Online Lewis Weight Sampling

David P. Woodruff
Carnegie Mellon University
dwoodruf@cs.cmu.edu

Taisuke Yasuda
Carnegie Mellon University
taisukey@cs.cmu.edu

Abstract

The seminal work of Cohen and Peng [CP15] (STOC 2015) introduced *Lewis weight sampling* to the theoretical computer science community, which yields fast row sampling algorithms for approximating d-dimensional subspaces of ℓ_p up to $(1+\varepsilon)$ relative error. Several works have extended this important primitive to other settings, including the online coreset model and sliding window models [BDM+20] (FOCS 2020) as well as the adversarial streaming model [BHM+21] (NeurIPS 2021). However, these results are only for $p \in \{1,2\}$, and results for $p = 1$ require a suboptimal $\tilde{O}(d^2/\varepsilon^2)$ samples.

In this work, we design the first nearly optimal ℓ_p subspace embeddings for all $p \in (0,\infty)$ in the online coreset, sliding window, and the adversarial streaming models. In all three models, our algorithms store $\tilde{O}(d/\varepsilon^2)$ rows for $p \in (0,2)$ and $\tilde{O}(d^{p/2}/\varepsilon^2)$ rows for $p \in (2,\infty)$. This answers a substantial generalization of the main open question of [BDM+20], and gives the first results for all $p \notin \{1,2\}$ and achieves nearly optimal sample complexities for all p.

Towards our result, we give the first analysis of "one-shot" Lewis weight sampling of sampling rows proportionally to their Lewis weights, which gives a sample complexity of $\tilde{O}(d^{p/2}/\varepsilon^2)$ rows for $p > 2$. Previously, such a sampling scheme was only known to have a sample complexity of $\tilde{O}(d^{p/2}/\varepsilon^5)$ [CP15], whereas a bound of $\tilde{O}(d^{p/2}/\varepsilon^2)$ is known if a more sophisticated recursive sampling algorithm is used [MMWY21, LT91]. Note that the recursive sampling strategy cannot be implemented in an online setting, thus necessitating an analysis of one-shot Lewis weight sampling. Perhaps surprisingly, our analysis crucially uses a novel connection to online numerical linear algebra, *even for offline Lewis weight sampling.*

As an application, we obtain the first one-pass streaming coreset algorithms for $(1+\varepsilon)$ approximation of important generalized linear models, such as logistic regression and p-probit regression. Our upper bounds are parameterized by a complexity parameter μ introduced by [MSSW18], and we also provide the first lower bounds showing that a linear dependence on μ is necessary.

1 Introduction

We consider the problem of computing ℓ_p *subspace embeddings* in the setting of big data analysis. In this problem, we are given a large input matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ where $n \gg d$, and we seek an approximation $\mathbf{SA} \in \mathbb{R}^{r\times d}$ with $d \leq r \ll n$ such that

$$\text{for all } \mathbf{x} \in \mathbb{R}^d, \qquad \|\mathbf{SAx}\|_p = (1 \pm \varepsilon)\|\mathbf{Ax}\|_p. \tag{1}$$

That is, we seek a small summary of $\mathbf{A}$ which approximates every vector in its column space in the ℓ_p norm, where the summary takes the form $\mathbf{SA}$ for some $r \times n$ matrix $\mathbf{S} \in \mathbb{R}^{r\times n}$. Such a dimensionality reduction result is of fundamental importance to machine learning and theoretical computer science, and the utility of such a result has been proven through a long line of work on this problem [Sar06, DDH+09, SW11, WZ13, MM13, CP15, WW19, LWW21, WY22]. While $\mathbf{S}$ can in principle be any linear map, including a dense matrix (see, e.g., the dense Cauchy sketches of [SW11]), the best known results for ℓ_p subspace embeddings proceed by a sampling approach [DDH+09, CP15, MMWY21], in which $\mathbf{S}$ only has one nonzero entry per row. We focus on such an approach in this work, and refer to the number r of rows of $\mathbf{S}$ as the *sample complexity*.

This problem is also central to the functional analysis literature, and nearly optimal upper bounds have been known since the works of [Lew78, BLM89, Tal90, LT91, Tal95, SZ01, Sch11], which are complemented by recently discovered lower bounds due to [LWW21]. These results were then turned algorithmic by a result of [CP15], which showed that it was possible to compute a sketch $\mathbf{SA}$ satisfying (1) in nearly input sparsity time. These results gave $r = \tilde{O}(d/\varepsilon^2)$ rows for $p \in [1, 2]$ and $r = \tilde{O}(d^{p/2}/\varepsilon^5)$ for $p > 2$. Furthermore, the algorithm of [CP15] had a simple two-step procedure consisting of (1) approximating Lewis weights and then (2) sampling each row proportionally to these weights, making it an attractive algorithm in various settings. It was recently shown how to extend the results to $r = \tilde{O}(d/\varepsilon^2)$ for $p \in (0, 2]$ and $r = \tilde{O}(d^{p/2}/\varepsilon^2)$ by [MMWY21], through the use of a more sophisticated recursive sampling strategy due to [Tal90, LT91].

Despite this recent progress in the algorithmic theory of ℓ_p subspace embeddings, there are important problems that remain. Driven by increasingly challenging practical problems associated with the analysis of modern data sets, recent work in algorithmic data science has focused on more and more restrictive requirements and models of computation. These include:

- the *streaming model*, in which the data set can only be accessed through one pass through a stream of the rows of the data set
- the *online coreset model*, which further restricts the streaming model by only allowing for a small number of rows to be irrevocably stored (i.e., cannot be thrown away to save space)
- the *sliding window model*, in which only the W most recent rows in a stream are considered as the input at any time
- the *adversarial streaming model*, in which the algorithm is allowed to be randomized, but must succeed against an adversary that can specify inputs that can depend on previous outputs of the algorithm

These models of computation address important practical requirements for applications, and we refer to a rich line of previous work, and references therein, on the motivations for studying the online [CMP20, BDM+20], sliding window [BDM+20, UU21, EMMZ22], and adversarial streaming [BJWY22, BHM+21] models.

The streaming model can be addressed quite straightforwardly by employing a standard merge-and-reduce procedure to convert the offline ℓ_p subspace embedding algorithms of [CP15, MMWY21] into streaming algorithms. However, the latter three extensions of the streaming model described above are more challenging to handle. For $p = 2$, [CMP20] gave nearly optimal results for the online coreset model, showing that one can maintain approximately $r = \tilde{O}(d/\varepsilon^2)$ rows in an online manner, while essentially only losing a $\log \kappa^{\mathsf{OL}}$ factor, where $\kappa^{\mathsf{OL}} = \kappa^{\mathsf{OL}}(\mathbf{A})$ is a natural quantity known as the *online condition number of* $\mathbf{A}$ (Definition 3.2). The same work also shows that such a condition number dependence is required [CMP20, Theorem 5.1]. The work of [BDM+20] then gave an elegant reduction from sliding windows to online coreset algorithms, thus achieving similar guarantees in this model. They also showed a deterministic algorithm for ℓ_2 subspace embeddings, which automatically has guarantees in the adversarial streaming model. Thus, for $p = 2$, all of these questions are nearly settled.

On the other hand, for $p \neq 2$, the landscape is far worse, even for $p = 1$. The work of [BDM+20] obtained a bound of $r = \tilde{O}(d^2/\varepsilon^2)$ rows in both the online coreset and sliding window models, which is loose by a factor of d compared to the optimal sample complexity in the offline model. They also give a deterministic sliding window algorithm for $p = 1$ sampling $\tilde{O}(d/\varepsilon^2)$ rows, but this algorithm runs in exponential time. They leave the following as their main open question:

Question 1.1. *Can the sample complexity of ℓ_1 subspace embeddings in the online coreset model be improved from $\tilde{O}(d^2/\varepsilon^2)$ to $\tilde{O}(d/\varepsilon^2)$?*

The work of [BDM+20] was further extended to guarantees in the adversarial streaming model by [BHM+21], also leading to a sample complexity of $\tilde{O}(d^2/\varepsilon^2)$ rows. However, all other $p \in (0, \infty) \setminus \{1, 2\}$ are absent altogether, in both the sliding window model and the adversarial streaming model. While it seems possible to extend the techniques of [BDM+20, BHM+21] to $p \in (0, 2)$, the problems of suboptimal sample complexity or exponential running time would still remain. Additionally, the behavior of Lewis weights changes substantially from $p < 2$ to $p > 2$ (e.g., lack of monotonicity), which breaks many parts of existing approaches.

Furthermore, the consideration of the online model and its variants brings up a natural question on Lewis weight sampling, even in the offline setting:

Question 1.2. *Is it possible to obtain a sample complexity bound of $\tilde{O}(d^{p/2}/\varepsilon^2)$ rows for $p > 2$ using the simple strategy of sampling proportionally to ℓ_p Lewis weights?*

Aside from being a more aesthetically pleasing result than the recursive sampling strategy of [MMWY21], in certain situations such as the online settings which we consider, the application of Lewis weight sampling would only work if the algorithm is a simple scheme of sampling proportionally to weights; the recursive sampling strategy cannot work without knowledge of all of the rows of the matrix. Thus, Question 1.2 is a central unresolved question in the study of Lewis weight sampling. We also note that Lewis weights and their sampling guarantees have been central to many recent advances in machine learning and theoretical computer science [CD21, PPP21, MMWY21] even beyond ℓ_p losses [CWW19, LWYZ20, MRM21, MMWY21], making it even more important to gain an improved understanding of Lewis weight sampling.

1.1 Our Contributions

1.1.1 Online ℓ_p Lewis Weight Sampling

As our first contribution, we answer Question 1.1 affirmatively, achieving an online coreset for ℓ_1 subspace embeddings with

$$O\left(\frac{d}{\varepsilon^2}(\log n)\log(n\kappa^{\mathsf{OL}})\right)$$

rows[1]. In fact, we show much more than this, by obtaining the first online coresets for ℓ_p subspace embeddings for all $p \in (0, \infty) \setminus \{1, 2\}$ with $(1 + \varepsilon)$ error. Our dependence on the dimension d is optimal for all $p \in (0, \infty)$ up to polylogarithmic factors due to known lower bounds for ℓ_p subspace embeddings [LWW21], and our dependence on ε is quadratic for $p \in (0, \infty)$, which is also optimal [LWW21]. Thus, we in fact answer a substantial generalization of Question 1.1. Our results are summarized in Table 1.

	Sample Size	
$p = 1$	d^2/ε^2	[BDM+20, Theorem 4.1]
$p = 2$	d/ε^2	[CMP20, BDM+20]
$0 < p < 2$	d/ε^2	Theorem 6.6
$2 < p < \infty$	$d^{p/2}/\varepsilon^2$	Theorem 6.6

Table 1: Our results for online Lewis weight sampling. We suppress polylogarithmic factors in n, κ^{OL}, ε^{-1}.

[1] Note that one can compose this algorithm with itself, in an online fashion, so that n here can be replaced by $O\left(\frac{d}{\varepsilon^2}(\log n)\log(n\kappa^{\mathsf{OL}})\right)$. For simplicity of presentation, we state our results without this optimization.

In order to obtain our result, we make a key change over prior approaches towards online coresets for ℓ_p subspace embeddings [CMP20, BDM$^+$20]: we decouple the problem of approximating the importance of a row and approximating the matrix itself. That is, we maintain two sketches, one for approximating an online generalization of Lewis weights which we call *online Lewis weights* (see Section 3 for definitions and properties), and one which uses the online Lewis weights as importance scores in order to obtain an ℓ_p subspace embedding. This has two advantages: (1) we can build on prior work for spectral approximation to approximate the Lewis quadratic, and (2) by conditioning on the success of the approximation of Lewis weights, we can simply treat the online sampling process for the ℓ_p subspace embedding exactly as an offline sampling process, which significantly simplifies the analysis. In particular, we avoid complex sequential chaining arguments such as those considered in, e.g., [RST10, BDR21]. This decoupled approach may be of independent interest for future work on matrix approximation and importance sampling, especially in online models and other restricted models of computation.

1.1.2 Offline ℓ_p Lewis Weight Sampling

En route to obtaining the result of Table 1, we answer Question 1.2 in the affirmative, thereby closing a long-standing gap in the study of ℓ_p Lewis weight sampling since [CP15]. We refer to Table 2 for a summary of this result alongside prior results.

Sample Size	Sampling Algorithm	
$d^{p/2}/\varepsilon^5$	One-Shot	[CP15, BLM89]
$d^{p/2}/\varepsilon^2$	Recursive	[MMWY21, LT91]
$d^{p/2}/\varepsilon^2$	One-Shot	Theorems 1.3, 5.2

Table 2: Our results for offline Lewis weight sampling. We suppress polylogarithmic factors in n and ε^{-1}.

Perhaps surprisingly, our result crucially relies on a novel connection to online leverage scores [CMP20], which allows us to circumvent the problem of non-monotonicity of ℓ_p Lewis weights for $p > 2$, which was the major barrier towards achieving this result. We also give the first results for Lewis weight sampling which simultaneously achieve an optimal dependence on d and ε along with a polylogarithmic dependence on the failure rate δ for any $p \neq 2$, which is crucial for certain applications such as ℓ_p subspace embeddings in sliding windows. Prior results had at least one problem of only achieving constant probability of success [CP15], suboptimal dependence on ε [BLM89], or suboptimal dependence on d [Sch87]. Our analysis also allows for the use of ℓ_p Lewis weight approximations which satisfy weaker guarantees than the requirement of upper bounding the true ℓ_p Lewis weights, which can be computed in $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^\omega)$ time [Lee16, JLS21], rather than $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^{O(p)})$ time [CP15]. Thus, our Lewis weight sampling result gives "the best of all worlds" in terms of running time and dependencies on d, ε, and δ, up to logarithmic factors. Altogether, our results show that there is no need to sacrifice sample complexity when using ℓ_p Lewis weight sampling, other than logarithmic factors.

Theorem 1.3. *["Best of All Worlds" ℓ_p Lewis Weight Sampling] Let $p > 2$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\delta \in (0,1)$ be a failure rate parameter and let $\varepsilon \in (0,1)$ be an accuracy parameter. Let $\mathbf{w} \in \mathbb{R}^n$ be one-sided ℓ_p Lewis weights (Definition 2.2) with $\|\mathbf{w}\|_1 \leq O(d)$, which can be computed in*

$$\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^\omega)$$

time [Lee16, Theorem 5.3.1], [JLS21, Lemma 2.5]. Let

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

be an oversampling parameter. Suppose that weights $\mathbf{s} \in \mathbb{R}^n$ are sampled by independently setting $\mathbf{s}_i = 1/\mathbf{p}_i^{1/p}$ with probability $\mathbf{p}_i \coloneqq \min\{\alpha\mathbf{w}_i, 1\}$ and $\mathbf{s}_i = 0$ otherwise. Let $\mathbf{S} = \operatorname{diag}(\mathbf{s})$. Then, with probability at least $1-\delta$,

$$\text{for all } \mathbf{x} \in \mathbb{R}^d,\ \|\mathbf{SA}\|_p = (1 \pm \varepsilon)\|\mathbf{Ax}\|_p$$

and the sample complexity of $\mathbf{S}$ *is at most*

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

We also give similar high probability Lewis weight sampling results for $0 < p < 2$ in Appendix A, which we need for our high probability online coresets for $0 < p < 2$, as well as applications to sliding windows. We give a discussion of our techniques in Section 1.2.2, and our proofs are contained in Section 5.

1.1.3 Applications: Sliding Windows and Adversarial Streams

As an application of our results for online coresets for ℓ_p subspace embeddings, we obtain significantly improved results for two other important related problems: ℓ_p subspace embeddings in the *sliding window model* and in the *adversarial streaming model*.

Sliding Window Model. In the sliding window model of ℓ_p subspace embeddings, we are given a stream of rows $\{\mathbf{a}_i\}_{i=1}^n$ as well as a parameter $W \in \mathbb{N}$, which specifies the size of a window. Then, at each time $i \in [n]$, we consider the matrix $\mathbf{A}_i^W$ which denotes the $W \times d$ matrix formed by rows $i, i-1, i-2, \ldots, i-W+1$, that is, the W most recent rows at time i. Our goal is to output an ℓ_p subspace embedding for $\mathbf{A}_i^W$ at time i, for each $i \in [n]$.

A simple observation from the work of [BDM+20] shows how to convert algorithms for online coresets for ℓ_p subspace embeddings into algorithms for sliding windows, by running the online coreset algorithm "in reverse". That is, suppose that we have maintained an online coreset $\mathbf{S}_i^W\mathbf{A}_i^W$ for $\mathbf{A}_i^W$ such that for $j \in [W]$, the last j rows of $\mathbf{S}_i^W\mathbf{A}_i^W$ are a subspace embedding for the last j rows of $\mathbf{A}_i^W$. Then, we can update this sketch by throwing away the first row of $\mathbf{S}_i^W\mathbf{A}_i^W$ (which could be a zero row that is maintained implicitly), appending the new row, and recomputing an online coreset for this new matrix if necessary to save space. Although we lose a factor of $(1+\varepsilon)$ in the distortion each time we recompute the online coreset, by carrying out this process in a binary tree fashion, we only compose this approximation at most $\log n$ times per each row and thus we can set ε to be $\varepsilon/\log n$ instead so that the total distortion is only $(1+\varepsilon)$. Although this reduction is stated for deterministic algorithms in [BDM+20], because we have a logarithmic dependence on δ for our coresets, we can afford to union bound over all blocks of the merge-and-reduce tree to obtain the following:

Theorem 1.4 (ℓ_p Lewis Weight Sampling in Sliding Windows)**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $p \in (0,\infty)$. *Let* $\delta \in (0,1)$ *be a failure rate parameter and let* $\varepsilon \in (0,1)$ *be an accuracy parameter. Let* $W \in \mathbb{N}$ *be a window size parameter. Let* κ *be the stream condition number of* $\mathbf{A}$, *that is, any submatrix of consecutive rows of* $\mathbf{A}$ *has condition number at most* κ. *Then, there is a sliding window coreset algorithm* $\mathcal{A}$ *such that, with probability at least* $1-\delta$, $\mathcal{A}$ *outputs a weighted subset of* m *rows with sampling matrix* $\mathbf{S}$ *such that*

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

for all $\mathbf{x} \in \mathbb{R}^d$ *and every* $i \in [n]$, *and*

$$m = \begin{cases} O\left(\frac{d^{p/2}}{\varepsilon^2}\cdot(\log(n\kappa))^{p/2+1}(\log n)^2\left[(\log d)^2(\log n) + \log\frac{1}{\delta}\right]\right) & p \in (2,\infty) \\ O\left(\frac{d}{\varepsilon^2}\cdot\log(n\kappa)(\log n)^2\left[(\log d)^2\log n + \log\frac{1}{\delta}\right]\right) & p \in (1,2) \\ O\left(\frac{d}{\varepsilon^2}\cdot\log(n\kappa)(\log n)^2\log\frac{n}{\delta}\right) & p = 1 \\ O\left(\frac{d}{\varepsilon^2}\cdot\log(n\kappa)(\log n)^2\left[(\log d)^3 + \log\frac{1}{\delta}\right]\right) & p \in (0,1) \end{cases}$$

Proof. Our merge-and-reduce algorithm is exactly as described in Section 5 of [BDM+20], except that we use a randomized algorithm with failure rate δ set to $O(\delta/n)$ rather than a deterministic algorithm as the "reduce" algorithm. This allows us to union bound over all $O(n)$ reduce operations used in the merge-and-reduce algorithm. The result then follows from applying Theorems 7.5 and 6.6 to get the guarantees for the "reduce" algorithm. □

Adversarial Streaming Model. Our results also give the first nearly optimal results for online coresets for ℓ_p subspace embeddings in the *adversarial streaming model*, for any $p \neq 2$. In this model of streaming, when an adversary chooses an update to the stream, in this case a new row $\mathbf{a}_i$ of the input matrix, this update is allowed to depend on all previous outputs of the algorithm as well as the previous updates to the stream. Upon seeing this update, the streaming algorithm must process this new row $\mathbf{a}_i$, output a subspace embedding which is visible to the adversary, and the process repeats. The challenge of this model is that, for randomized algorithms, it is possible for the adversary to gradually learn the internal random bits used by the randomized algorithm and eventually break the algorithm. Indeed, many classical streaming algorithms fail in this model because of this problem [HW13].

To combat the adversarial streaming model, it is known that many algorithms based on the idea of random sampling are quite successful in this model [BY20], with a small additional overhead. In the context of matrix approximation, this theme was further explored and extended to importance sampling algorithms by work of [BHM$^+$21], who gave adversarially robust importance sampling algorithms for ℓ_2 subspace embeddings, ℓ_1 subspace embeddings, and other problems. However, these results lose a d factor over known results for offline ℓ_p subspace embeddings, and even a polynomial factor in κ^{OL}. Furthermore, [BHM$^+$21] discuss the difficulties in adapting the matrix martingale proofs of [CMP20] for spectral approximation, which hinders the success of the existing theory of online ℓ_p approximation in this setting.

We overcome these problems in a simple manner, by leveraging our decoupled approach to online importance sampling. Indeed, we give a *deterministic* online coreset algorithm (Algorithm 2) which can give approximate online Lewis weights for any $p \in (0, \infty)$. Note that deterministic algorithms are automatically adversarially robust. Furthermore, these satisfy the one-sided Lewis property by Lemma 6.2, and bound the true Lewis weights for $p \in (0, 2)$ by Lemma 6.4, and have a small sum by Lemma 6.5. Thus, simply sampling from these weights independently gives an ℓ_p subspace embedding, which is also automatically adversarially robust, since the algorithm uses fresh randomness after each adversarial row input. Thus, for $p > 2$, our Theorem 6.6 is already adversarially robust, and for $p \in (0, 2)$, a minor modification to Theorem 7.5 of replacing the online Lewis weight approximation by Algorithm 2 yields an adversarially robust streaming algorithm with the same guarantees.

1.1.4 Applications: Online Coresets for Generalized Linear Models

As applications of our online approximation of Lewis weights, we obtain the first online coresets, and in fact the first one-pass streaming algorithms, for a variety of generalized linear models.

The work of [MSSW18] gave one of the first investigations of coresets for unregularized logistic regression in the worst case, and showed that in general, logistic regression does not admit coresets with sublinear in n memory. To get around this problem, they defined a natural complexity parameter $\mu(\mathbf{A})$ which characterizes the complexity of the dataset, and gave sensitivity sampling algorithms for the logistic regression problem with sample complexity $\operatorname{poly}(\mu(\mathbf{A}), d, \log n, \varepsilon^{-1})$, but required multiple passes through the data. This work was later extended to an oblivious one-pass streaming algorithm by [MOW21], which achieved an $O(1)$-approximation using $\operatorname{poly}(\mu(\mathbf{A}), d, \log n)$ bits of space. The results of [MRM21] also gave further results in this direction, which generalized the results to handle a broad class of "nice hinge functions" which includes the logistic, ReLU, and hinge losses, and further improved the polynomial dependencies in the coreset size using ℓ_1 Lewis weights. The recent work of [MOP22] showed that similar ideas can handle the p-generalized probit regression model, a generalization of the probit regression model which they introduce using the p-generalized Gaussian distribution [DBPS18]. They introduce a generalization of the $\mu(\mathbf{A})$ parameter to the degree p analogue $\mu_p(\mathbf{A})$[2] and give a two-pass streaming coreset algorithm for $p \neq 2$ and a one-pass streaming coreset algorithm for $p = 2$.

In the important one-pass streaming setting, the only results that apply from the above line of work are the oblivious $O(1)$-approximation algorithm of [MOW21] for logistic regression, as well as the $(1+\varepsilon)$-approximation for the probit regression model. We remedy this situation by providing analogues of [MRM21, MOP22] in the one-pass streaming setting via our online coresets for Lewis weights.

Online Coresets for Nice Hinge Functions. Recall first the definitions of μ_p complexity [MOP22] and nice hinge functions [MRM21]:

[2] The parameter $\mu_p(\mathbf{A})$ [MOP22] for $p = 1$ corresponds to $\mu(\mathbf{A})$ of [MSSW18, MOW21, MRM21]. See Definition 1.5.

