will be a key ingredient in bounding the sampling error in our sensitivity sampling analysis.

2.3. Leverage Score and Sensitivity Bounds

In our final argument, Lemma 2.1 will be applied with the matrix $\mathbf{A}$ set to be (a modified version of) the sampled matrix $\mathbf{SA}$. Thus, we require a bound on the leverage scores and sensitivities τ and σ of $\mathbf{SA}$. In fact, σ is naturally bounded by the ℓ_p sensitivity sampling algorithm: indeed, if we a priori assume that $\|\mathbf{SAx}\|_p^p \geq (1/2)\|\mathbf{Ax}\|_p^p$ for every $\mathbf{x} \in \mathbb{R}^d$, then $\boldsymbol{\sigma}_i^p(\mathbf{SA})$ is at most

$$\sup_{\mathbf{SAx}\neq 0} \frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{SAx}\|_p^p} \leq 2 \sup_{\mathbf{Ax}\neq 0} \frac{1}{p_i} \frac{|[\mathbf{Ax}](i)|^p}{\|\mathbf{Ax}\|_p^p} \leq \frac{2\boldsymbol{\sigma}_i^p(\mathbf{A})}{p_i} \tag{4}$$

which is at most 2α if $p_i \geq \boldsymbol{\sigma}_i^p(\mathbf{A})/\alpha$. While this is only an informal argument, this intuition can be formalized, as we discuss later in this section.

Next, we require a bound on the leverage scores τ of $\mathbf{SA}$, which is more challenging, as sensitivity sampling does not directly bound this quantity. To address this problem, we show how to bound the leverage scores of an arbitrary matrix by the ℓ_p sensitivities of the matrix. In particular, we show that for $q \geq p$, the largest ℓ_q sensitivity bounds the largest ℓ_p sensitivity.

Lemma 2.2 (Monotonicity of Max ℓ_p Sensitivity)**.** *Let* $q \geq p > 0$ *and* $\mathbf{y} \in \mathbb{R}^n$. *Then,*

$$\frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_p^p} \leq \frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_q^q}.$$

We also use an "approximate converse" of the above result:

Lemma 2.3 (Reverse Monotonicity of Max ℓ_p Sensitivity)**.** *Let* $q \geq p > 0$ *and* $\mathbf{y} \in \mathbb{R}^n$. *Then,*

$$\frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_q^q} \leq \left(\frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_p^p}\right)^{q/p} n^{q/p-1}.$$

While these lemmas only apply to the max ℓ_p sensitivity and are quite loose when the max ℓ_p sensitivity can be arbitrary, we use the crucial fact that the ℓ_p sensitivities of $\mathbf{SA}$ are essentially "flat", that is, the maximum ℓ_p sensitivity will be within a small factor of the *average* ℓ_p sensitivity $\mathfrak{S}^p(\mathbf{A})/n$. Thus, Lemma 2.2 and Lemma 2.3 allow us to bound the leverage scores by the ℓ_p sensitivities for $p > 2$ and $p < 2$, respectively. This idea also allows us to prove Theorem 1.7.

2.4. Gaussianization Reduction for Sampling Algorithms

In the works of (Bourgain et al., 1989; Ledoux & Talagrand, 1991), a version of Lemma 2.1 tailored to ℓ_p Lewis weight sampling is used as a part of a recursive sampling algorithm, where the Rademacher process represents the sampling error of a process that samples each row $i \in [n]$ with probability $1/2$ and scales the result by 2. Indeed, if $\mathbf{S}_{i,i}$ takes the value 0 or $2^{1/p}$ with probability $1/2$ each and $\|\mathbf{Ax}\|_p^p = 1$, then $(\mathbf{S}_{i,i}^p - 1)$ is a Rademacher variable and

$$\|\mathbf{SAx}\|_p^p - 1 = \sum_{i=1}^n (\mathbf{S}_{i,i}^p - 1)|[\mathbf{Ax}](i)|^p,$$

and thus Lemma 2.1 bounds (3). While this only reduces the number of rows by a factor of 2, this process can be applied recursively for $O(\log n)$ rounds to reduce the number of rows to $\mathrm{poly}(d)$.

However, we instead primarily use Lemma 2.1 in a reduction based on (Cohen & Peng, 2015) for an algorithm with *one* round of sampling. In this reduction, we bound the sampling error (3) by introducing an independent copy $\mathbf{S}'$ of $\mathbf{S}$ and estimate

$$\Lambda \leq \mathbf{E} \sup_{\|\mathbf{Ax}\|_p=1} \left| \|\mathbf{SAx}\|_p^p - \|\mathbf{S}'\mathbf{Ax}\|_p^p \right|$$

Because $\mathbf{S}$ and $\mathbf{S}'$ are identically distributed, multiplying each coordinate $i \in [n]$ by a Rademacher variable $\boldsymbol{\varepsilon}_i \sim \{\pm 1\}$ does not change the distribution. Then by applying the triangle inequality, it follows that

$$\Lambda \leq 2 \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{Ax}\|_p=1} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i |[\mathbf{SAx}](i)|^p \right|, \tag{5}$$

which closely resembles Lemma 2.1.

At this point, for each fixing of $\mathbf{S}$, we wish to apply Lemma 2.1 with $\mathbf{A}$ replaced by $\mathbf{SA}$, with sensitivities bounded using the idea of (4). However, we cannot a priori assume that $\|\mathbf{SAx}\|_p^p \geq (1/2)\|\mathbf{Ax}\|_p^p$. To fix this, the idea of (Cohen & Peng, 2015) is to introduce an *auxiliary subspace embedding* $\mathbf{S}'$ such that $\mathbf{S}'\mathbf{A}$ also has ℓ_p sensitivities bounded by α, and *does* satisfy $\|\mathbf{S}'\mathbf{Ax}\|_p^p = \Theta(1)\|\mathbf{Ax}\|_p^p$ for every $\mathbf{x} \in \mathbb{R}^d$. Then, we can apply the result of

Lemma 2.1 to the concatenated matrix

$$\mathbf{A}' \coloneqq \begin{pmatrix} \mathbf{SA} \\ \mathbf{S}'\mathbf{A} \end{pmatrix}.$$

It can then be shown that the quantity E bounded in Lemma 2.1 indeed bounds the quantity in (5), and furthermore, we can fix the argument of (4) by bounding

$$\sup_{\mathbf{A}'\mathbf{x}\neq 0} \frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{A}'\mathbf{x}\|_p^p} \leq \sup_{\mathbf{Ax}\neq 0} \frac{1}{p_i} \frac{|[\mathbf{Ax}](i)|^p}{\|\mathbf{S}'\mathbf{Ax}\|_p^p} \leq \frac{O(\boldsymbol{\sigma}_i^p(\mathbf{A}))}{p_i}.$$

Importantly, we only need to use the existence of $\mathbf{S}'$ for the analysis and thus $\mathbf{S}'$ can be constructed in any way, rather than just by sensitivity sampling.

2.5. Construction of Auxiliary ℓ_p Subspace Embeddings

We now delve into further detail about the other key piece of our analysis, which is the construction of auxiliary ℓ_p subspace embeddings $\mathbf{S}'$ which are compatible with the reduction argument in Section 2.4.

2.5.1. Sensitivity Sampling, $p < 2$

For our sensitivity sampling result for $p < 2$ in Theorem 1.5, we aim to sample roughly $m = \varepsilon^{-2}\mathfrak{S}^p(\mathbf{A})^{2/p}$ rows. We first briefly sketch the intuition behind this bound. If we sample each row $i \in [n]$ with probability roughly $\boldsymbol{\sigma}_i^p(\mathbf{A})/\alpha$ for an oversampling parameter $\alpha > 0$, then we expect the ℓ_p sensitivities of $\mathbf{SA}$ to be bounded by α according to the informal reasoning of (4). We then wish to bound the sampling error using Lemma 2.1, which, combined with the bound on the leverage scores using the sensitivities in Lemma 2.3, gives a bound of

$$\tau \leq \alpha^{2/p} m^{2/p-1} = \alpha \mathfrak{S}^p(\mathbf{A}))^{2/p-1}$$

by using that m is $\mathfrak{S}^p(\mathbf{A})/\alpha$ in expectation. Since we want this to be at most ε^2, we set $\alpha = \varepsilon^2/\mathfrak{S}^p(\mathbf{A})^{2/p-1}$, which gives a bound of $m = \varepsilon^{-2}\mathfrak{S}^p(\mathbf{A})^{2/p}$ as claimed.

Now coming back to the formal argument using the auxiliary ℓ_p subspace embedding $\mathbf{S}'$, we wish to construct a subspace embedding $\mathbf{S}'$ that preserves ℓ_p norms of $\mathbf{Ax}$ up to a constant factor, but also has ℓ_p sensitivities at most α, since this is the bound that $\mathbf{SA}$ will satisfy. To construct this $\mathbf{S}'$, we proceed by first using Lewis weight sampling (Theorem 1.8) to construct a $\Theta(1)$-approximate ℓ_p subspace sampling with $\tilde{O}(d)$ rows. Importantly, we show that this sampling procedure can only increase the total ℓ_p sensitivity by a constant factor (see Lemma A.1). We then apply a flattening procedure (see Lemma A.4), which yields an ℓ_p isometry with at most $\mathfrak{S}^p(\mathbf{A})/\alpha$ rows such that the sensitivity of each row is bounded by α. Furthermore, because the number of rows $\mathfrak{S}^p(\mathbf{A})/\alpha$ of this auxiliary subspace embedding $\mathbf{S}'$ is at most the bound m that we seek, we can apply the same reasoning as before using Lemma 2.3 to bound the leverage scores of $\mathbf{S}'\mathbf{A}$ to recover the same bound as that for $\mathbf{SA}$. Thus, we obtain our desired construction of the auxiliary ℓ_p subspace embedding, allowing the reduction argument as described in Section 2.4 to go through.

2.5.2. Sensitivity Sampling, $p > 2$

For our result on sensitivity sampling when $p > 2$, the intuition and reasoning roughly follows the case of $p < 2$. However, we must be more careful with the construction of the auxiliary ℓ_p subspace embedding, since we cannot use Lewis weight sampling to construct it; this is because the sample complexity of $\tilde{O}(d^{p/2})$ for Lewis weight sampling is larger than our sample complexity bound of $m = \varepsilon^{-2}\mathfrak{S}^p(\mathbf{A})^{2-2/p}$ that we aim for. Thus, in order to obtain our auxiliary ℓ_p subspace embedding, we must essentially achieve the same results that we claim in Theorem 1.5, using an *alternate construction* that does not directly use one-shot sensitivity sampling. For this, we instead revisit the *recursive sampling* process mentioned earlier in Section 2.4, which originates from the functional analysis literature. Here, we first flatten our input matrix using Lemma A.4 to a matrix with at most $(4/3)n$ rows and ℓ_p sensitivities at most $O(\mathfrak{S}^p(\mathbf{A})/n)$, and then use Lemma 2.1 directly as the sampling error bound for a uniform sampling algorithm which samples each row with probability $1/2$. Note then that overall, we retain $(4/3)(1/2)n = (2/3)n$ rows altogether in expectation after this result, and furthermore, we have a sampling error of at most ε. Then, because we sample only a constant fraction of rows after each application of the procedure, it can be shown that recursively applying this result for $O(\log n)$ iterations accumulates a total sampling error of $O(\varepsilon \log n)$, while reducing the number of rows down to $\varepsilon^{-2}\mathfrak{S}^p(\mathbf{A})^{2-2/p}$, which matches the number of rows that we claim for sensitivity sampling in Theorem 1.5. By rescaling ε by an $O(\log n)$ factor, we can obtain a total error of ε while only losing polylogarithmic factors in n. This is carried out in Lemma D.3. Thus, we may again carry out the reduction argument as described in Section 2.4.

2.5.3. Root Leverage Score Sampling, $p < 2$

For our root leverage score sampling theorem Theorem 1.9, we take a conceptually different approach for obtaining the leverage score bound required in Lemma 2.1. For sensitivity sampling, our idea was to bound the leverage scores by relating them to the ℓ_p sensitivities, which were controlled by the sensitivity sampling process. For root leverage score sampling, however, the idea is that by sampling by the root leverage scores, we directly control the *leverage scores* of the sampled matrix $\mathbf{SA}$, rather than the sensitivity scores. Indeed, one can show that if the sampling probabilities p_i satisfy $p_i \geq \boldsymbol{\tau}_i(\mathbf{A})^{p/2}/\alpha$, and if $\|\mathbf{SAx}\|_2^2 \geq (1/2)\|\mathbf{Ax}\|_2^2$ for

every $\mathbf{x} \in \mathbb{R}^d$, then the leverage scores of $\mathbf{SA}$ are bounded by

$$\sup_{\mathbf{SAx}\neq 0} \frac{|[\mathbf{SAx}](i)|^2}{\|\mathbf{SAx}\|_2^2} \leq 2 \sup_{\mathbf{Ax}\neq 0} \frac{1}{p_i^{2/p}} \frac{|[\mathbf{Ax}](i)|^2}{\|\mathbf{Ax}\|_2^2} \leq 2\frac{\boldsymbol{\tau}_i(\mathbf{A})}{p_i^{2/p}} \tag{6}$$

which is at most $2\alpha^{2/p}$. However, it is not clear that $\|\mathbf{SAx}\|_2^2 \geq (1/2)\|\mathbf{Ax}\|_2^2$ should hold at all, since $\mathbf{S}$ is only a subspace embedding for ℓ_p and not ℓ_2. For this intuition to go through, we require an auxiliary ℓ_p subspace embedding that preserves both ℓ_p norms and ℓ_2 norms, with flat ℓ_2 leverage scores.

To make this argument work, we crucially use the fact that the bound in Lemma 2.1 does not depend polynomially on the number of rows of the input matrix. Thus, we can afford to construct $\mathbf{S}'\mathbf{A}$ to have many rows, as long as its leverage scores are controlled. Thus, in Lemma D.5, we take the approach of constructing an ℓ_p isometry of $\mathbf{A}$ by splitting every row $\mathbf{a}_i$ into k copies of $\mathbf{a}_i/k^{1/p}$, where we will set $k = 1/\alpha$. Note then that after splitting, every row contains at most a $1/k$ fraction of the ℓ_2 mass, so the ℓ_2 leverage scores are all at most α (in fact, all ℓ_q sensitivities for any q are at most α). Furthermore, crucially, the ℓ_2 norm will be preserved up to a factor of $\alpha^{1/p-1/q}$. Then altogether, we can fix (6) and instead bound

$$\sup_{\mathbf{SAx}\neq 0} \frac{|[\mathbf{SAx}](i)|^2}{\|\mathbf{A}'\mathbf{x}\|_2^2} \leq 2 \sup_{\mathbf{Ax}\neq 0} \frac{1}{p_i^{2/p}} \frac{|[\mathbf{Ax}](i)|^2}{\alpha^{2/p-1}\|\mathbf{Ax}\|_2^2}$$

which is at most 2α, where $\mathbf{A}'$ is the concatenation of $\mathbf{SA}$ and $\mathbf{S}'\mathbf{A}$. These ideas lead to our Theorem 1.9.

2.5.4. Leverage Score + Sensitivity Sampling, $p > 2$

Finally, for our Theorem 1.10, we take essentially the same approach as the recursive form of our sensitivity sampling result for $p > 2$ in Theorem 1.5, except that we improve our flattening approach by flattening *leverage scores* as well, which is formalized in Lemma D.9. Note that when we only flatten ℓ_p sensitivities using Lemma A.4, then the resulting bound on the leverage scores is just roughly the average ℓ_p sensitivity, which is $\mathfrak{S}^p(\mathbf{A})/n$. We show that by also flattening the leverage scores, we can improve this bound to

$$\left(\frac{d}{n}\right)^{2/p} \left(\frac{\mathfrak{S}^p(\mathbf{A})}{n}\right)^{1-2/p}.$$

Because $\mathfrak{S}^p(\mathbf{A})$ is always at least $d/2$ for $p > 2$ due to our Theorem 1.7, this is always better than the previous bound of $\mathfrak{S}^p(\mathbf{A})/n$, which ultimately leads to our improved sampling algorithm Theorem 1.10.

