Optimal bounds for ℓ_p sensitivity sampling via ℓ_2 augmentation

Alexander Munteanu [1] **Simon Omlor** [2 3]

Abstract

Data subsampling is one of the most natural methods to approximate a massively large data set by a small representative proxy. In particular, sensitivity sampling received a lot of attention, which samples points proportional to an individual importance measure called sensitivity. This framework reduces in very general settings the size of data to roughly the VC dimension d times the total sensitivity $\mathfrak{S}$ while providing strong $(1 \pm \varepsilon)$ guarantees on the quality of approximation. The recent work of Woodruff & Yasuda (2023c) improved substantially over the general $\tilde{O}(\varepsilon^{-2}\mathfrak{S}d)$ bound for the important problem of ℓ_p subspace embeddings to $\tilde{O}(\varepsilon^{-2}\mathfrak{S}^{2/p})$ for $p \in [1,2]$. Their result was subsumed by an earlier $\tilde{O}(\varepsilon^{-2}\mathfrak{S}d^{1-p/2})$ bound which was implicitly given in the work of Chen & Derezinski (2021). We show that their result is tight when sampling according to plain ℓ_p sensitivities. We observe that by augmenting the ℓ_p sensitivities by ℓ_2 sensitivities, we obtain better bounds improving over the aforementioned results to optimal linear $\tilde{O}(\varepsilon^{-2}(\mathfrak{S}+d)) = \tilde{O}(\varepsilon^{-2}d)$ sampling complexity for all $p \in [1,2]$. In particular, this resolves an open question of Woodruff & Yasuda (2023c) in the affirmative for $p \in [1,2]$ and brings sensitivity subsampling into the regime that was previously only known to be possible using Lewis weights (Cohen & Peng, 2015). As an application of our main result, we also obtain an $\tilde{O}(\varepsilon^{-2}\mu d)$ sensitivity sampling bound for logistic regression, where μ is a natural complexity measure for this problem. This improves over the previous $\tilde{O}(\varepsilon^{-2}\mu^2 d)$ bound of Mai et al. (2021) which was based on Lewis weights subsampling.

[*]Equal contribution [1]Dortmund Data Science Center, Faculties of Statistics and Computer Science, TU Dortmund University, Dortmund, Germany [2]Faculty of Statistics, TU Dortmund University, Dortmund, Germany [3]Lamarr-Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany. Correspondence to: Alexander Munteanu <alexander.munteanu@tu-dortmund.de>, Simon Omlor <simon.omlor@tu-dortmund.de>.

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

1. Introduction

Massive data sets have become ubiquitous in recent years and standard machine learning approaches reach the limits of tractability when these large data need to be analyzed. Given a data matrix $A \in \mathbb{R}^{n\times d}$ where the number of data points exceeds the number of features by a large margin, i.e., $n \gg d$, a popular approach to address the computational limitations is to subsample the data (Munteanu, 2023). While uniform sampling is widely used in practice, it can be associated with large loss of approximation accuracy for various machine learning models. Thus, a lot of work has been dedicated to the design of importance sampling schemes, to subsample points proportional to some sort of importance measure for the contribution of individual points, such that important points become more likely to be in the sample, while less important or redundant points are sampled with lower probability (Munteanu & Schwiegelshohn, 2018).

General setting. The arguably most general and popular importance sampling approach is called the sensitivity framework (Feldman & Langberg, 2011; Feldman et al., 2020). It aims at approximating functions applied to data that can be expressed in the following way: consider an individual non-negative loss function $h_i\colon \Omega \to \mathbb{R}_{\geq 0}$ for each row vector $a_i, i \in [n] := \{1,2,\dots,n\}$ of the data matrix A, more precisely we set $h_i(x) = h(a_i x)$ for some function $h\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$. The problem we consider is approximating

$$f(x) := \sum_{i\in[n]} h_i(x) \tag{1}$$

over all x in the domain Ω, by subsampling and reweighting the individual contributions. Formally, we obtain $S \subset [n]$ with $|S| \ll n$, and $w_i > 0$ for all $i \in S$ and define

$$\tilde{f}(x) := \sum_{i\in S} w_i h_i(x) \tag{2}$$

to be the surrogate loss, that given an approximation parameter $\varepsilon \in (0,1)$ should be a pointwise $(1\pm\varepsilon)$-approximation for the original loss function. More specifically, we require that

$$\forall x \in \Omega\colon \tilde{f}(x) = (1\pm\varepsilon) f(x). \tag{3}$$

The surrogate function (2) can then be used in downstream machine learning tasks, such as optimization, with very little

bounded errors, and can be processed much more efficiently than the original loss. Note, that most empirical risk minimization problems or negative log-likelihood functions can be expressed as in Equation (1).