Definition 1.5 (μ_p Complexity (Definition 2.2 [MOP22], Definition 2 [MSSW18]))**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $p \in (0, \infty)$*. Then,*

$$\mu_p(\mathbf{A}) \coloneqq \sup_{\mathbf{A}\mathbf{x}\neq 0} \frac{\sum_{\mathbf{a}_i^\top \mathbf{x}>0} \left|\mathbf{a}_i^\top \mathbf{x}\right|^p}{\sum_{\mathbf{a}_i^\top \mathbf{x}<0} \left|\mathbf{a}_i^\top \mathbf{x}\right|^p}.$$

We say $\mathbf{A}$ *is* μ_p*-complex if* $\mu_p(\mathbf{A}) \leq \mu_p < \infty$*. If* $p = 1$*, we drop the subscript and simply say that* $\mathbf{A}$ *is* μ*-complex.*

Definition 1.6 (Nice Hinge Functions (Definition 7, [MRM21]))**.** *We say that* $\varphi : \mathbb{R} \to \mathbb{R}^+$ *is an* (L, a_1, a_2)-nice hinge function *if there exist universal constants* $L > 0$*,* $a_1 \geq 0$*, and* $a_2 \geq 0$ *such that*

- $x \mapsto \varphi(x)$ *is L-Lipschitz.*
- $|\mathrm{ReLU}(x) - \varphi(x)| \leq a_1$ *for all* $x \in \mathbb{R}$*.*
- $\varphi(x) \geq a_2$ *for all* $x \geq 0$*.*

Note that nice hinge functions include the ReLU, hinge loss $\max\{0, 1 + x\}$, and the logistic loss. The work of [MRM21] shows that oversampling from Lewis weights by a factor of $\mu(\mathbf{A})^2/\varepsilon^2$ yields relative error coresets for any nice hinge function. By using our ℓ_1 Lewis weight overestimates of either Theorem 6.6 or Theorem 7.5, we immediately extend the results of [MRM21] to the online setting:

Theorem 1.7. *Let* φ *be a nice hinge function (Definition 1.6) with* $a_2 > 0$ *or the* ReLU *loss. Let* $\tilde{\mathbf{w}} \in \mathbb{R}^n$ *be online* ℓ_1 *Lewis weight estimates obtained by either Theorem 6.6 or Theorem 7.5 and let* $T = \|\tilde{\mathbf{w}}\|_1$*. Let* $\mathbf{p}_i \geq C(\mu(\mathbf{A})/\varepsilon)^2 \max\{\tilde{\mathbf{w}}_i, 1/n\}$ *for some* $C = O(\log \frac{nT}{\varepsilon\delta})$*. For each* $i \in [n]$*, let*

$$\mathbf{s}_i = \begin{cases} 1/\mathbf{p}_i & \text{with probability } \mathbf{p}_i \\ 0 & \text{otherwise.} \end{cases}$$

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

$$\left|\sum_{i=1}^n \mathbf{s}_i \varphi([\mathbf{A}\mathbf{x}](i)) - \varphi([\mathbf{A}\mathbf{x}](i))\right| \leq \varepsilon \sum_{i=1}^n \varphi([\mathbf{A}\mathbf{x}](i)).$$

Furthermore, with probability at least $1 - \delta$*, there are at most*

$$O\left(\frac{d\mu(\mathbf{A})^2}{\varepsilon^2} \log(n\kappa^{\mathsf{OL}}) \log \frac{nT}{\varepsilon\delta}\right)$$

nonzero entries $\mathbf{s}_i$*.*

Proof. This is essentially a direct application of [MRM21, Theorem 5, Corollary 6, Theorem 8, Corollary 9]. There are slight differences – the sampling method in [MRM21] draws without replacement from the Lewis weight sampling distribution, while we sample each row independently. These differences are minor, and can be handled as we do in the proofs of Theorem 6.6 or Theorem 7.5. □

Streaming Coresets for p-Probit Regression. The probit model is a popular generalized linear model which uses the Gaussian cdf as the link function to model binary data. This model was recently generalized to the p-generalized probit model (or p-probit model for short) in [MOP22], which models binary data $Y_i \in \{0, 1\}$ for $i \in [n]$ as

$$\mathbf{Pr}\{Y_i = 1\} = \mathbf{E}[Y_i] = \Phi_p(\mathbf{z}_i^\top \mathbf{x}),$$

where

$$\Phi_p(r) = \frac{p^{1-1/p}}{2\Gamma(1/p)} \int_{-\infty}^{r} \exp(-|t|^p/p)\ dt$$

is the cdf of the p-generalized normal distribution, $\mathbf{z}_i \in \mathbb{R}^d$ is the feature vector for the label y_i, and $\mathbf{x} \in \mathbb{R}^d$ is a parameter vector. We then define

$$\psi_p(x) = -\log(\Phi_p(-x))$$

to be the p-probit loss, and the negative log likelihood of the dataset under the parameter vector $\mathbf{x} \in \mathbb{R}^d$ can be written as

$$\mathcal{L}_p(\mathbf{x}) \coloneqq \sum_{i=1}^n \psi_p([\mathbf{A}\mathbf{x}](i))$$

where the ith row of $\mathbf{A}$ is $\mathbf{a}_i = -(2y_i - 1)\mathbf{z}_i$ [MOP22]. We give the following streaming coreset theorem. Note that due to the reservoir sampler, this result is not an online coreset, in the sense that the rows are not selected irrevocably.

Theorem 1.8 (One-Pass Streaming Coresets for p-Probit Regression)**.** *Let $\tilde{\mathbf{w}} \in \mathbb{R}^n$ be online ℓ_1 Lewis weight estimates obtained by either Theorem 6.6 or Theorem 7.5 and let $T = \|\tilde{\mathbf{w}}\|_1$. Then, there is a one-pass streaming algorithm which computes a coreset $\mathbf{s} \in \mathbb{R}^n$ such that with probability at least $1 - \delta$, for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\left| \sum_{i=1}^n \mathbf{s}_i \psi_p([\mathbf{A}\mathbf{x}](i)) - \psi_p([\mathbf{A}\mathbf{x}](i)) \right| \leq \varepsilon \sum_{i=1}^n \psi_p([\mathbf{A}\mathbf{x}](i)).$$

Furthermore, with probability at least $1 - \delta$, there are at most

$$O\left(\frac{Sd}{\varepsilon^2}(\log S) \log \frac{\mu_p(\mathbf{A})}{\varepsilon}\right)$$

nonzero entries $\mathbf{s}_i$, for

$$S = O(\mu_p(\mathbf{A}))(d \log(n\kappa^{\mathsf{OL}}))^{1 \vee (p/2)}.$$

Proof. It is shown in Lemma 2.10 that the ψ_p-sensitivities are bounded as

$$\sup_{\mathbf{A}\mathbf{x} \neq 0} \frac{\psi_p([\mathbf{A}\mathbf{x}](i))}{\mathcal{L}_p(\mathbf{x})} \leq O(\mu_p(\mathbf{A}))\left(\frac{1}{n} + \mathbf{s}_i^p(\mathbf{A})\right),$$

where $\mathbf{s}_i^p(\mathbf{A})$ are the ℓ_p sensitivities (Equation (2)). We show that our approximate online Lewis weights bound the ℓ_p sensitivities, which allow us to sample from them. For $0 < p < 2$, the approximate online Lewis weights $\tilde{\mathbf{w}}_i$ upper bound the true Lewis weights by Lemma 6.4, which means that they upper bound the ℓ_p sensitivities (note that we do not use our sampling-based Lewis weight estimation algorithm, which requires splitting of rows which can affect the p-probit loss). For $p > 2$, we use Lemma 6.2 and Lemma 2.3 to argue that $\|\tilde{\mathbf{w}}\|_1^{p/2-1}\tilde{\mathbf{w}}_i$ is an upper bound on the ℓ_p sensitivities. Finally, combining the result with the VC dimension bound in Lemma 2.9 of [MOP22] and the sensitivity framework results of [FSS20] (see also Appendix B of [MOP22] for a discussion of the sensitivity framework in this context), as well as the sampling scheme using weighted reservoir sampling using these sensitivities as in [MOP22] yields the desired result. □

We note that our coreset size also substantially improves the polynomial factors of those in [MOP22], who gave bounds of

$$S = \begin{cases} O(\mu_p(\mathbf{A})d) & p = 2 \\ O(\mu_p(\mathbf{A})d^p(d \log d)^2 & p \in [1, 2) \\ O(\mu_p(\mathbf{A})d^{2p}(d \log d)^2 & p \in (2, \infty) \end{cases}$$

for S as used in Theorem 1.8.

New Lower Bounds. To complement our improved algorithmic results, we provide the first lower bounds on the size of mergeable[3] coresets for p-probit regression and logistic regression, for instances with bounded μ_p-complexity. Despite a line of work on guarantees for generalized linear models parameterized by μ-complexity, lower bounds depending on μ were previously unknown [MSSW18, MOW21, MRM21, MOP22]. Our work shows that the linear dependence on μ in Theorem 1.8 as well as the results of [MOP22] are tight.

As done in previous work [MSSW18], we consider the following communication game: Alice has a dataset $\mathbf{A} \in \mathbb{R}^{n_1 \times d}$ and Bob has a dataset $\mathbf{B} \in \mathbb{R}^{n_2 \times d}$, and Alice must send a single message M to Bob such that

[3] By mergeable, we mean that a coreset for $\mathbf{A}$ and a coreset for $\mathbf{B}$ can be combined into a coreset for $\mathbf{A} \circ \mathbf{B}$.

Bob can output an approximate solution to the logistic regression problem for the concatenated dataset $\mathbf{A} \circ \mathbf{B}$ using just M and $\mathbf{B}$. We then show lower bounds on the number of bits required by M. This in particular includes mergeable coreset algorithms, since M here can be taken to be the coreset, which can approximately solve logistic regression on $\mathbf{A} \circ \mathbf{B}$ in combination with $\mathbf{B}$. Because our hard instances are only in polylogarithmically many dimensions, if the weights are specified using polylogarithmically many bits, then M is a lower bound on the size of the coreset, up to polylogarithmic factors.

Theorem 1.9 (Informal Version of Theorems 8.9, 8.10)**.** *There exists $\mathbf{A} \in \mathbb{R}^{m \times d}$ with $d = O(\log^2 m)$ and μ_p-complexity at most $O(m)$ for any $1 \leq p < \infty$ such that for any $1 \leq \Delta \leq O(m^{1/3})$, a mergeable coreset which approximates the optimal p-probit cost or the logistic regression cost up to a Δ relative error must have size at least $\tilde{\Omega}(m/\Delta)$. In particular, a constant factor approximation to the optimal p-probit regression cost or logistic regression cost for a μ_p-complex dataset requires $\tilde{\Omega}(\mu_p)$ points in the coreset.*

1.2 Our Techniques

1.2.1 Online Coresets for ℓ_p Subspace Embeddings

We first discuss previous approaches towards online coresets for ℓ_p subspace embeddings as well as their shortcomings, and then discuss our ideas which allow us to overcome them.

For ℓ_2 [CMP20], a spectral approximation to $\mathbf{A}$ simultaneously yields both an ℓ_2 subspace embedding as well as estimates to the online leverage scores of $\mathbf{A}$, and yields nearly optimal sample complexities. A naïve generalization of this to ℓ_p leads to the idea of using an ℓ_p subspace embedding in order to estimate sampling probabilities. This strategy naturally lends itself to the technique of *sensitivity sampling*, in which rows are selected with probability proportional to the sensitivity score defined by

$$\mathbf{s}_i^p(\mathbf{A}) \coloneqq \sup_{\mathbf{x} \in \mathbb{R}^d} \frac{|\langle \mathbf{a}_i, \mathbf{x} \rangle|^p}{\|\mathbf{A}\mathbf{x}\|_p^p}, \tag{2}$$

or its online variants. Indeed, given an ℓ_p subspace embedding for $\mathbf{A}$, one can approximate the denominator of the above term, as well as the numerator, given access to that row, and this is how [BDM+20] proceeds to obtain their online ℓ_1 subspace embedding result. However, the problem is that sensitivity sampling is not known to admit efficient chaining arguments, and leads to a loss of a factor of d in the sample complexity compared to Lewis weight sampling. Although one would ideally have estimates to Lewis weights (or an appropriate online generalization), the challenge faced by [BDM+20] is that they cannot estimate these Lewis weights given an ℓ_p subspace embedding of $\mathbf{A}$.

One of the main observations necessary towards achieving our results is the *decoupling of the Lewis weight approximation step and the sampling step*. That is, we maintain two different sketches of $\mathbf{A}$, one for approximating online Lewis weights, and one for getting an ℓ_p subspace embedding. By taking this approach, we also get the additional benefit of a much simpler analysis: by using fresh randomness for the subspace embedding sampling, we are able to avoid using a martingale argument for the subspace embedding, and only use such arguments for approximating the Lewis weights. With Lewis weight estimates in hand, we simply appeal to standard offline analyses of Lewis weight sampling. That is, *the only "online" aspect of online ℓ_p subspace embedding is in the Lewis weight estimation.* We view the simplicity of the analysis as one of our main strengths of this work.

With this idea in hand, we show how to *deterministically* estimate online ℓ_p Lewis weights, by using deterministic online coresets for spectral approximation developed by [BDM+20]. For $p < 2$, we show that these weights in fact bound the true ℓ_p Lewis weights, and also have a small sum. This is enough to use these weights as sampling weights, using existing results on ℓ_p Lewis weight sampling. For $p > 2$, while these approximated weights do not necessarily bound the true ℓ_p Lewis weights, we show that they satisfy a one-sided generalization of the property of Lewis weights, which were recently shown to be sufficient to make the Lewis weight sampling arguments work [WY22]. Our result follows.

For $p \in (0, 2)$, we also analyze a sampling-based Lewis weight estimation algorithm analogous to the online leverage score estimation algorithm of [CMP20], in which randomly sampled rows are used to estimate the ℓ_p Lewis weights. While quite similar to the analysis of [CMP20], a straight adaptation does not work due to complications that arise due to introducing the Lewis weights, and we use a variation on the idea of splitting rows in order to control the matrix martingale. This alternate algorithm may be simpler to implement, and may be more attractive in certain cases.

Our deterministic Lewis weight approximation algorithm is discussed in Section 6, while our sampling-based Lewis weight approximation algorithm for $0 < p < 2$ is discussed in Section 7. In Section 4, we describe a variant of our algorithm to work in batches of rows, which allows the algorithm to run in input sparsity time.

1.2.2 Improved Analysis of Lewis Weight Sampling

Our basic framework for the analysis of one-shot Lewis weight sampling is based around a reduction idea of [CP15], as well as an adaptation of this technique by [CD21]. This reduction shows that for $p \in (0, 2)$, in order to bound the expected error of the one-shot Lewis sampling procedure, it suffices to bound the expected error of a sampling procedure which samples each row with probability $1/2$, under the condition that the input matrix has uniformly bounded Lewis weights.

The [CP15] reduction. The idea of [CP15] is roughly as follows. The expected error of the one-shot Lewis sampling procedure can be written as

$$\mathop{\mathbf{E}}_{\mathbf{S}}\left[\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p - 1\right|\right].$$

We wish to show that this quantity is bounded by ε. Now one can note that the quantity inside the absolute is a zero mean random variable. This fact, combined with a standard symmetrization argument, bounds this quantity by

$$\mathop{\mathbf{E}}_{\mathbf{S}}\mathop{\mathbf{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\sum_{i=1}^{n}\boldsymbol{\sigma}_i|[\mathbf{S}\mathbf{A}\mathbf{x}](i)|^p\right|\right], \tag{3}$$

where $\boldsymbol{\sigma}_i$ are independent Rademacher variables. This expression is, in fact, exactly an expression that is bounded by previous works on Lewis weight sampling from the functional analysis literature [Tal90, LT91, Tal95, SZ01], which bounds this expectation by roughly ε, whenever the ℓ_p Lewis weights of the matrix are at most ε^2. The idea now is that, heuristically, we would expect the Lewis weights of the matrix $\mathbf{SA}$ to be at most ε^2 if $\mathbf{S}$ samples each row i with probability roughly $\mathbf{p}_i = \mathbf{w}_i^p(\mathbf{A})/\varepsilon^2$ and scale the result by $1/\mathbf{p}_i^{1/p}$. Indeed, if two rows of $\mathbf{A}$ are the same up to a scaling, then the Lewis weights would simply differ by that scaling factor, raised to the pth power. Thus, the Lewis weight of $\mathbf{a}_i/\mathbf{p}_i^{1/p}$ would be expected to be $1/\mathbf{p}_i$ times larger than that of $\mathbf{a}_i$, which would just be ε^2. The problem here is that $\mathbf{SA}$ and $\mathbf{A}$ do not belong to the same matrix. To address this problem, [CP15] make the observation that, using the monotonicity of Lewis weights for $p \in (0, 2)$ [CP15, Lemma 5.5], one can bound (3) by the corresponding quantity using the matrix $\mathbf{A}''$, which concatenates $\mathbf{SA}$ with $\mathbf{A}$. This allows one to argue that the rows of $\mathbf{A}''$ corresponding to $\mathbf{SA}$ indeed do have Lewis weights bounded by ε^2. Furthermore, one can replace $\mathbf{A}$ in $\mathbf{A}''$ by a version $\mathbf{A}'$ of $\mathbf{A}$ which splits rows with large Lewis weight into multiple copies, so that every row of $\mathbf{A}''$ in fact has Lewis weight bounded by ε^2. The result then follows by applying existing bounds from [Tal90, LT91, Tal95, SZ01].

Circumventing non-monotonicity via online Lewis weights. The part of the above argument which breaks for $p > 2$ is precisely in the lack of monotonicity for $p > 2$. That is, if we add a row to a matrix $\mathbf{A}$, then the Lewis weights of the existing rows may in fact increase. What we would really like in order for the above proof to go through is to define a monotonic version of ℓ_p Lewis weights which behaves well with respect to row additions. This motivates a connection to ideas from *online numerical linear algebra* [CMP20, BDM+20, WY22].

The change we make to the above argument is the following: rather than considering the ℓ_p Lewis weights for the matrix $\mathbf{A}''$, we define a version of *online* Lewis weights for $\mathbf{A}''$ as follows. We first use the existing weights for the split up version $\mathbf{A}'$ of $\mathbf{A}$. Then, we define weights for $\mathbf{SA}$ by treating them as a batch of rows which arrive after the rows $\mathbf{A}'$. We show that such weights exist and thus satisfy a batch monotonicity, which is sufficient to show that the weights are bounded by ε^2. Furthermore, we can show that these weights also have a small sum. We also show that the weights defined in this way satisfy a one-sided Lewis weight property, which we show is sufficient to make the result of [LT91] still go through. Our result follows.

Our full proof of this result can be found in Section 5.

2 Preliminaries

2.1 Lewis Weights

Lewis weights were initially discovered in the functional analysis community by [Lew78], who used them to obtain optimal bounds on distances between subspaces of ℓ_2 and ℓ_p, in the sense of Banach–Mazur distance. The use of Lewis weights as sampling probabilities for approximating d-dimensional subspaces of L_p was first introduced by [Sch87], whose results were subsequently refined and extended by [BLM89, Tal90, LT91, Tal95, SZ01]. This technique was then brought to the algorithms community [CP15], whose definition of Lewis weights we give below:

Definition 2.1 (ℓ_p Lewis weights [CP15])**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $p \in (0,\infty)$*. Then, the* ℓ_p *Lewis weights are the unique weights* $\mathbf{w} \in \mathbb{R}^n$ *which satisfy*

$$\mathbf{w}_i = \left[\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A})^- \mathbf{a}_i\right]^{p/2},$$

where $\mathbf{W} = \operatorname{diag}(\mathbf{w})$*. We denote these weights as* $\mathbf{w}_i^p(\mathbf{A})$*, or* $\mathbf{W}_i^p(\mathbf{A})$ *for its diagonal matrix.*

While this definition is recursive since $\mathbf{w}$ appears on both sides of the equation, the existence of such weights is nonetheless proven by [Lew78, SZ01, CP15]. Furthermore, [CP15] give efficient, and in fact input sparsity time, algorithms for approximating these weights. Algorithms for approximating Lewis weights have subsequently been improved by works such as [Lee16, FLPS22, JLS21].

It has since been shown that relaxations of Definition 2.1 can be more useful, which admit more efficient approximation algorithms while still leading to similar guarantees for applications [JLS21, WY22]. We refer to this as the γ-one-sided ℓ_p Lewis property:

Definition 2.2 (One-sided ℓ_p Lewis weights and bases [WY22])**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $p \in (0,\infty)$*. Let* $\gamma \in (0,1]$*. Then, weights* $\mathbf{w} \in \mathbb{R}^n$ *are* γ-one-sided ℓ_p Lewis weights *if*

$$\mathbf{w}_i \geq \gamma \cdot \boldsymbol{\tau}_i(\mathbf{W}^{1/2-1/p}\mathbf{A}),$$

where $\mathbf{W} \coloneqq \operatorname{diag}(\mathbf{w})$*, or equivalently,*

$$\mathbf{w}_i \geq \gamma^{p/2}\left[\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A})\mathbf{a}_i\right]^{p/2}.$$

If $\gamma = 1$*, we just say that* $\mathbf{w}$ *are* one-sided ℓ_p Lewis weights *Let* $\mathbf{R} \in \mathbb{R}^{d\times d}$ *be a change of basis matrix such that* $\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}$ *has orthonormal columns. Then,* $\mathbf{A}\mathbf{R}$ *is a* one-sided ℓ_p Lewis basis*.*

We note that [JLS21, WY22] used this definition only with $\gamma = 1$. The flexibility of allowing for $\gamma = \Theta(1) < 1$ will, as we show, not affect any of the bounds for sampling, while allowing for more convenient approximation when handling rows in a batched online setting Section 4.

We will also frequently refer to the associated quadratic form, which is $\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A}$. Several other properties similar to those of Lewis weights carry over to one-sided Lewis weights, which will be useful to us. The following properties of γ-one-sided ℓ_p Lewis weights are proven in [WY22] for $\gamma = 1$. The extension to $\gamma \in (0,1]$ is straightforward from their proof.

Lemma 2.3 (Lemmas 2.8 and 2.10 of [WY22])**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $p \in (0,\infty)$*. Let* $\mathbf{w}$ *be* γ*-one-sided* ℓ_p *Lewis weights for* $\mathbf{A}$ *and let* $\mathbf{R}$ *be a one-sided* ℓ_p *Lewis basis. Then,*

- *for every* $i \in [n]$*,*

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

- *for all* $\mathbf{x} \in \mathbb{R}^d$,
$$\left\|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{x}\right\|_2 \leq \begin{cases} \left\|\mathbf{w}\right\|_1^{1/2-1/p}\left\|\mathbf{A}\mathbf{x}\right\|_p & \text{if } p \geq 2 \\ \frac{1}{\gamma^{1/p-1/2}}\left\|\mathbf{A}\mathbf{x}\right\|_p & \text{if } p < 2 \end{cases}$$

- *for every* $i \in [n]$,
$$\sup_{\mathbf{x}\in\mathrm{rowspan}(\mathbf{A})\setminus\{0\}} \frac{\left|\langle \mathbf{a}_i, \mathbf{x}\rangle\right|^p}{\left\|\mathbf{A}\mathbf{x}\right\|_p^p} \leq \left\|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{x}\right\|_2 \leq \begin{cases} \left\|\mathbf{w}\right\|_1^{1/2-1/p}\mathbf{w}_i & \text{if } p \geq 2 \\ \frac{1}{\gamma^{1/p-1/2}}\mathbf{w}_i & \text{if } p < 2 \end{cases}$$

Similarly, we have the following simple modification of a result from [JLS21]:

Lemma 2.4 (Lemma 2.6, [JLS21])**.** *Let* $p > 2$*. Let* $\mathbf{w} \in \mathbb{R}^n$ *be* γ*-one-sided* ℓ_p *Lewis weights for* $\mathbf{A} \in \mathbb{R}^{n\times d}$*. Then, for all* $\mathbf{x} \in \mathbb{R}^d$,
$$\left\|\mathbf{A}\mathbf{x}\right\|_p \leq \frac{1}{\gamma^{1/p-1/2}}\left\|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{x}\right\|_2.$$

2.1.1 Lemmas from Linear Algebra

We record a couple of linear algebraic lemmas that we will use repeatedly.

Lemma 2.5. *Let* $\mathbf{R} = \mathbf{V}\tilde{\mathbf{R}}\mathbf{V}^\top \in \mathbb{R}^{d\times d}$ *where* $\tilde{\mathbf{R}} \in \mathbb{R}^{r\times r}$ *is a symmetric positive definite matrix and* $\mathbf{V} \in \mathbb{R}^{d\times r}$ *has orthonormal columns. Then,*
$$\mathbf{R}^- = \mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top.$$

Proof. Note that $\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top$ is an inverse for the column space of $\mathbf{R}$, i.e.,
$$\mathbf{R}(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top)\mathbf{R} = \mathbf{V}\tilde{\mathbf{R}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{R}}\mathbf{V}^\top = \mathbf{R}$$
and a weak inverse, i.e.,
$$(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top)\mathbf{R}(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top) = \mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top.$$
One can also easily check that both $\mathbf{R}(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top)$ and $(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top)\mathbf{R}$ are Hermitian. Thus, $\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top$ is uniquely determined to be the pseudoinverse of $\mathbf{R}$. □

Lemma 2.6. *Let* $0 \preceq \mathbf{R} \preceq \mathbf{S} \in \mathbb{R}^{d\times d}$ *by symmetric positive semidefinite matrices. Let* $\mathbf{a} \in \mathrm{rowspan}(\mathbf{R})$*. Then,*
$$\mathbf{a}^\top\mathbf{R}^-\mathbf{a} \geq \mathbf{a}^\top\mathbf{S}^-\mathbf{a}.$$

Proof. Let $\mathbf{V} \in \mathbb{R}^{d\times r}$ be an orthonormal basis for $V := \mathrm{rowspan}(\mathbf{R})$, where $r = \dim(V)$. Let $\mathbf{P} = \mathbf{V}\mathbf{V}^\top$ be the projection matrix onto V. Write $\mathbf{a} = \mathbf{V}\mathbf{b}$ for $\mathbf{b} \in \mathbb{R}^r$ and $\mathbf{R} = \mathbf{V}\tilde{\mathbf{R}}\mathbf{V}^\top$, $\mathbf{P}\mathbf{S}\mathbf{P} = \mathbf{V}\tilde{\mathbf{S}}\mathbf{V}^\top$ for $\tilde{\mathbf{R}}, \tilde{\mathbf{S}} \in \mathbb{R}^{r\times r}$. Then, we have that
$$\mathbf{a}^\top\mathbf{R}^-\mathbf{a} = \mathbf{b}^\top\mathbf{V}^\top(\mathbf{V}\tilde{\mathbf{R}}\mathbf{V}^\top)^-\mathbf{V}\mathbf{b} = \mathbf{b}^\top\tilde{\mathbf{R}}^{-1}\mathbf{b}$$
and
$$\mathbf{a}^\top\mathbf{S}^-\mathbf{a} = \mathbf{b}^\top\mathbf{V}^\top(\mathbf{V}\tilde{\mathbf{S}}\mathbf{V}^\top)^-\mathbf{V}\mathbf{b} = \mathbf{b}^\top\tilde{\mathbf{S}}^{-1}\mathbf{b}.$$
Furthermore, for all $\mathbf{x} \in \mathbb{R}^r$, we have that
$$\mathbf{x}^\top\mathbf{V}^\top\mathbf{R}\mathbf{V}\mathbf{x} \leq \mathbf{x}^\top\mathbf{V}^\top\mathbf{S}\mathbf{V}\mathbf{x} = \mathbf{x}^\top\mathbf{V}^\top\mathbf{P}\mathbf{S}\mathbf{P}\mathbf{V}\mathbf{x}$$
so $\tilde{\mathbf{R}} \preceq \tilde{\mathbf{S}}$, meaning that $\tilde{\mathbf{R}}^{-1} \succeq \tilde{\mathbf{S}}^{-1}$. Thus,
$$\mathbf{a}^\top\mathbf{R}^-\mathbf{a} = \mathbf{b}^\top\tilde{\mathbf{R}}^{-1}\mathbf{b} \geq \mathbf{b}^\top\tilde{\mathbf{S}}^{-1}\mathbf{b} = \mathbf{a}^\top\mathbf{S}^-\mathbf{a}.$$ □

3 Online Lewis Weights

In this section, we introduce both known and new results in online numerical linear algebra, especially pertaining to online Lewis weights.