We note that we do not obtain a corresponding one-shot sampling algorithm, where the main difficulty is in constructing an auxiliary ℓ_p subspace embedding that has few rows, flat leverage scores, does not increase ℓ_p norms, and does not decrease ℓ_2 norms. Nonetheless, the recursive sampling procedure still leads to an efficient algorithm.

3. Conclusion and Future Directions

Our work introduces a new analysis for sensitivity sampling for ℓ_p subspace embeddings, which breaks a previous general sampling barrier of $\tilde{O}(\varepsilon^{-2}\mathfrak{S}^p(\mathbf{A})d)$ samples via a simple union bound argument, to obtain an improved bound of $\tilde{O}(\varepsilon^{-2}\mathfrak{S}^p(\mathbf{A})^{2/p})$ samples for $p < 2$ and $\tilde{O}(\varepsilon^{-2}\mathfrak{S}^p(\mathbf{A})^{2-2/p})$ samples for $p > 2$. We also present other novel results for sampling algorithms for ℓ_p subspace embeddings based on our techniques, showing that the popular root leverage score sampling algorithm yields a bound of $\tilde{O}(\varepsilon^{-4/p}d)$ for $p < 2$, as well as an improved $\tilde{O}(\varepsilon^{-2}d^{2/p}\mathfrak{S}^p(\mathbf{A})^{2-4/p})$ bound for $p > 2$ using a recursive sampling algorithm that combines ℓ_p sensitivity flattening with leverage score flattening. Our improved analyses of sensitivity sampling as well as our novel leverage score and sensitivity flattening algorithm give the best known sampling guarantees for a number of structured regression problems with small arbitrary noise.

We conclude with several open questions. Perhaps the most natural is to completely resolve Question 1.4 by characterizing the sample complexity of sensitivity sampling. We conjecture that a sample complexity of $\tilde{O}(\varepsilon^{-2}(\mathfrak{S}^p(\mathbf{A}) + d))$ is possible for ℓ_p subspace embeddings, and perhaps for more broad settings where sensitivity sampling applies as well. Furthermore, for $p > 2$, we believe it is of interest to obtain this bound even without the use of sensitivity sampling via other methods. Finally, we raise the question of obtaining sampling algorithms for subspace embeddings for the Huber loss with nearly optimal sample complexity, for which our results may be useful.

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. and Sachdeva, S. Faster p-norm minimizing flows, via smoothed q-norm problems. In Chawla, S. (ed.), *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pp. 892–910. SIAM, 2020.

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, 2019a.

Adil, D., Peng, R., and Sachdeva, S. Fast, provably convergent IRLS algorithm for p-norm linear regression. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14166–14177, 2019b.

Avron, H., Sindhwani, V., and Woodruff, D. P. Sketching structured matrices for faster nonlinear regression. In Burges, C. J. C., Bottou, L., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2994–3002, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/9683cc5f89562ea48e72bb321d9f03fb-Abstract.html.

Bourgain, J., Lindenstrauss, J., and Milman, V. Approximation of zonoids by zonotopes. *Acta Math.*, 162(1-2): 73–141, 1989.

Braverman, V., Feldman, D., and Lang, H. New frameworks for offline and streaming coreset constructions. *CoRR*, abs/1612.00889, 2016. URL http://arxiv.org/abs/1612.00889.

Braverman, V., Drineas, P., Musco, C., Musco, C., Upadhyay, J., Woodruff, D. P., and Zhou, S. Near optimal linear algebra in the online and sliding window models. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, pp. 517–528. IEEE, 2020.

Braverman, V., Hassidim, A., Matias, Y., Schain, M., Silwal, S., and Zhou, S. Adversarial robustness of streaming algorithms through importance sampling. 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. 3544–3557, 2021.

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.

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. 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, 2015a.

Clarkson, K. L. and Woodruff, D. P. Sketching for M-estimators: A unified approach to robust regression. In Indyk, P. (ed.), *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pp. 921–939. SIAM, 2015b.

Cohen, M. B. and Peng, R. $\mathrm{L_p}$ 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.

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.

Fazel, M., Lee, Y. T., Padmanabhan, S., and Sidford, A. Computing lewis weights to high precision. In Naor, J. S. and Buchbinder, N. (eds.), *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pp. 2723–2742. SIAM, 2022.

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.

Feldman, D., Schmidt, M., and Sohler, C. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.*, 49 (3):601–657, 2020. doi: 10.1137/18M1209854. URL https://doi.org/10.1137/18M1209854.

Ghadiri, M., Peng, R., and Vempala, S. S. Faster p-norm regression using sparsity. *arXiv preprint arXiv:2109.11537*, 2021.

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.

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.

Johnson, T. B. and Guestrin, C. Training deep models faster with robust, approximate importance sampling. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7276–7286, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/967990de5b3eac7b87d49a13c6834978-Abstract.html.

Katharopoulos, A. and Fleuret, F. Biased importance sampling for deep neural network training. *CoRR*, abs/1706.00043, 2017. URL http://arxiv.org/abs/1706.00043.

Langberg, M. and Schulman, L. J. Universal epsilon-approximators for integrals. In Charikar, M. (ed.), *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pp. 598–607. SIAM, 2010.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Lee, Y. T. *Faster algorithms for convex and combinatorial optimization*. PhD thesis, Massachusetts Institute of Technology, 2016.

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

Mahoney, M. W. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, 2011.

Meyer, R. A., Musco, C., Musco, C., Woodruff, D. P., and Zhou, S. Fast regression for structured inputs. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Munteanu, A., Schwiegelshohn, C., Sohler, C., and Woodruff, D. P. On coresets for logistic regression. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6562–6571, 2018.

Munteanu, A., Omlor, S., and Peters, C. p-generalized probit regression and scalable maximum likelihood estimation via sketching and coresets. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2073–2100. PMLR, 2022.

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. URL https://doi.org/10.1109/FOCS54457.2022.00076.

Nelson, J. Chaining introduction with some computer science applications. *Bull. EATCS*, 120, 2016. URL http://eatcs.org/beatcs/index.php/beatcs/article/view/450.

Schechtman, G. More on embedding subspaces of L_p in l_r^n. *Compositio Math.*, 61(2):159–169, 1987.

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

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.

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.

Tukan, M., Maalouf, A., and Feldman, D. Coresets for near-convex functions. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual*

Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

Varadarajan, K. R. and Xiao, X. On the sensitivity of shape fitting problems. In D'Souza, D., Kavitha, T., and Radhakrishnan, J. (eds.), *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2012, December 15-17, 2012, Hyderabad, India*, volume 18 of *LIPIcs*, pp. 486–497. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2012.

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

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. doi: 10.1109/FOCS54457.2022.00075. URL https://doi.org/10.1109/FOCS54457.2022.00075.

Woodruff, D. P. and Yasuda, T. Online Lewis weight sampling. In *Proceedings of the 34th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2023*. SIAM, 2023.

A. Properties of ℓ_p Sensitivities

A.1. Monotonicity of Max ℓ_p Sensitivity

We first provide proofs of Lemma 2.2 and Lemma 2.3. The results are similar to results used in (Braverman et al., 2020; Musco et al., 2022). In particular, it generalizes Lemma 4.6 of (Braverman et al., 2020) and is a simplification of a specific instance of Lemma C.3 of (Musco et al., 2022).

Proof of Lemma 2.2. We have that

$$\|\mathbf{y}\|_q^q = \sum_{i=1}^n |\mathbf{y}(i)|^q \leq \|\mathbf{y}\|_\infty^{q-p} \sum_{i=1}^n |\mathbf{y}(i)|^p = \|\mathbf{y}\|_\infty^{q-p} \|\mathbf{y}\|_p^p,$$

so

$$\frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_p^p} \leq \frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_q^q / \|\mathbf{y}\|_\infty^{q-p}} = \frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_q^q}.$$

□

Proof of Lemma 2.3. Since $\|\mathbf{y}\|_p \leq \|\mathbf{y}\|_q n^{1/p-1/q}$, we have that

$$\frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_q^q} \leq \frac{|\mathbf{y}(i)|^q}{\|\mathbf{y}\|_p^q \cdot n^{1-q/p}} \leq \frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_p^q \cdot n^{1-q/p}} = \left(\frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_p^p}\right)^{q/p} n^{q/p-1}.$$

□

A.2. Total Sensitivity

We now derive bounds on the total ℓ_p sensitivity.

A.2.1. Sampling Preserves Total Sensitivity

Lemma A.1 (Sampling Preserves Total Sensitivity)**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $1 \leq p < \infty$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix such that with probability at least $3/4$,*

$$\|\mathbf{SAx}\|_p = (1 \pm 1/2)\|\mathbf{Ax}\|_p$$

simultaneously for every $\mathbf{x} \in \mathbb{R}^d$. Then, with probability at least $1/2$,

$$\mathbf{Pr}\{\mathfrak{S}^p(\mathbf{SA}) \leq 8\mathfrak{S}^p(\mathbf{A})\} \geq \frac{1}{2}.$$

Proof. We have that

$$\mathfrak{S}^p(\mathbf{SA}) = \sum_{i=1}^n \sup_{\mathbf{SAx} \neq 0} \frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{SAx}\|_p^p} = \sum_{i=1}^n \mathbf{S}_{i,i}^p \sup_{\mathbf{SAx} \neq 0} \frac{|[\mathbf{Ax}](i)|^p}{\|\mathbf{Ax}\|_p^p} \frac{\|\mathbf{Ax}\|_p^p}{\|\mathbf{SAx}\|_p^p} \leq \sum_{i=1}^n \mathbf{S}_{i,i}^p \boldsymbol{\sigma}_i^p(\mathbf{A}) \sup_{\mathbf{SAx} \neq 0} \frac{\|\mathbf{Ax}\|_p^p}{\|\mathbf{SAx}\|_p^p}.$$

We are guaranteed that

$$\mathbf{Pr}\left\{\sup_{\mathbf{SAx} \neq 0} \frac{\|\mathbf{Ax}\|_p^p}{\|\mathbf{SAx}\|_p^p} \leq 2\right\} \geq \frac{3}{4}.$$

On the other hand, we have that

$$\mathbf{E}\left[\sum_{i=1}^n \mathbf{S}_{i,i}^p \boldsymbol{\sigma}_i^p(\mathbf{A})\right] = \sum_{i=1}^n \mathbf{E}[\mathbf{S}_{i,i}^p] \boldsymbol{\sigma}_i^p(\mathbf{A}) = \mathfrak{S}^p(\mathbf{A})$$

so by Markov's inequality,

$$\mathbf{Pr}\left\{\sum_{i=1}^{n}\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})\leq 4\mathfrak{S}^p(\mathbf{A})\right\}\geq\frac{3}{4}.$$

By a union bound,

$$\mathbf{Pr}\{\mathfrak{S}^p(\mathbf{SA})\leq 8\mathfrak{S}^p(\mathbf{A})\}\geq\frac{1}{2}.$$

□

We also prove a high probability and high accuracy version of Lemma A.1.

Lemma A.2 (Sensitivity Sampling Preserves Total Sensitivity: High Probability and Accuracy)**.** *Let* $\mathbf{A}\in\mathbb{R}^{n\times d}$ *and* $1\leq p<\infty$*. Let* $0<\varepsilon,\delta<1$*. Let* $\mathbf{S}$ *be a random* ℓ_p *sampling matrix such that with probability at least* $1-\delta$*,*

$$\|\mathbf{SAx}\|_p=(1\pm\varepsilon)\|\mathbf{Ax}\|_p$$

simultaneously for every $\mathbf{x}\in\mathbb{R}^d$*. Furthermore, suppose that*

$$\frac{\boldsymbol{\sigma}_i}{q_i}\leq M:=\frac{\varepsilon^2\mathfrak{S}^p(\mathbf{A})}{3\log\frac{2}{\delta}}$$

for every $i\in[n]$*. Then, with probability at least* $1-2\delta$*,*

$$\mathbf{Pr}\{\mathfrak{S}^p(\mathbf{SA})=(1\pm O(\varepsilon))\mathfrak{S}^p(\mathbf{A})\}\geq 1-2\delta.$$

Proof. The proof follows Lemma A.1. Just as in Lemma A.1, we have that

$$\mathfrak{S}^p(\mathbf{SA})\leq\sum_{i=1}^{n}\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})\sup_{\mathbf{SAx}\neq 0}\frac{\|\mathbf{Ax}\|_p^p}{\|\mathbf{SAx}\|_p^p}.$$

Similarly,

$$\mathfrak{S}^p(\mathbf{SA})\geq\sum_{i=1}^{n}\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})\inf_{\mathbf{SAx}\neq 0}\frac{\|\mathbf{Ax}\|_p^p}{\|\mathbf{SAx}\|_p^p}.$$

Furthermore, since $\boldsymbol{\sigma}_i/q_i\leq M$, $\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})/M$ is a random variable bounded by 1, with

$$\mathbf{E}\left[\sum_{i=1}^{n}\frac{\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})}{M}\right]=\frac{\mathfrak{S}^p(\mathbf{A})}{M}\geq\frac{3}{\varepsilon^2}\log\frac{2}{\delta}.$$

Thus by Chernoff bounds, we have that

$$\mathbf{Pr}\left\{\sum_{i=1}^{n}\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})=(1\pm\varepsilon)\mathfrak{S}^p(\mathbf{A})\right\}\geq 1-\delta.$$

We conclude by a union bound as in Lemma A.1. □

A.2.2. Total Sensitivity Lower Bounds

We start with the classical result that the total ℓ_2 sensitivity is exactly d:

Lemma A.3. *Let* $\mathbf{A}\in\mathbb{R}^{n\times d}$ *and let* $\mathbf{U}\in\mathbb{R}^{n\times d}$ *be an orthnormal basis for the column space of* $\mathbf{A}$*. Then,*

$$\boldsymbol{\tau}_i(\mathbf{A})=\left\|\mathbf{e}_i^\top\mathbf{U}\right\|_2^2$$

and

$$\sum_{i=1}^{n}\boldsymbol{\tau}_i(\mathbf{A})=\|\mathbf{U}\|_F^2=d.$$

Proof. We have that

$$\boldsymbol{\tau}_i(\mathbf{A}) = \sup_{\mathbf{x}\in\mathbb{R}^d, \mathbf{A}\mathbf{x}\neq 0} \frac{|[\mathbf{A}\mathbf{x}](i)|^2}{\|\mathbf{A}\mathbf{x}\|_2^2} = \sup_{\mathbf{x}\in\mathbb{R}^d, \mathbf{U}\mathbf{x}\neq 0} \frac{|[\mathbf{U}\mathbf{x}](i)|^2}{\|\mathbf{U}\mathbf{x}\|_2^2} = \sup_{\mathbf{x}\in\mathbb{R}^d, \mathbf{x}\neq 0} \frac{|[\mathbf{U}\mathbf{x}](i)|^2}{\|\mathbf{x}\|_2^2} = \left\|\mathbf{e}_i^\top \mathbf{U}\right\|_2^2.$$

□

We now use Lemma A.3 together with Lemma 2.2 and Lemma 2.3 to derive lower bounds on $\mathfrak{S}^p(\mathbf{A})$.