Importance subsampling. The sensitivity framework provides a way of obtaining the guarantee of Equation (3) by first computing individual sensitivities for $i \in [n]$

$$\varsigma_i = \sup_{x \in \Omega} \frac{h_i(x)}{\sum_{j \in [n]} h_j(x)}, \tag{4}$$

where we let $\varsigma_i = 0$ when the denominator is zero. We note that computing the sensitivity is usually intractable, but it suffices to obtain close approximations, which can be done efficiently for many important problems.

Then, subsampling $S \subset [n]$ with probabilities p_i proportional to ς_i and reweighting the individual contributions by $w_i = 1/(p_i|S|)$, preserves the objective function in expectation and when $|S| = \tilde{O}(\varepsilon^{-2}\mathfrak{S}d)$, we obtain by a concentration result that the $(1 \pm \varepsilon)$-approximation defined in Equation (3) holds with constant probability. Here, d is the VC dimension of a set system associated with the functions h_i, and $\mathfrak{S} = \sum_{i \in [n]} \varsigma_i$ denotes their total sensitivity.

Subspace embeddings for ℓ_p. In this paper, we focus on applying the above framework to the more specific, yet very versatile problem of constructing a so-called ℓ_p subspace embedding via subsampling. Given a data matrix $A \in \mathbb{R}^{n \times d}$ with *row-vectors* $a_i, i \in [n]$, and a norm parameter $p \in [1, \infty]$, our goal is to calculate probabilities p_i, and weights w_i for each $i \in [n]$, such that a small random subsample $S \subset [n]$ according to this distribution satisfies

$$\forall x \in \mathbb{R}^d \colon \sum_{i \in S} w_i |a_i x|^p = (1 \pm \varepsilon) \sum_{i \in [n]} |a_i x|^p \tag{5}$$

with probability at least $1 - \delta$.

We remark at this point that since the ℓ_p norms are absolutely homogeneous, previous work usually includes the weights by folding them into the data, i.e., $w_i |a_i x|^p$ becomes $|(w^{1/p} a_i)x|^p$. This usually allows to simplify large parts of the technical analysis by reducing to the unweighted case. We explicitly *do not* use this trick for reasons to be discussed later.

With the definition of our main problem in place, we next define the sensitivity measure associated with this loss function and note that they are also called ℓ_p leverage scores in our scope.

Definition 1.1 (ℓ_p-sensitivities/-leverage scores)**.** Let $A \in \mathbb{R}^{n \times d}$, and $1 \leq p < \infty$. We define the i-th ℓ_p-sensitivity $\varsigma_i^{(p)}$ or ℓ_p-leverage score $l_i^{(p)}$ of A to be

$$\varsigma_i^{(p)} = l_i^{(p)} = \sup_{x \in \mathbb{R}^d \setminus \{0\}} \frac{|a_i x|^p}{\|Ax\|_p^p},$$

and the total ℓ_p sensitivity as $\mathfrak{S}^{(p)} \colon = \sum_{i \in [n]} \varsigma_i^{(p)}$.

It can be shown in some settings, that the total sensitivity is a lower bound on the required number of samples (cf. Tolochinsky et al., 2022), and d is a natural lower bound for our problem because this number of samples is required to even preserve the rank of A under subsampling. It is also known that $\mathfrak{S}^{(p)} \leq d$ for all $p \in [1, 2]$ (cf. Woodruff & Yasuda, 2023c). Thus, a natural question to ask is whether it is possible to obtain the guarantee of Equation (5) within roughly $\mathfrak{S}^{(p)} + d = O(d)$ samples.