For a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{A}_j \in \mathbb{R}^{j \times d}$ denotes the submatrix of $\mathbf{A}$ formed by the first j rows. The following notion of online leverage scores was introduced by [CMP20, BDM+20]:

Definition 3.1 (Online Leverage Scores)**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then, for each $i \in [n]$, the ith online leverage score is defined as*

$$\boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) \coloneqq \begin{cases} \min\{\mathbf{a}_i^\top (\mathbf{A}_{i-1}^\top \mathbf{A}_{i-1})^- \mathbf{a}_i, 1\} & \text{if } \mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1}) \\ 1 & \text{otherwise} \end{cases}$$

It is not hard to see that the online leverage scores are at least the standard leverage scores. It can also be shown that the sum of online leverage scores is not much more than the sum of the standard leverage scores.

Definition 3.2. *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Define the* online condition number *of $\mathbf{A}$ to be*

$$\kappa^{\mathsf{OL}} = \kappa^{\mathsf{OL}}(\mathbf{A}) \coloneqq \|\mathbf{A}\|_2 \max_{i=1}^{n} \|\mathbf{A}_i^-\|_2.$$

Lemma 3.3 (Sum of Online Leverage Scores [CMP20])**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then,*

$$\sum_{i=1}^{n} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) \leq O(d \log \kappa^{\mathsf{OL}})$$

Proof. This follows from setting $\lambda = (\max_{i=1}^{n} \|\mathbf{A}_i^-\|_2)^{-1}$, i.e., the minimum singular value for any $\mathbf{A}_i$ for $i \in [n]$, in Theorem 2.2 of [CMP20]. The result follows by noticing that with this choice of λ, the online ridge leverage scores used in [CMP20] are the same as the online leverage scores, up to constant factors. □

We now give the definition for the online ℓ_p Lewis weights, which are defined analogously to online leverage scores (Definition 3.1). Similar definitions have appeared in [BDM+20, WY22].

Definition 3.4 (Online ℓ_p Lewis Weights)**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p < \infty$. Then, for each $i \in [n]$, the ith online ℓ_p Lewis weight is defined as*

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) \coloneqq \begin{cases} \min\left\{\left[\mathbf{a}_i^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- \mathbf{a}_i\right]^{p/2}, 1\right\} & \text{if } \mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1}) \\ 1 & \text{otherwise} \end{cases}$$

where $\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_j$ is the $j \times j$ diagonal matrix with $\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_j(i,i) = \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})$.

We first show that for $0 < p < 2$, the online Lewis weights upper bound Lewis weights.

Lemma 3.5. *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p < 2$. Then, for each $i \in [n]$,*

$$\mathbf{w}_i^p(\mathbf{A}) \leq \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})$$

Proof. We proceed by induction. It suffices to consider the case when $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) < 1$, since $\mathbf{w}_i^p(\mathbf{A}) \leq 1$ for every $i \in [n]$. In particular, $\mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1})$ and

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \left[\mathbf{a}_i^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- \mathbf{a}_i\right]^{p/2}.$$

Then, since $1 - \frac{2}{p} < 0$, we have that

$$\begin{aligned} \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1} \succeq \mathbf{W}^p(\mathbf{A})_{i-1} \succ 0 &\implies \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \preceq \mathbf{W}^p(\mathbf{A})_{i-1}^{1-2/p} \\ &\implies \mathbf{A}_{i-1}^\top (\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} - \mathbf{W}^p(\mathbf{A})_{i-1}^{1-2/p}) \mathbf{A}_{i-1} \preceq 0 \end{aligned}$$

$$\implies \mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1} \preceq \mathbf{A}_{i-1}^\top \mathbf{W}^{p}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1}.$$

By Lemma 2.6, it follows that for every $\mathbf{a} \in \operatorname{rowspan}(\mathbf{A}_{i-1})$,

$$\mathbf{a}^\top(\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-\mathbf{a} \geq \mathbf{a}^\top(\mathbf{A}_{i-1}^\top \mathbf{W}^{p}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-\mathbf{a}.$$

Similarly, we have that

$$\mathbf{a}^\top(\mathbf{A}_{i-1}^\top \mathbf{W}^{p}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-\mathbf{a} \geq \mathbf{a}^\top(\mathbf{A}^\top \mathbf{W}^{p}(\mathbf{A})^{1-2/p}\mathbf{A})^-\mathbf{a}$$

for every $\mathbf{a} \in \operatorname{rowspan}(\mathbf{A}_{i-1})$. The result follows by taking $p/2$-th roots on the chain of inequalities. □

Note that for $p > 2$, the above proof fails since $1 - \frac{2}{p} > 0$, which causes the inequalities to go the wrong way. Nevertheless, we show that these weights satisfy the *one-sided Lewis property*, which is in fact sufficient to make the chaining argument go through.

Lemma 3.6 (One-Sided Lewis Property of Online Lewis Weights). *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $0 < p < \infty$. *Then, for each* $i \in [n]$,

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) \geq \boldsymbol{\tau}_i(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A}).$$

Proof. We already have the result when $\mathbf{a}_i \notin \operatorname{rowspan}(\mathbf{A}_{i-1})$, so we assume $\mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1})$. Similarly, we can assume that $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) < 1$. In this case,

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \left[\mathbf{a}_i^\top(\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-\mathbf{a}_i\right]^{p/2}$$

which rearranges to

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = (\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{a}_i)^\top(\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-(\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{a}_i).$$

By Lemma 2.6, this is bounded below by

$$(\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{a}_i)^\top(\mathbf{A}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1-2/p}\mathbf{A})^-(\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{a}_i) = \boldsymbol{\tau}_i(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A}),$$

which is the claimed result. □

3.1 The Sum of Online Lewis Weights

Finally, we bound the sum of online Lewis weights, using bounds on the sum of online leverage scores. Our proof substantially simplifies the proofs of [BDM+20, Lemma 4.7, Lemma 5.15], which relied on an elaborate argument involving recursive applications of a "whack-a-mole" lemma of [CLM+15], and also slightly improves the bound by logarithmic factors.

Lemma 3.7 (Sum of Online Lewis Weights). *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $0 < p < \infty$. *Then,*

$$\sum_{i=1}^n \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) \leq O(d)\log(n\kappa^{\mathsf{OL}}(\mathbf{A})).$$

Proof. Our analysis is similar to those given by [CMP20] and [WY22]. For $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) < 1$, we have that

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \left[\mathbf{a}_i^\top(\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-\mathbf{a}_i\right]^{p/2}.$$

This rearranges to

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = (\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{a}_i)^\top(\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-(\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{a}_i),$$

which is exactly the ith online leverage score of $\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A}$. Similar reasoning for $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = 1$ shows that $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A})$. Thus,

$$\sum_{i=1}^n \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \sum_{i=1}^n \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A}) \leq O(d \log \kappa^{\mathsf{OL}}(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A}))$$

by Lemma 3.3. Furthermore, we have for any $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$ that

$$\|\mathbf{A}_i\mathbf{x}\|_2 \leq \left\|\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A}_i)^{1/2-1/p}\mathbf{A}_i\mathbf{x}\right\|_2 \leq \left\|\mathbf{W}^{p}(\mathbf{A}_i)^{1/2-1/p}\mathbf{A}_i\mathbf{x}\right\|_2 \leq d^{|1/2-1/p|}\|\mathbf{A}_i\mathbf{x}\|_p \leq (nd)^{|1/2-1/p|}\|\mathbf{A}_i\mathbf{x}\|_2,$$

so $\kappa^{\mathsf{OL}}(\mathbf{A}) = \mathrm{poly}(n)\kappa^{\mathsf{OL}}(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A})$. Thus,

$$\sum_{i=1}^n \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) \leq O(d)\log(n\kappa^{\mathsf{OL}}(\mathbf{A})).$$

□

By using the fact that one-sided ℓ_p Lewis weights bound the ℓ_p sensitivities (Lemma 2.3) as well as the above bound on the online ℓ_p Lewis weights, we obtain a bound on the sum of ℓ_p online sensitivities. This significantly generalizes Lemma 4.7 of [BDM$^+$20] and shaves a log factor.

Corollary 3.8 (Sum of Online Sensitivities)**.** *Let $p \in (0,\infty)$ and $\mathbf{A} \in \mathbb{R}^{n\times d}$. Define the ith online ℓ_p sensitivity as*

$$\mathbf{s}_i^{p,\mathsf{OL}}(\mathbf{A}) := \begin{cases} \min\Big\{1, \sup_{\|\mathbf{A}_{i-1}\mathbf{x}\|_p=1}|[\mathbf{A}\mathbf{x}](i)|^p\Big\} & \text{if } \mathbf{a}_i \in \mathrm{rowspan}(\mathbf{A}_{i-1}) \\ 1 & \text{otherwise} \end{cases}$$

Then,

$$\sum_{i=1}^n \mathbf{s}_i^{p,\mathsf{OL}}(\mathbf{A}) \leq O(d\log(n\kappa^{\mathsf{OL}}(\mathbf{A}))^{1\vee(p/2)}$$

Proof. We instead bound the sum of scores

$$\sup_{\|\mathbf{A}_i\mathbf{x}\|_p=1}|[\mathbf{A}\mathbf{x}](i)|^p.$$

Note that this is within a factor of 2 of $\mathbf{s}_i^{p,\mathsf{OL}}$, and is just $\mathbf{s}_i^p(\mathbf{A}_i)$. Indeed, for an ℓ_p unit vector $\mathbf{A}_i\mathbf{x}$ with $|[\mathbf{A}\mathbf{x}](i)|^p = a$, if $a \geq 1/2$, then we are done, and otherwise,

$$\frac{a}{1-a} - a = \frac{a^2}{1-a} \leq \frac{1}{2}\frac{a}{1-a} \implies \frac{a}{1-a} \leq 2a.$$

Then, by Lemma 2.10 of [WY22], we have that

$$\sup_{\|\mathbf{A}_i\mathbf{x}\|_p=1}|[\mathbf{A}\mathbf{x}](i)|^p \leq \left\|\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}_i)\right\|_1^{0\vee(p/2-1)}\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}_i)$$

so summing over i yields the desired result. □

The above proof applies to any set of weights $\mathbf{w}$ that satisfy an *online one-sided Lewis weights* property; we simply need to replace Lemma 2.10 of [WY22] with Lemma 2.3. This flexibility is useful when one desires an algorithmic approximation on the online sensitivities.

Definition 3.9 (Online γ-One-sided ℓ_p Lewis Weights)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p < \infty$. Let $\gamma \in (0,1]$. Then, $\mathbf{w} \in \mathbb{R}^n$ are* one-sided online ℓ_p Lewis weights *if for each $i \in [n]$, $\mathbf{w}$ restricted to the first i rows are γ-one-sided ℓ_p Lewis weights (Definition 2.2) of $\mathbf{A}_i$.*

Corollary 3.10. *Let $\mathbf{w}$ be online γ-one-sided ℓ_p Lewis weights for $\mathbf{A}$. Then,*

$$\sum_{i=1}^n \mathbf{s}_i^{p,\mathsf{OL}}(\mathbf{A}) \leq O\left(\frac{\|\mathbf{w}\|_1^{1\vee(p/2)}}{\gamma^{0\vee(1/p-1/2)}}\right)$$

4 Batch Processing of Rows

In many practical scenarios, it may be convenient to consider variants of the online coreset model in which multiple row may be processed in batches at a time, in order to save on running time. For example, this may corresponding to data arriving in packets over the internet. Such improvements are considered in [CMP20, BDM+20] for online leverage scores. We show how this can be done, even for online Lewis weights.

Recall that highly efficient algorithms for approximating leverage scores are known, running in time $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^\omega)$ for $(1+\varepsilon)$-factor approximations to the leverage scores [SS11, DMMW12, CLM+15]. We let $\textsc{ApproxLev}(\mathbf{A}, \varepsilon)$ refer to such a routine.

4.1 Batch Online Lewis Weights, $0 < p < 4$

Algorithm 1 Batch Online Lewis Weights, $p \in (0, 4)$

```
input: Previous Lewis quadratic $\mathbf{M}$, new rows $\mathbf{A} \in \mathbb{R}^{n \times d}$, $p \in (0, 4)$.
output: Batch online Lewis weight $\{\tilde{\mathbf{w}}_i\}_{i=1}^n$.
$\tilde{\mathbf{w}}_i^{(0)} \leftarrow 1$ for all $i \in [n]$
for $t \in [T]$ do
    $\mathbf{B} \leftarrow \begin{bmatrix} \operatorname{diag}(\tilde{\mathbf{w}}^{(t-1)})^{1/2-1/p}\mathbf{A} \\ \mathbf{M}^{1/2} \end{bmatrix}$
    $\tilde{\mathbf{w}}_i^{(t)} \leftarrow \textsc{ApproxLev}(\mathbf{B}, 1/10)$ for $i \in [n]$
return $\tilde{\mathbf{w}}^{(T)}$
```

We recall the following notation of [CP15]. For two nonnegative numbers v, w, we denote $v \approx_\alpha w$ to mean $v/\alpha \leq w \leq \alpha v$. We extend this to nonnegative vectors $\mathbf{v}, \mathbf{w}$, as well as to symmetric PSD matrices via the Löwner order.

We first adapt Lemma 3.2 of [CP15] to the batch online setting:

Lemma 4.1. *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $\mathbf{M} \in \mathbb{R}^{d \times d}$ be a symmetric PSD matrix. Suppose that nonnegative vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ satisfy $\mathbf{v} \approx_\alpha \mathbf{w}$ for $\alpha \geq 1$. Then,*

$$[\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{V}^{1-2/p}\mathbf{A} + \mathbf{M})^- \mathbf{a}_i]^{p/2} \approx_{\alpha^{|p/2-1|}} [\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A} + \mathbf{M})^- \mathbf{a}_i]^{p/2}.$$

Proof. For $\mathbf{v} \approx_\alpha \mathbf{w}$, we have $\mathbf{V}^{1-2/p} \approx_{\alpha^{|1-2/p|}} \mathbf{W}^{1-2/p}$. Then, $\mathbf{A}^\top \mathbf{V}^{1-2/p}\mathbf{A} \approx_{\alpha^{|1-2/p|}} \mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A}$, so $\mathbf{A}^\top \mathbf{V}^{1-2/p}\mathbf{A} + \mathbf{M} \approx_{\alpha^{|1-2/p|}} \mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A} + \mathbf{M}$. By using Lemma 2.6 to apply $\mathbf{a}_i$ to the pseudoinverse quadratic form and taking $p/2$-th powers, we find that

$$[\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{V}^{1-2/p}\mathbf{A} + \mathbf{M})^- \mathbf{a}_i]^{p/2} \approx_{\alpha^{|p/2-1|}} [\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A} + \mathbf{M})^- \mathbf{a}_i]^{p/2}. \qquad \square$$

By the Banach fixed point theorem, Lemma 4.1 implies the existence of the following weights:

Corollary 4.2 (Batch Online ℓ_p Lewis Weights)**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$, let $\mathbf{M} \in \mathbb{R}^{d \times d}$ be a symmetric PSD matrix, and let $0 < p < 4$. There exists weights $\mathbf{w}_i^p(\mathbf{A}; \mathbf{M})$ such that*

$$\mathbf{w}_i^p(\mathbf{A}; \mathbf{M}) = \left[\mathbf{a}_i^\top (\mathbf{M} + \mathbf{A}^\top \mathbf{W}^p(\mathbf{A}; \mathbf{M})^{1-p/2}\mathbf{A})^- \mathbf{a}_i\right]^{p/2}.$$

We now show that Algorithm 1 returns multiplicative approximations to the batch online ℓ_p Lewis weights of Corollary 4.2. We start by showing that after one iteration, we obtain a good approximation within $\mathrm{poly}(n)$ factors, by adapting Lemma 3.5 of [CP15].

Lemma 4.3. *After $t = 1$ in Algorithm 1, we have that $\tilde{\mathbf{w}}_i \approx_{\beta^{p/2} n^{|p/2-1|}} \mathbf{w}_i^p(\mathbf{A}; \mathbf{M})$ for $\beta = 1 + 1/10$.*

Proof. By rearranging the guarantee of $\mathbf{w}_i^p(\mathbf{A}; \mathbf{M})$ in Corollary 4.2, we see that $\mathbf{w}_i^p(\mathbf{A}; \mathbf{M})$ satisfies

$$\mathbf{w}_i = (\mathbf{w}_i^{1/2-1/p}\mathbf{a}_i)^\top (\mathbf{M} + \mathbf{A}^\top \mathbf{W}^p(\mathbf{A}; \mathbf{M})^{1-p/2}\mathbf{A})^- (\mathbf{w}_i^{1/2-1/p}\mathbf{a}_i).$$

That is, $\mathbf{w}_i^p(\mathbf{A};\mathbf{M})$ are exactly the first n leverage scores of the matrix

$$\mathbf{B} = \begin{bmatrix} \mathbf{W}^p(\mathbf{A};\mathbf{M})^{1/2-1/p}\mathbf{A} \\ \mathbf{M}^{1/2} \end{bmatrix}.$$

Thus, there exists a change of basis $\mathbf{R}$ such that $\mathbf{BR}$ has orthonormal columns. We then rename $\mathbf{A}$ to $\mathbf{AR}$ and $\mathbf{M}^{1/2}$ to $\mathbf{M}^{1/2}\mathbf{R}$ and proceed by assuming that $\mathbf{B}$ has orthonormal columns. Then, $\mathbf{B}^\top\mathbf{B} = \mathbf{I}_d$. We then claim that $\mathbf{A}^\top\mathbf{A} + \mathbf{M} \approx_{n^{|1-2/p|}} \mathbf{I}_d$.

Note that $\|\mathbf{b}_i\|_2^2 = \mathbf{w}_i^p(\mathbf{A};\mathbf{M})$. Then for any unit vector $\mathbf{u}$, we have that

$$1 = \mathbf{u}^\top\mathbf{u} = \mathbf{u}^\top\mathbf{B}^\top\mathbf{B}\mathbf{u} = \sum_{i=1}^n (\mathbf{u}^\top\mathbf{b}_i)^2 + \mathbf{u}^\top\mathbf{M}\mathbf{u} = \sum_{i=1}^n \mathbf{w}_i^p(\mathbf{A};\mathbf{M})[(\mathbf{w}_i^p(\mathbf{A};\mathbf{M})^{-1}\mathbf{u}^\top\mathbf{b}_i)^2] + \mathbf{u}^\top\mathbf{M}\mathbf{u}$$

while

$$\begin{aligned}\mathbf{u}^\top\mathbf{A}^\top\mathbf{A}\mathbf{u} + \mathbf{u}^\top\mathbf{M}\mathbf{u} &= \sum_{i=1}^n \mathbf{w}_i^p(\mathbf{A};\mathbf{M})^{2/p-1}(\mathbf{u}^\top\mathbf{b}_i)^2 + \mathbf{u}^\top\mathbf{M}\mathbf{u} \\ &= \sum_{i=1}^n \mathbf{w}_i^p(\mathbf{A};\mathbf{M})^{2/p}[\mathbf{w}_i^p(\mathbf{A};\mathbf{M})^{-1}(\mathbf{u}^\top\mathbf{b}_i)^2] + \mathbf{u}^\top\mathbf{M}\mathbf{u}.\end{aligned}$$

Then just as reasoned in [CP15], the worst case distortion is $n^{|p/2-1|}$ between these two quantities. □

Corollary 4.4. *Let $T = O(\log\log n)$ and let $\tilde{\mathbf{w}}$ be the output of Algorithm 1. Then, $\tilde{\mathbf{w}} \approx_{O(1)} \mathbf{w}_i^p(\mathbf{A};\mathbf{M})$.*

Proof. The total multiplicative contribution of the blowups from β is at most $\beta^{\frac{p/2}{1-|p/2-1|}}$ for $\beta = 1 + 1/10$. The contribution from the starting error is at most $n^{|p/2-1|^T}$. Thus, for $T = O(\log\log n)$, we obtain an $O(1)$-approximation to the batch online ℓ_p Lewis weights. □

Using these, we obtain that batch processing yields weights that are still one-sided ℓ_p Lewis weights, and are bounded by the online leverage scores of a reweighted matrix.

Lemma 4.5. *Let $p \in (0,4)$, let $\mathbf{A} \in \mathbb{R}^{n\times d}$, and let $\mathbf{B} \in \mathbb{R}^{m\times d}$. Let*

$$\mathbf{C} = \begin{bmatrix} \mathbf{B} \\ \mathbf{A} \end{bmatrix}$$

Let $\tilde{\mathbf{w}}_i$ for $i \in [m]$ be γ-one-sided ℓ_p Lewis weights of $\mathbf{B}$. Let $\mathbf{M} = \mathbf{B}^\top\tilde{\mathbf{W}}^{1-2/p}\mathbf{B}$, and let $\hat{\mathbf{w}}$ satisfy $\mathbf{w}_i^p(\mathbf{A};\mathbf{M}) \leq \hat{\mathbf{w}} \leq \lambda\mathbf{w}_i^p(\mathbf{A};\mathbf{M})$. Then, the concatenation $\mathbf{w} = \tilde{\mathbf{w}} \circ \hat{\mathbf{w}}$ are γ-one-sided ℓ_p Lewis weights of $\mathbf{C}$ if $p \in (0,2]$ and $\min\{\gamma, \lambda^{1-p/2}\}$-one-sided ℓ_p Lewis weights of $\mathbf{C}$ if $p \in (2,4)$. Furthermore, if

$$\mathbf{w}_i \leq O(1)\boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{W}^{1/2-1/p}\mathbf{C})$$

for the first m rows of $\mathbf{C}$, then this is also true for all the rows. The weights $\hat{\mathbf{w}}$ can be computed in time $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^\omega)$ given $\mathbf{M}$.

Proof. For the last n rows, if $0 < p \leq 2$, we have by Lemma 2.6 that

$$\begin{aligned}\hat{\mathbf{w}}_i \geq \mathbf{w}_i^p(\mathbf{A};\mathbf{M}) &= \left[\mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{W}^p(\mathbf{A};\mathbf{M})^{1-2/p}\mathbf{A} + \mathbf{M})^-\mathbf{a}_i\right]^{p/2} \\ &\geq \left[\mathbf{a}_i^\top(\mathbf{A}^\top\hat{\mathbf{W}}^{1-2/p}\mathbf{A} + \mathbf{M})^-\mathbf{a}_i\right]^{p/2} \\ &= \left[\mathbf{a}_i^\top(\mathbf{C}^\top\mathbf{W}^{1-2/p}\mathbf{C})^-\mathbf{a}_i\right]^{p/2}\end{aligned}$$

and if $2 < p < 4$, that

$$\hat{\mathbf{w}}_i \geq \mathbf{w}_i^p(\mathbf{A};\mathbf{M}) = \lambda^{1-p/2}\left[\mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{W}^p(\mathbf{A};\mathbf{M})^{1-2/p}\mathbf{A} + \mathbf{M})^-\mathbf{a}_i\right]^{p/2}$$

$$
\begin{aligned}
&\geq \lambda^{1-p/2}\Big[\mathbf{a}_i^\top(\mathbf{A}^\top\hat{\mathbf{W}}^{1-2/p}\mathbf{A}+\mathbf{M})^-\mathbf{a}_i\Big]^{p/2} \\
&= \lambda^{1-p/2}\Big[\mathbf{a}_i^\top(\mathbf{C}^\top\mathbf{W}^{1-2/p}\mathbf{C})^-\mathbf{a}_i\Big]^{p/2}
\end{aligned}
$$

For the first m rows, we have by Lemma 2.6 that

$$
\begin{aligned}
\tilde{\mathbf{w}}_i &\geq \gamma\big[\mathbf{b}_i^\top(\mathbf{M})^-\mathbf{b}_i\big]^{p/2} \\
&\geq \gamma\Big[\mathbf{b}_i^\top(\mathbf{A}^\top\hat{\mathbf{W}}^{1-2/p}\mathbf{A}+\mathbf{M})^-\mathbf{b}_i\Big]^{p/2} \\
&= \gamma\Big[\mathbf{b}_i^\top(\mathbf{C}^\top\mathbf{W}^{1-2/p}\mathbf{C})^-\mathbf{b}_i\Big]^{p/2}.
\end{aligned}
$$

These rearrange to the statement that $\mathbf{w}$ are γ-one-sided ℓ_p Lewis weights.

Furthermore, we have that

$$
\begin{aligned}
\hat{\mathbf{w}}_i \leq O(1)\mathbf{w}_i^p(\mathbf{A};\mathbf{M}) &= O(1)\Big[\mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{W}^p(\mathbf{A};\mathbf{M})^{1-2/p}\mathbf{A}+\mathbf{M})^-\mathbf{a}_i\Big]^{p/2} \\
&\leq O(1)\Big[\mathbf{a}_i^\top(\mathbf{A}^\top\hat{\mathbf{W}}^{1-2/p}\mathbf{A}+\mathbf{M})^-\mathbf{a}_i\Big]^{p/2} \\
&\leq O(1)\Big[\mathbf{a}_i^\top(\mathbf{C}_{i-1}^\top\mathbf{W}_{i-1}^{1-2/p}\mathbf{C}_{i-1})^-\mathbf{a}_i\Big]^{p/2}
\end{aligned}
$$

which rearranges to the statement that $\mathbf{w}_i \leq O(1)\boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{W}^{1/2-1/p}\mathbf{C})$. □

4.2 Tensor Trick for Batch Online Lewis Weights, $4 \leq p < \infty$

For $p \geq 4$, the iterative algorithm of Algorithm 1 does not work directly. Instead, we use a trick of [MMM+22] to reduce $p \geq 4$ to the case of $p < 4$ as follows. Let k be an integer large enough so that $2 \leq p/k < 4$ (i.e., $k = \lfloor p/4 \rfloor + 1$). Then, we define the Khatri–Rao power $\mathbf{A}^{\otimes k} \in \mathbb{R}^{n\times d^k}$, where the ith row is the k-fold tensor product $\mathbf{a}_i^{\otimes k} = \mathbf{a}_i \otimes \mathbf{a}_i \otimes \cdots \otimes \mathbf{a}_i$ of $\mathbf{a}_i$ with itself. Note then that for any $\mathbf{x} \in \mathbb{R}^d$,

$$
\big\langle \mathbf{a}_i^{\otimes k}, \mathbf{x}^{\otimes k}\big\rangle = \langle \mathbf{a}_i, \mathbf{x}\rangle^k,
$$

so

$$
\big\|\mathbf{A}^{\otimes k}\mathbf{x}^{\otimes k}\big\|_{p/k}^{p/k} = \sum_{i=1}^n \big|\big\langle \mathbf{a}_i^{\otimes k}, \mathbf{x}^{\otimes k}\big\rangle\big|^{p/k} = \sum_{i=1}^n |\langle \mathbf{a}_i, \mathbf{x}\rangle|^p = \|\mathbf{A}\mathbf{x}\|_p^p.
$$

Thus, it suffices to compute sensitivity upper bounds and subspace embeddings for $\ell_{p/k}$ on the Khatri–Rao power matrix $\mathbf{A}^{\otimes k}$ in order to handle ℓ_p for $\mathbf{A}$. Note that the $\ell_{p/k}$ Lewis weights of $\mathbf{A}^{\otimes k}$ sum to $(d^k)^{p/2k} = d^{p/2}$, so the upper bound on the sum of sensitivities is as desired. We may then just apply Lemma 4.5 on the Khatri–Rao power matrix.