By using a simple argument based on "splitting rows" (see, e.g., (Ledoux & Talagrand, 1991; Cohen & Peng, 2015; Chen & Derezinski, 2021; Musco et al., 2022)), it is possible to assume without loss of generality that the maximum ℓ_p sensitivity is related to the average ℓ_p sensitivity, up to a factor of 2:

Lemma A.4 (ℓ_p Sensitivity Flattening)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $1 \le p < \infty$. Let $C \ge 1$. Then, there exists a $\mathbf{A}' \in \mathbb{R}^{m\times d}$ for $m = (1 + 1/C)n$ such that $\|\mathbf{A}\mathbf{x}\|_p = \|\mathbf{A}'\mathbf{x}\|_p$ for every $\mathbf{x} \in \mathbb{R}^d$, $\mathfrak{S}^p(\mathbf{A}) = \mathfrak{S}^p(\mathbf{A}')$, and $\boldsymbol{\sigma}_{i'}^p(\mathbf{A}') \le C\mathfrak{S}^p(\mathbf{A})/n$ for every $i' \in [m]$.*

Proof of Lemma A.4. Suppose that for any row $\mathbf{a}_i \in \mathbb{R}^d$ of $\mathbf{A}$ for $i \in [n]$ with $\boldsymbol{\sigma}_i^p(\mathbf{A}) \ge C\mathfrak{S}^p(\mathbf{A})/n$, we replace the row with $k := \lceil \boldsymbol{\sigma}_i^p(\mathbf{A})/(C\mathfrak{S}^p(\mathbf{A})/n)\rceil$ copies of $\mathbf{a}_i/k^{1/p}$ to form a new matrix $\mathbf{A}'$. Then, we add at most

$$\sum_{i:\boldsymbol{\sigma}_i^p(\mathbf{A})\ge\mathfrak{S}^p(\mathbf{A})/n} \left\lceil \frac{\boldsymbol{\sigma}_i^p(\mathbf{A})}{C\mathfrak{S}^p(\mathbf{A})/n} \right\rceil - 1 \le \sum_{i:\boldsymbol{\sigma}_i^p(\mathbf{A})\ge\mathfrak{S}^p(\mathbf{A})/n} \frac{\boldsymbol{\sigma}_i^p(\mathbf{A})}{C\mathfrak{S}^p(\mathbf{A})/n} = \frac{\mathfrak{S}^p(\mathbf{A})}{C\mathfrak{S}^p(\mathbf{A})/n} = \frac{n}{C}$$

rows. Furthermore, we clearly have that $\|\mathbf{A}\mathbf{x}\|_p = \|\mathbf{A}'\mathbf{x}\|_p$ for every $\mathbf{x} \in \mathbb{R}^d$, and also for any row $i' \in [m]$ that comes from row $i \in [n]$ in the original matrix,

$$\frac{|[\mathbf{A}'\mathbf{x}](i')|^p}{\|\mathbf{A}'\mathbf{x}\|_p^p} \le \frac{C\mathfrak{S}^p(\mathbf{A})/n}{\boldsymbol{\sigma}_i^p(\mathbf{A})} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \le \frac{C\mathfrak{S}^p(\mathbf{A})}{n}.$$

Finally, it is also clear that the sum of the sensitivities is also preserved, since the sum of the sensitivities of the k copies of each row $i \in [n]$ in the original matrix is $\boldsymbol{\sigma}_i^p(\mathbf{A})$. □

We can now prove Theorem 1.7:

Proof of Theorem 1.7. Let $\mathbf{A}' \in \mathbb{R}^{2n\times d}$ be the matrix given by Lemma A.4 applied with $C = 1$. Then for $p > 2$, we have by Lemma 2.2 that

$$\frac{d}{n} \le \max_{i=1}^n \boldsymbol{\sigma}_i^2(\mathbf{A}') \le \max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A}') \le \frac{2\mathfrak{S}^p(\mathbf{A})}{n}$$

and for $p < 2$, we have by Lemma 2.3 that

$$\frac{d}{n} \le \max_{i=1}^n \boldsymbol{\sigma}_i^2(\mathbf{A}') \le \left(\max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A}')\right)^{2/p} n^{2/p-1} \le \left(\frac{2\mathfrak{S}^p(\mathbf{A})}{n}\right)^{2/p} n^{2/p-1} = \frac{2^{2/p}\mathfrak{S}^p(\mathbf{A})^{2/p}}{n}$$

which yield the claimed results. □

A.2.3. Random Matrices Have Small Total Sensitivity

We show that the above lower bounds can be tight, up to logarithmic factors. We will use Dvoretzky's theorem, which can be found in, e.g., Fact 15 in (Sohler & Woodruff, 2018):

Theorem A.5 (Dvoretzky's Theorem)**.** *Let $1 \le p < 2$. Let $n = (d/\varepsilon)^{O(p)}$ be sufficiently large, and let $\mathbf{A}$ be a suitably scaled random $n \times d$ Gaussian matrix. Then, with probability at least $99/100$, we have for every $\mathbf{x} \in \mathbb{R}^d$ that*

$$\|\mathbf{A}\mathbf{x}\|_p = (1 \pm \varepsilon)\|\mathbf{x}\|_2.$$

This gives a proof of Theorem 1.6:

Proof of Theorem 1.6. By applying Theorem A.5 with $\varepsilon = 1/2$, we have that $\|\mathbf{A}\mathbf{x}\|_p^p = \Theta(n)\|\mathbf{x}\|_2^p$ with probability at least $99/100$. Note also that

$$\max_{i=1}^{n} \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2 \leq O(\sqrt{d \log n}) = O(\sqrt{d \log d})$$

which also happens with probability at least $99/100$. By a union bound, both events happen with probability at least $98/100$.

Now for any $\mathbf{x} \in \mathbb{R}^d$ with unit ℓ_2 norm, we have that

$$|[\mathbf{A}\mathbf{x}](i)|^p \leq \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^p \cdot \|\mathbf{x}\|_2^p = \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^p \leq O(d \log d)^{p/2}.$$

Thus,

$$\mathfrak{S}^p(\mathbf{A}) \leq n \cdot \max_{i=1}^{n} \sup_{\|\mathbf{x}\|_2=1} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \leq n \cdot \max_{i=1}^{n} \sup_{\|\mathbf{x}\|_2=1} \frac{O(d \log d)^{p/2}}{\Theta(n)} = O(d \log d)^{p/2}.$$

□

A.3. Structured Matrices with Small Sensitivity, $p > 2$

Proof of Lemma 1.11. Let r be an integer such that $2^r \leq p < 2^{r+1}$. Then, for each $i \in [n]$, we may write

$$\mathbf{a}_i = \mathbf{k}_i + \mathbf{s}_i = \sum_{j=1}^{k} \alpha_{i,j} \mathbf{v}_j + \sum_{j=1}^{s} \beta_{i,j} \mathbf{e}_{i_j}$$

where $\mathbf{v}_j \in \mathbb{R}^d$ for $j \in [k]$. Then, the tensor product $\mathbf{a}_i^{\otimes 2^r}$ of $\mathbf{a}_i$ with itself 2^r times can be written as a linear combination of tensor products $\mathbf{y}_1 \otimes \cdots \otimes \mathbf{y}_{2^r}$, where each $\mathbf{y}_q$ for $q \in [2^r]$ is one of $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k, \mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_s}\}$. Thus, $\mathbf{a}_i^{\otimes r}$ lies in the span of at most $(k+s)^{2^r}$ vectors, for a fixed choice of $\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_s}$. Since there are at most d^s possible choices of the sparsity pattern, every $\mathbf{a}_i^{\otimes 2^r}$ for $i \in [n]$ lies in the span of at most $d' := d^s(k+s)^{2^r}$ vectors. That is, if $\mathbf{A}^{\otimes 2^r}$ is the Khatri-Rao 2^rth power of $\mathbf{A}$, then $\mathbf{A}^{\otimes 2^r}$ is a rank d' matrix. Then, we have that

$$|[\mathbf{A}\mathbf{x}](i)|^p = (|[\mathbf{A}\mathbf{x}](i)|^{2^r})^{p/2^r} = (\langle \mathbf{a}_i, \mathbf{x}\rangle^{2^r})^{p/2^r} = \left|\left\langle \mathbf{a}_i^{\otimes 2^r}, \mathbf{x}^{\otimes 2^r}\right\rangle\right|^{p/2^r}$$

so

$$\sup_{\mathbf{A}\mathbf{x}\neq 0} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} = \sup_{\mathbf{A}\mathbf{x}\neq 0} \frac{\left|[\mathbf{A}^{\otimes 2^r}\mathbf{x}^{\otimes 2^r}](i)\right|^{p/2^r}}{\|\mathbf{A}^{\otimes 2^r}\mathbf{x}^{\otimes 2^r}\|_{p/2^r}^{p/2^r}} \leq \sup_{\mathbf{A}^{\otimes 2^r}\mathbf{x}\neq 0} \frac{\left|[\mathbf{A}^{\otimes 2^r}\mathbf{x}](i)\right|^{p/2^r}}{\|\mathbf{A}^{\otimes 2^r}\mathbf{x}\|_{p/2^r}^{p/2^r}}$$

that is, the $\ell_{p/2^r}$ sensitivities of $\mathbf{A}^{\otimes 2^r}$ upper bound the ℓ_p sensitivities of $\mathbf{A}$. Since $p/2^r \leq 2$, the total $\ell_{p/2^r}$ sensitivity of $\mathbf{A}^{\otimes 2^r}$ is bounded by its rank, which is d'. □

Proof of Lemma 1.14. Let r be an integer such that $2^r \leq p < 2^{r+1}$. Fix some $\mathbf{x} \in \mathbb{R}^{k(q+1)}$. Now consider the vector $\langle \mathbf{a}, \mathbf{x}\rangle$, where $\mathbf{a}$ is a $k(q+1)$-dimensional vector of monomials of degree 0 through q of the indeterminate variables $a_1, a_2, \ldots, a_k$, that is,

$$\mathbf{a} = (1, a_1, a_1^2, \ldots, a_1^q, \quad 1, a_2, a_2^2, \ldots, a_2^q, \quad \ldots, \quad 1, a_k, a_k^2, \ldots, a_k^q).$$

Then, $\langle \mathbf{a}, \mathbf{x}\rangle$ is a degree q polynomial in the indeterminates $a_1, a_2, \ldots, a_k$ with coefficients specified by $\mathbf{x}$, so $\langle \mathbf{a}, \mathbf{x}\rangle^{2^r}$ is a polynomial in the indeterminates $a_1, a_2, \ldots, a_k$, such that every monomial term is at most degree $2^r q$ in each variable. Note that there are at most k variables, so there can be at most $(2^r q + 1)^k$ possible monomials, by choosing the degree of each of the monomials. Let $\mathbf{x}'$ denote the coefficients of this polynomial in the monomial basis, for a given set of original coefficients $\mathbf{x}$.

Now consider the matrix $V^q(\mathbf{A})$. Then, for a fixed $\mathbf{x} \in \mathbb{R}^{k(q+1)}$, $[V^q(\mathbf{A})\mathbf{x}](i)^{2^r}$ is the evaluation of $\langle \mathbf{a}, \mathbf{x}\rangle^{2^r}$ at the ith row $\mathbf{a}_i$ of $\mathbf{A}$ for the indeterminates $a_1, a_2, \ldots, a_k$, so it can be written as the linear combination of at most $(2^r q + 1)^k$ monomials evaluated at $\mathbf{a}_i$, with coefficients $\mathbf{x}'$. Thus, $[V^q(\mathbf{A})\mathbf{x}](i)^{2^r} = \mathbf{A}'\mathbf{x}'$ for some $\mathbf{A}'$ with rank at most $(2^r q + 1)^k$.

Finally, note that

$$|[V^q(\mathbf{A})\mathbf{x}](i)|^p = (|[V^q(\mathbf{A})\mathbf{x}](i)|^{2^r})^{p/2^r} = |[\mathbf{A}'\mathbf{x}'](i)|^{p/2^r}.$$

Thus, the total ℓ_p sensitivity of $V^q(\mathbf{A})$ is bounded by the total $\ell_{p/2^r}$ sensitivity of $\mathbf{A}'$, which is at most $(2^r q + 1)^k \leq (pq+1)^k$. □

A.4. Total ℓ_p Sensitivity Under Perturbations

Proof of Lemma 1.15. For any $\mathbf{x} \in \mathbb{R}^d$, we have that

$$\begin{aligned}\|(\mathbf{A}+\mathbf{E})\mathbf{x}\|_p &= \|\mathbf{A}\mathbf{x}\|_p \pm \|\mathbf{E}\mathbf{x}\|_p \\ &= \|\mathbf{A}\mathbf{x}\|_p \pm \sqrt{n}\|\mathbf{E}\mathbf{x}\|_2 \\ &= \|\mathbf{A}\mathbf{x}\|_p \pm \frac{\sigma_{\min}}{\sqrt{n}}\|\mathbf{x}\|_2 \\ &= \|\mathbf{A}\mathbf{x}\|_p \pm \frac{\sigma_{\min}}{\sqrt{n}}\frac{1}{\sigma_{\min}}\|\mathbf{A}\mathbf{x}\|_2 \\ &= \|\mathbf{A}\mathbf{x}\|_p \pm \frac{1}{2}\|\mathbf{A}\mathbf{x}\|_p \\ &= (1 \pm 1/2)\|\mathbf{A}\mathbf{x}\|_p\end{aligned}$$

so

$$\frac{|[(\mathbf{A}+\mathbf{E})\mathbf{x}](i)|^p}{\|(\mathbf{A}+\mathbf{E})\mathbf{x}\|_p^p} \le 2^{p-1}\frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|(\mathbf{A}+\mathbf{E})\mathbf{x}\|_p^p} + 2^{p-1}\frac{|[\mathbf{E}\mathbf{x}](i)|^p}{\|(\mathbf{A}+\mathbf{E})\mathbf{x}\|_p^p} \le 2^p\frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} + 2^p\frac{|[\mathbf{E}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p}.$$

The first term is clearly bounded by $2^p\boldsymbol{\sigma}_i^p(\mathbf{A})$ for any $\mathbf{x}$. On the other hand, the second term is bounded by

$$2^p\frac{|[\mathbf{E}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \le 2^p\frac{\|\mathbf{E}\mathbf{x}\|_p^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \le 2^p n^{p/2}\frac{\|\mathbf{E}\mathbf{x}\|_2^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \le 2^p\frac{\sigma_{\min}^p}{n^{p/2+1}}\frac{\|\mathbf{x}\|_2^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \le 2^p\frac{1}{n^{p/2+1}}\frac{\|\mathbf{A}\mathbf{x}\|_2^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \le 2^p\frac{\|\mathbf{A}\mathbf{x}\|_p^p}{n\|\mathbf{A}\mathbf{x}\|_p^p} = \frac{2^p}{n}.$$

Thus, the total sensitivity is bounded by

$$2^p\sum_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A}) + \frac{1}{n} = 2^p(\mathfrak{S}^p(\mathbf{A})+1).$$

□

B. Entropy Estimates

In this section, we collect our results on estimates on various metric entropies, which are needed for our chaining arguments. Our results here are based on similar results given by (Bourgain et al., 1989). However, we modify their arguments to only depend on leverage scores and ℓ_p sensitivities, rather than using Lewis weights (Lewis, 1978; Bourgain et al., 1989).

B.1. Preliminaries

We first recall general definitions from convex geometry that are relevant to this section.

Definition B.1 (d_X-balls)**.** Let d_X be a metric on $\mathbb{R}^d$. Then, for $\mathbf{x} \in \mathbb{R}^d$ and $t \ge 0$, we define the d_X-ball of radius t $B_d(\mathbf{x}, t)$ to be

$$B_X(\mathbf{x}, t) \coloneqq \left\{\mathbf{x}' \in \mathbb{R}^d : d_X(\mathbf{x}, \mathbf{x}') \le t\right\}.$$

Definition B.2 (Covering numbers and metric entropy)**.** Let $K, T \subseteq \mathbb{R}^d$ be two convex bodies. Then, the *covering number* $E(K, T)$ is defined as

$$E(K, T) \coloneqq \min\left\{k \in \mathbb{N} : \exists\{\mathbf{x}_i\}_{i=1}^k, K \subseteq \bigcup_{i=1}^k(\mathbf{x}_i + T)\right\}.$$

If d_X is a metric and $t > 0$ a radius, then $E(K, d_X, t)$ is defined as

$$E(K, d_X, t) \coloneqq E(K, B_X(0, t))$$

(see Definition B.1). The *metric entropy* is the logarithm of the covering number.

Next, we introduce some notation that is specific to our setting of ℓ_p subspace embeddings.