A plain application of the sensitivity framework requires $\tilde{O}(\varepsilon^{-2}\mathfrak{S}d)$ samples, and a lower bound of $\tilde{\Omega}(d/\varepsilon^2)$ was given by Li et al. (2021) in a broader setting against any data structure that answers ℓ_p subspace queries. The sensitivity sampling upper bound thus seems off by a factor d, and the recent work of Woodruff & Yasuda (2023c) has made significant progress by showing that one can do better for all $p \in (1, 4)$. They obtained near optimal bounds when p becomes close to $p = 2$. In this particular case, a $\tilde{O}(\varepsilon^{-2}\mathfrak{S}) = \tilde{O}(\varepsilon^{-2}d)$ bound was known before to be possible by ℓ_2 leverage score sampling (Mahoney, 2011). However, near the boundary $p = 1$ (and for $p > 4$) the bounds of Woodruff & Yasuda (2023c) are again off by a factor d. In the case $p < 2$, the previous work of Chen & Derezinski (2021) gave better results that were achieved implicitly by relating ℓ_p leverage scores to Lewis weights. Again, these bounds become worse when $1 \leq p \ll 2$ and are off by a $\sqrt{d}$ factor for $p = 1$. See below for a more detailed comparison between these works, and our work.

This brings us to the main open question in this line of research, which is also the central question studied in our paper:

Question 1.2. *Is it possible to achieve a sampling complexity of $\tilde{O}(\varepsilon^{-2}(\mathfrak{S} + d))$ for constructing an ℓ_p subspace embedding, see Equation (5), via sensitivity sampling?*

1.1. Our contribution

Our first contribution is an $\tilde{\Omega}(d^{2-p/2})$ lower bound against *pure* ℓ_p leverage score sampling, i.e., when the sampling probabilities are proportional to $l_i^{(p)}$.

Theorem 1.3 (Informal restatement of Theorem B.1)**.** *There exists a matrix $A \in \mathbb{R}^{m \times 2d}$, for sufficiently large $m \gg 2d$, such that if we sample each row $i \in [n]$ with probability $p_i := \min\{1, kl_i^{(p)}\}$ for some $k \in \mathbb{N}$, then with high probability, the ℓ_p subspace embedding guarantee (see Equation (5)) does not hold unless $k = \Omega(d^{2-p/2}/(\log d)^{p/2})$.*

The proof in the appendix is conducted by constructing a matrix with two parts. Subsampling at least d rows of each part is necessary to even preserve the rank. However, the total ℓ_p sensitivity of one part is significantly larger than

the total sensitivity of the other part, roughly by a factor of $d^{1-p/2}$. This implies that ℓ_p leverage score sampling requires oversampling by that factor to collect the required rows from both parts, which yields the lower bound.

In particular, Theorem 1.3 proves that the upper bounds implied by Chen & Derezinski (2021) for pure ℓ_p leverage score sampling are optimal in the worst case, settling the complexity of pure ℓ_p sampling up to polylogarithmic factors. It also proves that we necessarily need to change the sampling probabilities to get below the lower bound towards optimal linear d dependence. We note that augmentation by uniform sampling does not help, since the size of the high sensitivity part can be increased to give the same imbalance between the number of rows of the two parts as between their sampling probabilities in pure ℓ_p sampling.

Our main result is the following sensitivity sampling bound that is achieved by augmenting ℓ_p with ℓ_2 sensitivities, a new technique that we call "ℓ_2 augmentation". The theorem answers Question 1.2 in the affirmative for all $p \in [1, 2]$, improving over Chen & Derezinski (2021); Woodruff & Yasuda (2023c), and settling the complexity of ℓ_p subspace embeddings via sensitivity sampling in this regime. We leave the case $p > 2$ as an open question, where our work resembles the same bounds as Woodruff & Yasuda (2023c).

Our theorem is generalized to also hold for the ℓ_p variant of the ReLU function where $h(r) = \max\{0, t\}^p$. For this loss function, we also require a data dependent mildness parameter $\mu = \mu(A)$ to be bounded.

Definition 1.4 (μ-complexity, Munteanu et al., 2018; 2022, slightly modified)**.** Given a data set $A \in \mathbb{R}^{n\times d}$ we define the parameter $\mu := \mu(A)$ as

$$\mu(A) = \sup_{\|Ax\|_p=1} \frac{\|Ax\|_p^p}{\|(Ax)^-\|_p^p} = \sup_{\|Ax\|_p=1} \frac{\|Ax\|_p^p}{\|(Ax)^+\|_p^p},$$

where $(Ax)^+$, and $(Ax)^-$ are the vectors comprising only the positive resp. negative entries of Ax and all others set to 0.[1]

For the main problem where $h(r) = |t|^p$, we can remove this parameter or equivalently set $\mu = 1$.