4.3 Batch Online Lewis Weights, $2 \leq p < \infty$

While the tensor trick of Section 4.2 is sufficient for many cases, it is still desirable to avoid this and directly get an analogue of Corollary 4.2, for example in our Section 5. We will obtain this by modifying the classical convex program argument for the existence of ℓ_p Lewis weights [Lew78, SZ01, CP15]. This also immediately leads to polynomial time algorithms [CP15].

Lemma 4.6 (Batch Online ℓ_p Lewis Weights, $2 \leq p < \infty$)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $\mathbf{M} = \mathbf{L}^\top\mathbf{L} \in \mathbb{R}^{d\times d}$ be a symmetric PSD matrix, and let $2 \leq p < \infty$. There exists weights $\mathbf{w} \in \mathbb{R}^n$ such that for $i \in [n]$,*

$$
\mathbf{w}_i = \left(\frac{p}{2}\right)^{\frac{p/2}{1-2/p}}(\mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{W}^{1-2/p}\mathbf{A}+\mathbf{M})^{-1}\mathbf{a}_i)^{p/2}
$$

and

$$
\sum_{i=1}^n \mathbf{w}_i \leq \left(\frac{p}{2}\right)^{\frac{1}{1-2/p}} d.
$$

Proof. Consider the following optimization problem over symmetric PSD matrices $\mathbf{Q}$:

$$\begin{aligned} \text{maximize} \quad & \det(\mathbf{Q}) \\ \text{subject to} \quad & \sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{Q} \mathbf{a}_i)^{p/2} + \sum_{j=1}^d \mathbf{l}_j^\top \mathbf{Q} \mathbf{l}_j \leq d \\ & \mathbf{Q} \succeq 0 \end{aligned}$$

where $\mathbf{a}_i$ is the ith row of $\mathbf{A}$ and $\mathbf{l}_j$ is the jth row of $\mathbf{L}$. Let $\mathbf{Q}$ be any matrix which attains this maximum. Note then that

$$\sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{Q} \mathbf{a}_i)^{p/2} + \sum_{j=1}^d \mathbf{l}_j^\top \mathbf{Q} \mathbf{l}_j = d$$

since otherwise scaling $\mathbf{Q}$ up can increase the objective function. Furthermore, by considering Lagrange multipliers, the gradient of the constraint is some scalar C times the gradient of of the objective, so

$$\sum_{i=1}^n \frac{p}{2} (\mathbf{a}_i^\top \mathbf{Q} \mathbf{a}_i)^{p/2-1} \mathbf{a}_i \mathbf{a}_i^\top + \sum_{j=1}^d \mathbf{l}_j \mathbf{l}_j^\top = C \det(\mathbf{Q}) \mathbf{Q}^{-1}.$$

We now define

$$\mathbf{w}_i \coloneqq \left(\frac{p}{2}\right)^{\frac{1}{1-2/p}} (\mathbf{a}_i^\top \mathbf{Q} \mathbf{a}_i)^{p/2}.$$

Then, we have that

$$\mathbf{A}^\top \mathbf{W}^{1-2/p} \mathbf{A} + \mathbf{M} = C \det(\mathbf{Q}) \mathbf{Q}^{-1}$$

for $\mathbf{W} = \operatorname{diag}(\mathbf{w})$. Rearranging, we have that

$$\mathbf{Q} = C \det(\mathbf{Q}) (\mathbf{A}^\top \mathbf{W}^{1-2/p} \mathbf{A} + \mathbf{M})^{-1}$$

so

$$\mathbf{w}_i = \left(\frac{p}{2}\right)^{\frac{1}{1-2/p}} (\mathbf{a}_i \mathbf{Q} \mathbf{a}_i)^{p/2} = \left(\frac{p}{2}\right)^{\frac{1}{1-2/p}} (C \det(\mathbf{Q}))^{p/2} [\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p} \mathbf{A} + \mathbf{M})^{-1} \mathbf{a}_i]^{p/2}$$

and thus

$$\begin{aligned} \mathbf{w}_i &= \left(\frac{p}{2}\right)^{\frac{2/p}{1-2/p}} (C \det(\mathbf{Q})) [(\mathbf{w}_i^{1/2-1/p} \mathbf{a}_i)^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p} \mathbf{A} + \mathbf{M})^{-1} (\mathbf{w}_i^{1/2-1/p} \mathbf{a}_i)] \\ &= \left(\frac{p}{2}\right)^{\frac{2/p}{1-2/p}} (C \det(\mathbf{Q})) \boldsymbol{\tau}_i(\mathbf{B}) \end{aligned}$$

where $\mathbf{B}$ is the vertical concatenation of $\mathbf{W}^{1/2-1/p}\mathbf{A}$ and $\mathbf{L}$. Note also that for rows j corresponding to $\mathbf{L}$ in $\mathbf{B}$, we have that

$$(C \det(\mathbf{Q})) \boldsymbol{\tau}_j(\mathbf{B}) = (C \det(\mathbf{Q})) \mathbf{l}_j^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p} \mathbf{A} + \mathbf{M})^{-1} \mathbf{l}_j = \mathbf{l}_j^\top \mathbf{Q} \mathbf{l}_j.$$

Now by the normalization constraint, we have that

$$\sum_{i=1}^n \left(\frac{2}{p}\right)^{\frac{1}{1-2/p}} \mathbf{w}_i + \sum_{j=1}^d \mathbf{l}_j^\top \mathbf{Q} \mathbf{l}_j = \sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{Q} \mathbf{a}_i)^{p/2} + \sum_{j=1}^d \mathbf{l}_j^\top \mathbf{Q} \mathbf{l}_j = d.$$

However,

$$\left(\frac{2}{p}\right)^{\frac{1}{1-2/p}} \mathbf{w}_i = \left(\frac{p}{2}\right)^{\frac{-1}{1-2/p}} \left(\frac{p}{2}\right)^{\frac{2/p}{1-2/p}} (C \det(\mathbf{Q})) \boldsymbol{\tau}_i(\mathbf{B}) = \frac{2}{p} (C \det(\mathbf{Q})) \boldsymbol{\tau}_i(\mathbf{B})$$

so we must have that $p/2 = C \det(\mathbf{Q})$. The result follows. □

Remark 4.7. *Note that if we set $\mathbf{M} = 0$ and redefine $\mathbf{w}_i' \coloneqq \mathbf{w}_i / (p/2)^{\frac{1}{1-2/p}}$, then we will retrieve the usual definition of ℓ_p Lewis weights.*

5 Nearly Optimal One-Shot Lewis Weight Sampling, $2 < p < \infty$

The following is a generalization of what is known for Lewis weights to the setting of one-sided Lewis weights, with higher moments.

Theorem 5.1. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \in (2,\infty)$. Let $\mathbf{w}$ be γ-one-sided ℓ_p Lewis weights for $\mathbf{A}$. Suppose that*

$$\frac{\mathbf{w}_i}{d} \leq \beta$$

for all $i \in [n]$, for some $\beta > 0$. Define the quantity

$$\Lambda \coloneqq \sup_{\|\mathbf{A}\mathbf{x}\|_p = 1} \left| \sum_{i=1}^n \boldsymbol{\sigma}_i |\langle \mathbf{a}_i, \mathbf{x}\rangle|^p \right|$$

Then,

$$\mathop{\mathbf{E}}_{\boldsymbol{\sigma}}[\Lambda^l] \leq \left[O(1)\beta \cdot T_{\mathbf{w}}^{p/2}[\gamma^{-1}(\log d)^2(\log n) + l]\right]^{l/2}$$

for any $l \geq 1$, where $\boldsymbol{\sigma} = \{\boldsymbol{\sigma}_i\}_{i=1}^n$ are independent Rademacher variables.

We work out the proof of this result in detail in Appendix C. Using this, we obtain the following analysis for a one-shot Lewis weight sampling algorithm.

Theorem 5.2 (High probability one-shot Lewis weight sampling)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $2 < p < \infty$. Let $\delta \in (0,1)$ be a failure rate parameter and let $\varepsilon \in (0,1)$ be an accuracy parameter. Let $\mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights that sum to $T_{\mathbf{w}}$. Suppose that we set $\mathbf{s}_i = 1/\mathbf{p}_i^{1/p}$ with probability $\mathbf{p}_i$, where*

$$\mathbf{p}_i \geq \min\left\{\left(\frac{(p/2)^{\frac{1}{1-2/p}}}{\gamma}\right)^{p/2} \frac{\mathbf{w}_i}{d\beta}, 1\right\},$$

for

$$\beta = \frac{\varepsilon^2}{T^{p/2}[\gamma^{-1}(\log d)^2(\log n) + \log\frac{1}{\delta}]} \in (0,1)$$

with

$$T \coloneqq T_{\mathbf{w}} + O(d).$$

Then, with probability at least $1-\delta$,

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

for all $\mathbf{x} \in \mathbb{R}^d$, and $\mathbf{s}$ samples at most

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

rows of $\mathbf{A}$ with probability at least $1-\delta$.

Proof. We WLOG assume that $\mathbf{A}$ is just the subset of rows of the original matrix such that $\mathbf{p}_i < 1$. In particular, we may assume that $\mathbf{w}_i < 1$ for all $i \in [n]$.

Symmetrization. We wish to bound

$$\mathop{\mathbf{E}}_{\mathbf{s}}\left[\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\left(\sum_{i=1}^n |\langle \mathbf{s}_i\mathbf{a}_i, \mathbf{x}\rangle|^p\right) - 1\right|^l\right], \tag{4}$$

for $l = O(\log \frac{1}{\delta})$. Indeed, if we can bound this by $(C \cdot \varepsilon)^l$ for $C > 0$, then by Markov's inequality,

$$\mathbf{Pr}\left\{ \sup_{\|\mathbf{A}\mathbf{x}\|_p = 1} \left| \left(\sum_{i=1}^n |\langle \mathbf{s}_i \mathbf{a}_i, \mathbf{x} \rangle|^p \right) - 1 \right| \geq \frac{C \cdot \varepsilon}{\delta^{1/l}} \right\} = \mathbf{Pr}\left\{ \sup_{\|\mathbf{A}\mathbf{x}\|_p = 1} \left| \left(\sum_{i=1}^n |\langle \mathbf{s}_i \mathbf{a}_i, \mathbf{x} \rangle|^p \right) - 1 \right|^l \geq \frac{(C \cdot \varepsilon)^l}{\delta} \right\} \leq \delta,$$

which implies that

$$\mathbf{Pr}\left\{ \text{for all } \mathbf{x} \in \mathbb{R}^d,\ \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p = (1 \pm O(\varepsilon))\|\mathbf{A}\mathbf{x}\| \right\} \geq 1 - \delta.$$

By a standard symmetrization procedure [CP15, Lemma 7.4], (4) is bounded by

$$2^l \mathop{\mathbf{E}}_{\mathbf{s}} \mathop{\mathbf{E}}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{A}\mathbf{x}\|_p = 1} \left| \sum_{i=1}^n \boldsymbol{\sigma}_i |\langle \mathbf{s}_i \mathbf{a}_i, \mathbf{x} \rangle|^p \right|^l \right].$$

We now define a new matrix $\mathbf{A}''$ along with weights $\mathbf{w}_i''$ which we will show satisfy the following:

Condition 5.3.

- $\mathbf{w}''$ *are* $\min\{\gamma, \Omega(1)\}$*-one-sided Lewis weights for* $\mathbf{A}''$
- $\mathbf{w}_i''/d \leq \beta$
- $\sum_i \mathbf{w}_i'' = O(T)$

These three items will allow us to apply Theorem 5.1 on $\mathbf{A}''$ with weights $\mathbf{w}''$.

Flattening A. Let $\mathbf{A}'$ be the original matrix $\mathbf{A}$ with the same weights $\mathbf{w}$, except that whenever $\mathbf{w}_i \geq d\beta$, we replace $\mathbf{a}_i$ with k copies of $\mathbf{a}_i/k^{1/p}$ for $k = \lceil 1/d\beta \rceil$, and set $\mathbf{w}_j' = \mathbf{w}_i/k$ for each copy j of the original row i. Note then that

$$\mathbf{A}^\top \mathbf{W}^{1-2/p} \mathbf{A} = \mathbf{A}'^\top \mathbf{W}'^{1-2/p} \mathbf{A}'$$

since

$$\mathbf{w}_i^{1-2/p} \mathbf{a}_i \mathbf{a}_i^\top = k \cdot \left(\frac{\mathbf{w}_i}{k} \right)^{1-2/p} \left(\frac{\mathbf{a}_i}{k^{1/p}} \right) \left(\frac{\mathbf{a}_i}{k^{1/p}} \right)^\top.$$

Then,

$$\begin{aligned} \mathbf{w}_j' = \frac{\mathbf{w}_i}{k} &\geq \frac{\gamma \cdot \boldsymbol{\tau}_i(\mathbf{W}^{1/2-1/p}\mathbf{A})}{k} \\ &= \gamma \cdot \frac{(\mathbf{w}_i^{1/2-1/p}\mathbf{a}_i)^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p} \mathbf{A})^- (\mathbf{w}_i^{1/2-1/p}\mathbf{a}_i)}{k} \\ &= \gamma \cdot k^{1-2/p} \cdot k^{2/p} \frac{(\mathbf{w}_j'^{1/2-1/p}\mathbf{a}_i/k^{1/p})^\top (\mathbf{A}'^\top \mathbf{W}'^{1-2/p} \mathbf{A}')^- (\mathbf{w}_j'^{1/2-1/p}\mathbf{a}_i/k^{1/p})}{k} \\ &= \gamma \cdot \boldsymbol{\tau}_j(\mathbf{W}'^{1/2-1/p}\mathbf{A}') \end{aligned}$$

so $\mathbf{w}_j'$ are γ-one-sided ℓ_p Lewis weights for $\mathbf{A}'$. Clearly, we also have $\mathbf{w}_j' \leq d\beta$.

Extending w′ via Batch Online Lewis Weights. We now define $\mathbf{A}''$ to be the matrix

$$\mathbf{A}'' = \begin{bmatrix} \mathbf{A}' \\ \mathbf{S}\mathbf{A} \end{bmatrix}$$

where $\mathbf{S}\mathbf{A}$ is the sampled matrix. We then set $\mathbf{w}_i''$ to be $\mathbf{w}_i'$ for rows corresponding to $\mathbf{A}'$, and we set $\mathbf{w}_i''$ to be the batch online ℓ_p Lewis weights of $\mathbf{S}\mathbf{A}$ with respect to $\mathbf{M} = \mathbf{A}'^\top \mathbf{W}'^{1-2/p} \mathbf{A}' = \mathbf{A}\mathbf{W}^{1-2/p}\mathbf{A}$, as given by Lemma 4.6, for rows corresponding to $\mathbf{S}\mathbf{A}$.

We now show that $\mathbf{w}''$ satisfy Condition 5.3. We start with the first item. For rows i corresponding to $\mathbf{A}'$, this follows from the fact that

$$\mathbf{w}''_i = \mathbf{w}'_i \geq \gamma \cdot \boldsymbol{\tau}_i(\mathbf{W}'^{1/2-1/p}\mathbf{A}') \geq \gamma \cdot \boldsymbol{\tau}_i(\mathbf{W}''^{1/2-1/p}\mathbf{A}'')$$

by the monotonicity of leverage scores under row additions. For rows i corresponding to $\mathbf{SA}$, this follows from the guarantees of Lemma 4.6, which rearranges to the statement that

$$\mathbf{w}''_i = \left(\frac{p}{2}\right)^{\frac{1}{1-2/p}} \boldsymbol{\tau}_i(\mathbf{W}''^{1/2-1/p}\mathbf{A}'').$$

We now show the second item of Condition 5.3. This is immediate by definition of $\mathbf{w}''$ for rows corresponding to $\mathbf{A}'$. For rows corresponding to $\mathbf{SA}$, letting $\hat{\mathbf{w}}$ denote the weights $\mathbf{w}''$ restricted to the rows of $\mathbf{SA}$,

$$\begin{aligned}\hat{\mathbf{w}}_i &= \left(\frac{p}{2}\right)^{\frac{p/2}{1-2/p}} ((\mathbf{s}_i\mathbf{a}_i)^\top((\mathbf{SA})\hat{\mathbf{W}}^{1-2/p}(\mathbf{SA}) + \mathbf{A}^\top\mathbf{WA})^{-1}(\mathbf{s}_i\mathbf{a}_i))^{p/2} \\ &\leq \left(\frac{p}{2}\right)^{\frac{p/2}{1-2/p}} \mathbf{s}_i^p(\mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{WA})^{-1}\mathbf{a}_i)^{p/2} \\ &\leq \left(\frac{p}{2}\right)^{\frac{p/2}{1-2/p}} \frac{1}{\mathbf{p}_i}\frac{\mathbf{w}_i}{\gamma^{p/2}} \leq d\beta.\end{aligned}$$

The third item follows from the fact that the weights restricted to $\mathbf{SA}$ sum to at most $O(d)$. Then, by Theorem 5.1,

$$2^l \mathop{\mathbf{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{A}''\mathbf{x}\|_p=1}\left|\sum_{i=1}^n \boldsymbol{\sigma}_i|\langle\mathbf{a}''_i,\mathbf{x}\rangle|^p\right|^l\right] \leq \left[O(1)\beta\cdot T^{p/2}[\gamma^{-1}(\log d)^2(\log n)+l]\right]^{l/2} \leq O(\varepsilon)^l.$$

High probability error bounds. We now finish the argument as in [CP15, CD21]. For a given fixing of $\mathbf{s}$, let

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

Then for the corresponding $\mathbf{A}''$, we have for all $\mathbf{x}$ that

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

so

$$\sup_{\|\mathbf{Ax}\|_p=1}\left|\sum_{i=1}^{n'}\boldsymbol{\sigma}_i|\langle\mathbf{a}''_i,\mathbf{x}\rangle|^p\right| \leq (2+F_{\mathbf{s}})\sup_{\|\mathbf{A}''\mathbf{x}\|_p=1}\left|\sum_{i=1}^{n'}\boldsymbol{\sigma}_i|\langle\mathbf{a}''_i,\mathbf{x}\rangle|^p\right| \leq O(2+F_{\mathbf{s}})\varepsilon.$$

Altogether, we have that

$$\mathop{\mathbf{E}}_{\mathbf{s}}[F_{\mathbf{s}}^l] \leq \mathop{\mathbf{E}}_{\mathbf{s}}[O(2+F_{\mathbf{s}})^l\varepsilon^l] \leq (O(1)^l + \mathbf{E}[F_{\mathbf{s}}^l])O(\varepsilon)^l$$

which means that $\mathbf{E}_{\mathbf{s}}[F_{\mathbf{s}}] \leq O(\varepsilon)^l$. □

Theorem 1.3 is a simple corollary of this result by specializing to one-sided ℓ_p Lewis weights as given in [Lee16, JLS21].

6 Deterministic Online Lewis Weight Estimation

The following is the deterministic OnlineBSS algorithm of [BDM+20], which derandomizes an algorithm of [CMP20] with only logarithmic overheads.

Theorem 6.1 (Deterministic OnlineBSS, Appendix B.2, [BDM+20])**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\varepsilon \in (0,1)$ be an accuracy parameter. Then, there is a deterministic online coreset algorithm* DetOnlineBSS *which stores a weighted subset $\tilde{\mathbf{A}}$ of m rows for*

$$m = O\left(\frac{d}{\varepsilon^2}(\log n)(\log \kappa(\mathbf{A}))\right)$$

such that

$$(1-\varepsilon)\tilde{\mathbf{A}}_i^\top \tilde{\mathbf{A}}_i \preceq \mathbf{A}_i^\top \mathbf{A}_i \preceq (1+\varepsilon)\tilde{\mathbf{A}}_i^\top \tilde{\mathbf{A}}_i$$

for all $i \in [n]$.

Using this primitive, we give Algorithm 2, which is a deterministic algorithm for estimating weights $\tilde{\mathbf{w}}_i$.

Algorithm 2 Deterministic Online Lewis Weight Estimation

```
input: $\mathbf{A} \in \mathbb{R}^{n\times d}$, $p \in (0,\infty)$.
output: Online Lewis weight estimates $\{\tilde{\mathbf{w}}_i\}_{i=1}^n$.
1: Initialize DetOnlineBSS with $\varepsilon = 1/2$ (Theorem 6.1)
2: for $i \in [n]$ do
3:     if $\mathrm{rank}(\mathbf{A}_i) > \mathrm{rank}(\mathbf{A}_{i-1})$ then
4:         $\tilde{\mathbf{w}}_i \leftarrow 1$
5:     else
6:         Let $\tilde{\mathbf{M}}_{i-1}$ be the approximate quadratic form provided by DetOnlineBSS
7:         $\tilde{\mathbf{w}}_i \leftarrow \min\left\{[2\cdot \mathbf{a}_i^\top \tilde{\mathbf{M}}_{i-1}^- \mathbf{a}_i]^{p/2}, 1\right\}$
8:     Update DetOnlineBSS with the row $\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i$
9: return $\{\tilde{\mathbf{w}}_i\}_{i=1}^n$
```

We first show that the weights generated by Algorithm 2 are one-sided ℓ_p Lewis weights.

Lemma 6.2. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $0 < p < \infty$. Let $\tilde{\mathbf{w}}_i$ be weights generated by Algorithm 2, and let $\tilde{\mathbf{W}} = \mathrm{diag}(\tilde{\mathbf{w}})$. Then, for all $i \in [n]$,*

$$\tilde{\mathbf{w}}_i \geq \boldsymbol{\tau}_i\left(\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A}\right).$$

Proof. If $\tilde{\mathbf{w}}_i \geq 1$, then we already have the claim since leverage scores are always at most 1, so assume that $\tilde{\mathbf{w}}_i < 1$. In particular, we have that $\mathbf{a}_i \in \mathrm{rowspan}(\mathbf{A}_{i-1})$. Then, with $\tilde{\mathbf{M}}_{i-1}$ defined as in Line 6,

$$\tilde{\mathbf{M}}_{i-1} \preceq \frac{3}{2}\tilde{\mathbf{A}}_{i-1}^\top \tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1}$$

by the deterministic guarantee of DetOnlineBSS. Then by Lemma 2.6,

$$\tilde{\mathbf{w}}_i^{2/p} = 2\cdot \mathbf{a}_i^\top \tilde{\mathbf{M}}_{i-1}^- \mathbf{a}_i \geq 2\cdot\frac{2}{3}\mathbf{a}_i^\top(\tilde{\mathbf{A}}_{i-1}^\top \tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1})^-\mathbf{a}_i > \mathbf{a}_i^\top(\tilde{\mathbf{A}}_{i-1}^\top \tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1})^-\mathbf{a}_i$$

so

$$\tilde{\mathbf{w}}_i \geq (\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i)^\top(\mathbf{A}_{i-1}^\top \tilde{\mathbf{W}}_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-(\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i) = \boldsymbol{\tau}_i^{\mathsf{OL}}(\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A}) \geq \boldsymbol{\tau}_i(\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A}),$$

which is the desired statement. □

Next, we show that these weights have a small sum. For this, we first show that these weights are bounded by the online leverage scores of a reweighted matrix:

Lemma 6.3. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p < \infty$. Let $\tilde{\mathbf{w}}_i$ be weights generated by Algorithm 2, and let $\tilde{\mathbf{W}} = \mathrm{diag}(\tilde{\mathbf{w}})$. Then,*

$$\tilde{\mathbf{w}}_i \leq 4\cdot \boldsymbol{\tau}_i^{\mathsf{OL}}(\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A})$$

Proof. We have equality for $\mathbf{a}_i \notin \operatorname{rowspan}(\mathbf{A}_{i-1})$, with both being 1. It remains to bound the $\tilde{\mathbf{w}}_i$ for $\mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1})$. Then, with $\tilde{\mathbf{M}}_{i-1}$ defined as in Line 6,

$$\tilde{\mathbf{M}}_{i-1} \succeq \frac{1}{2}\tilde{\mathbf{A}}_{i-1}^\top \tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1}$$

by the deterministic guarantee of DetOnlineBSS. Then by Lemma 2.6,

$$\tilde{\mathbf{w}}_i^{2/p} \leq 2\cdot \mathbf{a}_i^\top \tilde{\mathbf{M}}_{i-1}^- \mathbf{a}_i \leq 2\cdot 2\cdot \mathbf{a}_i^\top (\tilde{\mathbf{A}}_{i-1}^\top \tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1})^- \mathbf{a}_i = 4\cdot \mathbf{a}_i^\top (\tilde{\mathbf{A}}_{i-1}^\top \tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1})^- \mathbf{a}_i$$

so

$$\tilde{\mathbf{w}}_i \leq 4\cdot (\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i)^\top (\mathbf{A}_{i-1}^\top \tilde{\mathbf{W}}_{i-1}^{1-2/p}\mathbf{A}_{i-1})^- (\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i) = 4\cdot \boldsymbol{\tau}_i^{\mathsf{OL}}(\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A}).$$

□

It remains to bound the sum of the online leverage scores of $\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A}$, which involves bounding the condition number of $\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A}$. For $p > 2$, this is related to the condition number of $\mathbf{A}$ up to a $\operatorname{poly}(n)$ factor by Lemmas 2.3 and 2.4. For $0 < p < 2$, we directly compare $\tilde{\mathbf{w}}_i$ to the true ℓ_p Lewis weights in order to bound the condition number to within a $\operatorname{poly}(n)$ factor of the condition number of $\mathbf{A}$.