Definition B.3. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \ge 1$, we define the ball

$$B^p(\mathbf{A}) \coloneqq \left\{\mathbf{A}\mathbf{x} \in \mathbb{R}^n : \|\mathbf{A}\mathbf{x}\|_p \le 1\right\}.$$

We simply write B^p if $\mathbf{A}$ is clear from context.

B.2. Dual Sudakov Minoration

One powerful tool for bounding covering numbers for covers of the Euclidean ball is the *dual Sudakov minoration* theorem, which bounds covering numbers in terms of the so-called *Levy mean*:

Definition B.4 (Levy mean)**.** Let $\|\cdot\|_X$ be a norm. Then, the *Levy mean of* $\|\cdot\|_X$ is defined to be

$$M_X := \frac{\mathbf{E}_{\mathbf{g}\in\mathcal{N}(0,\mathbf{I}_d)}\|\mathbf{g}\|_X}{\mathbf{E}_{\mathbf{g}\in\mathcal{N}(0,\mathbf{I}_d)}\|\mathbf{g}\|_2}.$$

Bounds on the Levy mean imply bounds for covering the Euclidean ball by $\|\cdot\|_X$-balls via the following result:

Theorem B.5 (Dual Sudakov minoration, Proposition 4.2 of (Bourgain et al., 1989))**.** *Let $\|\cdot\|_X$ be a norm, and let $B \subseteq \mathbb{R}^d$ denote the Euclidean ball in d dimensions. Then,*

$$\log E(B, \|\cdot\|_X, t) \leq O(d)\frac{M_X^2}{t^2}$$

B.3. Entropy Estimates for $p > 2$

We now use the preceding results to obtain the entropy estimates necessary to prove our main result for $p > 2$. We start by bounding the Levy mean for the norm defined by $\mathbf{x} \mapsto \|\mathbf{A}\mathbf{x}\|_q$ for some matrix $\mathbf{A}$.

Lemma B.6. *Let $q \geq 2$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Then,*

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_d)}\|\mathbf{A}\mathbf{g}\|_q \leq n^{1/q}\sqrt{q\cdot\tau}.$$

Proof. We have for every $i \in [n]$ that

$$\mathbf{E}|[\mathbf{A}\mathbf{g}](i)|^q = \frac{2^{q/2}\Gamma(\frac{q+1}{2})}{\sqrt{\pi}}\|\mathbf{e}_i^\top \mathbf{A}\|_2^q \leq q^{q/2}\cdot\tau^{q/2}$$

since $[\mathbf{A}\mathbf{g}](i)$ is distributed as a Gaussian random variable. Then by Jensen's inequality and linearity of expectation,

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_d)}\|\mathbf{A}\mathbf{g}\|_q \leq \left(\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_d)}\|\mathbf{A}\mathbf{g}\|_q^q\right)^{1/q} = \left(n\cdot q^{q/2}\cdot\tau^{q/2}\right)^{1/q} = n^{1/q}\sqrt{q\cdot\tau}$$

□

By combining the above calculation with Theorem B.5, we obtain the following:

Corollary B.7. *Let $2 \leq q < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Then,*

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

Proof. For $\mathbf{A}$ orthonormal, $B^2(\mathbf{A}) = B^2$ is isometric to the Euclidean ball in d dimensions, and thus Theorem B.5 applies. □

We also get a similar result for $q = \infty$, by applying Corollary B.7 with $q = O(\log n)$.

Corollary B.8. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Then,*

$$\log E(B^2, B^\infty, t) \leq O(1)\frac{(\log n)\cdot\tau}{t^2}$$

Proof. This follows from the fact that for $\mathbf{y} \in \mathbb{R}^n$,

$$\|\mathbf{y}\|_\infty \leq \|\mathbf{y}\|_q \leq n^{1/q}\|\mathbf{y}\|_\infty = O(1)\|\mathbf{y}\|_\infty$$

for $q = O(\log n)$. □

B.4. Entropy Estimates for $p < 2$

By interpolation, we can improve the bound in Corollary B.7, which is needed for our results for $p < 2$:

Lemma B.9. *Let $2 < r < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Let $1 \leq t \leq \mathrm{poly}(d)$. Then,*

$$\log E(B^2, B^r, t) \leq O(1)\frac{1}{(t/2)^{2r/(r-2)}} \cdot \left(\frac{r}{r-2}\log d + \log n\right)\tau$$

Proof. Let $q > r$, and let $0 < \theta < 1$ satisfy

$$\frac{1}{r} = \frac{1-\theta}{2} + \frac{\theta}{q}$$

Then by Hölder's inequality, we have for any $\mathbf{y} \in \mathbb{R}^n$ that

$$\|\mathbf{y}\|_r = \left(\sum_{i=1}^n |\mathbf{y}(i)|^{r(1-\theta)}|\mathbf{y}(i)|^{r\theta}\right)^{1/r} \leq \left(\sum_{i=1}^n |\mathbf{y}(i)|^2\right)^{(1-\theta)/2}\left(\sum_{i=1}^n |\mathbf{y}(i)|^q\right)^{\theta/q} = \|\mathbf{y}\|_2^{1-\theta}\|\mathbf{y}\|_q^\theta$$

Then for any $\mathbf{y}, \mathbf{y}' \in B^2$, we have

$$\|\mathbf{y}-\mathbf{y}'\|_r \leq \|\mathbf{y}-\mathbf{y}'\|_2^{1-\theta}\|\mathbf{y}-\mathbf{y}'\|_q^\theta \leq 2\|\mathbf{y}-\mathbf{y}'\|_q^\theta$$

so

$$\log E(B^2, B^r, t) \leq \log E(B^2, B^q, (t/2)^{1/\theta}) \leq O(1)\frac{n^{2/q}q\cdot\tau}{(t/2)^{2/\theta}}$$

by Corollary B.7. Now, we have

$$\frac{2}{\theta} = 2\frac{\frac{1}{2}-\frac{1}{q}}{\frac{1}{2}-\frac{1}{r}} = \frac{q-2}{q}\frac{2r}{r-2}$$

so by taking $q = O(\frac{r}{r-2}\log d + \log n)$, we have that $n^{2/q} = O(1)$ and $(t/2)^{1/\theta} = \Theta(1)(t/2)^{2r/(r-2)}$, so we conclude as claimed. □

Using Lemma B.9, we obtain the following analogue of Corollary B.7 for $p < 2$.

Lemma B.10. *Let $1 \leq p < 2$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Then,*

$$\log E(B^p, B^\infty, t) \leq O(1)\frac{1}{t^p}\left(\frac{\log d}{2-p} + \log n\right)\tau.$$

Proof. In order to bound a covering of B^p by B^∞, we first cover B^p by B^2, and then use Corollary B.8 to cover B^2 by B^∞.

We will first bound $E(B^p, B^2, t)$ using Lemma B.9. For each $k \geq 0$, let $\mathcal{E}_k \subseteq B^p$ be a maximal subset of B^p such that for each distinct $\mathbf{y}, \mathbf{y}' \in \mathcal{E}_k$, $\|\mathbf{y}-\mathbf{y}'\|_2 > 8^k t$, with $\mathcal{E}_k := \{0\}$ for $8^{k+1}t > n^{1/p-1/q}$. Note then that

$$|\mathcal{E}_k| \geq E(B^p, B^2, 8^k t).$$

By averaging, for each k, there exists $\mathbf{y}^{(k)} \in \mathcal{E}_k$ such that if

$$\mathcal{F}_k := \left\{\mathbf{y} \in \mathcal{E}_k : \|\mathbf{y}-\mathbf{y}^{(k)}\|_2 \leq 8^{k+1}t\right\},$$

then

$$|\mathcal{F}_k| \geq \frac{|\mathcal{E}_k|}{E(B^p, B^2, 8^{k+1}t)} \geq \frac{E(B^p, B^2, 8^k t)}{E(B^p, B^2, 8^{k+1}t)}$$

We now use this observation to construct an $\ell_{p'}$-packing of B^2, where p' is the Hölder conjugate of p. Let

$$\mathcal{G}_k := \left\{\frac{1}{8^{k+1}t}(\mathbf{y}-\mathbf{y}^{(k)}) : \mathbf{y} \in \mathcal{F}_k\right\}.$$

Then, $\mathcal{G}_k \subseteq B^2$ and $\mathcal{G}_k \subseteq B^p \cdot 2/8^{k+1}t$, and $\|\mathbf{y}-\mathbf{y}'\|_2 > 1/8$ for every distinct $\mathbf{y}, \mathbf{y}' \in \mathcal{G}_k$. Then by Hölder's inequality,

$$\frac{1}{8^2} \leq \|\mathbf{y}-\mathbf{y}'\|_2^2 \leq \|\mathbf{y}-\mathbf{y}'\|_p \|\mathbf{y}-\mathbf{y}'\|_{p'} \leq \frac{4}{8^{k+1}t}\|\mathbf{y}-\mathbf{y}'\|_{p'}$$

so $\|\mathbf{y}-\mathbf{y}'\|_{p'} \geq 2 \cdot 8^{k-2}t$. Thus, $\mathcal{G}_k$ is an $\ell_{p'}$-packing of B^2, so

$$\log E(B^2, B^{p'}, 8^{k-2}t) \geq \log|\mathcal{G}_k| = \log|\mathcal{F}_k| \geq \log E(B^p, B^2, 8^k t) - \log E(B^p, B^2, 8^{k+1}t). \tag{7}$$

Summing over k gives

$$\begin{aligned}
\log E(B^p, B^2, t) &= \sum_{k\geq 0} \log E(B^p, B^2, 8^k t) - \log E(B^p, B^2, 8^{k+1}t) \\
&\leq \sum_{k\geq 0} \log E(B^2, B^{p'}, 8^{k-2}t) && (7) \\
&\leq O(1)\frac{1}{(t/2)^{2p'/(p'-2)}} \cdot \left(\frac{p'}{p'-2}\log d + \log n\right)\tau && \text{Lemma B.9 and Corollary B.8} \\
&= O(1)\frac{1}{(t/2)^{2p/(2-p)}} \cdot \left(\frac{p}{2-p}\log d + \log n\right)\tau
\end{aligned}$$

where we take $p'/(p'-2) = 1$ for $p' = \infty$. Using this and Corollary B.8, we now bound

$$\begin{aligned}
\log E(B^p, B^\infty, t) &\leq \log E(B^p, B^2, \lambda) + \log E(B^2, B^\infty, t/\lambda) \\
&\leq O(1)\frac{1}{(\lambda/2)^{2p/(2-p)}} \cdot \left(\frac{p}{2-p}\log d + \log n\right)\tau + O(1)\frac{(\log n)\cdot\tau}{(t/\lambda)^2}
\end{aligned}$$

for any $\lambda \in [1, t]$. We choose λ satisfying

$$\frac{1}{(\lambda/2)^{2p/(2-p)}} = \frac{(\lambda/2)^2}{t^2},$$

which gives

$$(\lambda/2)^{2p/(2-p)} = \left(t^2\right)^{\frac{2p/(2-p)}{2+2p/(2-p)}} = t^p$$

so we obtain a bound of

$$O(1)\frac{1}{t^p}\left(\frac{1}{2-p}\log d + \log n\right)\tau.$$

□

C. Bounding a Gaussian Process

Using our estimates from Appendix B, we study estimates on the Gaussian process given by

$$X : \mathbf{y} \mapsto \sum_{i=1}^n \mathbf{g}_i |\mathbf{y}(i)|^p, \qquad \mathbf{y} \in B^p(\mathbf{A})$$

for $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$, and in particular, tail bounds and moment bounds on the quantity

$$\sup_{\mathbf{y}\in B^p(\mathbf{A})} |X(\mathbf{y})| = \sup_{\|\mathbf{A}\mathbf{x}\|_p \leq 1} \left|\sum_{i=1}^n \mathbf{g}_i |[\mathbf{A}\mathbf{x}](i)|^p\right|$$

As we show later, this Gaussian process bounds the error of the sensitivity sampling estimate. Our techniques here are based on similar results obtained by (Ledoux & Talagrand, 1991).

The main tool is Dudley's tail inequality for Gaussian processes:

Theorem C.1 (Theorem 8.1.6, (Vershynin, 2018)). *Let $(X(t))_{t\in T}$ be a Gaussian process with pseudo-metric $d_X(s,t) \coloneqq \|X(s)-X(t)\|_2 = \sqrt{\mathbf{E}(X(s)-X(t))^2}$ and let*

$$\operatorname{diam}(T) \coloneqq \sup\{d_X(s,t) : s,t \in T\}$$

Then, there is a constant $C = O(1)$ such that for every $z \geq 0$,

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

The expectation bound version of the theorem will also be useful:

Theorem C.2 (Theorem 8.1.3, (Vershynin, 2018)). *Let $(X(t))_{t\in T}$ be a Gaussian process with pseudo-metric $d_X(s,t) \coloneqq \|X(s)-X(t)\|_2 = \sqrt{\mathbf{E}(X(s)-X(t))^2}$. Then, there is a constant $C = O(1)$ such that*

$$\mathbf{E}\sup_{t\in T} X_t \leq C\int_0^\infty \sqrt{\log E(T, d_X, u)}\, du.$$

C.1. Bounds on d_X

We first bound d_X as well as the d_X-diameter of $B^p(\mathbf{A})$.

Lemma C.3. *Let $1 \leq p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Define the pseudo-metric*

$$d_X(\mathbf{y},\mathbf{y}') \coloneqq \left(\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)}\left|\sum_{i=1}^n \mathbf{g}_i|\mathbf{y}(i)|^p - \sum_{i=1}^n \mathbf{g}_i|\mathbf{y}'(i)|^p\right|^2\right)^{1/2}$$

Let $\sigma \geq \max_{i\in S}^n \boldsymbol{\sigma}_i^p(\mathbf{A})$. Then, for any $\mathbf{y},\mathbf{y}' \in B_p(\mathbf{A})$,

$$d_X(\mathbf{y},\mathbf{y}') \leq \begin{cases} 2\|\mathbf{y}-\mathbf{y}'\|_\infty^{p/2} & p<2 \\ 2p\cdot\sigma^{1/2-1/p}\cdot\|\mathbf{y}-\mathbf{y}'\|_\infty & p>2 \end{cases}$$

Proof. Note first that by expanding out the square and noting that $\mathbf{E}[\mathbf{g}_i\mathbf{g}_j] = \mathbb{1}(i=j)$, we have

$$d_X(\mathbf{y},\mathbf{y}') = \left(\sum_{i=1}^n (|\mathbf{y}(i)|^p - |\mathbf{y}'(i)|^p)^2\right)^{1/2}$$

For $p<2$, we bound this as

$$\begin{aligned} d_X(\mathbf{y},\mathbf{y}')^2 &= \sum_{i=1}^n (|\mathbf{y}(i)|^p - |\mathbf{y}'(i)|^p)^2 \\ &= \sum_{i=1}^n (|\mathbf{y}(i)|^{p/2} - |\mathbf{y}'(i)|^{p/2})^2(|\mathbf{y}(i)|^{p/2} + |\mathbf{y}'(i)|^{p/2})^2 \\ &\leq \sum_{i=1}^n (|\mathbf{y}(i)-\mathbf{y}'(i)|^{p/2})^2(|\mathbf{y}(i)|^{p/2} + |\mathbf{y}'(i)|^{p/2})^2 \\ &\leq 2\|\mathbf{y}-\mathbf{y}'\|_\infty^p \sum_{i=1}^n |\mathbf{y}(i)|^p + |\mathbf{y}'(i)|^p \\ &\leq 4\|\mathbf{y}-\mathbf{y}'\|_\infty^p. \end{aligned}$$