Theorem 1.5 (Informal restatement of Theorem G.1)**.** *Let $A \in \mathbb{R}^{n\times d}$. Let $\varepsilon, \delta \in (0, 3/10)$, $p \in [1, 2]$. Consider $f(Ax) = \sum_{i=1}^n h(a_i x)$ for $h(r) = |r|^p$, where we set $\mu = 1$, or $h(r) = \max\{0, r\}^p$, in which case $\mu = \mu(A)$, see Definition 1.4. Set $\alpha = O\left(\frac{(\log(d\mu\log(1/\delta)/\varepsilon)\log^2 d + \log(1/\delta)}{\varepsilon^2}\right)$. Let $S \subset [n]$ be a sample of size $m = O(d\mu\alpha)$*

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

where the probability for any $i \in [n]$ satisfies

$$p_i := \mathbf{Pr}(i \in S) \geq \min\left\{1, \alpha\left(\mu l_i^{(p)} + l_i^{(2)} + \frac{1}{n}\right)\right\},$$

and the corresponding weight is set to $w_i = 1/p_i$. Then with probability at least $1 - \delta$ it holds for all $x \in \mathbb{R}^d$ simultaneously that

$$\left|\sum_{i\in S} w_i h(a_i x) - \sum_{i\in[n]} h(a_i x)\right| \leq \varepsilon \sum_{i\in[n]} h(a_i x).$$

As an application, we extend our main result to hold for the logistic loss function $h(t) = \ln(1 + \exp(t))$ as well. This improves over the previous $\tilde{O}(\varepsilon^{-2}\mu^2 d)$ bound of Mai et al. (2021) which was based on Lewis weights subsampling.

Theorem 1.6 (Informal restatement of Theorem H.4)**.** *Let $\varepsilon, \delta \in (0, 3/10)$. Let $A \in \mathbb{R}^{n\times d}$ with $\mu = \mu(A)$, see Definition 1.4. Consider $f(Ax) = \sum_{i=1}^n h(a_i x)$ with the logistic loss $h(t) = \ln(1+\exp(t))$. Further assume that we sample with probabilities $p_i \geq \min\{1, \alpha(\mu l_i^{(1)} + l_i^{(2)} + \frac{\mu d}{n})\}$ for $\alpha \geq O((\log^3(d\mu\log(\delta^{-1})/\varepsilon) + \ln(\delta^{-1}))/\varepsilon^2)$, where the number of samples is*

$$m = O\left(\frac{d\mu}{\varepsilon^2}\left(\log^3(d\mu\log(\delta^{-1})/\varepsilon) + \log(\delta^{-1})\right)\right).$$

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

$$\left|\sum_{i\in S} w_i h(a_i x) - \sum_{i\in[n]} h(a_i x)\right| \leq \varepsilon \sum_{i\in[n]} h(a_i x).$$

1.2. Comparison to related work

For ℓ_p subspace embeddings, the total sensitivity is bounded by d, for $p \in [1, 2]$, and by $d^{p/2}$ for $p \in (2, \infty)$, (cf. Woodruff & Yasuda, 2023c). It is known that using so-called ℓ_p Lewis weights, we can subsample a nearly optimal amount $\tilde{O}(\varepsilon^{-2}d)$ respectively $\tilde{O}(\varepsilon^{-2}d^{p/2})$ of rows to obtain the ℓ_p subspace embedding guarantee of Equation (5), (see Cohen & Peng, 2015; Woodruff & Yasuda, 2023a). Recent work of Jambulapati et al. (2023) recovers matching bounds via a novel sampling distribution, and for a broad array of semi-norms. On the other hand, using ℓ_p sensitivity sampling in a plain application of the sensitivity framework requires $\tilde{O}(\varepsilon^{-2}\mathfrak{S}^{(p)}d)$, which is off by a factor d in the worst case for any value of $p \in [1, \infty)$.

Recently improved sensitivity sampling bounds of Woodruff & Yasuda (2023c) are $\tilde{O}(\varepsilon^{-2}\mathfrak{S}^{2/p})$ for $p \in [1, 2]$ and $\tilde{O}(\varepsilon^{-2}\mathfrak{S}^{2-2/p})$ for $p \in [2, \infty)$. These bounds are much better than the standard bounds when $p \in (1, 4)$ is close

[1] To see that the suprema are equal, one can flip signs: for any x the suprema range over, we have that $\frac{\|Ax\|_p^p}{\|(Ax)^+\|_p^p} = \frac{\|A(-x)\|_p^p}{\|(A(-x))^-\|_p^p}$.

to the case $p = 2$, but they deteriorate towards $p = 1$ and $p = 4$, where the gap is again a factor of d. For $p > 4$ their worst case bounds are even worse than the plain sensitivity framework. We note that Woodruff & Yasuda (2023c) gave an improved bound for this regime as well, albeit not with a direct sensitivity sampling approach. Instead, they gave an algorithm that recursively 'flattens and samples' heavy rows with respect to their sensitivities for ℓ_2 and ℓ_p.