Lemma 6.4. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p < 2$. Let $\tilde{\mathbf{w}}_i$ be weights generated by Algorithm 2, and let $\tilde{\mathbf{W}} = \operatorname{diag}(\tilde{\mathbf{w}})$. Then, for all $i \in [n]$,*

$$\mathbf{w}_i^p(\mathbf{A}) \leq \tilde{\mathbf{w}}_i.$$

Furthermore, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{A}\mathbf{x}\|_2 \leq \left\|\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A}\mathbf{x}\right\|_2 \leq d^{1/p-1/2}\|\mathbf{A}\mathbf{x}\|_p.$$

Proof. By an inductive proof almost identical to that of Lemma 6.2, we have that

$$\mathbf{w}_i^p(\mathbf{A}) \leq \tilde{\mathbf{w}}_i$$

for every $i \in [n]$. Then, since $1-2/p < 0$ for $p \in (0,2)$, we have

$$\left\|\tilde{\mathbf{W}}^{1-2/p}\mathbf{A}\mathbf{x}\right\|_2 \leq \left\|\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A}\mathbf{x}\right\|_2 \leq d^{1/p-1/2}\|\mathbf{A}\mathbf{x}\|_p$$

and

$$\left\|\tilde{\mathbf{W}}^{1-2/p}\mathbf{A}\mathbf{x}\right\|_2 \geq \|\mathbf{A}\mathbf{x}\|_2.$$

□

We thus have

$$\left\|(\tilde{\mathbf{W}}_i^{1/2-1/p}\mathbf{A}_i)^-\right\|_2 \leq \operatorname{poly}(n)\left\|\mathbf{A}_i^-\right\|_2,$$

for every $i \in [n]$, which results in the following sum of approximate online Lewis weights:

Lemma 6.5. *Let $p \in (0,\infty)$ and $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\tilde{\mathbf{w}} \in \mathbb{R}^n$ be generated from Algorithm 2. Then,*

$$\sum_{i=1}^n \tilde{\mathbf{w}}_i \leq O(d)\log(n\kappa^{\mathsf{OL}}(\mathbf{A})).$$

Proof. We have that

$$\begin{aligned}\sum_{i=1}^n \tilde{\mathbf{w}}_i &\leq \sum_{i=1}^n 4\cdot \boldsymbol{\tau}_i^{\mathsf{OL}}(\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A})\\ &\leq O(d)\log\kappa^{\mathsf{OL}}(\tilde{\mathbf{W}}^{1/2-1/p}\mathbf{A})\\ &\leq O(d)\log(n\kappa^{\mathsf{OL}}(\mathbf{A})).\end{aligned}$$

□

As a result, of our Lewis weight estimates as well as our new one-shot Lewis weight sampling result of Theorem 5.2 for $p > 2$ or Theorem A.2 for $p < 2$, we obtain our main online sampling result:

Theorem 6.6. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \in (0,\infty)$. Let $\delta \in (0,1)$ be a failure rate parameter and let $\varepsilon \in (0,1)$ be an accuracy parameter. Then there is an online coreset algorithm $\mathcal{A}$ such that, with probability at least $1-\delta$, $\mathcal{A}$ outputs a weighted subset of m rows with sampling matrix $\mathbf{S}$ such that*

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

for all $\mathbf{x} \in \mathbb{R}^d$, for every $i \in [n]$, and

$$m = \begin{cases} O\left(\frac{d^{p/2}}{\varepsilon^2}\right)(\log(n\kappa^{\mathsf{OL}}))^{p/2+1}\left[(\log d)^2(\log n) + \log\frac{1}{\delta}\right] & p \in (2,\infty) \\ O\left(\frac{d}{\varepsilon^2}\right)\log(n\kappa^{\mathsf{OL}})\left[(\log d)^2\log n + \log\frac{1}{\delta}\right] & p\in(1,2) \\ O\left(\frac{d}{\varepsilon^2}\right)\log(n\kappa^{\mathsf{OL}})\log\frac{n}{\delta} & p = 1 \\ O\left(\frac{d}{\varepsilon^2}\right)\log(n\kappa^{\mathsf{OL}})\left[(\log d)^3 + \log\frac{1}{\delta}\right] & p \in (0,1) \end{cases}$$

7 Sampling-Based Online Lewis Weight Estimation, $0 < p < 2$

7.1 Flattening Online Lewis Weights

We first show a reduction to the case where all online Lewis weights that do not increase the rank of the rows are uniformly bounded, using the idea of splitting rows. This will be useful for the sampling-based method, by controlling the worst-case behavior of the matrix martingale for use in a matrix Freedman's inequality. Note that we only need to split rows which $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) < 1$, since otherwise, these rows are sampled with probability 1 and thus do not affect the matrix Freedman. Note also that such a splitting is isometric in ℓ_p (i.e., $\|\mathbf{A}'\mathbf{x}\|_p = \|\mathbf{A}\mathbf{x}\|_p$ for all $\mathbf{x}\in\mathbb{R}^d$), so the pseudo condition number can only change by at most $\mathrm{poly}(n)$ factors, and thus the bound on the sum of online ℓ_p Lewis weights via Lemma 3.7 is unaffected.

Lemma 7.1 (Online Lewis Weight Flattening)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p < 2$. Let $\beta \in (0,1)$ be a cutoff parameter. Let $\mathbf{A}'$ be the matrix formed by replacing $\mathbf{a}_i$ by $k := \lceil 1/\beta\rceil$ copies of $\mathbf{a}_i/k^{1/p}$ whenever $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) < 1$. Then, for every row j for the new matrix $\mathbf{A}'$ which comes from such a row,*

$$\mathbf{w}_j^{p,\mathsf{OL}}(\mathbf{A}') \leq \beta.$$

Proof. For the rows i considered in this lemma, we have that

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{2/p} = \left[\mathbf{a}_i^\top(\mathbf{A}_{i-1}^\top\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-\mathbf{a}_i\right]^{p/2}$$

We will inductively show the desired result, along with the invariant that

$$\mathbf{A}_i^\top\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_i^{1-2/p}\mathbf{A}_i \preceq \mathbf{A}_j'^\top\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A}')_j^{1-2/p}\mathbf{A}_j' \tag{5}$$

whenever j is the last row formed as one of the k copies of row $i\in[n]$ in the original matrix.

Let j be a row formed as one of the k copies of row $i \in [n]$ in the original matrix $\mathbf{A}$. Then,

$$\mathbf{A}_{i-1}^\top\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1} \preceq \mathbf{A}_{j-1}'^\top\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A}')_{j-1}^{1-2/p}\mathbf{A}_{j-1}'$$

since we only add rows over the first such index j, so by Lemma 2.5, we have that

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{2/p} = \mathbf{a}_i^\top(\mathbf{A}_{i-1}^\top\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-\mathbf{a}_i \geq \mathbf{a}_i^\top(\mathbf{A}_{j-1}'^\top\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A}')_{j-1}^{1-2/p}\mathbf{A}_{j-1}')\mathbf{a}_i.$$

Then dividing both sides by $k^{2/p}$, we have that

$$\left(\frac{\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})}{k}\right)^{2/p} \geq \left(\frac{\mathbf{a}_i}{k^{1/p}}\right)^\top(\mathbf{A}_{j-1}'^\top\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A}')_{j-1}^{1-2/p}\mathbf{A}_{j-1}')\left(\frac{\mathbf{a}_i}{k^{1/p}}\right) = \mathbf{w}_j^{p,\mathsf{OL}}(\mathbf{A}')^{2/p},$$

that is, $\frac{1}{k}\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) \geq \mathbf{w}_j^{p,\mathsf{OL}}(\mathbf{A}')$. It follows that $\mathbf{w}_j^{p,\mathsf{OL}}(\mathbf{A}') \leq 1/k \leq \beta$. Finally, if S is the set of all k rows formed from row $i \in [n]$ in the original matrix, then

$$\sum_{j\in S} \mathbf{w}_j^{p,\mathsf{OL}}(\mathbf{A}')^{1-2/p}\Big(\frac{\mathbf{a}_i}{k^{1/p}}\Big)\Big(\frac{\mathbf{a}_i}{k^{1/p}}\Big)^\top \succeq \sum_{j\in S}\left(\frac{\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})}{k}\right)^{1-2/p}\Big(\frac{\mathbf{a}_i}{k^{1/p}}\Big)\Big(\frac{\mathbf{a}_i}{k^{1/p}}\Big)^\top = \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})\mathbf{a}_i\mathbf{a}_i^\top$$

where we have used that $1 - 2/p < 0$. This establishes the invariant (5). □

7.2 Online Lewis Weight Estimation

We now show how to obtain Lewis weight overestimates $\tilde{\mathbf{w}}_i$ that have a small sum.

Algorithm 3 Sampling-Based Online Lewis Weight Estimation

```
input: $\mathbf{A} \in \mathbb{R}^{n\times d}$, $p \in (0,2)$, oversampling parameter $\alpha \in (0,1)$.
output: Online Lewis weight estimates $\{\tilde{\mathbf{w}}_i\}_{i=1}^n$.
$\tilde{\mathbf{A}}_0 \leftarrow \varnothing$          ▷ Matrix estimate for sketching Lewis quadratic
for $i \in [n]$ do
    $\tilde{\mathbf{w}}_i \leftarrow \Big[\mathbf{a}_i^\top(\tilde{\mathbf{A}}_{i-1}\tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1})^-\mathbf{a}_i\Big]^{p/2}$
    $\mathbf{p}_i \leftarrow \min\{\frac{1}{\alpha}\tilde{\mathbf{w}}_i, 1\}$
    $\tilde{\mathbf{A}}_i \coloneqq \begin{cases}\begin{bmatrix}\tilde{\mathbf{A}}_{i-1}\\ \mathbf{a}_i/\sqrt{\mathbf{p}}_i\end{bmatrix} & \text{with probability } \mathbf{p}_i\\ \tilde{\mathbf{A}}_{i-1} & \text{otherwise}\end{cases}$
return $\{\tilde{\mathbf{w}}_i\}_{i=1}^n$
```

7.2.1 Lower Bound on Online Lewis Weight Estimates

We first show that the approximate online Lewis weights are bounded below by the true Lewis weights, analogously to Lemma 3.3 of [CMP20], which shows how to approximate online leverage scores. We will need the matrix Freedman's inequality [Tro11]:

Theorem 7.2 (Matrix Freedman's Inequality [Tro11]). *Let $\mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a matrix whose values are self-adjoint matrices with dimension d, and let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be the difference sequence. Assume that the difference sequence is uniformly bounded in the sense that*

$$\|\mathbf{X}_k\|_2 \leq R \textit{ almost surely, for } k = 1, \ldots, n.$$

Define the predictable quadratic variation process of the martingale:

$$\mathbf{W}_k \coloneqq \sum_{j=1}^k \mathop{\mathbf{E}}_{j-1}\big[\mathbf{X}_j^2\big], \textit{ for } k = 1, \ldots, n.$$

Then, for all $\varepsilon > 0$ and $\sigma^2 > 0$,

$$\mathbf{Pr}\{\|\mathbf{Y}_n\|_2 \geq \varepsilon \textit{ and } \|\mathbf{W}_n\|_2 \leq \sigma^2\} \leq d \cdot \exp\left(-\frac{-\varepsilon^2/2}{\sigma^2 + R\varepsilon/3}\right).$$

Theorem 7.3 (Online Lewis weight estimates bound Lewis weights). *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p < 2$. Furthermore, suppose that all the online Lewis weights of $\mathbf{A}$ are bounded by β whenever i does not increase the rank, that is, $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) \leq \beta$ for each $i \in [n]$ with $\mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1})$. Furthermore, suppose that*

$$\alpha + \beta \leq \frac{\varepsilon^2}{\log\frac{d}{\delta}}$$

for $\delta \in (0,1)$. Then, with probability at least $1-\delta$, we have for all $i \in [n]$ that

$$\tilde{\mathbf{w}}_i^p \geq \frac{1}{1+\varepsilon}\mathbf{w}_i^p(\mathbf{A}),$$

and that

$$\tilde{\mathbf{A}}\tilde{\mathbf{W}}^{1-2/p}\tilde{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A},$$

Proof. For simplicity of notation, let $\mathbf{w} = \mathbf{w}^{p,\mathsf{OL}}(\mathbf{A})$ and $\mathbf{W} = \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})$.

Our proof closely follows [CMP20]. Let

$$\mathbf{G} \coloneqq \mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A}$$

where $\mathbf{W}^p(\mathbf{A})$ is the diagonal matrix of the offline Lewis weights, and let

$$\mathbf{u}_i \coloneqq (\mathbf{G}^-)^{1/2}\mathbf{a}_i.$$

Note that

$$\mathbf{u}_i^\top\mathbf{u}_i = \mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A})^-\mathbf{a}_i = \mathbf{w}_i^p(\mathbf{A})^{2/p}. \tag{6}$$

We consider the matrix martingale $0 = \mathbf{Y}_0, \mathbf{Y}_1, \dots, \mathbf{Y}_n \in \mathbb{R}^{d\times d}$ with difference sequence $\mathbf{X}_1, \dots, \mathbf{X}_n$. If $\|\mathbf{Y}_{i-1}\|_2 \geq \varepsilon$, then we set $\mathbf{X}_i \coloneqq 0$, otherwise

$$\mathbf{X}_i \coloneqq \begin{cases} \left(\frac{1}{\mathbf{p}_i}\tilde{\mathbf{w}}_i^{1-2/p} - \mathbf{w}_i^{1-2/p}\right)\mathbf{u}_i\mathbf{u}_i^\top & \text{if } \mathbf{a}_i \text{ is sampled in } \tilde{\mathbf{A}} \\ -\mathbf{w}_i^{1-2/p}(\mathbf{u}_i\mathbf{u}_i^\top) & \text{otherwise} \end{cases}$$

This gives

$$\mathbf{Y}_{i-1} = (\mathbf{G}^-)^{1/2}\left(\tilde{\mathbf{A}}_{i-1}^\top\tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1} - \mathbf{A}_{i-1}^\top\mathbf{W}_{i-1}^{1-2/p}\mathbf{A}_{i-1}\right)(\mathbf{G}^-)^{1/2}$$

Bounds on the Difference Sequence. We now bound $\|\mathbf{X}_j\|_2$ and $\mathbf{W}_i \coloneqq \sum_{j=1}^i \mathbf{E}_{j-1}[\mathbf{X}_j^2]$ for use in the matrix Freedman inequality. These bounds are trivial when $\|\mathbf{Y}_{j-1}\|_2 \geq \varepsilon$, so suppose that $\|\mathbf{Y}_{j-1}\|_2 < \varepsilon$.

We will first show that $\mathbf{u}_i^\top\mathbf{u}_i \leq (1+\varepsilon)\tilde{\mathbf{w}}_i^{2/p}$. Because $\tilde{\mathbf{A}}_{i-1}^\top\tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1}$ and $\mathbf{A}_{i-1}^\top\mathbf{W}_{i-1}^{1-2/p}\mathbf{A}_{i-1}$ always have the same row space and are symmetric, we can write them as $\mathbf{V}\tilde{\mathbf{R}}\mathbf{V}^\top$ and $\mathbf{V}\mathbf{R}\mathbf{V}^\top$, respectively, where $\mathbf{V} \in \mathbb{R}^{d\times r}$ is an orthonormal basis of $\operatorname{rowspan}(\mathbf{A}_{i-1})$. If $\mathbf{a}_i \notin \operatorname{rowspan}(\mathbf{A}_{i-1})$, then $\mathbf{X}_i = 0$, so suppose that $\mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1})$. Note then that $\mathbf{a}_i$ can be written as $\mathbf{a}_i = \mathbf{V}\mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^r$. Then,

$$\begin{aligned} \tilde{\mathbf{w}}_i^{2/p} &= \mathbf{a}_i^\top(\tilde{\mathbf{A}}_{i-1}^\top\tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1})^-\mathbf{a}_i \\ &= \mathbf{b}^\top\mathbf{V}^\top(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top)\mathbf{V}\mathbf{b} && \text{Lemma 2.5} \\ &= \mathbf{b}^\top\tilde{\mathbf{R}}^{-1}\mathbf{b} \\ &= \mathbf{b}^\top(\mathbf{R} + (\tilde{\mathbf{R}} - \mathbf{R}))^{-1}\mathbf{b}. \end{aligned} \tag{7}$$

Now let $\mathbf{P} = \mathbf{V}\mathbf{V}^\top$ be the projection matrix onto $\operatorname{rowspan}(\mathbf{A}_{i-1})$. Let $\mathbf{E} \in \mathbb{R}^{r\times r}$ be such that $\mathbf{P}\mathbf{G}\mathbf{P} = \mathbf{V}\mathbf{E}\mathbf{V}^\top$. Then, we have that

$$\mathbf{V}\mathbf{E}^{-1/2}(\tilde{\mathbf{R}} - \mathbf{R})\mathbf{E}^{-1/2}\mathbf{V}^\top = \mathbf{P}\mathbf{Y}_{i-1}\mathbf{P} \preceq \mathbf{Y}_{i-1} \preceq \varepsilon\mathbf{I}_d.$$

so for every $\mathbf{x} \in \mathbb{R}^r$,

$$\mathbf{x}^\top\mathbf{E}^{-1/2}(\tilde{\mathbf{R}} - \mathbf{R})\mathbf{E}^{-1/2}\mathbf{x} \leq \varepsilon\mathbf{x}^\top\mathbf{x}.$$

By a change of variable into $\mathbf{y} = \mathbf{E}^{-1/2}\mathbf{x}$, this gives

$$\mathbf{y}^\top(\tilde{\mathbf{R}} - \mathbf{R})\mathbf{y} \leq \varepsilon\mathbf{y}^\top\mathbf{E}\mathbf{y}$$

for every $\mathbf{y} \in \mathbb{R}^r$, that is, that $\tilde{\mathbf{R}} - \mathbf{R} \preceq \varepsilon\mathbf{E}$. Also note that by Lemma 3.5 and the fact that $1 - 2/p < 0$,

$$\mathbf{W}^p(\mathbf{A}) \preceq \mathbf{W} \implies \mathbf{W}^p(\mathbf{A})^{1-2/p} \succeq \mathbf{W}^{1-2/p}$$

$$\begin{aligned}
&\implies \mathbf{G} = \mathbf{A}^\top \mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A} \succeq \mathbf{A}_{i-1}^\top \mathbf{W}^p(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1} \succeq \mathbf{A}_{i-1}^\top \mathbf{W}_{i-1}^{1-2/p}\mathbf{A}_{i-1} \\
&\implies \mathbf{PGP} \succeq \mathbf{P}(\mathbf{A}_{i-1}^\top \mathbf{W}_{i-1}^{1-2/p}\mathbf{A}_{i-1})\mathbf{P} = \mathbf{A}_{i-1}^\top \mathbf{W}_{i-1}^{1-2/p}\mathbf{A}_{i-1} \\
&\implies \mathbf{E} \succeq \mathbf{R}.
\end{aligned}$$

Thus, $\mathbf{R} + (\tilde{\mathbf{R}} - \mathbf{R}) \preceq \mathbf{E} + \varepsilon\mathbf{E}$. Then, continuing the calculation from (7), we have

$$\begin{aligned}
\tilde{\mathbf{w}}_i^{2/p} &= \mathbf{b}^\top(\mathbf{R} + (\tilde{\mathbf{R}} - \mathbf{R}))^{-1}\mathbf{b} \\
&\geq \mathbf{b}^\top(\mathbf{E} + \varepsilon\mathbf{E})^{-1}\mathbf{b} \\
&= \frac{1}{1+\varepsilon}\mathbf{b}^\top\mathbf{E}^{-1}\mathbf{b} \\
&= \frac{1}{1+\varepsilon}\mathbf{u}_i^\top\mathbf{u}_i \\
&= \frac{1}{1+\varepsilon}\mathbf{w}_i^p(\mathbf{A})^{2/p} && \text{Equation (6)}
\end{aligned} \tag{8}$$

Given our bound on $\mathbf{u}_i^\top\mathbf{u}_i$ in (6) and (8), as well as Lemma 3.5 we can now bound $\mathbf{X}_i$. We first have that

$$\begin{aligned}
\|\mathbf{X}_i\|_2 &\leq \frac{\tilde{\mathbf{w}}_i^{1-2/p}}{\mathbf{p}_i}\left\|\mathbf{u}_i\mathbf{u}_i^\top\right\|_2 + \mathbf{w}_i^{1-2/p}\left\|\mathbf{u}_i\mathbf{u}_i^\top\right\|_2 \\
&\leq \frac{(1+\varepsilon)\tilde{\mathbf{w}}_i}{\mathbf{p}_i} + \mathbf{w}_i \\
&\leq O(\alpha+\beta)
\end{aligned} \tag{9}$$

almost surely. Furthermore, we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{i-1}[\mathbf{X}_i^2] &\preceq \mathbf{p}_i \cdot \left(\frac{\tilde{\mathbf{w}}_i^{1-2/p}}{\mathbf{p}_i} - \mathbf{w}_i^{1-2/p}\right)^2 (\mathbf{u}_i\mathbf{u}_i^\top)^2 + (1-\mathbf{p}_i)\mathbf{w}_i^{2(1-2/p)}(\mathbf{u}_i\mathbf{u}_i^\top)^2 \\
&\preceq 2\frac{\tilde{\mathbf{w}}_i^{2(1-2/p)}}{\mathbf{p}_i}(\mathbf{u}_i\mathbf{u}_i^\top)^2 + 2\mathbf{w}_i^{2(1-2/p)}(\mathbf{u}_i\mathbf{u}_i^\top)^2 \\
&\preceq 2(1+\varepsilon)\frac{\tilde{\mathbf{w}}_i}{\mathbf{p}_i}\cdot\tilde{\mathbf{w}}_i^{1-2/p}\mathbf{u}_i\mathbf{u}_i^\top + 2\mathbf{w}_i\cdot\mathbf{w}_i^{1-2/p}\mathbf{u}_i\mathbf{u}_i^\top \\
&\preceq O(\alpha+\beta)(\tilde{\mathbf{w}}_i^{1-2/p} + \mathbf{w}_i^{1-2/p})\mathbf{u}_i\mathbf{u}_i^\top \\
&\preceq O(\alpha+\beta)\mathbf{w}_i^p(\mathbf{A})^{1-2/p}\mathbf{u}_i\mathbf{u}_i^\top
\end{aligned}$$

where the last inequality uses that $1 - 2/p < 0$. Then, for the predictable quadratic variation process $\mathbf{W}_i \coloneqq \sum_{k=1}^i \mathbf{E}_{k-1}[\mathbf{X}_k^2]$ of the martingale $\{\mathbf{Y}_i\}$, we have

$$\begin{aligned}
\|\mathbf{W}_i\|_2 &= O(\alpha+\beta)\left\|(\mathbf{G}^-)^{1/2}\left(\sum_{k=1}^i \mathbf{w}_k^p(\mathbf{A})^{1-2/p}\mathbf{a}_k\mathbf{a}_k^\top\right)(\mathbf{G}^-)^{1/2}\right\|_2 \\
&= O(\alpha+\beta)\left\|((\mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A})^-)^{1/2}\left(\mathbf{A}_k^\top\mathbf{W}^p(\mathbf{A})_k^{1-2/p}\mathbf{A}_k\right)((\mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A})^-)^{1/2}\right\|_2 \\
&\leq O(\alpha+\beta)\left\|((\mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A})^-)^{1/2}\left(\mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A}\right)((\mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A})^-)^{1/2}\right\|_2 \\
&= O(\alpha+\beta)
\end{aligned} \tag{10}$$

using Lemma 2.6.

The Matrix Freedman's Inequality. We may now apply Theorem 7.2 along with the bounds of (9) and (10), which gives that

$$\mathbf{Pr}\{\|\mathbf{Y}_n\|_2 \geq \varepsilon\} \leq d\cdot\exp\left(-\Omega(1)\cdot\frac{\varepsilon^2}{\alpha+\beta}\right) \leq d\cdot\exp\left(-\log\frac{d}{\delta}\right) = \delta.$$

Then under this event, (8) holds for every $i \in [n]$. Furthermore, we also have under this event that

$$\tilde{\mathbf{A}}\tilde{\mathbf{W}}^{1-2/p}\tilde{\mathbf{A}} \preceq \mathbf{A}\mathbf{W}^{1-2/p}\mathbf{A} + \varepsilon \mathbf{A}\mathbf{W}^{p}(\mathbf{A})^{1-2/p}\mathbf{A} \preceq (1+\varepsilon)\mathbf{A}^\top \mathbf{W}^{p}(\mathbf{A})^{1-2/p}\mathbf{A},$$

again using that $1 - 2/p < 0$. □

7.2.2 Upper Bound on the Sum of Online Lewis Weight Estimates

Unlike in the case for ℓ_2, the proof in Theorem 7.3 is not sufficient for upper bounding the estimated online Lewis weights for every $i \in [n]$. This is due to the fact that the spectral error in our Lewis quadratic is $\mathbf{A}\mathbf{W}^{p}(\mathbf{A})^{1-2/p}\mathbf{A}$, which is in fact spectrally greater than $\mathbf{A}\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1-2/p}\mathbf{A}$ since $1 - 2/p < 0$. We thus appeal to the strategy of Lemma 3.4 of [CMP20], which uses a different proof to bound the sum of online Lewis weights.