For $p>2$, we have by convexity that

$$|\mathbf{y}(i)|^p - |\mathbf{y}'(i)|^p \leq p|\mathbf{y}(i)-\mathbf{y}'(i)|(|\mathbf{y}(i)|^{p-1} + |\mathbf{y}'(i)|^{p-1})$$

and that $\|\mathbf{y}\|_\infty \le \sigma^{1/p}$, so we have

$$\begin{aligned}
d_X(\mathbf{y},\mathbf{y}')^2 &= \sum_{i=1}^n (|\mathbf{y}(i)|^p - |\mathbf{y}'(i)|^p)^2 \\
&\le p^2 \sum_{i=1}^n |\mathbf{y}(i) - \mathbf{y}'(i)|^2 (|\mathbf{y}(i)|^{p-1} + |\mathbf{y}'(i)|^{p-1})^2 \\
&\le 2p^2 \|\mathbf{y} - \mathbf{y}'\|_\infty^2 \sum_{i=1}^n |\mathbf{y}(i)|^{2p-2} + |\mathbf{y}'(i)|^{2p-2} \\
&\le 2p^2 \max\{\|\mathbf{y}\|_\infty, \|\mathbf{y}'\|_\infty\}^{p-2} \|\mathbf{y} - \mathbf{y}'\|_\infty^2 \sum_{i=1}^n |\mathbf{y}(i)|^p + |\mathbf{y}'(i)|^p \\
&\le 4p^2 \sigma^{1-2/p} \|\mathbf{y} - \mathbf{y}'\|_\infty^2
\end{aligned}$$

□

Lemma C.4. *Let $1 \le p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\sigma \ge \max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A})$. Then, the diameter of $B^p(\mathbf{A})$ with respect to d_X is bounded by*

$$\operatorname{diam}(B^p(\mathbf{A})) \le \begin{cases} 4 \cdot \sigma^{1/2} & p < 2 \\ 4p \cdot \sigma^{1/2} & p > 2 \end{cases}$$

Proof. For any $\mathbf{y} \in B^p(\mathbf{A})$, we have that $\|\mathbf{y}\|_\infty \le \sigma^{1/p}$, so combining the triangle inequality and Lemma C.3 yields the result. □

C.2. Computing the Entropy Integral

We may now evaluate the entropy integral required in Theorem C.1. We use the following calculus lemma:

Lemma C.5. *Let $0 < \lambda \le 1$. Then,*

$$\int_0^\lambda \sqrt{\log \frac{1}{t}}\, dt = \lambda\sqrt{\log(1/\lambda)} + \frac{\sqrt{\pi}}{4} \operatorname{erfc}(\sqrt{\log(1/\lambda)}) \le \lambda\left(\sqrt{\log(1/\lambda)} + \frac{\sqrt{\pi}}{2}\right)$$

Proof. We calculate

$$\begin{aligned}
\int_0^\lambda \sqrt{\log \frac{1}{t}}\, dt &= 2\int_{\sqrt{\log(1/\lambda)}}^\infty x^2 \exp(-x^2)\, dx && x = \sqrt{\log(1/t)} \\
&= -\int_{\sqrt{\log(1/\lambda)}}^\infty x \cdot -2x \exp(-x^2)\, dx \\
&= -\left(x\exp(-x^2)\Big|_{\sqrt{\log(1/\lambda)}}^\infty - \int_{\sqrt{\log(1/\lambda)}}^\infty \exp(-x^2)\, dx \right) && \text{integration by parts} \\
&= \lambda\sqrt{\log\frac{1}{\lambda}} + \frac{\sqrt{\pi}}{2} \operatorname{erfc}\left(\sqrt{\log\frac{1}{\lambda}}\right)
\end{aligned}$$

□

Lemma C.6 (Entropy integral bound for $p < 2$)**.** *Let $1 \le p < 2$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be orthonormal. Let $\tau \ge \max_{i=1}^n \|\mathbf{e}_i^\top \mathbf{A}\|_2^2$ and let $\sigma \ge \max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A})$. Then,*

$$\int_0^\infty \sqrt{\log E(B^p, d_X, t)}\, dt \le O(\tau^{1/2}) \left(\frac{\log d}{2-p} + \log n\right)^{1/2} \log \frac{d\sigma}{\tau}$$

Proof. Note that it suffices to integrate the entropy integral to $\operatorname{diam}(B^p(\mathbf{A}))$ rather than ∞, which is at most $4\sigma^{1/2}$ for $p < 2$ and $4p\sigma^{1/2}$ for $p > 2$ by Lemma C.4.

By Lemma C.3, we have that

$$\log E(B^p, d_X, t) \leq \log E(B^p, 2\|\cdot\|_\infty^{p/2}, t) = \log E(B^p, B^\infty, (t/2)^{2/p})$$

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

$$\log E(B^p, B^\infty, t) \leq O(d) \log \frac{n}{t}$$

so

$$\begin{aligned}
\int_0^\lambda \sqrt{\log E(B^p, B^\infty, t)}\, dt = \int_0^\lambda \sqrt{d \log \frac{n}{t}}\, dt &\leq \lambda\sqrt{d \log n} + \sqrt{d} \int_0^\lambda \sqrt{\log \frac{1}{t}}\, dt \\
&\leq \lambda\sqrt{d \log n} + \sqrt{d}\left(\lambda\sqrt{\log \frac{1}{\lambda}} + \frac{\sqrt{\pi}}{2}\lambda\right) \qquad \text{Lemma C.5} \\
&\leq O(\lambda)\sqrt{d \log \frac{n}{\lambda}}
\end{aligned}$$

On the other hand, for large radii larger than λ, we use the bounds of Lemma B.10, which gives

$$\log E(B^p, B^\infty, (t/2)^{2/p}) \leq O(1)\frac{1}{t^2}\left(\frac{\log d}{2-p} + \log n\right)\tau$$

so the entropy integral gives a bound of

$$O(1)\left[\left(\frac{\log d}{2-p} + \log n\right)\tau\right]^{1/2} \int_\lambda^{4p\sigma^{1/2}} \frac{1}{t}\, dt = O(1)\left[\left(\frac{\log d}{2-p} + \log n\right)\tau\right]^{1/2} \log \frac{4p\sigma^{1/2}}{\lambda}.$$

We choose $\lambda = \sqrt{\tau/d}$, which yields the claimed conclusion. □

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

Lemma C.7 (Entropy integral bound for $p > 2$)**.** *Let $2 < p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$ and let $\sigma \geq \max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A})$. Then,*

$$\int_0^\infty \sqrt{\log E(B^p, d_X, t)}\, dt \leq O(p\tau^{1/2}) \cdot (\sigma n)^{1/2-1/p} (\log n)^{1/2} \cdot \log \frac{p^2 d\sigma}{\tau}$$

Proof. The proof is similar to the case of $p < 2$. We again introduce a parameter λ. For radii below λ, the bound is the same as Lemma C.6. For radii above λ, we use Lemma C.3 to bound

$$\log E(B^p, d_X, t) \leq \log E(B^p, 2p \cdot \sigma^{1/2-1/p} \cdot \|\cdot\|_\infty, t) \leq \log E(B^p, B^\infty, t/2p \cdot \sigma^{1/2-1/p})$$

Then by Corollary B.8,

$$\begin{aligned}
\log E(B^p, B^\infty, t/2p \cdot \sigma^{1/2-1/p}) &\leq \log E(B^2, B^\infty, t/2p \cdot (\sigma n)^{1/2-1/p}) \\
&\leq O(p^2)\frac{(\log n) \cdot \tau}{t^2} \cdot (\sigma n)^{1-2/p}
\end{aligned}$$

so the entropy integral gives a bound of

$$O(p\tau^{1/2}) \cdot (\sigma n)^{1/2-1/p} (\log n)^{1/2} \cdot \int_\lambda^{\operatorname{diam}(B^p(\mathbf{A}))} \frac{1}{t}\, dt \leq O(p\tau^{1/2}) \cdot (\sigma n)^{1/2-1/p} (\log n)^{1/2} \cdot \log \frac{p\sigma^{1/2}}{\lambda}$$

Choosing $\lambda = \sqrt{\tau/d}$ yields the claimed conclusion. □

C.3. Moment Bounds

Finally, using our tail bound from combining Theorem C.1 with the entropy bounds of Appendix C.2 and Lemma C.4, we obtain the following moment bounds:

Lemma C.8. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $1 \le p < \infty$. Let $\tau \ge \max_{i=1}^n \boldsymbol{\tau}_i(\mathbf{A})$ and let $\sigma \ge \max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A})$. Let*

$$\Lambda \coloneqq \sup_{\|\mathbf{A}\mathbf{x}\|_p \le 1} \left| \sum_{i=1}^n \mathbf{g}_i |[\mathbf{A}\mathbf{x}](i)|^p \right|$$

Let $\mathcal{E} \coloneqq \int_0^\infty \sqrt{\log E(B^p(\mathbf{A}), d_X, u)}\, du$ and $\mathcal{D} = \operatorname{diam}(B^p(\mathbf{A}))$. Then,

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

Proof. By Theorem C.1, we have for $T = B^p(\mathbf{A})$ that

$$\mathbf{Pr}\left\{\Lambda \ge C\left[\int_0^\infty \sqrt{\log E(T, d_X, u)}\, du + z \cdot \operatorname{diam}(T)\right]\right\} \le 2\exp(-z^2)$$

for a constant $C = O(1)$. Then,

$$\begin{aligned}
\mathbf{E}[(\Lambda/\mathcal{D})^l] &= l\int_0^\infty z^l\, \mathbf{Pr}\{\Lambda \ge z\mathcal{D}\}\, dz \\
&\le (2\mathcal{E}/\mathcal{D})^{l+1} + l\int_{2\mathcal{E}/\mathcal{D}}^\infty z^l\, \mathbf{Pr}\{\Lambda \ge z\mathcal{D}\}\, dz \\
&\le (2\mathcal{E}/\mathcal{D})^{l+1} + l\int_{2\mathcal{E}/\mathcal{D}}^\infty z^l\, \mathbf{Pr}\{\Lambda \ge \mathcal{E} + (z/2)\mathcal{D}\}\, dz \\
&\le (2\mathcal{E}/\mathcal{D})^{l+1} + 2l\int_0^\infty z^l \exp(-z^2/4)\, dz \\
&\le (2\mathcal{E}/\mathcal{D})^{l+1} + O(l)^{l/2}
\end{aligned}$$

so

$$\mathbf{E}[\Lambda^l] \le (2\mathcal{E})^l(\mathcal{E}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l.$$

□

D. Sampling Guarantees

We first reduce our proof of sampling guarantees to the problem of bounding a Gaussian process:

Lemma D.1 (Reduction to Gaussian processes)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $1 \le p < \infty$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix (Definition 1.1). Then,*

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{A}\mathbf{x}\|_p=1} \left| \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p - 1 \right|^l \le (2\pi)^{l/2} \mathop{\mathbf{E}}_{\mathbf{S}} \mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)} \sup_{\|\mathbf{A}\mathbf{x}\|_p=1} \left| \sum_{i\in S} \mathbf{g}_i |[\mathbf{S}\mathbf{A}\mathbf{x}](i)|^p \right|^l,$$

where $S \subseteq [n]$ is the set of rows with sampling probability $q_i < 1$.

Proof. By a standard symmetrization argument (Cohen & Peng, 2015; Chen & Derezinski, 2021), we have that

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{A}\mathbf{x}\|_p=1} \left| \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p - 1 \right|^l \le 2^l \mathop{\mathbf{E}}_{\mathbf{S},\boldsymbol{\varepsilon}} \sup_{\|\mathbf{A}\mathbf{x}\|_p\le 1} \left| \sum_{i\in S} \boldsymbol{\varepsilon}_i |[\mathbf{S}\mathbf{A}\mathbf{x}](i)|^p \right|^l,$$

where $\boldsymbol{\varepsilon} \sim \{\pm 1\}^n$ are independent Rademacher variables. In turn, the right hand side is bounded by

$$2^l(\pi/2)^{l/2} \mathop{\mathbf{E}}_{\mathbf{S},\mathbf{g}} \sup_{\|\mathbf{A}\mathbf{x}\|_p\le 1} \left| \sum_{i\in S} \mathbf{g}_i |[\mathbf{S}\mathbf{A}\mathbf{x}](i)|^p \right|^l$$

via a Rademacher-Gaussian comparison theorem (see, e.g., Equation 4.8 of (Ledoux & Talagrand, 1991)). □

D.1. Sensitivity Sampling, $p < 2$

Our first result is a sensitivity sampling guarantee for $p < 2$.

Theorem D.2 (Sensitivity Sampling for $p < 2$)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $1 \le p < 2$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix with sampling probabilities $q_i = \min\{1, 1/n + \boldsymbol{\sigma}_i^p(\mathbf{A})/\alpha\}$ for an oversampling parameter α set to*

$$\begin{aligned}\frac{1}{\alpha} &= \frac{\mathfrak{S}^p(\mathbf{A})^{2/p-1}}{\varepsilon^2}\left[O(l\log n)^{2/p-1}\left(\frac{\log d}{2-p} + \log\frac{l\log n}{\varepsilon}\right)(\log d)^2 + l\right]\\ &= \frac{\mathfrak{S}^p(\mathbf{A})^{2/p-1}}{\varepsilon^2}\operatorname{poly}\left(\log n, \log\frac{1}{\delta}, \frac{1}{2-p}\right)\end{aligned}$$

for

$$l = O\left(\log\frac{1}{\delta} + \log\log n + \log\frac{1}{2-p} + \log\frac{d}{\varepsilon}\right).$$

Then, with probability at least $1-\delta$, simultaneously for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{SAx}\|_p^p = (1 \pm \varepsilon)\|\mathbf{Ax}\|_p^p.$$

Furthermore, with probability at least $1-\delta$, $\mathbf{S}$ samples

$$\frac{\mathfrak{S}^p(\mathbf{A})^{2/p}}{\varepsilon^2}\operatorname{poly}\left(\log n, \log\frac{1}{\delta}, \log\frac{1}{2-p}\right)$$

rows.

Proof. Our approach is to bound

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{Ax}\|_p=1}\left|\|\mathbf{SAx}\|_p^p - 1\right|^l$$

for a large even integer l. Using Lemma D.1, we first bound

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{Ax}\|_p=1}\left|\|\mathbf{SAx}\|_p^p - 1\right|^l \le (2\pi)^{l/2}\mathop{\mathbf{E}}_{\mathbf{S}}\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)}\sup_{\|\mathbf{Ax}\|_p=1}\left|\sum_{i\in S}\mathbf{g}_i|[\mathbf{SAx}](i)|^p\right|^l$$

where $S = \{i \in [n] : q_i < 1\}$. For simplicity of presentation, we assume $S = [n]$, which will not affect our proof.

By Theorem 1.8, there exists a matrix $\mathbf{A}' \in \mathbb{R}^{m_1\times d}$ with $m_1 = O(d(\log d)^3)$ such that

$$\|\mathbf{A}'\mathbf{x}\|_p^p = (1 \pm 1/2)\|\mathbf{Ax}\|_p^p$$

for all $\mathbf{x} \in \mathbb{R}^d$. Furthermore, because $\mathbf{A}'$ in Theorem 1.8 is constructed by random sampling, Lemma A.1 shows that $\mathfrak{S}^p(\mathbf{A}') \le 8\mathfrak{S}^p(\mathbf{A})$ (note that we only need existence of this matrix). We then construct a matrix $\mathbf{A}'' \in \mathbb{R}^{m_2\times d}$ with $m_2 = O(\alpha^{-1}\mathfrak{S}^p(\mathbf{A}) + d(\log d)^3) = O(\alpha^{-1}\mathfrak{S}^p(\mathbf{A}))$ such that

$$\sigma \coloneqq \max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A}'') \le \alpha,$$

$\mathfrak{S}^p(\mathbf{A}') = \mathfrak{S}^p(\mathbf{A}'')$, and $\|\mathbf{A}'\mathbf{x}\|_p = \|\mathbf{A}''\mathbf{x}\|_p$ for all $\mathbf{x} \in \mathbb{R}^d$ by viewing $\mathbf{A}'$ as an $(m_1 + \alpha^{-1}\mathfrak{S}^p(\mathbf{A})) \times d$ matrix with all zeros except for the first m_1 rows and then applying Lemma A.4.

Now let

$$\mathbf{A}''' \coloneqq \begin{pmatrix}\mathbf{A}''\\ \mathbf{SA}\end{pmatrix}$$

be the $(m_2 + n_{\mathbf{S}}) \times d$ matrix formed by the vertical concatenation of $\mathbf{A}''$ with $\mathbf{SA}$, where $n_{\mathbf{S}}$ is the number of rows sampled by $\mathbf{S}$.