Although improving over the standard bounds for $p < 2$ as well, the main improvement of Woodruff & Yasuda (2023c) lies in the case $p > 2$. This is because prior to their results, Chen & Derezinski (2021) implicitly showed a bound of $\mathfrak{S}d^{1-p/2}$ by relating ℓ_p sensitivities to Lewis weights up to an additional factor of $d^{1-p/2}$. Oversampling the sensitivity scores by this amount guarantees that the increased scores exceed the Lewis weights, which in turn implies a sampling complexity of $\tilde{O}(\varepsilon^{-2}\mathfrak{S}d^{1-p/2})$ to be sufficient. We show in Theorem 1.3 that their bound is tight up to polylogarithmic factors in the worst case, when $\mathfrak{S} = \Theta(d)$, if we sample according to pure ℓ_p leverage scores.

This leads us to our main result, Theorem 1.5, which samples for any value of $p \in [1, 2]$ a number of $\tilde{O}(\varepsilon^{-2}(\mathfrak{S}^{(p)} + \mathfrak{S}^{(2)})) = \tilde{O}(\varepsilon^{-2}(\mathfrak{S}^{(p)} + d))$ many samples according to a mixture of ℓ_p, and ℓ_2 sensitivities with a uniform distribution $1/n$. This technique relies only on ℓ_p sensitivity sampling and in the worst case, the number of samples amounts to $\tilde{O}(\varepsilon^{-2}d)$ for all $p \in [1, 2]$. We note that this matches up to polylogarithmic factors the optimal complexities obtained by Lewis weights, and by other novel sampling probabilities (Cohen & Peng, 2015; Woodruff & Yasuda, 2023a; Jambulapati et al., 2023). See Figure 1 for a visual comparison of our bounds with previous sensitivity sampling bounds.

In particular, note that our bounds improve the d^2 respectively $d^{3/2}$ dimension dependence of previous bounds for $p = 1$ to linear. This allows an application to logistic regression (Theorem 1.6), obtaining up to polylogarithmic factors a sampling bound of $\tilde{O}(\varepsilon^{-2}d\mu)$. Previously, the linear dependence on d was only known to be possible using ℓ_1 Lewis weights (Mai et al., 2021; Woodruff & Yasuda, 2023a). In fact our result even improves over their $\tilde{O}(\varepsilon^{-2}d\mu^2)$ bound by a factor of μ, a complexity parameter introduced by Munteanu et al. (2018; 2022) for compressing data in the scope of logistic regression and other asymmetric functions. We note that linear μ and (near-)linear d dependence was recently achieved in the sketching regime, though at the cost of constant approximation factors (Munteanu et al., 2023). We remark that our polylogarithmic dependencies hidden in the $\tilde{O}$ notation are only $\mathrm{polylog}(\mu, d, \varepsilon^{-1}, \delta^{-1})$ and do not depend on n, which is also a minor improvement compared to almost all mentioned previous works.

Our paper assumes that we have access to ℓ_p sensitivity scores, without giving details on how to compute or approximate them. We refer to (Dasgupta et al., 2009; Woodruff & Zhang, 2013; Clarkson et al., 2016; Munteanu et al., 2022) for classic techniques such as ellipsoidal rounding and well-conditioned bases, as well as to recent advances in constructing improved well-conditioned bases (Wang & Woodruff, 2022), novel ℓ_p spanning sets (Woodruff & Yasuda, 2023b; Bhaskara et al., 2023), or direct ℓ_p sensitivity approximations (Padmanabhan et al., 2023). We also refer to (Mahabadi et al., 2020; Munteanu & Omlor, 2024) for ℓ_p sensitivity sampling in data streams.