Theorem 7.4. *With probability at least $1-\delta$, we have that*

$$\sum_{i=1}^{n} \tilde{\mathbf{w}}_i \leq O(d)\log(n\kappa^{\mathsf{OL}}) + O\left(\log\frac{1}{\delta}\right)$$

Proof. We closely follow [CMP20]. Let $\lambda = (\max_{i=1}^{n}\left\|\mathbf{A}_i^-\right\|_2)^{-1}/n^C$, where C will be a sufficiently large constant exponent which can change from line to line. We then let

$$\Delta_i \coloneqq \log\det(\tilde{\mathbf{A}}_i^\top\tilde{\mathbf{W}}_i^{1-2/p}\tilde{\mathbf{A}}_i + \lambda\mathbf{I}_d) - \log\operatorname{pdet}(\tilde{\mathbf{A}}_{i-1}^\top\tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I}_d)$$

We first show that $\mathbf{E}_{i-1}[\exp(\tilde{\mathbf{w}}_i/8 - \Delta_i]$ is always at most 1 whenever $\mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1})$. Then by the pseudodeterminant lemma, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{i-1}\left[\exp\left(\frac{\tilde{\mathbf{w}}_i}{8} - \Delta_i\right)\right] = \mathbf{p}_i \cdot &\frac{\exp(\tilde{\mathbf{w}}_i/8)}{1 + (\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i)^\top(\tilde{\mathbf{A}}_{i-1}^\top\tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I}_d)^{-1}(\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i)/\mathbf{p}_i} \\ &+ (1-\mathbf{p}_i)\exp(\tilde{\mathbf{w}}_i/8).\end{aligned}$$

Note that since $\tilde{\mathbf{w}}_i \leq 1$, we have $\exp(\tilde{\mathbf{w}}_i/8) \leq 1 + \tilde{\mathbf{w}}_i/4$. Now if $\tilde{\mathbf{w}}_i/\alpha < 1$, then $\mathbf{p}_i = \tilde{\mathbf{w}}_i/\alpha$ and

$$\begin{aligned}\tilde{\mathbf{w}}_i &= (\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i)^\top(\tilde{\mathbf{A}}_{i-1}^\top\tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1})^-(\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i) \\ &= (1 \pm n^{-C})(\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i)^\top(\tilde{\mathbf{A}}_{i-1}^\top\tilde{\mathbf{W}}_{i-1}^{1-2/p}\tilde{\mathbf{A}}_{i-1} + \lambda\mathbf{I}_d)^{-1}(\tilde{\mathbf{w}}_i^{1/2-1/p}\mathbf{a}_i)\end{aligned}$$

by rearranging and taking $p/2$th powers. Then,

$$\begin{aligned}\mathop{\mathbf{E}}_{i-1}\left[\exp\left(\frac{\tilde{\mathbf{w}}_i}{8} - \Delta_i\right)\right] &\leq \mathbf{p}_i\frac{1 + \tilde{\mathbf{w}}_i/4}{1 + (1-n^{-C})\alpha} + (1-\mathbf{p}_i)(1 + \tilde{\mathbf{w}}_i/4) \\ &= (1 + \tilde{\mathbf{w}}_i/4)\left(1 - \mathbf{p}_i\frac{(1-n^{-C})\alpha}{1 + (1-n^{-C})\alpha}\right) \\ &= \left(1 + \mathbf{p}_i\frac{\alpha}{4}\right)\left(1 - \mathbf{p}_i\frac{(1-n^{-C})\alpha}{1 + (1-n^{-C})\alpha}\right) \leq 1\end{aligned}$$

for α sufficiently small. Otherwise, if $\mathbf{p}_i = 1$, then

$$\mathop{\mathbf{E}}_{i-1}\left[\exp\left(\frac{\tilde{\mathbf{w}}_i}{8} - \Delta_i\right)\right] \leq \frac{1 + \tilde{\mathbf{w}}_i/4}{1 + (1-n^{-C})\tilde{\mathbf{w}}_i} \leq 1.$$

Next, we analyze the expected product of $\exp(\tilde{\mathbf{w}}_i/8 - \Delta_i)$ over the first k steps. If $\mathbf{a}_k \notin \operatorname{rowspan}(\mathbf{A}_{k-1})$, we have that

$$\mathbf{E}\left[\exp\left(\sum_{i=1}^{k}\frac{\tilde{\mathbf{w}}_i}{8} - \Delta_i\right)\right] = \mathop{\mathbf{E}}_{\text{first } k-1 \text{ steps}}\left[\exp\left(\sum_{i=1}^{k-1}\frac{\tilde{\mathbf{w}}_i}{8} - \Delta_i\right)\mathop{\mathbf{E}}_{k-1}\left[\frac{\tilde{\mathbf{w}}_k}{8} - \Delta_k\right]\right] \leq \mathbf{E}\left[\exp\left(\sum_{i=1}^{k-1}\frac{\tilde{\mathbf{w}}_i}{8} - \Delta_i\right)\right].$$

Inductively, we have that

$$\mathbf{E}\left[\exp\left(\sum_{i=1}^{n}\frac{\tilde{\mathbf{w}}_i}{8}-\Delta_i\right)\right]\leq 1$$

By Markov's inequality, we then have that

$$\mathbf{Pr}\left\{\sum_{i=1}^{n}\tilde{\mathbf{w}}_i > 8\log\frac{1}{\delta}+8\sum_{i=1}^{n}\Delta_i\right\}\leq\delta.$$

Note that

$$\begin{aligned}\sum_{i=1}^{n}\Delta_i &= \log\det(\tilde{\mathbf{A}}^\top\tilde{\mathbf{W}}^{1-2/p}\tilde{\mathbf{A}}+\lambda\mathbf{I}_d)-\log\det(\lambda\mathbf{I}_d)\\ &\leq \log\det((1+O(\varepsilon))(\mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A}))-\log\det(\lambda\mathbf{I}_d)\end{aligned}$$

by Theorem 7.3. Furthermore, for any unit vector $\mathbf{x}\in\mathbb{R}^d$, we have by properties of Lewis weights that

$$\left\|\mathbf{W}^p(\mathbf{A})^{1/2-1/p}\mathbf{A}\mathbf{x}\right\|_2 = \mathrm{poly}(d)\|\mathbf{A}\mathbf{x}\|_p = \mathrm{poly}(n,d)\|\mathbf{A}\mathbf{x}\|_2,$$

so the operator norm of $\mathbf{A}^\top\mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A}$ is within a $\mathrm{poly}(n,d)$ factor of the operator norm of $\mathbf{A}^\top\mathbf{A}$. Thus,

$$\sum_{i=1}^{n}\tilde{\mathbf{w}}_i\leq O(d)\log\frac{\mathrm{poly}(n,d)\|\mathbf{A}\|_2}{\lambda}=O(d)\log(n\kappa^{\mathsf{OL}})+O\left(\log\frac{1}{\delta}\right).$$

This yields the desired result. □

7.3 ℓ_p Subspace Embeddings via Online Lewis Weight Sampling

Given our online Lewis weight estimates, we may now conclude by sampling proportionally to these weights. We use fresh randomness to sample proportionally to these weights, independently of the success of the random sampling process used to estimate the online Lewis weights. Thus, we can condition on the successes of Theorems 7.3 and 7.4 to obtain Lewis weight upper bounds with a small sum. We then get our main result:

Theorem 7.5. *Let $\mathbf{A}\in\mathbb{R}^{n\times d}$ and $p\in(0,2)$. Let $\delta\in(0,1)$ be a failure rate parameter and let $\varepsilon\in(0,1)$ be an accuracy parameter. Then there is an online coreset algorithm $\mathcal{A}$ such that, with probability at least $1-\delta$, $\mathcal{A}$ outputs a weighted subset of m rows with sampling matrix $\mathbf{S}$ such that*

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

for all $\mathbf{x}\in\mathbb{R}^d$ and every $i\in[n]$, and

$$m=\begin{cases}O\left(\frac{T}{\varepsilon^2}\left[(\log d)^2\log n+\log\frac{1}{\delta}\right]\right) & p\in(1,2)\\ O\left(\frac{T}{\varepsilon^2}\log\frac{n}{\delta}\right) & p=1\\ O\left(\frac{T}{\varepsilon^2}\left[(\log d)^3+\log\frac{1}{\delta}\right]\right) & p\in(0,1)\end{cases}$$

for $T=O(d)\log(n\kappa^{\mathsf{OL}})$.

Proof. We first use Lemma 7.1 to flatten the rows down to an online Lewis weight of $O(1/\log(d/\delta))$, which can be done in an online fashion, given knowledge of an upper bound on the online Lewis weight. We can then obtain upper bounds $\tilde{\mathbf{w}}_i$ on online Lewis weights using Algorithm 3 and Theorem 7.3 with $\varepsilon=O(1)$ so that the upper bounds sum to at most

$$T=O(d)\log(n\kappa^{\mathsf{OL}}).$$

This requires $O(T\log(d/\delta))$ samples. Using these Lewis weight estimates, we then sample using Theorem A.2. The number of samples used here dominates the samples used to approximate the Lewis weights, and is as given in the theorem statement. □

8 Applications: Online Coresets for Generalized Linear Models

8.1 Lower Bound for μ-Complex Datasets

We start with the following geometric lemma, which helps in constructing an input instance with bounded μ-complexity.

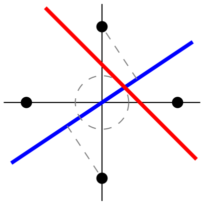

Figure 1: A line cannot be close to all four points and the origin. The blue line is close to the origin, but far from $(0, \pm 1)$. The red line is close to $(0, 1)$ and $(1, 0)$ but far from the origin.

Lemma 8.1. *Consider the four points $S = \{(\pm 1, 0), (0, \pm 1)\}$ in $\mathbb{R}^2$. Let $L = \left\{\mathbf{x} \in \mathbb{R}^2 : \langle \mathbf{u}, \mathbf{x} \rangle = b\right\}$ be an affine set in $\mathbb{R}^2$, for a unit vector $\mathbf{u} \in \mathbb{R}^2$ and offset b. Then, at least one of the following is true:*

- $|b| \geq \frac{1}{2\sqrt{2}}$
- *there are $\mathbf{s}_\pm \in S$ such that $\langle \mathbf{u}, \mathbf{s}_+ \rangle \geq b + \frac{1}{2\sqrt{2}}$ and $\langle \mathbf{u}, \mathbf{s}_- \rangle \leq b - \frac{1}{2\sqrt{2}}$*

Proof. Assume WLOG that $\mathbf{u}_1 \geq \mathbf{u}_2 \geq 0$. Then, $\langle \mathbf{u}, \mathbf{s} \rangle$ are $\mathbf{u}_1, -\mathbf{u}_1, \mathbf{u}_2, -\mathbf{u}_2$ for $\mathbf{s} \in S$. Since $\mathbf{u}$ is a unit vector, we have that $\mathbf{u}_1 \geq 1/\sqrt{2}$, so

$$\mathbf{u}_1 = \langle \mathbf{u}, (1, 0) \rangle \geq \frac{1}{\sqrt{2}}, \qquad -\mathbf{u}_1 = \langle \mathbf{u}, (-1, 0) \rangle \leq -\frac{1}{\sqrt{2}}.$$

Then if $|b| \leq \frac{1}{2\sqrt{2}}$, then

$$\mathbf{u}_1 = \langle \mathbf{u}, (1, 0) \rangle \geq b + \frac{1}{2\sqrt{2}}, \qquad -\mathbf{u}_1 = \langle \mathbf{u}, (-1, 0) \rangle \leq b - \frac{1}{2\sqrt{2}}.$$ □

A similar conclusion continues to hold if we replace the standard basis vectors by approximately orthogonal points:

Corollary 8.2. *Let $S \subseteq \mathbb{R}^2$ be a set of four points such that $|\langle \mathbf{s}, \mathbf{s}' \rangle| \leq 1/100$ for each $\mathbf{s} \neq \mathbf{s}' \in S$. Then, at least one of the following is true:*

- $|b| \geq \frac{1}{3\sqrt{2}}$
- *there are $\mathbf{s}_\pm \in S$ such that $\langle \mathbf{u}, \mathbf{s}_+ \rangle \geq b + \frac{1}{3\sqrt{2}}$ and $\langle \mathbf{u}, \mathbf{s}_- \rangle \leq b - \frac{1}{3\sqrt{2}}$*

Using the above lemma, we will construct a variant of the lower bound instance of [MSSW18] for logistic regression with bounded μ-complexity. Recall that the instance of [MSSW18] shows an $\Omega(n)$ lower bound for an instance with unbounded μ-complexity by arranging n points in a circle on the plane. Then, they reduce the INDEX problem to the problem of computing a coreset for logistic regression as follows. Alice's input point set is taken to be the points on the circle corresponding to $A \subseteq [n]$, where A is Alice's input set, where all of the labels are 1. Then, Alice computes a coreset for logistic regression and then sends the coreset to Bob. Bob then adds the circle point corresponding to his index $b \in [n]$ with label -1. If $b \notin A$, then there exists a hyperplane separating Alice's points and Bob's point, and thus the cost of logistic regression can be shown to be arbitrarily small, whereas if $b \in A$, then the cost is at least a constant. This shows an $\Omega(n)$ lower bound for any constant approximation.

In order to construct an input instance with bounded μ-complexity, we first consider adding the four points as in the configuration of Figure 1 and Lemma 8.1. We also add the origin twice, once with label 1 and once with label -1. This will turn out to be enough to argue that the μ-complexity will be bounded by $O(n)$. In order to prove our lower bound statement, we will in fact need a further modification and take a high-dimensional version of the circle instance, by using the following theorem from coding theory. This result has been used several times for related lower bounds [LWW21, MMWY21, WY22].

Theorem 8.3 (Theorem 7, [PTB13])**.** *Let m, D be integers. Then, there exists a set $\mathcal{F} \subseteq \{\pm 1\}^d$ of binary vectors of size $|\mathcal{F}| = n$ for*

$$d = 2^m - 1$$
$$n = \begin{cases} \frac{2^{(D-\lfloor D/4 \rfloor)m} - 1}{2^m - 1} & \textit{if } m \textit{ is odd} \\ 2^{(D-\lfloor D/4 \rfloor)m} - 1 & \textit{if } m \textit{ is even} \end{cases}$$

and for any $\mathbf{x}, \mathbf{x}' \in \mathcal{F}$ with $\mathbf{x} \neq \mathbf{x}'$,

$$|\langle \mathbf{x}, \mathbf{x}' \rangle| = \left| \sum_{i=1}^{n} (-1)^{\mathbb{1}(\mathbf{x}_i = \mathbf{x}'_i)} \right| \leq 1 + 2(D-1)2^{m/2}.$$

Definition 8.4 (Hard p-Probit Instance)**.** *We define a p-probit coreset instance as follows. Let $n \in \mathbb{N}$ and let $\Delta \in \mathbb{N}$. Let $A \subseteq [n]$ and let $b \in [n]$. Let $\mathcal{F} \subseteq \{\pm 1\}^d$ be as given in Theorem 8.3 with parameters $n + 4 \leq |\mathcal{F}| \leq O(n)$, $d = \Theta(\log^2 n)$, and $D = \Theta(\frac{\log n}{\log d}) = \Theta(\frac{\log n}{\log \log n})$. We associate each $i \in [n]$ with some $\mathbf{x}^{(i)} \in \mathcal{F}$. Note that we have four remaining points in $\mathcal{F}$ unused by those corresponding to $[n]$. We add these to the input dataset with label 1, calling these the "first four points".*

Now for each $a \in A$, we add Δ copies of the unit vector $\mathbf{x}'^{(a)} := \mathbf{x}^{(a)}/\sqrt{d}$ for $\mathbf{x}^{(a)} \in \mathcal{F}$ with label 1 to the input dataset. We also add $(0,0)$ with label 1 and $(0,0)$ with label -1. Finally, we add n copies of $\mathbf{x}'^{(b))} := \mathbf{x}^{(b)}/\sqrt{d}$ for $\mathbf{x}^{(b)} \in \mathcal{F}$ with label -1. We define the associated matrix $\mathbf{A} \in \mathbb{R}^{m \times (d+1)}$ for $m = \Delta|A| + n + 4 + 2$, where each row $\mathbf{a}_i$ is $y_i \cdot (\mathbf{b}_i, 1)$ where y_i is the label and $\mathbf{b}_i \in \mathbb{R}^d$ is the added point.

We first argue that this instance has bounded μ_p-complexity.

Lemma 8.5 (Bounded μ-Complexity)**.** *Consider the input instance of Definition 8.4. Then, for any $A \subseteq [n]$ and $b \in [n]$,*

$$\mu_p(\mathbf{A}) \leq O(\Delta n).$$

Proof. We consider any $\mathbf{x} \in \mathbb{R}^{d+1}$. Note that the definition of $\mu_p(\mathbf{A})$ is scale invariant, so we can scale $\mathbf{x}$ so that its first d coordinates form a unit vector. We refer to the first d coordinates as $\mathbf{u} \in \mathbb{R}^d$ and the last coordinate as b. We now case on b. First, if $|b| \geq 2$, then for $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\|_2 \leq 1$ and $y \in \{\pm 1\}$,

$$|\langle \mathbf{u}, \mathbf{x} \rangle| \leq 1 \leq \frac{|b|}{2}.$$

Thus, the sign of $\langle \mathbf{u}, \mathbf{x} \rangle + by$ is just the sign of by, and the magnitude is $\Theta(|b|)$. By construction, the dataset contains at least one point with label 1 and one point with label -1, so we have that

$$\frac{\|(\mathbf{A}\mathbf{x})^+\|_p^p}{\|(\mathbf{A}\mathbf{x})^-\|_p^p} \leq \frac{\Theta(\Delta n |b|^p)}{\Theta(|b|^p)} \leq O(\Delta n).$$

Next, suppose that $|b| \in [\frac{1}{3\sqrt{2}}, 2]$. In this case, at least one of the points with $\mathbf{u} = (0,0)$ will have $by < 0$, so we have that $\|(\mathbf{A}\mathbf{x})^-\|_p^p \geq \frac{1}{(3\sqrt{2})^p}$. On the other hand, we have

$$|\langle \mathbf{u}, \mathbf{x} \rangle + by| \leq 1 + 2 = 3$$

for any $\|\mathbf{x}\|_2 \leq 1$ and $y \in \{\pm 1\}$, so $\|(\mathbf{A}\mathbf{x})^+\|_p^p \leq 3^p \Delta n$. Thus, we again have that

$$\frac{\|(\mathbf{A}\mathbf{x})^+\|_p^p}{\|(\mathbf{A}\mathbf{x})^-\|_p^p} \leq O(\Delta n).$$

Finally, suppose that $|b| < \frac{1}{3\sqrt{2}}$. Note that for any $\mathbf{x}, \mathbf{x}' \in \mathcal{F}$, we have

$$|\langle \mathbf{x}, \mathbf{x}' \rangle| \leq O\left(\frac{\frac{\log n}{\log \log n}\sqrt{d}}{d}\right) = O\left(\frac{1}{\log \log n}\right) \leq \frac{1}{100}.$$

Then by Corollary 8.2, there is at least one point $\mathbf{x}$ of the "first four points" (see Definition 8.4) such that

$$\langle \mathbf{u}, \mathbf{x} \rangle \leq -b - \frac{1}{3\sqrt{2}} \implies \langle \mathbf{u}, \mathbf{x} \rangle + b \leq -\frac{1}{3\sqrt{2}}$$

Thus, we again have that $\|(\mathbf{A}\mathbf{x})^-\|_p^p \geq \frac{1}{(3\sqrt{2})^p}$ and $\|(\mathbf{A}\mathbf{x})^+\|_p^p \leq (1 + \frac{1}{3\sqrt{2}})^p \Delta n$ so we have

$$\frac{\|(\mathbf{A}\mathbf{x})^+\|_p^p}{\|(\mathbf{A}\mathbf{x})^-\|_p^p} \leq O(\Delta n)$$

again. This covers all cases, so we conclude. □

To do cost calculations, we need several approximations on the p-probit cost function:

Lemma 8.6. *Let $r \geq 1$. Then,*

$$c \exp(-r^p/p) \leq \Phi_p(-r) \leq \exp(-r^p/p)$$

for some sufficiently small constant $c > 0$, and

$$\Phi_p(r) = \Theta(r^p)$$

Proof. The lower bound is given in [MOP22, Lemma 2.6]. The upper bound is given by the following calculation:

$$\begin{aligned}\int_{-\infty}^{-r} \exp(-|t|^p/p)\, dt &= \int_r^\infty \exp(-t^p/p)\, dt \\ &\leq -\int_r^\infty -t^{p-1} \exp(-t^p/p)\, dt \leq -\exp(-t^p/p)|_r^\infty = \exp(-r^p/p).\end{aligned}$$

The second item follows from [MOP22, Lemma C.3]. □

Next, we lower bound the cost for any instance such that $b \in A$.

Lemma 8.7 (Cost Lower Bound)**.** *Suppose that $b \in A \subseteq [n]$. Then, the instance of Definition 8.4 has cost at least*

$$\sum_{i=1}^n \psi_p([\mathbf{A}\mathbf{x}](i)) \geq \Omega\left(\Delta \log \frac{n}{\Delta}\right)$$

for any $\mathbf{x} \in \mathbb{R}^{d+1}$.

Proof. If $b \in A$, then we show that the cost on the n copies of Bob's points and the Δ copies of Alice's points already incurs a cost of at least $\Delta \log \frac{n}{\Delta}$. Note that the cost on these $n + \Delta$ points is at least

$$\begin{aligned}\min_{x \geq 0} n \log(\Phi_p(-x)) + \Delta \log(\Phi_p(x)) &\geq \min_{x \geq 0} c(n \exp(-x^p) + \Delta x^p) \\ &= \min_{y \geq 0} c(n \exp(-y) + \Delta y)\end{aligned}$$

for some sufficiently small constant c, by Lemma 8.6. This is a convex function with critical point y which satisfies $n \exp(y) = \Delta$, or $y = \log(n/\Delta)$, which has a cost of

$$\Omega\left(\Delta \log \frac{n}{\Delta}\right).$$ □

On the other hand, if $b \notin A$, we show that we can upper bound the cost.

Lemma 8.8 (Cost Upper Bound)**.** *Suppose that $b \notin A \subseteq [n]$. Then, there exists a $\mathbf{x} \in \mathbb{R}^{d+1}$ such that the instance of Definition 8.4 has cost at most*

$$\sum_{i=1}^{n} \psi_p([\mathbf{A}\mathbf{x}](i)) \leq O(\log(\Delta n)).$$

Proof. Because $b \notin A$, if $\mathbf{u} = \mathbf{x}'^{(b)} = \mathbf{x}^{(b)}/\sqrt{d}$, then

$$\left\langle \mathbf{u}, \mathbf{x}'^{(b)} \right\rangle = 1$$

while for any other $\mathbf{x}^{(i)} \in \mathcal{F}$,

$$\left|\left\langle \mathbf{u}, \mathbf{x}'^{(i)} \right\rangle\right| \leq \Theta\left(\frac{\frac{\log n}{\log\log n}\sqrt{d}}{d}\right) = \Theta\left(\frac{1}{\log\log n}\right).$$

Thus, we may set $b = \Theta(\log(\Delta n)^{1/p})$ and $\lambda = \Theta(\log(\Delta n)^{1/p})$ such that with the hyperplane in the scaled direction $\lambda\mathbf{u}$ and offset b, the cost from Bob's points is at most

$$n \log\left(1 + \exp\left(-\left(\left\langle \lambda\mathbf{u}, \mathbf{x}'^{(b)} \right\rangle + b\right)\right)\right) = n \log(\Theta(-\log(\Delta n))) = O(1),$$

the cost of all of Alice's points and the first four points is at most

$$O(\Delta n) \log\left(1 + \exp\left(-\left(\left\langle \lambda\mathbf{u}, \mathbf{x}'^{(b)} \right\rangle + b\right)\right)\right) = O(\Delta n) \log(\Theta(-\log(\Delta n))) = O(1),$$

and the cost from the two points at the origin is at most $O(\log(\Delta n))$. □

By combining Lemmas 8.5, 8.7, and 8.7, we obtain the following hardness theorem:

Theorem 8.9 (Coreset Lower Bound for p-Probit Regression)**.** *There exists $\mathbf{A} \in \mathbb{R}^{m\times d}$ with $d = O(\log^2 m)$ and μ-complexity at most $O(m)$ such that for any $1 \leq \Delta \leq O(m^{1/3})$, a mergeable coreset which approximates the optimal p-probit cost up to a Δ relative error must use $\Omega(m/\Delta)$ bits of space. In particular, a constant factor approximation to the optimal p-probit regression cost for a μ_p-complex dataset requires $\Omega(\mu_p)$ bits of space.*

Proof. Consider the instance of Definition 8.4. We set $m = O(\Delta n)$, so that the μ-complexity is $O(m)$ and the dataset consists of $O(m)$ points in $d = O(\log^2 n) = O(\log^2 m)$ dimensions. Since $\Delta \leq O(m^{1/3})$, the lower bound on the cost when $b \in A$ is $\Delta \log(n/\Delta) = \Omega(\Delta \log m)$ while the upper bound on the cost when $b \notin A$ is $\log(n\Delta) = O(\log m)$. Thus, a Δ-approximation can differentiate between these two instances. Thus, such a coreset can solve the INDEX problem on $n = O(m/\Delta)$ items, so the total number of bits used must be at least $\Omega(m/\Delta)$. □

Because the asymptotics of the logistic regression loss is the same as that of p-probit regression for $p = 1$, up to constant factors, so we get a similar statement for logistic regression:

Theorem 8.10 (Coreset Lower Bound for Logistic Regression)**.** *There exists $\mathbf{A} \in \mathbb{R}^{m\times d}$ with $d = O(\log^2 m)$ and μ-complexity at most $O(m)$ such that for any $1 \leq \Delta \leq O(m^{1/3})$, a mergeable coreset which approximates the optimal p-probit cost up to a Δ relative error must use at least $\Omega(m/\Delta)$ bits of space. In particular, a constant factor approximation to the optimal p-probit regression cost for a μ-complex dataset requires $\tilde{\Omega}(\mu)$ bits of space.*

9 Acknowledgements

David P. Woodruff and Taisuke Yasuda were supported by ONR grant N00014-18-1-2562 and a Simons Investigator Award.