Sensitivity Bounds for $\mathbf{A}'''$. We will first bound the ℓ_p sensitivities of $\mathbf{A}'''$. For any row i corresponding to a row of $\mathbf{A}''$, the ℓ_p sensitivities are already bounded by α, and furthermore, ℓ_p sensitivities can clearly only decrease with row additions. For any row i corresponding to a row of $\mathbf{SA}$ that is sampled with probability $q_i < 1$, we have that

$$\frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{A}'''\mathbf{x}\|_p^p} \leq 2\frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{Ax}\|_p^p} \leq 2\frac{1}{q_i}\frac{|[\mathbf{Ax}](i)|^p}{\|\mathbf{Ax}\|_p^p} \leq 2\frac{\boldsymbol{\sigma}_i^p(\mathbf{A})}{q_i} = 2\alpha.$$

Thus, we have that $\boldsymbol{\sigma}_i^p(\mathbf{A}''') \leq 2\alpha$ for every row i of $\mathbf{A}'''$.

With a bound on the ℓ_p sensitivities of $\mathbf{A}'''$ in hand, we may then convert this into a bound on the leverage scores of $\mathbf{A}'''$ using Lemma 2.3, which gives

$$\tau \coloneqq \max_{i=1}^{n} \boldsymbol{\tau}_i(\mathbf{A}''') \leq (2\alpha)^{2/p}(m_2 + n_{\mathbf{S}})^{2/p-1}$$

where $n_{\mathbf{S}}$ is the number of nonzero entries of $\mathbf{S}$.

Moment Bounds on the Sampling Error. We now fix a choice of $\mathbf{S}$, and define

$$F_{\mathbf{S}} \coloneqq \sup_{\|\mathbf{Ax}\|_p=1} \Big| \|\mathbf{SAx}\|_p^p - 1 \Big|.$$

Note that the event that $n_{\mathbf{S}}$ is at least

$$n_{\text{thresh}} \coloneqq O(l \log n)\, \mathbf{E}[n_S] = O(l \log n)\alpha^{-1}\mathfrak{S}^p(\mathbf{A}),$$

occurs with probability at most $\text{poly}(n)^{-l}$ by Chernoff bounds over the randomness of $\mathbf{S}$, and

$$F_{\mathbf{S}}^l \leq \left[1 + \sum_{i=1}^{n} \frac{1}{q_i}\right]^l \leq (n+1)^{2l},$$

and thus this event contributes at most $\text{poly}(n)^{-l}$ to the moment bound $\mathbf{E}\, F_{\mathbf{S}}^l$. Thus, we focus on bounding $\mathbf{E}\, F_{\mathbf{S}}^l$ conditioned on $n_{\mathbf{S}} \leq n_{\text{thresh}}$. Define

$$G_{\mathbf{S}} \coloneqq \sup_{\|\mathbf{A}'''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|$$

for $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$. Then,

$$\|\mathbf{A}'''\mathbf{x}\|_p^p \leq (1 + 2 + F_{\mathbf{S}})\|\mathbf{Ax}\|_p^p$$

so

$$\begin{aligned} F_{\mathbf{S}}^l &\leq 2^l \sup_{\|\mathbf{Ax}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|^l \\ &\leq 2^l (1 + 2 + F_{\mathbf{S}})^l \sup_{\|\mathbf{A}'''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|^l \\ &\leq 2^{2l-1}(3^l + F_{\mathbf{S}}^l) G_{\mathbf{S}}^l. \end{aligned} \tag{8}$$

We then take expectations on both sides with respect to $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$, and bound the right hand side using Lemma C.8, which gives

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \leq (2\mathcal{E})^l \frac{\mathcal{E}}{\mathcal{D}} + O(\sqrt{l}\mathcal{D})^l$$

where $\mathcal{E}$ is the entropy integral and $\mathcal{D} = 4\sigma^{1/2}$ is the diameter by Lemma C.4. We have by Lemma C.6 that

$$\begin{aligned} \mathcal{E} &\leq O(\tau^{1/2})\left(\frac{\log d}{2-p} + \log(m_2 + n_{\mathbf{S}})\right)^{1/2} \log \frac{d\sigma}{\tau} \\ &\leq O(\alpha^{1/p}(m_2 + n_{\mathbf{S}})^{1/p-1/2})\left(\frac{\log d}{2-p} + \log(m_2 + n_{\mathbf{S}})\right)^{1/2} \log \frac{d\sigma}{\tau} \end{aligned}$$

Thus, conditioned on $n_{\mathbf{S}} \leq n_{\text{thresh}}$, we have that

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \leq \left[O(\alpha^{1/p} n_{\text{thresh}}^{1/p-1/2})\left(\frac{\log d}{2-p} + \log n_{\text{thresh}}\right)^{1/2} \log d\right]^l + O(\sqrt{l}\sqrt{\alpha})^l.$$

Note that

$$\alpha^{1/p} n_{\text{thresh}}^{1/p-1/2} = O(l \log n)^{1/p-1/2} \alpha^{1/p} (\alpha^{-1}\mathfrak{S}^p(\mathbf{A}))^{1/p-1/2} = O(l \log n)^{1/p-1/2} \alpha^{1/2} \mathfrak{S}^p(\mathbf{A})^{1/p-1/2},$$

which shows that

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \leq \varepsilon^l \delta$$

due to our choice of α and l.

Now if we take conditional expectations on both sides of (8) conditioned on the event $\mathcal{F}$ that $n_{\mathbf{S}} \leq n_{\text{thresh}}$, then we have

$$\mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}] \leq 2^{2l-1}(3^l + \mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}])\varepsilon^l \delta \leq (3^l + \mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}])(4\varepsilon)^l \delta$$

which means

$$\mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}] \leq \frac{(12\varepsilon)^l \delta}{1-(4\varepsilon)^l \delta} \leq 2(12\varepsilon)^l \delta$$

for $(4\varepsilon)^l \delta \leq 1/2$. We thus have

$$\mathbf{E}[F_{\mathbf{S}}^l] \leq \frac{(12\varepsilon)^l \delta}{1-(4\varepsilon)^l \delta} \leq 2(12\varepsilon)^l \delta + \text{poly}(n)^{-l}$$

altogether. Finally, we have by a Markov bound that

$$F_{\mathbf{S}}^l \leq 2(12\varepsilon)^l + \frac{1}{\delta}\,\text{poly}(n)^l \leq 3(12\varepsilon)^l$$

with probability at least $1-\delta$, which means that

$$F_{\mathbf{S}} \leq 3 \cdot 12\varepsilon = 36\varepsilon$$

with probability at least $1-\delta$. Rescaling ε by constant factors yields the claimed result. □

D.2. Sensitivity Sampling, $p > 2$

For $p > 2$, we first need a construction of a matrix with a small number of rows and small sensitivity. While this construction can be made to be a randomized algorithm succeeding with high probability, it uses a sophisticated recursive sampling strategy. While this is necessary for our results later in Theorem D.10, such a complicated algorithm may be undesirable. In Theorem D.4, we use this result to show that a more direct one-shot sensitivity sampling can in fact achieve a similar guarantee.

Lemma D.3 (Recursive Sensitivity Sampling)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $2 < p < \infty$. Let $0 < \varepsilon < 1$. Then, there exists a matrix $\mathbf{A}' \in \mathbb{R}^{m\times d}$ for*

$$m = O(p^2)\frac{\mathfrak{S}^p(\mathbf{A})^{2-2/p}}{\varepsilon^2}\log(pd)^2 \log\frac{pd}{\varepsilon}$$

such that

$$\|\mathbf{A}'\mathbf{x}\|_p^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p$$

for every $\mathbf{x} \in \mathbb{R}^d$ and $\mathfrak{S}^p(\mathbf{A}') \leq (1+O(\varepsilon))\mathfrak{S}^p(\mathbf{A})$.

Proof. Let $\mathbf{A}' \in \mathbb{R}^{m\times d}$ be the flattened isometric matrix given by Lemma A.4 with $C = 4$, where $m \leq (5/4)n$. Then for all $i \in [m]$, we have that

$$\boldsymbol{\sigma}_i^p(\mathbf{A}') \leq 4\frac{\mathfrak{S}^p(\mathbf{A})}{n} \leq 5\frac{\mathfrak{S}^p(\mathbf{A}')}{m}.$$

Now consider the random sampling matrix $\mathbf{S}$ with sampling probabilities $q_i = 1/2$. Note then that sampling with probability $q_i = 1/2$ and scaling by $1/q_i = 2$ corresponds to muliplying by the random variable $\boldsymbol{\varepsilon}_i + 1$, where $\boldsymbol{\varepsilon}_i$ is a Rademacher variable. Thus,

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{A}'\mathbf{x}\|_p=1} \left| \|\mathbf{S}\mathbf{A}'\mathbf{x}\|_p^p - 1 \right| = \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}} \sup_{\|\mathbf{A}'\mathbf{x}\|_p=1} \left| \sum_{i=1}^n (\boldsymbol{\varepsilon}_i + 1)|[\mathbf{A}'\mathbf{x}](i)|^p - |[\mathbf{A}'\mathbf{x}](i)|^p \right| = \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}} \sup_{\|\mathbf{A}'\mathbf{x}\|_p=1} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i |[\mathbf{A}'\mathbf{x}](i)|^p \right|.$$

By Lemma D.1 and Theorem C.2, this is bounded by

$$O(1) \int_0^\infty \sqrt{\log E(T, d_X, u)}\, du \leq O(p\tau^{1/2}) \cdot (\sigma n)^{1/2-1/p} (\log n)^{1/2} \cdot \log \frac{p^2 d\sigma}{\tau}$$

where τ is an upper bound on the leverage scores of $\mathbf{A}'$ and σ is an upper bound on the ℓ_p sensitivities of $\mathbf{A}'$. By Lemma 2.2, we have that $\tau \leq \sigma$, and furthermore, we can take $\sigma = 5\mathfrak{S}^p(\mathbf{A}')/m$. Thus, the resulting bound on the expected sampling error is at most

$$\varepsilon_{\mathbf{A}} \coloneqq O(p) \frac{\mathfrak{S}^p(\mathbf{A})^{1-1/p}}{\sqrt{n}} (\log n)^{1/2} \log(pd)$$

so with probability at least $99/100$, the same bound holds up to a factor of 100. Furthermore, $\mathbf{S}$ samples $m/2 \leq (5/8)n$ rows in expectation, so by Markov's inequality, it samples at most $(3/2)m/2 \leq (15/16)n$ rows with probability at least $1/3$. We also have that

$$\frac{\boldsymbol{\sigma}_i^p(\mathbf{A})}{q_i} = 2\boldsymbol{\sigma}_i^p(\mathbf{A}') \leq 10 \frac{\mathfrak{S}^p(\mathbf{A}')}{m}$$

so by Lemma A.2, we have that

$$\mathbf{Pr}\{\mathfrak{S}^p(\mathbf{S}\mathbf{A}') = (1 \pm O(\varepsilon_{\mathbf{A}}))\mathfrak{S}^p(\mathbf{A})\} \geq \frac{99}{100}.$$

By a union bound, $\mathbf{S}\mathbf{A}'$ samples at most $(15/16)n$ rows, has sampling error at most ε_n, and has ℓ_p total sensitivity at most $(1 + O(\varepsilon_{\mathbf{A}}))\mathfrak{S}^p(\mathbf{A})$ with probability at least $1/3 - 1/100 - 1/100 > 0$. Thus, such an instantiation of $\mathbf{S}\mathbf{A}'$ exists.

We now recursively apply our reasoning, by repeatedly applying the flattening and sampling operation. Note that each time we repeat this procedure, the number of rows goes down by a factor of $15/16$, while the total sensitivity and total sampling error accumulates. Let $\mathbf{A}_l$ denote the matrix obtained after l recursive applications of this procedure and let n_l denote the number of rows of $\mathbf{A}_l$. Then,

$$\begin{aligned}
\varepsilon_{\mathbf{A}_{l+1}} &= O(p) \frac{\mathfrak{S}^p(\mathbf{A}_{l+1})^{1-1/p}}{\sqrt{n}_{l+1}} (\log n_{l+1})^{1/2} \log(pd) \\
&\geq (1 - O(\varepsilon_{\mathbf{A}_l}))O(p) \frac{\mathfrak{S}^p(\mathbf{A}_l)^{1-1/p}}{\sqrt{n}_{l+1}} (\log n_{l+1})^{1/2} \log(pd) \\
&\geq \sqrt{\frac{16}{15}}(1 - O(\varepsilon_{\mathbf{A}_l}))O(p) \frac{\mathfrak{S}^p(\mathbf{A}_l)^{1-1/p}}{\sqrt{n}_l} (\log n_l)^{1/2} \log(pd) \\
&\geq \frac{101}{100} \cdot \varepsilon_{\mathbf{A}_l}
\end{aligned}$$

as long as $\varepsilon_{\mathbf{A}_l}$ is less than some absolute constant. Thus, the sum of the $\varepsilon_{\mathbf{A}_l}$ are dominated by the last $\varepsilon_{\mathbf{A}_l}$, up to a constant factor. Now let L be the smallest integer l such that $\varepsilon_{\mathbf{A}_l} \leq \varepsilon$. Then, we have that

$$\mathfrak{S}^p(\mathbf{A}_L) \leq (1 + O(\varepsilon))\mathfrak{S}^p(\mathbf{A})$$

and thus

$$\|\mathbf{A}_L\mathbf{x}\|_p^p = (1 \pm O(\varepsilon))\|\mathbf{A}\mathbf{x}\|_p^p$$

for every $\mathbf{x} \in \mathbb{R}^d$. Furthermore, n_L satisfies

$$\varepsilon = O(p) \frac{\mathfrak{S}^p(\mathbf{A})^{1-1/p}}{\sqrt{n}_L} (\log n_L)^{1/2} \log(pd)$$

or

$$n_L = O(p^2)\frac{\mathfrak{S}^p(\mathbf{A})^{2-2/p}}{\varepsilon^2}\log(pd)^2\log\frac{pd}{\varepsilon}.$$

□

Theorem D.4 (Sensitivity Sampling for $p > 2$)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $2 < p < \infty$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix with sampling probabilities $q_i = \min\{1, 1/n + \boldsymbol{\sigma}_i^p(\mathbf{A})/\alpha\}$ for an oversampling parameter α set to*

$$\frac{1}{\alpha} = O(p^2)\mathfrak{S}^p(\mathbf{A})^{1-2/p}(l\log n)^{1-2/p}\log(pd)\log\frac{l\log n}{\varepsilon} + O(p^2)l$$

for

$$l = O\left(\log\frac{1}{\delta} + \log\log n + \log p + \log\frac{\mathfrak{S}^p(\mathbf{A})}{\varepsilon}\right).$$

Then, with probability at least $1-\delta$, simultaneously for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{SAx}\|_p^p = (1 \pm \varepsilon)\|\mathbf{Ax}\|_p^p.$$

Furthermore, with probability at least $1-\delta$, $\mathbf{S}$ samples

$$\frac{\mathfrak{S}^p(\mathbf{A})^{2-2/p}}{\varepsilon^2}\operatorname{poly}\left(\log n, \log\frac{1}{\delta}, p\right)$$

rows.

Proof. Our approach is to bound

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{Ax}\|_p=1}\left|\|\mathbf{SAx}\|_p^p - 1\right|^l$$

for a large even integer l. Using Lemma D.1, we first bound

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{Ax}\|_p=1}\left|\|\mathbf{SAx}\|_p^p - 1\right|^l \le (2\pi)^{l/2}\mathop{\mathbf{E}}_{\mathbf{S}}\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)}\sup_{\|\mathbf{Ax}\|_p=1}\left|\sum_{i\in S}\mathbf{g}_i|[\mathbf{SAx}](i)|^p\right|^l$$

where $S = \{i \in [n] : q_i < 1\}$. For simplicity of presentation, we assume $S = [n]$, which will not affect our proof.