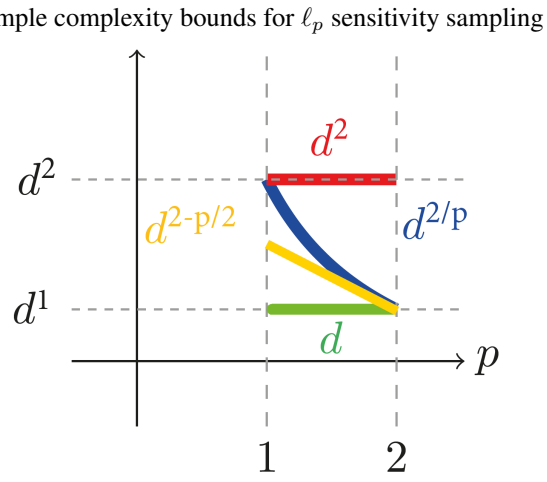

Figure 1. Leading dependence on d for ℓ_p sensitivity sampling for $p \in [1, 2]$ in the worst case, i.e., when $\mathfrak{S}^{(p)} = d$. The horizontal axis represents p. The vertical axis indicates the exponent on d in the respective sample complexity results. The red line indicates the standard bounds obtained from a plain application of the sensitivity framework (Feldman et al., 2020), blue indicates the result of Woodruff & Yasuda (2023c), yellow indicates the result of Chen & Derezinski (2021), and green indicates our new main result.

One might argue that the sensitivity sampling approach is not very interesting for $p \in [1, 2]$, since Lewis weights, among others, already obtain optimal bounds in this regime. However, leverage scores are usually simpler to calculate or to approximate. For instance Cohen & Peng (2015); Mai et al. (2021) calculate an approximation to Lewis weights by recursively reweighting the data and computing ℓ_2 leverage scores $O(\log\log(n))$ times over and over again. While the factor $O(\log\log(n))$ overhead seems minor from a theoretical perspective, this slows down computations by a non-negligible amount. We refer to (Mai et al., 2021; Munteanu et al., 2022) for details, where computational issues have been discussed and demonstrated in experiments.

The experiments in (Mai et al., 2021; Munteanu et al., 2022) also suggest that sensitivity sampling works much better than indicated by upper bounds, sometimes even better than Lewis weights. It is thus very important to find a theoretical explanation for the success of sensitivity sampling and to find out whether they also achieve the optimal complexity

or if there are lower bounds preventing them from achieving optimality. This is the motivation behind our work.

We would like to mention that very similar findings have been observed in the center-based clustering regime, where group sampling was known to produce subsamples of optimal size (Cohen-Addad et al., 2021; 2022; Huang et al., 2022). But group sampling was often outperformed by the conceptually and computationally simpler sensitivity sampling approach on practical and on hard instances (Schwiegelshohn & Sheikh-Omar, 2022). Independently of our work, this was explained in (Bansal et al., 2024) by proving that sensitivity sampling also achieves optimal subsample size for k-means and k-median clustering.

Further, ℓ_p sensitivity sampling has already been studied for a plethora of problems such as ℓ_p regression (Dasgupta et al., 2009), M-estimators (Clarkson & Woodruff, 2015a;b), near-convex functions (Tukan et al., 2020), logistic regression (Munteanu et al., 2018), other generalized linear models such as probit regression (Munteanu et al., 2022), Poisson and graphical models (Molina et al., 2018), IRT models in educational sciences and psychometrics (Frick et al., 2024). Also, some seemingly more distant works have strong connections to sensitivity sampling, such as graph sparsification using effective resistances (Spielman & Srivastava, 2011). Our new optimal bounds for all $p \in [1, 2]$ cover the most common regime encountered in all of these works and will thus be useful in obtaining improved bounds for a broad array of applications as well.

2. New sensitivity subsampling bounds

Our analysis uses several results of Woodruff & Yasuda (2023c), and our main argumentation follows their general outline. Since our analysis is in wide parts a strict generalization, the worst possible outcome of our investigations would simply resemble their exact same bounds. As we have indicated previously, this is actually the case for $p > 2$ in which our techniques do not improve over their bounds. The corresponding part of the analysis is therefore not contained in our paper and we focus on the $p \in [1, 2]$ regime in the remainder.

One main technical argument in this regime is that by monotonicity of maximal ℓ_p sensitivities, the largest ℓ_2 leverage score upper bounds the largest ℓ_p leverage score. To leverage this fact, previous analyses relied on an auxiliary subspace embedding for obtaining a constant factor subspace embedding that required $\mathrm{poly}(d)$ overhead.