References

[BDM$^+$20] Vladimir Braverman, Petros Drineas, Cameron Musco, Christopher Musco, Jalaj Upadhyay, David P. Woodruff, and Samson Zhou. 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*, pages 517–528. IEEE, 2020. (document), 1, 1, 1, 1, 1.1.3, 1.1.3, 1.2.1, 1.2.2, 3, 3, 3.1, 3.1, 4, 6, 6.1

[BDR21] Adam Block, Yuval Dagan, and Alexander Rakhlin. Majorizing measures, sequential complexities, and online learning. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 587–590. PMLR, 2021. 1

[BHM$^+$21] Vladimir Braverman, Avinatan Hassidim, Yossi Matias, Mariano Schain, Sandeep Silwal, and Samson Zhou. Adversarial robustness of streaming algorithms through importance sampling. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021. (document), 1, 1, 1.1.3

[BJWY22] Omri Ben-Eliezer, Rajesh Jayaram, David P. Woodruff, and Eylon Yogev. A framework for adversarially robust streaming algorithms. *J. ACM*, 69(2):17:1–17:33, 2022. 1

[BLM89] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Math.*, 162(1-2):73–141, 1989. 1, 1.1.2, 1.1.2, 2.1, B.4

[BY20] Omri Ben-Eliezer and Eylon Yogev. The adversarial robustness of sampling. In Dan Suciu, Yufei Tao, and Zhewei Wei, editors, *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2020, Portland, OR, USA, June 14-19, 2020*, pages 49–62. ACM, 2020. 1.1.3

[CD21] Xue Chen and Michal Derezinski. Query complexity of least absolute deviation regression via robust uniform convergence. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 1144–1179. PMLR, 2021. 1, 1.2.2, 5, A, A, A

[CLM$^+$15] Michael B. Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In Tim Roughgarden, editor, *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, ITCS 2015, Rehovot, Israel, January 11-13, 2015*, pages 181–190. ACM, 2015. 3.1, 4

[CMP20] Michael B. Cohen, Cameron Musco, and Jakub Pachocki. Online row sampling. *Theory Comput.*, 16:1–25, 2020. 1, 1, 1, 1.1.2, 1.1.3, 1.2.1, 1.2.1, 1.2.2, 3, 3.3, 3, 3.1, 4, 6, 7.2.1, 7.2.1, 7.2.2, 7.2.2

[CP15] Michael B. Cohen and Richard Peng. $\mathrm{L_p}$ row sampling by lewis weights. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 183–192. ACM, 2015. (document), 1, 1.1.2, 1.1.2, 1.2.2, 1.2.2, 1.2.2, 2.1, 2.1, 4.1, 4.1, 4.1, 4.3, 5, 5, A, A, A

[CWW19] Kenneth L. Clarkson, Ruosong Wang, and David P. Woodruff. Dimensionality reduction for tukey regression. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1262–1271. PMLR, 2019. 1

[DBPS18] Alex Dytso, Ronit Bustin, H Vincent Poor, and Shlomo Shamai. Analytical properties of generalized gaussian distributions. *Journal of Statistical Distributions and Applications*, 5(1):1–40, 2018. 1.1.4

[DDH$^+$09] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009. 1

[DMMW12] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13:3475–3506, 2012. 4

[EMMZ22] Alessandro Epasto, Mohammad Mahdian, Vahab S. Mirrokni, and Peilin Zhong. Improved sliding window algorithms for clustering and coverage via bucketing-based sketches. In Joseph (Seffi) Naor and Niv Buchbinder, editors, *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pages 3005–3042. SIAM, 2022. 1

[FLPS22] Maryam Fazel, Yin Tat Lee, Swati Padmanabhan, and Aaron Sidford. Computing lewis weights to high precision. In Joseph (Seffi) Naor and Niv Buchbinder, editors, *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pages 2723–2742. SIAM, 2022. 2.1

[FSS20] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.*, 49(3):601–657, 2020. 2

[HW13] Moritz Hardt and David P. Woodruff. How robust are linear sketches to adaptive inputs? In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 121–130. ACM, 2013. 1.1.3

[JLS21] Arun Jambulapati, Yang P. Liu, and Aaron Sidford. Improved iteration complexities for overconstrained p-norm regression. *CoRR*, abs/2111.01848, 2021. 1.1.2, 1.3, 2.1, 2.1, 2.1, 2.4, 5

[Lee16] Yin Tat Lee. *Faster algorithms for convex and combinatorial optimization.* PhD thesis, Massachusetts Institute of Technology, 2016. 1.1.2, 1.3, 2.1, 5

[Lew78] D. R. Lewis. Finite dimensional subspaces of L_p. *Studia Mathematica*, 63(2):207–212, 1978. 1, 2.1, 2.1, 4.3

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991. (document), 1, 1.1.2, 1.2.2, 1.2.2, 2.1, A, A, C, C, C, C, C

[LWW21] Yi Li, Ruosong Wang, and David P. Woodruff. Tight bounds for the subspace sketch problem with applications. *SIAM J. Comput.*, 50(4):1287–1335, 2021. 1, 1, 8.1

[LWYZ20] Yi Li, Ruosong Wang, Lin Yang, and Hanrui Zhang. Nearly linear row sampling algorithm for quantile regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5979–5989. PMLR, 2020. 1

[MM13] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 91–100. ACM, 2013. 1

[MMM$^+$22] Raphael A. Meyer, Cameron N. Musco, Christopher P. Musco, David P. Woodruff, and Samson Zhou. Fast regression for structured inputs. In *International Conference on Learning Representations*, 2022. 4.2

[MMWY21] Cameron Musco, Christopher Musco, David P. Woodruff, and Taisuke Yasuda. Active sampling for linear regression beyond the ℓ_2 norm. *CoRR*, abs/2111.04888, 2021. (document), 1, 1, 1.1.2, 8.1, C

[MOP22] Alexander Munteanu, Simon Omlor, and Christian Peters. p-generalized probit regression and scalable maximum likelihood estimation via sketching and coresets. *CoRR*, 2022. 1.1.4, 2, 2, 2, 1.5, 2, 2, 3, 8.1

[MOW21] Alexander Munteanu, Simon Omlor, and David P. Woodruff. Oblivious sketching for logistic regression. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7861–7871. PMLR, 2021. 1.1.4, 2, 2, 3

[MRM21] Tung Mai, Anup B. Rao, and Cameron Musco. Coresets for classification - simplified and strengthened. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021. 1, 1.1.4, 2, 2, 2, 1.6, 2, 2, 3

[MSSW18] Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6562–6571, 2018. (document), 1.1.4, 2, 1.5, 3, 8.1

[Pan03] Dmitry Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Ann. Probab.*, 31(4):2068–2081, 2003. C.1

[PPP21] Aditya Parulekar, Advait Parulekar, and Eric Price. L1 regression with lewis weights subsampling. In Mary Wootters and Laura Sanità, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2021, August 16-18, 2021, University of Washington, Seattle, Washington, USA (Virtual Conference)*, volume 207 of *LIPIcs*, pages 49:1–49:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. 1

[PTB13] Udaya Parampalli, Xiaohu Tang, and Serdar Boztas. On the construction of binary sequence families with low correlation and large sizes. *IEEE Trans. Inf. Theory*, 59(2):1082–1089, 2013. 8.3

[RST10] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 1984–1992. Curran Associates, Inc., 2010. 1

[Sar06] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 143–152. IEEE Computer Society, 2006. 1

[Sch87] Gideon Schechtman. More on embedding subspaces of L_p in l_r^n. *Compositio Math.*, 61(2):159–169, 1987. 1.1.2, 2.1

[Sch11] Gideon Schechtman. Tight embedding of subspaces of L_p in ℓ_p^n for even p. *Proc. Amer. Math. Soc.*, 139(12):4419–4421, 2011. 1

[SS11] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011. 4

[SW11] Christian Sohler and David P. Woodruff. Subspace embeddings for the l_1-norm with applications. In Lance Fortnow and Salil P. Vadhan, editors, *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 755–764. ACM, 2011. 1

[SZ01] Gideon Schechtman and Artem Zvavitch. Embedding subspaces of l_p into l_p^n, $0 < p < 1$. *Mathematische Nachrichten*, 227(1):133–142, 2001. 1, 1.2.2, 2.1, 2.1, 4.3, A, A

[Tal90] Michel Talagrand. Embedding subspaces of L_1 into l_1^N. *Proc. Amer. Math. Soc.*, 108(2):363–369, 1990. 1, 1.2.2, 2.1

[Tal95] Michel Talagrand. 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.*, pages 311–325. Birkhäuser, Basel, 1995. 1, 1.2.2, 2.1

[Tro11] Joel A. Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab.*, 16:262–270, 2011. 7.2.1, 7.2

[UU21] Jalaj Upadhyay and Sarvagya Upadhyay. A framework for private matrix analysis in sliding window model. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10465–10475. PMLR, 2021. 1

[Ver18] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. C.2

[VH14] Ramon Van Handel. Probability in high dimension. Technical report, Princeton University, 2014. C

[WW19] Ruosong Wang and David P. Woodruff. Tight bounds for ℓ_p oblivious subspace embeddings. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 1825–1843. SIAM, 2019. 1

[WY22] David P. Woodruff and Taisuke Yasuda. High-dimensional geometric streaming in polynomial space. *CoRR*, 2022. 1, 1.2.1, 1.2.2, 2.1, 2.2, 2.1, 2.3, 3, 3.1, 3.1, 8.1

[WZ13] David P. Woodruff and Qin Zhang. Subspace embeddings and ℓ_p-regression using exponential random variables. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 546–567. JMLR.org, 2013. 1

A High Probability ℓ_p Lewis Weight Sampling, $0 < p < 2$

Note that the Lewis weight sampling algorithm of [CP15] samples from Lewis weights with replacement, while we sample each row once with probability proportional to the Lewis weight estimate. We thus carry out a similar analysis for this sampling process in the following discussion. A similar analysis has been carried out by [CD21, Lemma 3.2] for the case of ℓ_1 Lewis weights, in the context of active ℓ_1 linear regression. In order to obtain a $\log\frac{1}{\delta}$ dependence on the failure rate δ, we provide generalizations of the analysis conducted in [LT91, SZ01] by analyzing higher moments.

We first show the following moment bounds, which are a modification of results by [LT91] in a similar way as we did for the case of $p > 2$ in Theorem 5.1.

Theorem A.1 (Rademacher moment bounds). *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \in (0,2)$. Let $\mathbf{w}$ be ℓ_p Lewis weights for $\mathbf{A}$. Suppose that*

$$\frac{\mathbf{w}_i}{d} \leq \beta$$

for all $i \in [n]$, for some $\beta > 0$. Define the quantity

$$\Lambda \coloneqq \sup_{\|\mathbf{A}\mathbf{x}\|_p = 1} \left| \sum_{i=1}^n \boldsymbol{\sigma}_i |\langle \mathbf{a}_i, \mathbf{x}\rangle|^p \right|$$

Let $l \geq 1$ and let $\boldsymbol{\sigma} = \{\boldsymbol{\sigma}_i\}_{i=1}^n$ be independent Rademacher variables. Then,

$$\mathop{\mathbf{E}}_{\boldsymbol{\sigma}}[\Lambda^l] \leq \begin{cases} \left[C(p)\beta \cdot d[(\log d)^2(\log n) + l]\right]^{l/2} & p \in (0,1) \\ n[C(p)\beta \cdot d \cdot l]^{l/2} & p = 1 \\ \left[C(p)\beta \cdot d[(\log d)^3 + l]\right]^{l/2} & p \in (1,2) \end{cases}$$

where $C(p)$ is a constant depending only on p.

Proof. For $p = 1$, this is Lemma 8.4 of [CP15] with a slight change in the normalization. For $p \in (0,1) \cup (1,2)$, we only briefly sketch this result, since a similar result is worked out in detail in Section C for $p > 2$. As in this result, we use Lemma C.1 to instead bound the same Gaussian process considered in [LT91]. The only difference is then that we apply a tail version of Dudley's entropy integral, which requires a diameter bound, which is easily seen to be $O(\sqrt{d})$ from [LT91] for $1 < p < 2$ and [SZ01] for $0 < p < 1$. Then by integrating this tail bound, we attain the claimed moment bounds. □

We use these results to bound the distortion of the subspace embedding, as done in [CP15, CD21].

Theorem A.2 (High probability one-shot Lewis weight sampling)**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p < 2$. Let $\delta \in (0,1)$ be a failure rate parameter and let $\varepsilon \in (0,1)$ be an accuracy parameter. Let $\mathbf{w} \in \mathbb{R}^n$ be the ℓ_p Lewis weights. Suppose that we set $\mathbf{s}_i = 1/\mathbf{p}_i^{1/p}$ with probability $\mathbf{p}_i$, where $\mathbf{p}_i \geq \min\{\mathbf{w}_i/(d\beta), 1\}$, for*

$$\beta = \begin{cases} \dfrac{\varepsilon^2}{d[(\log d)^2(\log n) + \log\frac{1}{\delta}]} & p \in (1,2) \\ \dfrac{\varepsilon^2}{d\log\frac{n}{\delta}} & p = 1 \\ \dfrac{\varepsilon^2}{d[(\log d)^3 + \log\frac{1}{\delta}]} & p \in (0,1) \end{cases}$$

Then, with probability at least $1 - \delta$,

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

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

Proof. In what follows C will be a constant which can change from line to line. Consider the lth moment of the error, i.e.,

$$\mathop{\mathbf{E}}_{\mathbf{s}}\left[\sup_{\|\mathbf{Ax}\|_p = 1}\left|\left(\sum_{i=1}^n |\langle \mathbf{s}_i \cdot \mathbf{a}_i, \mathbf{x}\rangle|^p\right) - 1\right|^l\right].$$

We will use $l = O(\log\frac{1}{\delta}))$ for $p \in (0,2) \setminus \{1\}$ and $l = O(\log\frac{n}{\delta})$ for $p = 1$. By a standard symmetrization argument [CP15, CD21], this is bounded above by

$$2^l \mathop{\mathbf{E}}_{\mathbf{s}}\left[\sup_{\|\mathbf{Ax}\|_p = 1}\left|\sum_{i=1}^n \sigma_i |\langle \mathbf{s}_i \cdot \mathbf{a}_i, \mathbf{x}\rangle|^p\right|^l\right],$$

where $\sigma = \{\sigma_i\}_{i=1}^n$ are independent Rademacher variables.

Let β be the Lewis weight upper bound required by the statement of the current theorem. Then let $\mathbf{A}' \in \mathbb{R}^{r \times d}$ be a matrix with Lewis weights uniformly bounded by β such that

$$\mathbf{A}'^\top \mathbf{W}^p(\mathbf{A}')^{1-2/p}\mathbf{A}' \succeq \mathbf{A}^\top \mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A}$$

and $\|\mathbf{A}'\mathbf{x}\|_p = O(\|\mathbf{Ax}\|_p)$ for all $\mathbf{x} \in \mathbb{R}^d$. Such a $\mathbf{A}'$ exists with $r = \tilde{O}(d/\beta)$ rows by Lemma B.1 of [CP15] (see also [CD21, Lemma B.1]) for $p \in (1,2)$, while we can take $r = O(n)$ for $p \in (0,1)$ by splitting rows of $\mathbf{A}$, as suggested in [CP15] for results that are independent of n. We then define the vertical concatenation

$$\mathbf{A}'' := \begin{bmatrix} \mathbf{SA} \\ \mathbf{A}' \end{bmatrix}$$

where $\mathbf{S} = \operatorname{diag}(\mathbf{s})$. By the monotonicity of Lewis weights for $p \in (0,2)$ [CP15, Lemma 5.5], the Lewis weights of a row in $\mathbf{A}''$ are at most the Lewis weights of the row in the original matrix (either $\mathbf{SA}$ or $\mathbf{A}'$). We have that

$$\mathbf{A}''^\top \mathbf{W}^p(\mathbf{A}'')^{1-2/p}\mathbf{A}'' \succeq \mathbf{A}'^\top \mathbf{W}^p(\mathbf{A}')^{1-2/p}\mathbf{A}' \succeq \mathbf{A}^\top \mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A}$$

by construction of $\mathbf{A}'$. Then for any row i in $\mathbf{A}''$ corresponding to $\mathbf{SA}$,

$$\begin{aligned}\mathbf{w}_i^p(\mathbf{A}'')^{2/p} &= (\mathbf{s}_i\mathbf{a}_i)^\top(\mathbf{A}''^\top \mathbf{W}^p(\mathbf{A}'')^{1-2/p}\mathbf{A}'')^-(\mathbf{s}_i\mathbf{a}_i) \\ &\leq (\mathbf{s}_i\mathbf{a}_i)^\top(\mathbf{A}^\top \mathbf{W}^p(\mathbf{A})^{1-2/p}\mathbf{A})^-(\mathbf{s}_i\mathbf{a}_i) \\ &\leq \frac{\mathbf{w}_i^p(\mathbf{A})^{2/p}}{\mathbf{p}_i^{2/p}} \\ &\leq \beta^{2/p}\end{aligned}$$

by Lemma 2.6, so $\mathbf{w}_i^p(\mathbf{A}'') \leq \beta$. For any row $i \in \mathbf{A}''$ corresponding to $\mathbf{A}'$, we immediately have that $\mathbf{w}_i^p(\mathbf{A}'') \leq \mathbf{w}_i^p(\mathbf{A}') \leq \beta$ by monotonicity of Lewis weights for $p \in (0,2)$. Thus, $\mathbf{A}''$ has Lewis weights uniformly bounded by β. Applying Theorem A.1, we find that

$$\mathop{\mathbf{E}}_{\mathbf{s}}\left[\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\left(\sum_{i=1}^n |\langle \mathbf{s}_i\cdot\mathbf{a}_i,\mathbf{x}\rangle|^p\right)-1\right|^l\right] \leq O(\varepsilon)^l$$

for $p \in (0,2)\setminus\{1\}$ and

$$\mathop{\mathbf{E}}_{\mathbf{s}}\left[\sup_{\|\mathbf{A}\mathbf{x}\|_1=1}\left|\left(\sum_{i=1}^n |\langle \mathbf{s}_i\cdot\mathbf{a}_i,\mathbf{x}\rangle|\right)-1\right|^l\right] \leq nO(\varepsilon)^l$$

for $p = 1$, by our choice of l and β. By Markov's inequality, we then have that

$$\mathbf{Pr}\left\{\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\|\mathbf{SA}\|_p^p - 1\right|^l \geq \frac{1}{\delta}(C\cdot\varepsilon)^l\right\} \leq \delta$$

for $p \in (0,2)\setminus\{1\}$ and

$$\mathbf{Pr}\left\{\sup_{\|\mathbf{A}\mathbf{x}\|_1=1}\left|\|\mathbf{SA}\|_1 - 1\right|^l \geq \frac{n}{\delta}(C\cdot\varepsilon)^l\right\} \leq \delta.$$

Then taking lth roots for $l = O(\log\frac{1}{\delta})$ for $p \in (0,2)\setminus\{1\}$ and $l = O(\log\frac{n}{\delta})$ for $p = 1$, we have that

$$\mathbf{Pr}\left\{\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\|\mathbf{SA}\|_p^p - 1\right| \geq C\cdot\varepsilon\right\} \leq \delta. \qquad \square$$

B Preliminaries from Probability in Banach Spaces

We introduce generalizations from the theory of probability in Banach spaces that we will need for our purposes.

B.1 Entropy Bounds

Definition B.1 (Covering numbers). *Let $B_1, B_2 \subseteq \mathbb{R}^d$ be two sets. Define $E(B_1, B_2)$ to be the minimum number of translates of B_2 required to cover B_1. For a metric d and radius t, define $E(B_1, d_X, t)$ to be the minimum number d_X-balls of radius t required to cover B_1.*

Definition B.2 (Levy mean). *The Levy mean is defined as*

$$M_X = \int_{\mathbb{S}^{d-1}} \|\mathbf{x}\|\,d\sigma(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathbb{S}^{d-1}}\|\mathbf{x}\|.$$

Remark B.3. *By noting that* $\mathbf{x} \sim \mathbb{S}^{d-1}$ *is the same as drawing a Gaussian vector and normalizing, this is*

$$M_X = \mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_d)} \left\| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \right\| = \frac{\mathbf{E}\|\mathbf{g}\|_2}{\mathbf{E}\|\mathbf{g}\|_2} \mathbf{E} \left\| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \right\| = \frac{1}{\mathbf{E}\|\mathbf{g}\|_2} \mathbf{E}\|\mathbf{g}\|$$

since the norm of the Gaussian is independent of its direction.

Lemma B.4 (Dual Sudakov minoration (Proposition 4.2, [BLM89]))**.** *Let* $(X, \|\cdot\|)$ *be Banach space on* $\mathbb{R}^d$ *and let be the Levy mean of* $\|\cdot\|$*. Then, for some constant* $C > 0$*, we have that*

$$\log E(B_2, t \cdot B_X) \leq C \cdot d \left(\frac{M_X}{t} \right)^2$$

where $B_2 = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$ *and* $B_X = \{\mathbf{x} : \|\mathbf{x}\| \leq 1\}$*.*

B.2 Subspaces of ℓ_p in Generalized Lewis' Position

Definition B.5 (One-sided $\mathbf{U}$-weights)**.** *Let* $\mathbf{U} \in \mathbb{R}^{n\times d}$ *be an orthonormal matrix and let* $\gamma \in (0, 1]$*. We then define* γ*-one-sided* $\mathbf{U}$*-weights as any weights* $\mathbf{v} \in \mathbb{R}^n$ *satisfying*

$$\mathbf{v}_i \geq \gamma \frac{\left\|\mathbf{e}_i^\top \mathbf{U}\right\|_2^2}{d}.$$

We define the normalized one-sided $\mathbf{U}$*-weights to be* $\bar{\mathbf{v}}_i \coloneqq \mathbf{v}_i/T$ *for* $T \coloneqq \sum_{i=1}^n \mathbf{v}_i$*.*

Definition B.6. *Let* $\mathbf{U} \in \mathbb{R}^{n\times d}$ *be an orthonormal matrix and let* $\bar{\mathbf{v}}$ *be normalized* γ*-one-sided* $\mathbf{U}$*-weights (Definition B.5). Let* $0 < q < \infty$*. We then define the following (quasi-)norm on* $\mathbb{R}^d$*:*

$$\|\mathbf{x}\|_{\bar{\mathbf{v}},q} \coloneqq \left[\sum_{i=1}^n \bar{\mathbf{v}}_i \left| [\bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x}](i) \right|^q \right]^{1/q} = \left\| \bar{\mathbf{V}}^{1/q-1/2}\mathbf{U}\mathbf{x} \right\|_q$$

where $\bar{\mathbf{V}} \coloneqq \operatorname{diag}(\bar{\mathbf{v}})$*. For* $q = \infty$*, define*

$$\|\mathbf{x}\|_{\bar{\mathbf{v}},q} \coloneqq \left\| \bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x} \right\|_\infty.$$

We note the following equivalence bounds of these norms:

Lemma B.7. *Let* $T = \sum_{i=1}^n \mathbf{v}_i$*. The following hold for all* $\mathbf{x} \in \mathbb{R}^d$*:*

- *For* $0 < p < q \leq \infty$*,* $\|\mathbf{x}\|_{\bar{\mathbf{v}},p} \leq \|\mathbf{x}\|_{\bar{\mathbf{v}},q}$
- *For* $0 < p < 2$*,* $\|\mathbf{x}\|_{\bar{\mathbf{v}},2} \leq (Td/\gamma)^{1/p-1/2} \|\mathbf{x}\|_{\bar{\mathbf{v}},p}$
- *For* $2 < p \leq \infty$*,* $\|\mathbf{x}\|_{\bar{\mathbf{v}},p} \leq (Td/\gamma)^{1/2-1/p} \|\mathbf{x}\|_{\bar{\mathbf{v}},2}$
- *For* $0 < p < 2 < q \leq \infty$*,* $\|\mathbf{x}\|_{\bar{\mathbf{v}},p} \leq (Td/\gamma)^{1/p-1/q} \|\mathbf{x}\|_{\bar{\mathbf{v}},q}$

Proof. For $0 < q < \infty$, we have by Jensen's inequality that

$$\left[\sum_{i=1}^n \bar{\mathbf{v}}_i \left| [\bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x}](i) \right|^p \right]^{q/p} \leq \sum_{i=1}^n \bar{\mathbf{v}}_i \left| [\bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x}](i) \right|^q$$

and taking qth roots on both sides gives the first inequality. For $q = \infty$, we have that

$$\sum_{i=1}^n \bar{\mathbf{v}}_i \left| [\bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x}](i) \right|^p \leq \left\| \bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x} \right\|_\infty^p \sum_{i=1}^n \bar{\mathbf{v}}_i = \left\| \bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x} \right\|_\infty^p$$

and taking pth roots on both sides gives the result.

We then have that

$$\begin{aligned}
\left\|\bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x}\right\|_\infty &= \max_{i=1}^{n}\left|\mathbf{e}_i^\top\bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x}\right| \\
&= \sqrt{T}\max_{i=1}^{n}\left|\mathbf{e}_i^\top\mathbf{V}^{-1/2}\mathbf{U}\mathbf{x}\right| \\
&\leq \sqrt{T}\max_{i=1}^{n}\frac{\sqrt{d}}{\sqrt{\gamma}\left\|\mathbf{e}_i^\top\mathbf{U}\right\|_2}\left|\mathbf{e}_i^\top\mathbf{U}\mathbf{x}\right| && \gamma\text{-one-sidedness} \\
&\leq \sqrt{T}\max_{i=1}^{n}\frac{\sqrt{d}}{\sqrt{\gamma}\left\|\mathbf{e}_i^\top\mathbf{U}\right\|_2}\left\|\mathbf{e}_i^\top\mathbf{U}\right\|_2\|\mathbf{x}\|_2 && \text{Cauchy–Schwarz} \\
&= \sqrt{Td/\gamma}\|\mathbf{x}\|_2 = \sqrt{Td/\gamma}\|\mathbf{U}\mathbf{x}\|_2 = \sqrt{Td/\gamma}\|\mathbf{x}\|_{\bar{\mathbf{v}},2}
\end{aligned}$$

which gives the bound on $\|\mathbf{x}\|_{\bar{\mathbf{v}},\infty}$. Finally, for $0 < p < 2$,

$$\|\mathbf{x}\|_{\bar{\mathbf{v}},2}^2 \leq \|\mathbf{x}\|_{\bar{\mathbf{v}},\infty}^{2-p}\|\mathbf{x}\|_{\bar{\mathbf{v}},p}^p \leq (Td/\gamma)^{(2-p)/2}\|\mathbf{x}\|_{\bar{\mathbf{v}},2}^{2-p}\|\mathbf{x}\|_{\bar{\mathbf{v}},p}^p \implies \|\mathbf{x}\|_{\bar{\mathbf{v}},2} \leq (Td/\gamma)^{1/p-1/2}\|\mathbf{x}\|_{\bar{\mathbf{v}},p}$$

and for $p \geq 2$,

$$\|\mathbf{x}\|_{\bar{\mathbf{v}},p}^p \leq \|\mathbf{x}\|_{\bar{\mathbf{v}},\infty}^{p-2}\|\mathbf{x}\|_{\bar{\mathbf{v}},2}^2 \leq (Td/\gamma)^{(p-2)/2}\|\mathbf{x}\|_{\bar{\mathbf{v}},2}^{p-2}\|\mathbf{x}\|_{\bar{\mathbf{v}},2}^2 \implies \|\mathbf{x}\|_{\bar{\mathbf{v}},p} \leq (Td/\gamma)^{1/2-1/p}\|\mathbf{x}\|_{\bar{\mathbf{v}},2}$$

The last inequality follows by combining the previous two inequalities. □

We now bound the Levy mean for ℓ_p norms induced by orthonormal matrices $\mathbf{U}$.