By Lemma D.3, there exists a matrix $\mathbf{A}' \in \mathbb{R}^{m_1\times d}$ with $m_1 = O(\mathfrak{S}^{2-2/p}\log(pd)^3)$ such that

$$\|\mathbf{A}'\mathbf{x}\|_p^p = (1 \pm 1/2)\|\mathbf{Ax}\|_p^p$$

for all $\mathbf{x} \in \mathbb{R}^d$, and $\mathfrak{S}^p(\mathbf{A}') \le O(1)\mathfrak{S}^p(\mathbf{A})$. Then for $m_2 = O(m_1 + \mathfrak{S}^p(\mathbf{A})\alpha^{-1})$, let $\mathbf{A}'' \in \mathbb{R}^{m_2\times d}$ be the matrix given by Lemma A.4 such that $\boldsymbol{\sigma}_i^p(\mathbf{A}'') \le \alpha$ for every $i \in [m_2]$ and $\|\mathbf{A}''\mathbf{x}\|_p = \|\mathbf{A}'\mathbf{x}\|_p$ for every $\mathbf{x} \in \mathbb{R}^d$. Now let

$$\mathbf{A}''' \coloneqq \begin{pmatrix}\mathbf{A}'' \\ \mathbf{SA}\end{pmatrix}$$

be the $(m_2 + n_{\mathbf{S}}) \times d$ matrix formed by the vertical concatenation of $\mathbf{A}''$ with $\mathbf{SA}$, where $n_{\mathbf{S}}$ is the number of rows sampled by $\mathbf{S}$.

Sensitivity Bounds for $\mathbf{A}'''$. We will first bound the ℓ_p sensitivities of $\mathbf{A}'''$. For any row i corresponding to a row of $\mathbf{A}''$, the ℓ_p sensitivities are already bounded by α, and furthermore, ℓ_p sensitivities can only decrease with row additions. For any row i corresponding to a row of $\mathbf{SA}$ that is sampled with probability $q_i < 1$, we have that

$$\frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{A}'''\mathbf{x}\|_p^p} \le \frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{A}''\mathbf{x}\|_p^p} \le 2\frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{Ax}\|_p^p} \le 2\alpha.$$

By Lemma 2.2, this immediately implies that the ℓ_2 sensitivities, or the leverage scores, are also bounded by 2α.

Moment Bounds on Sampling Error. We now fix a choice of $\mathbf{S}$, and define

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

Note that the event that $n_{\mathbf{S}}$ is at least

$$n_{\text{thresh}} \coloneqq O(l \log n)\, \mathbf{E}[n_S] = O(l \log n)\alpha^{-1}\mathfrak{S}^p(\mathbf{A}),$$

occurs with probability at most $\text{poly}(n)^{-l}$ by Chernoff bounds over the randomness of $\mathbf{S}$, and

$$F_{\mathbf{S}}^l \leq \left[1 + \sum_{i=1}^{n} \frac{1}{q_i}\right]^l \leq (n+1)^{2l},$$

and thus this event contributes at most $\text{poly}(n)^{-l}$ to the moment bound $\mathbf{E}\, F_{\mathbf{S}}^l$. Thus, we focus on bounding $\mathbf{E}\, F_{\mathbf{S}}^l$ conditioned on $n_{\mathbf{S}} \leq n_{\text{thresh}}$. Now define

$$G_{\mathbf{S}} \coloneqq \sup_{\|\mathbf{A}'''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|$$

for $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$. Then,

$$\|\mathbf{A}'''\mathbf{x}\|_p^p \leq (1 + 2 + F_{\mathbf{S}})\|\mathbf{A}\mathbf{x}\|_p^p$$

so

$$\begin{aligned} F_{\mathbf{S}}^l &\leq 2^l \sup_{\|\mathbf{A}\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|^l \\ &\leq 2^l (1+2+F_{\mathbf{S}})^l \sup_{\|\mathbf{A}'''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|^l \\ &\leq 2^{2l-1}(3^l + F_{\mathbf{S}}^l) G_{\mathbf{S}}^l. \end{aligned} \tag{9}$$

We then take expectations on both sides with respect to $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$, and bound the right hand side using Lemma C.8, which gives

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \leq (2\mathcal{E})^l \frac{\mathcal{E}}{\mathcal{D}} + O(\sqrt{l}\mathcal{D})^l$$

where $\mathcal{E}$ is the entropy integral and $\mathcal{D} = 4p\sigma^{1/2}$ is the diameter by Lemma C.4. We have by Lemma C.7 that

$$\begin{aligned} \mathcal{E} &\leq O(p\tau^{1/2}) \cdot (\sigma(m_2+n_{\mathbf{S}}))^{1/2-1/p} (\log(m_2+n_{\mathbf{S}}))^{1/2} \cdot \log \frac{p^2 d\sigma}{\tau} \\ &\leq O(p\alpha^{1/2}) \cdot (\alpha(m_2+n_{\mathbf{S}}))^{1/2-1/p} (\log(m_2+n_{\mathbf{S}}))^{1/2} \cdot \log(pd). \end{aligned}$$

Thus, conditioned on $n_{\mathbf{S}} \leq n_{\text{thresh}}$, we have that

$$\begin{aligned} \mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l &\leq \left[O(p\alpha^{1/2})(\alpha(m_2+n_{\text{thresh}}))^{1/2-1/p}(\log(m_2+n_{\text{thresh}}))^{1/2}\log(pd)\right]^l + O(\sqrt{l}p\sqrt{\alpha})^l \\ &\leq \left[O(p\alpha^{1-1/p}) n_{\text{thresh}}^{1/2-1/p} (\log n_{\text{thresh}})^{1/2} \log(pd)\right]^l + O(\sqrt{l}p\sqrt{\alpha})^l \end{aligned}$$

Note that

$$\alpha^{1-1/p} n_{\text{thresh}}^{1/2-1/p} = O(l\log n)^{1/2-1/p}\alpha^{1-1/p}(\alpha^{-1}\mathfrak{S}^p(\mathbf{A}))^{1/2-1/p} = O(l\log n)^{1/2-1/p}\alpha^{1/2}\mathfrak{S}^p(\mathbf{A})^{1/2-1/p},$$

which shows that

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \leq \varepsilon^l \delta$$

due to our choice of α and l.

Now if we take conditional expectations on both sides of (9) conditioned on the event $\mathcal{F}$ that $n_{\mathbf{S}} \leq n_{\text{thresh}}$, then we have

$$\mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}] \leq 2^{2l-1}(3^l + \mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}])\varepsilon^l\delta \leq (3^l + \mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}])(4\varepsilon)^l\delta$$

which means

$$\mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}] \leq \frac{(12\varepsilon)^l\delta}{1-(4\varepsilon)^l\delta} \leq 2(12\varepsilon)^l\delta$$

for $(4\varepsilon)^l\delta \leq 1/2$. We thus have

$$\mathbf{E}[F_{\mathbf{S}}^l] \leq \frac{(12\varepsilon)^l\delta}{1-(4\varepsilon)^l\delta} \leq 2(12\varepsilon)^l\delta + \text{poly}(n)^{-l}$$

altogether. Finally, we have by a Markov bound that

$$F_{\mathbf{S}}^l \leq 2(12\varepsilon)^l + \frac{1}{\delta}\,\text{poly}(n)^l \leq 3(12\varepsilon)^l$$

with probability at least $1-\delta$, which means that

$$F_{\mathbf{S}} \leq 3 \cdot 12\varepsilon = 36\varepsilon$$

with probability at least $1-\delta$. Rescaling ε by constant factors yields the claimed result. □

D.3. Root Leverage Score Sampling, $p < 2$

We start with a flattening lemma, which shows how to obtain an ℓ_p isometry that simultaneously flatten all ℓ_q sensitivities.

Lemma D.5 (Flattening All Sensitivities)**.** *Let $1 \leq p < \infty$ and $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $0 < \alpha < 1$. Then, there exists $\mathbf{A}' \in \mathbb{R}^{m\times d}$ for $m = O(n\alpha^{-1})$ such that*

$$\boldsymbol{\sigma}_i^q(\mathbf{A}') \leq \alpha$$

for every $i \in [m]$ and $1 \leq q < \infty$. Furthermore, for any $1 \leq q < \infty$ and $\mathbf{x} \in \mathbb{R}^d$, we have that $\|\mathbf{A}'\mathbf{x}\|_q = \Theta(\alpha^{1/p-1/q})\|\mathbf{A}\mathbf{x}\|_q$.

Proof. Let $k \coloneqq \lceil 1/\alpha \rceil$. Then, we construct $\mathbf{A}' \in \mathbb{R}^{m\times d}$ for $m = nk$ by replacing the ith row $\mathbf{a}_i$ of $\mathbf{A}$ for every $i \in [n]$ with k copies of $\mathbf{a}/k^{1/p}$. Then, for every row $j \in [m]$ that is a copy of row $i \in [n]$, we have that

$$\boldsymbol{\sigma}_j^q(\mathbf{A}) = \sup_{\mathbf{A}\mathbf{x}\neq 0} \frac{|[\mathbf{A}'\mathbf{x}](i)|^q}{\|\mathbf{A}'\mathbf{x}\|_q^q} \leq \sup_{\mathbf{A}\mathbf{x}\neq 0} \frac{\left|[k^{-1/p}\mathbf{A}\mathbf{x}](i)\right|^q}{k \cdot \left|[k^{-1/p}\mathbf{A}\mathbf{x}](i)\right|^q} \leq \frac{1}{k} \leq \alpha$$

as desired. The second conclusion holds since

$$\|\mathbf{A}'\mathbf{x}\|_q^q = k \cdot k^{-q/p}\|\mathbf{A}\mathbf{x}\|_q^q = k^{1-q/p}\|\mathbf{A}\mathbf{x}\|_q^q.$$

□

Theorem D.6 (Root Leverage Score Sampling)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $1 \leq p < 2$. Let $0 < \varepsilon, \delta < 1$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix with sampling probabilities $q_i = \min\{1, \boldsymbol{\tau}_i(\mathbf{A})^{p/2}/\alpha\}$ for an oversampling parameter α set to*

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

Then, with probability at least $1-\delta$, simultaneously for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p.$$

Furthermore, with probability at least $1-\delta$, $\mathbf{S}$ samples

$$\frac{n^{1-p/2}d^{p/2}}{\varepsilon^2}\,\text{poly}\left(\log n, \log\frac{1}{\delta}, \frac{1}{2-p}\right)$$

rows.

Proof. Our approach is to bound

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{Ax}\|_p=1} \left| \|\mathbf{SAx}\|_p^p - 1 \right|^l$$

for a large even integer l. Using Lemma D.1, we first bound

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{Ax}\|_p=1} \left| \|\mathbf{SAx}\|_p^p - 1 \right|^l \leq (2\pi)^{l/2} \mathop{\mathbf{E}}_{\mathbf{S}} \mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)} \sup_{\|\mathbf{Ax}\|_p=1} \left| \sum_{i\in S} \mathbf{g}_i |[\mathbf{SAx}](i)|^p \right|^l$$

where $S = \{i \in [n] : q_i < 1\}$. For simplicity of presentation, we assume $S = [n]$, which will not affect our proof.

By Lemma D.5, there exists a matrix $\mathbf{A}' \in \mathbb{R}^{m_1 \times d}$ with $m_1 = O(n/\alpha)$ such that $\|\mathbf{A}'\mathbf{x}\|_p = \|\mathbf{Ax}\|_p$ and $\|\mathbf{A}'\mathbf{x}\|_2 = \Theta(\alpha^{1/p-1/2})\|\mathbf{Ax}\|_2$ for all $\mathbf{x} \in \mathbb{R}^d$, and $\boldsymbol{\tau}_i(\mathbf{A}) = \boldsymbol{\sigma}_i^2(\mathbf{A}) \leq \alpha$ and $\boldsymbol{\sigma}_i^p(\mathbf{A}) \leq \alpha$ for all $i \in [n]$. Now let

$$\mathbf{A}'' \coloneqq \begin{pmatrix} \mathbf{A}' \\ \mathbf{SA} \end{pmatrix}$$

be the $(m_1 + n_{\mathbf{S}}) \times d$ matrix formed by the vertical concatenation of $\mathbf{A}'$ with $\mathbf{SA}$, where $n_{\mathbf{S}}$ is the number of rows sampled by $\mathbf{S}$.

Leverage Score Bounds for $\mathbf{A}''$. We will first bound the leverage scores (or ℓ_2 sensitivities) of $\mathbf{A}''$. For any row i corresponding to a row of $\mathbf{A}'$, the ℓ_2 sensitivities are already bounded by α, and furthermore, ℓ_2 sensitivities can clearly only decrease with row additions. For any row i corresponding to a row of $\mathbf{SA}$ that is sampled with probability $q_i < 1$, we have that

$$\frac{|[\mathbf{SAx}](i)|^2}{\|\mathbf{A}''\mathbf{x}\|_2^2} \leq \frac{|[\mathbf{SAx}](i)|^2}{\|\mathbf{A}'\mathbf{x}\|_2^2} = \frac{|[\mathbf{SAx}](i)|^2}{\Theta(\alpha^{2/p-1})\|\mathbf{Ax}\|_2^2} \leq \frac{1}{q_i^{2/p}} \frac{|[\mathbf{Ax}](i)|^2}{\Theta(\alpha^{2/p-1})\|\mathbf{Ax}\|_2^2} \leq \frac{\boldsymbol{\tau}_i(\mathbf{A})}{\Theta(\alpha^{2/p-1})q_i^{2/p}} = O(\alpha).$$

Thus, we have that $\boldsymbol{\tau}_i(\mathbf{A}'') = \boldsymbol{\sigma}_i^2(\mathbf{A}'') \leq O(\alpha)$ for every row i of $\mathbf{A}''$. By monotonicity of max sensitivity Lemma 2.2, we also have that $\boldsymbol{\sigma}_i^p(\mathbf{A}) \leq O(\alpha)$.

Moment Bounds on the Sampling Error. We now fix a choice of $\mathbf{S}$, and define

$$F_{\mathbf{S}} \coloneqq \sup_{\|\mathbf{Ax}\|_p=1} \left| \|\mathbf{SAx}\|_p^p - 1 \right|$$

$$G_{\mathbf{S}} \coloneqq \sup_{\|\mathbf{A}''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}''\mathbf{x}](i)|^p \right|$$

for $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$. Then,

$$\|\mathbf{A}''\mathbf{x}\|_p^p \leq (1 + 2 + F_{\mathbf{S}})\|\mathbf{Ax}\|_p^p$$

so

$$\begin{aligned} F_{\mathbf{S}}^l &\leq 2^l \sup_{\|\mathbf{Ax}\|_p=1} \left| \sum_{i=1}^{m_1+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}''\mathbf{x}](i)|^p \right|^l \\ &\leq 2^l (1 + 2 + F_{\mathbf{S}})^l \sup_{\|\mathbf{A}''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_1+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}''\mathbf{x}](i)|^p \right|^l \\ &\leq 2^{2l-1}(3^l + F_{\mathbf{S}}^l) G_{\mathbf{S}}^l. \end{aligned} \tag{10}$$

We then take expectations on both sides with respect to $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$, and bound the right hand side using Lemma C.8, which gives

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \leq (2\mathcal{E})^l \frac{\mathcal{E}}{\mathcal{D}} + O(\sqrt{l}\mathcal{D})^l$$

where $\mathcal{E}$ is the entropy integral and $\mathcal{D} = 4\sigma^{1/2} \leq 4\alpha^{1/2}$ is the diameter by Lemma C.4. We have by Lemma C.6 that

$$\begin{aligned}\mathcal{E} &\leq O(\tau^{1/2})\left(\frac{\log d}{2-p} + \log(m_1 + n_{\mathbf{S}})\right)^{1/2} \log \frac{d\sigma}{\tau} \\ &\leq O(\alpha^{1/2})\left(\frac{\log d}{2-p} + \log n\right)^{1/2} \log d\end{aligned}$$

By our choice of α and l, we have

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \leq \varepsilon^l \delta.$$

Now if we take conditional expectations on both sides of (10), then we have

$$\mathbf{E}[F_{\mathbf{S}}^l] \leq 2^{2l-1}(3^l + \mathbf{E}[F_{\mathbf{S}}^l])\varepsilon^l\delta \leq (3^l + \mathbf{E}[F_{\mathbf{S}}^l])(4\varepsilon)^l\delta$$

which means

$$\mathbf{E}[F_{\mathbf{S}}^l] \leq \frac{(12\varepsilon)^l\delta}{1-(4\varepsilon)^l\delta} \leq 2(12\varepsilon)^l\delta$$

for $(4\varepsilon)^l\delta \leq 1/2$. Finally, we have by a Markov bound that $F_{\mathbf{S}}^l \leq 2(12\varepsilon)^l$ with probability at least $1-\delta$, which means that $F_{\mathbf{S}} \leq 2 \cdot 12\varepsilon = 24\varepsilon$ with probability at least $1-\delta$. Rescaling ε by constant factors yields the claimed sampling error bound.