In our analysis, we bypass this problem by adding the ℓ_2 leverage scores to the sensitivity upper bound that defines our sampling probabilities. Intuitively, this allows us to obtain the subspace embedding guarantee for ℓ_2 and ℓ_p simultaneously: the ℓ_2 subspace embedding is known to hold already for small sample size $\tilde{O}(\varepsilon^{-2}d)$. Taking it from here, it enables the ℓ_p subspace embedding to work with little more samples. Fortunately, this overhead is negligible compared to the small sample already taken, and also smaller compared to the previous bounds to obtain the ℓ_p subspace embedding directly. Note that also reversely, the data matrix might have much smaller total ℓ_p sensitivity than d. In this case, augmenting the sample to at least the rank preserving lower bound can be accomplished with the least number of additional samples by ℓ_2 sensitivity sampling.

The main reason why the previous analyses do not admit a simultaneous ℓ_2 and ℓ_p subspace embedding, is that they tend to fold the weights into the data (or into the sampling matrix), rather than keeping the weights separately. This is a nice trick which simplifies wide parts of previous analyses by reducing the weighted case to the unweighted case. However, it prevents from our goals as we show in the following simple, yet instructive example for simultaneous ℓ_2 and ℓ_1 embedding, which requires to store the weights separately:

Take $A \in \mathbb{R}^{n\times 1}$ to be the matrix consisting of n copies of the row vector 1. Note that for $x = 1$ we have that $\|Ax\|_1 = n$ and $\|Ax\|_2^2 = n$. We wish to construct a subspace embedding, preserving both norms up to a factor of two. To this end, assume that we have a reduced and reweighted matrix $A' \in \mathbb{R}^{m\times 1}$ with $\|A'x\|_1 \geq n/2$. Then we also have that $\|A'x\|_2^2 \geq \|A'x\|_1^2/m \geq n^2/(4m)$. Now, we also require that $\|A'x\|_2^2 \leq 2n$. Combining both inequalities implies that $2n \geq n^2/(4m)$ which is equivalent to $m \geq n/8$. We conclude that any subspace embedding without auxiliary weights that preserves both, the ℓ_1, and the ℓ_2 norm up to a factor of 2 requires at least $\Omega(n)$ rows.

In stark contrast to this impossibility result, using probabilities $p_i \geq l_i^{(1)} + l_i^{(2)}$, standard sensitivity sampling allows us to take $\tilde{O}(\varepsilon^{-2}d^2)$ samples S and reweight them by w such that $\sum_{i\in S} w_i|a_ix| = (1 \pm \varepsilon)\|Ax\|_1$, and $\sum_{i\in S} w_i|a_ix|^2 = (1 \pm \varepsilon)\|Ax\|_2^2$ hold simultaneously.

To leverage this fact and improve the sampling complexity to linear, we need to open up and generalize large parts of the work of Woodruff & Yasuda (2023c) to deal with weighted ℓ_p norms, which we define as follows:

Definition 2.1. Given a vector $v \in \mathbb{R}^n$, $p \in [1, \infty)$ and weights $w \in \mathbb{R}^n_{\geq 1}$, we let

$$\|v\|_{w,p} := \left(\sum_{i=1}^{n} w_i|v_i|^p\right)^{1/p}.$$

For the ℓ_∞ norm, we also let

$$\|v\|_{w,\infty} := \lim_{p\to\infty} \|v\|_{w,p} = \max_{i\in[n]} |v_i|.$$

and

$$\|v\|_{w,\infty,p} := \max_{i\in[n]} |w_i^{1/p} v_i|$$

2.1. Outline of the analysis

The full formal details can be found in the appendix. Here we provide an outline for the proofs of our main results. We note that some definitions or notation might slightly differ from the appendix for the sake of a clean and concise presentation. The proof consists of several main steps for which we give high level descriptions:

2.1.1. Bounding by a Gaussian process

The first step is to bound the approximation error by a Gaussian process. Note that the subspace embedding guarantee of Equation (3) will be achieved when

$$\Lambda := \sup_{x\in\mathbb{R}^d} |f(Ax) - f_w(SAx)| \leq \varepsilon f(Ax),$$

where $f_w(Ax) = \sum_{i=1}^n w_i h(a_i x)$, and $S \in \{0,1\}^{m\times n}$ is a sampling matrix with exactly one non-zero entry in each row such that SA consists of m rows out of n from the data matrix A. By homogeneity of the loss functions, we can restrict the analysis to the case that $f(Ax) = 1$, and our goal is to bound the above term by ε.