Lemma B.8. *Let $\mathbf{U} \in \mathbb{R}^{n\times d}$ be an orthonormal matrix and let $\mathbf{v}$ be γ-one-sided $\mathbf{U}$-weights, let T be their sum, and let $\bar{\mathbf{v}}$ be their normalization. Let $1 \leq q < \infty$. The Levy mean for the (quasi-)norm $\|\cdot\|_{\bar{\mathbf{v}},q}$ (Definition B.6) is at most*

$$M_X \leq O((Tq)^{1/2}).$$

Proof. We have that

$$\begin{aligned}
M_X &= \frac{\mathbf{E}\|\mathbf{g}\|_{\bar{\mathbf{v}},q}}{\mathbf{E}\|\mathbf{g}\|_2} \\
&= \frac{1}{\mathbf{E}\|\mathbf{g}\|_2}\,\mathbf{E}\left[\sum_{i=1}^{n}\left|\bar{\mathbf{v}}_i^{-1/2}\cdot\mathbf{e}_i^\top\mathbf{U}\mathbf{g}\right|^q\bar{\mathbf{v}}_i\right]^{1/q} \\
&= \frac{1}{\mathbf{E}\|\mathbf{g}\|_2}\left[\sum_{i=1}^{n}\mathbf{E}\left|\bar{\mathbf{v}}_i^{-1/2}\cdot\mathbf{e}_i^\top\mathbf{U}\mathbf{g}\right|^q\bar{\mathbf{v}}_i\right]^{1/q} && \text{Jensen} \\
&\leq \frac{1}{\mathbf{E}\|\mathbf{g}\|_2}O(q^{1/2})\left[\sum_{i=1}^{n}\left\|\bar{\mathbf{v}}_i^{-1/2}\cdot\mathbf{e}_i^\top\mathbf{U}\right\|_2^q\bar{\mathbf{v}}_i\right]^{1/q} && \text{Gaussian moments} \\
&= \frac{1}{\mathbf{E}\|\mathbf{g}\|_2}O((Tq)^{1/2})\left[\sum_{i=1}^{n}\left\|\mathbf{v}_i^{-1/2}\cdot\mathbf{e}_i^\top\mathbf{U}\right\|_2^q\bar{\mathbf{v}}_i\right]^{1/q} \\
&\leq \frac{\sqrt{d}}{\mathbf{E}\|\mathbf{g}\|_2}O((Tq/\gamma)^{1/2})\left[\sum_{i=1}^{n}\left\|\frac{\mathbf{e}_i^\top\mathbf{U}}{\left\|\mathbf{e}_i^\top\mathbf{U}\right\|_2}\right\|_2^q\bar{\mathbf{v}}_i\right]^{1/q} && \gamma\text{-one-sidedness} \\
&= \frac{\sqrt{d}}{\mathbf{E}\|\mathbf{g}\|_2}O((Tq/\gamma)^{1/2})\left[\sum_{i=1}^{n}\bar{\mathbf{v}}_i\right]^{1/q} \\
&\leq O((Tq/\gamma)^{1/2}).
\end{aligned}$$

□

The above bound translates into entropy bounds in $\mathbb{R}^d$ by dual Sudakov minoration (Lemma B.4), which in turn implies entropy bounds in the subspace spanned by $\mathbf{V}^{-1/2}\mathbf{U}$:

Corollary B.9. *Let $\mathbf{U} \in \mathbb{R}^{n\times d}$ be an orthonormal matrix and let $1 \leq q < \infty$. Let $\mathbf{v} \in \mathbb{R}^n$ be γ-one-sided $\mathbf{U}$-weights, let T be their sum, and let $\bar{\mathbf{v}}$ be their normalization. Let $\bar{\mathbf{V}} = \operatorname{diag}(\bar{\mathbf{v}})$. Let $E \subseteq \mathbb{R}^n$ be the subspace spanned by $\bar{\mathbf{V}}^{-1/2}\mathbf{U}$ and for $0 < p < \infty$, and define the norm*

$$\|\mathbf{y}\|_{\bar{\mathbf{v}},p} \coloneqq \left[\sum_{i=1}^n \bar{\mathbf{v}}_i |\mathbf{y}(i)|^p\right]^{1/p}.$$

on E. Denote by $B_{\bar{\mathbf{v}},p}(E)$ the unit balls of $\|\cdot\|_{\bar{\mathbf{v}},p}$ in the subspace E. Then, for some constant $C > 0$, we have that

$$\log E(B_{\bar{\mathbf{v}},2}(E), t\cdot B_{\bar{\mathbf{v}},q}(E)) \leq C\cdot\frac{Tqd}{\gamma t^2}.$$

Proof. We have by Lemmas B.4 and B.8 that

$$\log E(B_2(\mathbb{R}^d), t\cdot B_{\bar{\mathbf{v}},q}(\mathbb{R}^d)) \leq C\cdot\frac{Tqd}{\gamma t^2}$$

where $B_2(\mathbb{R}^d) \subseteq \mathbb{R}^d$ is the unit ℓ_2 ball in d dimensions, and $B_{\bar{\mathbf{v}},q}(\mathbb{R}^d) \subseteq \mathbb{R}^d$ is the unit $\|\cdot\|_{\bar{\mathbf{v}},q}$ ball in d dimensions. Note that E equipped with the norm $\|\cdot\|_{\bar{\mathbf{v}},2}$ is isometric with $\mathbb{R}^d$ equipped with the usual ℓ_2 norm, since

$$\left\|\bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x}\right\|_{\mathbf{v},2} = \left[\sum_{i=1}^n \bar{\mathbf{v}}_i \left|[\bar{\mathbf{V}}^{-1/2}\mathbf{U}\mathbf{x}](i)\right|^2\right]^{1/2} = \|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2.$$

Furthermore, E equipped with the norm $\|\cdot\|_{\bar{\mathbf{v}},q}$ is isometric with $\mathbb{R}^d$ equipped with the $\|\cdot\|_{\bar{\mathbf{v}},q}$ norm. It follows that the covering numbers must then be the same. □

C One-Sided Lewis Weight Sampling, $2 < p < \infty$

We now work out the changes necessary to [LT91] to make the γ-one-sided weights sampling work. A similar proof for usual Lewis weights is worked out in detail in [MMWY21].

Theorem 5.1. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \in (2,\infty)$. Let $\mathbf{w}$ be γ-one-sided ℓ_p Lewis weights for $\mathbf{A}$. Suppose that*

$$\frac{\mathbf{w}_i}{d} \leq \beta$$

for all $i \in [n]$, for some $\beta > 0$. Define the quantity

$$\Lambda \coloneqq \sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\sum_{i=1}^n \boldsymbol{\sigma}_i |\langle \mathbf{a}_i, \mathbf{x}\rangle|^p\right|$$

Then,

$$\mathop{\mathbf{E}}_{\boldsymbol{\sigma}}[\Lambda^l] \leq \left[O(1)\beta\cdot T_{\mathbf{w}}^{p/2}[\gamma^{-1}(\log d)^2(\log n) + l]\right]^{l/2}$$

for any $l \geq 1$, where $\boldsymbol{\sigma} = \{\boldsymbol{\sigma}_i\}_{i=1}^n$ are independent Rademacher variables.

We first change the position of the subspace to the generalized Lewis's position associated with the orthonormal matrix $\mathbf{U} = \mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}$ (see Corollary B.9) and γ-one-sided $\mathbf{U}$-weights $\mathbf{v} = \mathbf{w}/d$ (note that $\mathbf{w}$ is a factor of d off from the normalization of Definition B.5). That is, we write

$$\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\sum_{i=1}^n \boldsymbol{\sigma}_i |\langle \mathbf{a}_i, \mathbf{x}\rangle|^p\right|^l = \sup_{\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p=1}\left|\sum_{i=1}^n \mathbf{w}_i\boldsymbol{\sigma}_i \left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p\right|^l$$

so that the subspace E of Corollary B.9 is the same as the one spanned by $\mathbf{W}^{-1/2}\cdot\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R} = \mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}$.

We now partition $[n]$ into two sets of coordinates, the coordinates J such that $\mathbf{w}_i \geq 1/\operatorname{poly}(n)$, and its complement. For coordinates $i \notin J$, note that

$$\sup_{\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p=1}\left|\sum_{i\notin J}\mathbf{w}_i\boldsymbol{\sigma}_i\left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p\right| \leq \frac{1}{\operatorname{poly}(n)}\sup_{\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p=1}\sum_{i\notin J}\left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p \leq \frac{1}{\operatorname{poly}(n)}$$

by Lemma 2.3, for any $\boldsymbol{\sigma}$, so it suffices to consider the coordinates in J.

Bounding by a Gaussian process. We first show that it suffices to bound a certain Gaussian process. To make this comparison, we will use the following lemma of Panchenko (see also Lemma 7.6 of [VH14]):

Lemma C.1 (Lemma 1, [Pan03]). *Let X, Y be random variables such that*

$$\mathbf{E}[\Phi(X)] \leq \mathbf{E}[\Phi(Y)]$$

for every increasing convex function Φ. If

$$\mathbf{Pr}\{Y \geq t\} \leq c_1\exp(-c_2t^\alpha) \qquad \text{for all } t \geq 0,$$

for some $c_1, \alpha \geq 1$ and $c_2 > 0$, then

$$\mathbf{Pr}\{X \geq t\} \leq c_1\exp(1-c_2t^\alpha) \qquad \text{for all } t \geq 0.$$

Let Φ be a convex increasing function. We first note that since $\mathbf{w}_i \leq d\beta$, $\mathbf{w}_i \leq (d\beta\mathbf{w}_i)^{1/2}$, so by the Rademacher contraction principle [LT91, Theorem 4.12], we have that

$$\mathop{\mathbf{E}}_{\boldsymbol{\sigma}}\Phi\left(\sup_{\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p=1}\left|\sum_{i\in J}\mathbf{w}_i\boldsymbol{\sigma}_i\left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p\right|\right) \leq \\ \mathop{\mathbf{E}}_{\boldsymbol{\sigma}}\Phi\left(2\beta^{1/2}\sup_{\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p=1}\left|\sum_{i\in J}(d\mathbf{w}_i)^{1/2}\boldsymbol{\sigma}_i\left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p\right|\right).$$

Then by a comparison theorem between Rademacher and Gaussian averages (see Equation 4.8 of [LT91]), we have that

$$\mathop{\mathbf{E}}_{\boldsymbol{\sigma}}\Phi\left(2\beta^{1/2}\sup_{\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p=1}\left|\sum_{i\in J}(d\mathbf{w}_i)^{1/2}\boldsymbol{\sigma}_i\left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p\right|\right) \leq \\ \mathop{\mathbf{E}}_{g}\Phi\left(\sqrt{2\pi}\beta^{1/2}\sup_{\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p=1}\left|\sum_{i\in J}(d\mathbf{w}_i)^{1/2}g_i\left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p\right|\right).$$

for independent standard Gaussians g_i. Thus, by Lemma C.1, it suffices to obtain tail bounds on

$$\sqrt{2\pi}\beta^{1/2}\sup_{\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p=1}\left|\sum_{i\in J}(d\mathbf{w}_i)^{1/2}g_i\left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p\right|. \tag{11}$$

Note that in the original proof of [LT91], the contraction principle and comparison theorems are directly used to bound the expected supremum, while we pass through Lemma C.1 to obtain tail bounds.

Bounding the Gaussian process. We now bound the Gaussian process of (11). We will obtain tail bounds via the following tail bound version of Dudley's inequality:

Theorem C.2 (Theorem 8.1.6, [Ver18]). *Let $(X_t)_{t\in T}$ be a Gaussian process with pseudo-metric $d_X(s,t) := \|X_s - X_t\|_2$. Let $E(T, d_X, u)$ denote the minimal number of d_X-balls of radius u required to cover T. Then, for every $u \geq 0$, we have that*

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

We thus need to bound the metric

$$d_X(\mathbf{y},\mathbf{y}') \coloneqq \left[\sum_{i\in J} d\mathbf{w}_i\left(\left|\mathbf{W}^{-1/p}\mathbf{y}(i)\right|^p - \left|\mathbf{W}^{-1/p}\mathbf{y}'(i)\right|^p\right)^2\right]^{1/2}$$

in order to obtain entropy diameter bounds. This is the exact same metric bounded by [LT91], with Lewis weights replaced by one-sided Lewis weights. Modifying their bound straightforwardly leads to the following:

Lemma C.3. *Define $\|\mathbf{y}\|_J \coloneqq \max_{i\in J}|\mathbf{y}(i)|$. Then,*

$$d_X(\mathbf{y},\mathbf{y}') \leq 2p\sqrt{d}(\|\mathbf{w}\|_1^{1/2-1/p})^{p/2-1}\left\|\mathbf{W}^{-1/p}\mathbf{y} - \mathbf{W}^{-1/p}\mathbf{y}'\right\|_J$$

Proof. We closely follow the proof of (15.18) of [LT91]. We first handle $p > 2$. For $a, b \geq 0$, we have the elementary

$$a^p - b^p \leq p(a^{p-1} + b^{p-1})|a-b|.$$

This gives that

$$\begin{aligned}
d_X(\mathbf{y},\mathbf{y}')^2 &\leq 2p^2\sum_{i\in J} d\mathbf{w}_i \max\left\{\left|\mathbf{W}^{-1/p}\mathbf{y}(i)\right|^{2p-2}, \left|\mathbf{W}^{-1/p}\mathbf{y}'(i)\right|^{2p-2}\right\}\left|\mathbf{W}^{-1/p}\mathbf{y}(i) - \mathbf{W}^{-1/p}\mathbf{y}'(i)\right|^2 \\
&\leq 2p^2\left\|\mathbf{W}^{-1/p}\mathbf{y} - \mathbf{W}^{-1/p}\mathbf{y}'\right\|_J^2 \sum_{i\in J} d\mathbf{w}_i \max\left\{\left|\mathbf{W}^{-1/p}\mathbf{y}(i)\right|^{2p-2}, \left|\mathbf{W}^{-1/p}\mathbf{y}'(i)\right|^{2p-2}\right\} \\
&\leq 2p^2(\|\mathbf{w}\|_1^{1/2-1/p})^{p-2}\left\|\mathbf{W}^{-1/p}\mathbf{y} - \mathbf{W}^{-1/p}\mathbf{y}'\right\|_J^2 \sum_{i\in J} d\mathbf{u}_i \max\left\{\left|\mathbf{W}^{-1/p}\mathbf{y}(i)\right|^{p}, \left|\mathbf{W}^{-1/p}\mathbf{y}'(i)\right|^{p}\right\} \\
&\leq 2p^2 d(\|\mathbf{w}\|_1^{1/2-1/p})^{p-2}\left\|\mathbf{W}^{-1/p}\mathbf{y} - \mathbf{W}^{-1/p}\mathbf{y}'\right\|_J^2(\|\mathbf{y}\|_p^p + \|\mathbf{y}'\|_p^p) \\
&\leq 4p^2 d(\|\mathbf{w}\|_1^{1/2-1/p})^{p-2}\left\|\mathbf{W}^{-1/p}\mathbf{y} - \mathbf{W}^{-1/p}\mathbf{y}'\right\|_J^2,
\end{aligned}$$

where we have used Lemma 2.3 to bound $\left|\mathbf{W}^{-1/p}\mathbf{y}(i)\right|$. Taking square roots yields the claim. □

As a consequence, we get diameter bounds.

Lemma C.4. *The d_X-diameter of B_p is at most $4p(\|\mathbf{w}\|_1^{p/2-1}d)^{1/2} \leq 4p(\|\mathbf{w}\|_1^{p/2})^{1/2}$.*

Proof. Let $\mathbf{y}, \mathbf{y}' \in B_p$. Then, by Lemma 2.3,

$$\begin{aligned}
\left\|\mathbf{W}^{-1/p}(\mathbf{y}-\mathbf{y}')\right\|_J^p &\leq \left\|\mathbf{W}^{-1/p}(\mathbf{y}-\mathbf{y}')\right\|_\infty^p \\
&\leq \max_{i=1}^n \mathbf{w}_i^{-1}\cdot\|\mathbf{w}\|_1^{p/2-1}\mathbf{w}_i\|\mathbf{y}-\mathbf{y}'\|_p^p \\
&\leq 2^p\|\mathbf{w}\|_1^{p/2-1}.
\end{aligned}$$

Then by Lemma C.3,

$$\begin{aligned}
\sup_{\mathbf{y},\mathbf{y}'\in B_p} d_X(\mathbf{y},\mathbf{y}') &\leq \sup_{\mathbf{y},\mathbf{y}'\in B_p} 2p\sqrt{d}(\|\mathbf{w}\|_1^{1/2-1/p})^{p/2-1}\left\|\mathbf{W}^{-1/p}\mathbf{y} - \mathbf{W}^{-1/p}\mathbf{y}'\right\|_J \\
&\leq \sup_{\mathbf{y},\mathbf{y}'\in B_p} 4p\sqrt{d}(\|\mathbf{w}\|_1^{1/2-1/p})^{p/2-1}\|\mathbf{w}\|_1^{1/2-1/p} \\
&\leq 4p\|\mathbf{w}\|_1^{p/4-1/2}\sqrt{d}.
\end{aligned}$$

□

Let $T_\mathbf{w}$ denote the sum of the one-sided Lewis weights. Then,

$$\|\mathbf{ARx}\|_p^p = \sum_{i=1}^n|[\mathbf{ARx}](i)|^p = \sum_{i=1}^n \mathbf{w}_i\left|[\mathbf{W}^{-1/p}\mathbf{ARx}](i)\right|^p$$

$$= d\sum_{i=1}^n \mathbf{v}_i \left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p = T_{\mathbf{w}} \sum_{i=1}^n \bar{\mathbf{v}}_i \left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p = T_{\mathbf{w}}\left\|\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}\right\|_{\bar{\mathbf{v}},p}^p$$

so the unit ℓ_p ball

$$B_p \coloneqq \left\{\mathbf{A}\mathbf{x} : \|\mathbf{A}\mathbf{x}\|_p \leq 1\right\}$$

is isometric to $T_{\mathbf{w}}^{-1/p} \cdot B_{\bar{\mathbf{v}},q}(E)$, which is the unit ball in the subspace $E = \operatorname{colspan}(\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R})$ from Corollary B.9, scaled down by $T_{\mathbf{w}}^{1/p}$.

Let

$$\Delta = 2p\sqrt{d}(T_{\mathbf{w}}^{1/2-1/p})^{p/2-1}.$$

Then for $q = O(\log n)$, note that

$$d_X(\mathbf{y},\mathbf{y}') \leq \Delta\left\|\mathbf{W}^{-1/p}\mathbf{y} - \mathbf{W}^{-1/p}\mathbf{y}'\right\|_J \leq O(\Delta)\left\|\mathbf{W}^{-1/p}\mathbf{y} - \mathbf{W}^{-1/p}\mathbf{y}'\right\|_{\bar{\mathbf{v}},q}$$

since $\mathbf{w}_i \geq 1/\operatorname{poly}(n)$ for $i \in J$. We may then bound the covering number of B_p by d_X (the metric associated with the Gaussian process) as

$$\begin{aligned}\log E(B_p, d_X, t) &\leq \log E(T_{\mathbf{w}}^{-1/p} \cdot B_{\bar{\mathbf{v}},q}(E), \|\cdot\|_{\bar{\mathbf{v}},q}, \Theta(t/\Delta)) \\ &= \log E(B_{\bar{\mathbf{v}},q}(E), \|\cdot\|_{\bar{\mathbf{v}},q}, \Theta(T_{\mathbf{w}}^{1/p}t/\Delta)) \\ &\leq \log E\left(B_{\bar{\mathbf{v}},q}(E), \|\cdot\|_{\bar{\mathbf{v}},q}, \Theta(1)\frac{t}{T_{\mathbf{w}}^{p/4-1/2}}\right).\end{aligned}$$

We now compute Dudley's entropy integral. We have that

$$\begin{aligned}&\int_0^\infty \sqrt{\log E(B_p, d_X, t)}\,dt \\ \leq\ & O(1)T_{\mathbf{w}}^{p/4-1/2}\int_0^\infty \sqrt{\log E(B_{\bar{\mathbf{v}},p}, \|\cdot\|_{\bar{\mathbf{v}},q}, t)}\,dt \\ \leq\ & O(1)T_{\mathbf{w}}^{p/4-1/2}\left[\int_0^1 \sqrt{\log E(B_{\bar{\mathbf{v}},p}, \|\cdot\|_{\bar{\mathbf{v}},q}, t)}\,dt + \int_1^\infty \sqrt{\log E(B_{\bar{\mathbf{v}},p}, \|\cdot\|_{\bar{\mathbf{v}},q}, t)}\,dt\right] \\ \leq\ & O(1)T_{\mathbf{w}}^{p/4-1/2}\left[\int_0^1 \sqrt{d\log\frac{d}{t}}\,dt + \int_1^{\operatorname{poly}(d)} \frac{\sqrt{T_{\mathbf{w}}\log n}}{\sqrt{\gamma}t}\,dt\right] \\ \leq\ & O(1)\frac{T_{\mathbf{w}}^{p/4}}{\sqrt{\gamma}}(\log d)\sqrt{\log n}\end{aligned}$$

where we have used a standard volume argument for the entropy bound for $t \in (0,1)$ and Corollary B.9 for the entropy bound for $t \in (1,\infty)$. Then by combining the above calculation with Theorem C.2 and the diameter calculation of Lemma C.4 that for some constant $C > 0$,

$$\mathbf{Pr}\left\{\Lambda' \geq C \cdot T_{\mathbf{w}}^{p/4}\left[\gamma^{-1/2}(\log d)\sqrt{\log n} + z\right]\right\} \leq 2\exp(-z^2)$$

for

$$\Lambda' = \sup_{\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p = 1}\left|\sum_{i\in J}(d\mathbf{w}_i)^{1/2} g_i \left|[\mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)\right|^p\right|.$$

Moment Bounds. We now piece together our work above to obtain moment bounds on Λ. We start by getting moment bounds for Λ'. We have that

$$\mathbf{E}[\Lambda'^l] = (C \cdot T_{\mathbf{w}}^{p/4})^l\,\mathbf{E}\left[\left(\frac{\Lambda'}{C \cdot T_{\mathbf{w}}^{p/4}}\right)^l\right]$$

$$
\begin{aligned}
&= (C \cdot T_{\mathbf{w}}^{p/4})^l \int_0^\infty z^l \cdot \mathbf{Pr}\Big\{ \frac{\Lambda'}{C \cdot T_{\mathbf{w}}^{p/4}} \geq z \Big\}\, dz \\
&= (C \cdot T_{\mathbf{w}}^{p/4})^l \Bigg[\int_0^{2\gamma^{-1/2}(\log d)\sqrt{\log n}} z^l \cdot \mathbf{Pr}\Big\{ \frac{\Lambda'}{C \cdot T_{\mathbf{w}}^{p/4}} \geq z \Big\}\, dz + \int_{2\gamma^{-1/2}(\log d)\sqrt{\log n}}^\infty z^l \cdot \mathbf{Pr}\Big\{ \frac{\Lambda'}{C \cdot T_{\mathbf{w}}^{p/4}} \geq z \Big\}\, dz \Bigg] \\
&\leq (C \cdot T_{\mathbf{w}}^{p/4})^l \Bigg[(4\gamma^{-1/2}(\log d)\sqrt{\log n})^l + \int_{2\gamma^{-1/2}(\log d)\sqrt{\log n}}^\infty z^l \cdot \mathbf{Pr}\Big\{ \Lambda' \geq C \cdot T_{\mathbf{w}}^{p/4} z \Big\}\, dz \Bigg] \\
&\leq (C \cdot T_{\mathbf{w}}^{p/4})^l \Bigg[(4\gamma^{-1/2}(\log d)\sqrt{\log n})^l + \int_{4\gamma^{-1/2}(\log d)\sqrt{\log n}}^\infty z^l \cdot \mathbf{Pr}\Big\{ \Lambda' \geq C \cdot T_{\mathbf{w}}^{p/4} [\gamma^{-1/2}(\log d)\sqrt{\log n} + 3z/4] \Big\}\, dz \Bigg] \\
&\leq (C \cdot T_{\mathbf{w}}^{p/4})^l \Bigg[(4\gamma^{-1/2}(\log d)\sqrt{\log n})^l + 2\int_{4\gamma^{-1/2}(\log d)\sqrt{\log n}}^\infty z^l \exp(-z^2/2)\, dz \Bigg] \\
&\leq (C \cdot T_{\mathbf{w}}^{p/4})^l \Bigg[(4\gamma^{-1/2}(\log d)\sqrt{\log n})^l + 2\frac{l!}{2^{l/2}(l/2)!} \Bigg] \\
&\leq (C \cdot T_{\mathbf{w}}^{p/4})^l \Big[(4\gamma^{-1/2}(\log d)\sqrt{\log n})^l + 2(l/2)^{l/2} \Big] \\
&= \Big[O(1) T_{\mathbf{w}}^{p/2} (\gamma^{-1}(\log d)^2 \log n + l) \Big]^{l/2}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathbf{E}[\Lambda^l] &\leq 2^l \Big((1/\operatorname{poly}(n))^l + \mathbf{E}[(\sqrt{2\pi}\beta^{1/2}\Lambda')^l] \Big) \\
&\leq \Big[O(1)\beta \cdot T_{\mathbf{w}}^{p/2} (\gamma^{-1}(\log d)^2 \log n + l) \Big]^{l/2}
\end{aligned}
$$

as desired.