Note that the expected number of rows sampled is at most

$$\frac{1}{\alpha}\sum_{i=1}^n \boldsymbol{\tau}_i(\mathbf{A})^{p/2} \leq \frac{1}{\alpha} n^{1-p/2}\left(\sum_{i=1}^n \boldsymbol{\tau}_i(\mathbf{A})\right)^{p/2} = \frac{1}{\alpha} n^{1-p/2} d^{p/2}$$

by Hölder's inequality. This implies the bound on number of sampled rows by Chernoff bounds. □

Finally, we show that by recursively applying Theorem D.6, we can reduce the number of rows to roughly $d/\varepsilon^{4/p}$. To bound the size, we need to solve a recursion for an upper bound on the number of rows. This is given by the following:

Lemma D.7 (Lemma 6.12, (Musco et al., 2022)). *Suppose $(a_i)_{i=0}^\infty$ satisfies the recurrence $a_{i+1} = \lambda a_i + b$ for some $b > 0$ and $\lambda \in (0,1)$. Then,*

$$a_i = \frac{1}{1-\lambda}\big(b - \lambda^i(b - (1-\lambda)a_0)\big).$$

This gives the following

Theorem D.8 (Recursive Root Leverage Score Sampling). *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $1 \leq p < 2$. Let $0 < \varepsilon, \delta < 1$. Let $\mathbf{S}$ be the result of recursively applying Theorem D.6 with failure probability $\delta/\Theta(\log\log n)$ and accuracy $\varepsilon/\Theta(\log\log n)$ recursively until the number of rows is at most*

$$m = \frac{d}{\varepsilon^{4/p}} \operatorname{poly}\left(\log n, \log\frac{1}{\delta}, \log\frac{1}{2-p}\right)$$

Then, with probability at least $1-\delta$, simultaneously for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p.$$

Proof. We apply Theorem D.6 with failure probability $\delta/\Theta(\log\log n)$ and accuracy $\varepsilon/\Theta(\log\log n)$ recursively for at most $R = O(\log\log n)$ rounds, until the number of rows is at most the claimed bound. By a union bound, we succeed at achieving $\varepsilon/\Theta(\log\log n)$ sampling error and sampling bound on all R rounds, that is, for any number of rows m, we reduce the number of rows to at most

$$\frac{m^{1-p/2}d^{p/2}}{\varepsilon^2} \operatorname{poly}\left(\log n, \log\frac{1}{\delta}, \frac{1}{2-p}\right).$$

We then apply the recurrence lemma Lemma D.7 on the logarithm of the above bound, so $\lambda = (1 - p/2)$ and

$$b = \log\left(\frac{d^{p/2}}{\varepsilon^2}\operatorname{poly}\left(\log n, \log\frac{1}{\delta}, \frac{1}{2-p}\right)\right).$$

Then, after $i = O(\log\log n)$ iterations, our log row count upper bound is

$$a_i = \log m_i = \frac{2}{p}\big(b - (1-p/2)^i(b - (p/2)\log n)\big) \le \frac{2}{p}\big(b + (1-p/2)^i(p/2)\log n\big) \le \frac{2}{p}(b+1)$$

or

$$m_i \le O\left(\frac{d^{p/2}}{\varepsilon^2}\operatorname{poly}\left(\log n, \log\frac{1}{\delta}, \frac{1}{2-p}\right)\right)^{2/p} = \frac{d}{\varepsilon^{4/p}}\operatorname{poly}\left(\log n, \log\frac{1}{\delta}, \frac{1}{2-p}\right)$$

□

D.4. Leverage Score + ℓ_p Sensitivity Sampling, $p > 2$

We start with a flattening lemma for $p > 2$, which shows how to slightly flatten leverage scores while preserving ℓ_p norms, and with only a small blow up in the number of rows.

Lemma D.9 (Flattening ℓ_p Sensitivities and Leverage Scores)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $2 < p < \infty$. Let $C \ge 1$. Then, there exists $\mathbf{A}' \in \mathbb{R}^{m\times d}$ for $m \le (1+1/C)n$ such that $\|\mathbf{A}\mathbf{x}\|_p = \|\mathbf{A}'\mathbf{x}\|_p$ and $\|\mathbf{A}'\mathbf{x}\|_2 \ge \|\mathbf{A}\mathbf{x}\|_2$ for every $\mathbf{x} \in \mathbb{R}^d$ and satisfies*

$$\max_{i\in[m]} \boldsymbol{\sigma}_i^p(\mathbf{A}') \le \max_{i\in[n]} \boldsymbol{\sigma}_i^p(\mathbf{A})$$
$$\max_{i\in[m]} \boldsymbol{\tau}_i(\mathbf{A}') \le (Cd/n)^{2/p} \max_{i\in[m]} \boldsymbol{\tau}_i(\mathbf{A}')^{1-2/p}$$
$$\mathfrak{S}^p(\mathbf{A}') = \mathfrak{S}^p(\mathbf{A})$$

Proof. The idea roughly follows Lemma A.4, except that we split rows that have large leverage score, rather than rows that have large ℓ_p sensitivity. For each $i \in [n]$, let $k_i = \lceil \boldsymbol{\tau}_i(\mathbf{A})/(Cd/n)\rceil$. Then for each $i \in [n]$ such that $k_i > 1$, we replace ith row $\mathbf{a}_i$ of $\mathbf{A}$ by k_i copies of $\mathbf{a}_i/k^{1/p}$. Clearly, we have that $\|\mathbf{A}'\mathbf{x}\|_p = \|\mathbf{A}\mathbf{x}\|_p$ for every $\mathbf{x} \in \mathbb{R}^d$. Furthermore, the number of rows added is at most

$$\sum_{i:\boldsymbol{\tau}_i(\mathbf{A})\ge Cd/n} \left\lceil \frac{\boldsymbol{\tau}_i(\mathbf{A})}{Cd/n}\right\rceil - 1 \le \sum_{i:\boldsymbol{\tau}_i(\mathbf{A})\ge Cd/n} \frac{\boldsymbol{\tau}_i(\mathbf{A})}{Cd/n} = \frac{d}{Cd/n} = \frac{n}{C}.$$

Next, note that for every $\mathbf{x} \in \mathbb{R}^d$,

$$k_i \cdot \left|k_i^{-2/p}[\mathbf{A}\mathbf{x}](i)\right|^2 \ge |[\mathbf{A}\mathbf{x}](i)|^2$$

since $k_i \ge 1$, so we have that $\|\mathbf{A}'\mathbf{x}\|_2 \ge \|\mathbf{A}\mathbf{x}\|_2$. Then, for any row $j \in [m]$ that is a copy of row $i \in [n]$ of $\mathbf{A}$, we have that

$$\boldsymbol{\tau}_j(\mathbf{A}') = \sup_{\mathbf{A}\mathbf{x}\ne 0} \frac{|[\mathbf{A}'\mathbf{x}](i)|^2}{\|\mathbf{A}'\mathbf{x}\|_2^2} \le \sup_{\mathbf{A}\mathbf{x}\ne 0} \frac{k_i^{-2/p}|[\mathbf{A}\mathbf{x}](i)|^2}{\|\mathbf{A}\mathbf{x}\|_2^2} \le \frac{(Cd/n)^{2/p}}{\boldsymbol{\tau}_i(\mathbf{A})^{2/p}}\boldsymbol{\tau}_i(\mathbf{A}) = (Cd/n)^{2/p}\boldsymbol{\tau}_i(\mathbf{A})^{1-2/p}.$$

Finally, it is clear that the ℓ_p sensitivities can only decrease and that the total ℓ_p sensitivity is preserved. □

Now using Lemma D.9, we first obtain a construction of a small ℓ_p approximate isometry in a way analogous to Lemma D.3.

Theorem D.10 (Recursive Leverage Score + ℓ_p Sensitivity Sampling)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $2 < p < \infty$. Let $0 < \varepsilon, \delta < 1$. Then, there exists an efficient algorithm producing a matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$ for*

$$m = O(p^2)\frac{d^{2/p}\mathfrak{S}^p(\mathbf{A})^{2-4/p}}{\varepsilon^2}\left(\log\frac{pd}{\delta}\right)^2 \log\frac{pd}{\varepsilon}.$$

such that

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p$$

for every $\mathbf{x} \in \mathbb{R}^d$ and $\mathfrak{S}^p(\mathbf{A}') \le (1 + O(\varepsilon))\mathfrak{S}^p(\mathbf{A})$.

Proof. Our proof is almost identical to Lemma D.3, but we will use high probability versions of the results, as we will directly use the recursive sampling procedure algorithmically. A similar algorithmic recursive sampling procedure is considered in (Musco et al., 2022). We first replace $\mathbf{A}'$ in Lemma D.3 with the matrix formed by first applying Lemma A.4 with $C = 4$, and then applying Lemma D.9 with $C = 4$. The resulting $\mathbf{A}'$ has at most $(5/4)^2 n = (25/16)n$ rows, preserves ℓ_p norms and ℓ_p total sensitivity, and has

$$\begin{aligned}\boldsymbol{\sigma}_i^p(\mathbf{A}') &\leq O(1)\frac{\mathfrak{S}^p(\mathbf{A})}{n}\\ \boldsymbol{\tau}_i(\mathbf{A}') &\leq O(1)\left(\frac{d}{n}\right)^{2/p}\left(\frac{\mathfrak{S}^p(\mathbf{A})}{n}\right)^{1-2/p} = O(1)\frac{d^{2/p}\mathfrak{S}^p(\mathbf{A})^{1-2/p}}{n}\end{aligned} \tag{11}$$

We then again consider sampling half of the rows of $\mathbf{A}'$ via an ℓ_p sampling matrix $\mathbf{S}$ with $q_i = 1/2$. By Lemma D.1 and Theorem C.1, we then have that

$$\mathbf{Pr}\left\{\sup_{\|\mathbf{A}'\mathbf{x}\|_p=1}\left|\|\mathbf{SA}'\mathbf{x}\|_p^p - 1\right| \leq O(p\tau^{1/2})\cdot(\sigma n)^{1/2-1/p}(\log n)^{1/2}\cdot\log\frac{p^2 d\sigma}{\tau} + 4p\sigma^{1/2}\cdot z\right\} \geq 1 - 2\exp(z^2)$$

where τ is an upper bound on the leverage scores of $\mathbf{A}'$ and σ is an upper bound on the ℓ_p sensitivities of $\mathbf{A}'$. These are bounded by (11), and thus applying these bounds and setting $z = O(\log(n/\delta))$ gives

$$\mathbf{Pr}\left\{\sup_{\|\mathbf{A}'\mathbf{x}\|_p=1}\left|\|\mathbf{SA}'\mathbf{x}\|_p^p - 1\right| \leq O(p)\frac{d^{1/p}\mathfrak{S}^p(\mathbf{A}')^{1-2/p}}{\sqrt{n}}\left((\log n)^{1/2}\cdot\log(pd) + \log\frac{\log n}{\delta}\right)\right\} \geq 1 - \frac{\delta}{\operatorname{poly}\log n}.$$

By a union bound, the same bound holds for the first $O(\log n)$ recursive calls to our recursive sampling algorithm, up to a different $\operatorname{poly}\log n$ factor in the denominator of the failure rate bound. Furthermore, by Chernoff bounds, we have that the number $n_{\mathbf{S}}$ of rows sampled by $\mathbf{S}$

$$\mathbf{Pr}\{n_{\mathbf{S}} \leq (25/32)(5/4)n < 0.98n\} \geq 1 - \exp\left(-\frac{1}{3}\cdot\frac{1}{16}\cdot\frac{25}{32}n\right) \geq 1 - \frac{\delta}{\operatorname{poly}\log n}$$

as long as $n \geq C\log\frac{1}{\delta}$ for a sufficiently large constant C. By a union bound, the same bound also holds for the first $O(\log n)$ recursive calls to our sampling algorithm. In this case, our sampling process terminates in at most $O(\log n)$ rounds, so both the bound on $n_{\mathbf{S}}$ and the sampling error bound hold for all $O(\log n)$ rounds.

We now apply the same reasoning as Lemma D.3 to bound the total sampling error. Let $\mathbf{A}_l$ denote the $n_l \times d$ sampled matrix after l rounds of the recursive sampling procedure, and let

$$\varepsilon_{\mathbf{A}_l} = O(p)\frac{d^{1/p}\mathfrak{S}^p(\mathbf{A}_l)^{1-2/p}}{\sqrt{n}_l}\left((\log n_l)^{1/2}\cdot\log(pd) + \log\frac{\log n_l}{\delta}\right).$$

Then,

$$\begin{aligned}\varepsilon_{\mathbf{A}_{l+1}} &= O(p)\frac{d^{1/p}\mathfrak{S}^p(\mathbf{A}_{l+1})^{1-2/p}}{\sqrt{n}_{l+1}}\left((\log n_{l+1})^{1/2}\log(pd) + \log\frac{\log n_{l+1}}{\delta}\right)\\ &\geq (1 - O(\varepsilon_{\mathbf{A}_l}))O(p)\frac{d^{1/p}\mathfrak{S}^p(\mathbf{A}_l)^{1-2/p}}{\sqrt{n}_{l+1}}\left((\log n_{l+1})^{1/2}\log(pd) + \log\frac{\log n_{l+1}}{\delta}\right)\\ &\geq \sqrt{\frac{1}{0.98}}(1 - O(\varepsilon_{\mathbf{A}_l}))O(p)\frac{d^{1/p}\mathfrak{S}^p(\mathbf{A}_l)^{1-2/p}}{\sqrt{n}_l}\left((\log n_l)^{1/2}\log(pd) + \log\frac{\log n_l}{\delta}\right)\\ &\geq \frac{101}{100}\cdot\varepsilon_{\mathbf{A}_l}\end{aligned}$$

so the sum of the $\varepsilon_{\mathbf{A}_l}$ are dominated by the last $\varepsilon_{\mathbf{A}_l}$, up to a constant factor. Now let L be the smallest integer l such that $\varepsilon_{\mathbf{A}_l} \leq \varepsilon$. Then, we have that

$$\mathfrak{S}^p(\mathbf{A}_L) \leq (1 + O(\varepsilon))\mathfrak{S}^p(\mathbf{A})$$

and thus

$$\|\mathbf{A}_L\mathbf{x}\|_p^p = (1 \pm O(\varepsilon))\|\mathbf{A}\mathbf{x}\|_p^p$$

for every $\mathbf{x} \in \mathbb{R}^d$. Furthermore, n_L satisfies

$$\varepsilon = O(p)\frac{d^{1/p}\mathfrak{S}^p(\mathbf{A})^{1-2/p}}{\sqrt{n}_L}\left((\log n_L)^{1/2}\log(pd) + \log\frac{\log n_L}{\delta}\right)$$

or

$$n_L = O(p^2)\frac{d^{2/p}\mathfrak{S}^p(\mathbf{A})^{2-4/p}}{\varepsilon^2}\left(\log\frac{pd}{\delta}\right)^2\log\frac{pd}{\varepsilon}.$$

□