To this end, we bound higher moments of the expected error using a standard symmetrization and Gaussianization argument by

$$\mathop{\mathbf{E}}_{S} |\Lambda|^l \leq (2\pi)^{l/2} \mathop{\mathbf{E}}_{S,g} \sup_{f(Ax)=1} \left| \sum_{i\in S} g_i w_i h(a_i x) \right|^l,$$

where the expectation is taken over the random subsample, represented by S, and on the right hand side additionally over i.i.d. standard Gaussians $g_i \sim N(0,1)$.

The sum on the right hand side is a Gaussian process that induces a pseudo metric, such that for all $y = SAx, y' = SAx'$, we have $d_X(y,y') :=$

$$\left(\mathop{\mathbf{E}}_{g\sim\mathcal{N}(0,I_m)} \left(\sum_{i\in S} g_i w_i h(y_i) - \sum_{i\in S} g_i w_i h(y_i')) \right)^2 \right)^{1/2}$$

acting in the reduced and reweighted space.

For bounding the Gaussian process, we use a slightly adapted moment bound of Woodruff & Yasuda (2023c)

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

which follows from a tail bound due to Sudakov, which is sometimes attributed to Dudley. See (Dudley, 2016; Ledoux & Talagrand, 1991) for bibliographical discussion and references.

Here, C is a sufficiently large absolute constant, $\mathcal{E}$ is an upper bound on the entropy of the Gaussian process, and $\mathcal{D}$ is an upper bound on the diameter of the set $T_S := T_S(A) = \{SAx \mid x \in \mathbb{R}^d, f(Ax) = 1\}$ according to the pseudo-metric, i.e., $\sup\{d_X(y,y') \mid y, y' \in T_S(A)\} \leq \mathcal{D}$.

Note, that we can accomplish our goal by using the moment bound for an appropriately large choice of l and applying Markov's inequality. Our remaining task thus reduces to bounding the diameter, and the entropy, and to quantify the required value of l, which will also determine a sufficient subsample size.

2.1.2. Bounding the diameter

We bound the diameter by relating it to the approximation error and to the largest possible coordinate in the reduced and reweighted ℓ_p norm vector among all vectors that satisfy $f(Ax) = 1$. More specifically, let

$$\sigma := \sup_{f(Ax)=1} \|SAx\|_{w,\infty,p}^p$$

Note that σ is similar to the largest ℓ_p leverage score. Also define

$$G := 1 + \sup_{f(Ax)=1} |f_w(SAx) - f(Ax)|.$$

Then we prove that the diameter with respect to the pseudo metric is bounded by

$$\operatorname{diam}(T_S(A)) \leq 4(G\sigma)^{1/2}.$$

We remark here, that this requires special care when treating the p-ReLU function $h(t) = \max\{0,t\}^p$ so not to loose additional μ factors unnecessarily. But in the final step, it can simply be upper bounded by the proper norm function since $\max\{0,t\}^p \leq |t|^p$ for all $t \in \mathbb{R}$. Thus, the same diameter bound applies to both functions without additional μ dependence.

2.1.3. Bounds on covering numbers

Before we can proceed with bounding the entropy of the Gaussian process, we first need to bound the smallest number of (weighted) ℓ_q balls of certain radius t that are required to cover ℓ_p balls, for various values of $p, q \in [1,\infty)$.

To this end, we define balls according to the weighted norms. Given a Matrix $A' \in \mathbb{R}^{m\times d}$, a weight vector $w \in \mathbb{R}^n$ and $q \geq 1$, we set

$$B_w^q := B_w^q(A') = \{x \in \mathbb{R}^d \mid \|A'x\|_{w,q} \leq 1\}.$$

For any $p, q \geq 1$ and $t > 0$, let the covering number $E(B_w^p, B_w^q, t)$ denotes the minimum cardinality of a set N of B_w^q balls of radius t required to cover the unit B_w^p ball. That is, N is chosen such that for any $x \in B_w^p$ there exists $x' \in N$ with $\|x - x'\|_{w,q} \leq t$. This enables chaining arguments to construct a sequence of t-nets at different scales, which can be smaller than using one single ε-net, i.e., one

fixed scale for the entire space. See Nelson (2016) for a survey on chaining techniques and applications.

To bound the covering numbers, we aim at applying a so-called Dual-Sudakov-minoration result (see Bourgain et al., 1989). Let $\|\cdot\|_X$ be a norm, and let $B \subseteq \mathbb{R}^d$ denote the Euclidean unit ball in d dimensions. Then,

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

where M_X denotes the Lévy mean

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

It is well known that the denominator is $\Theta(\sqrt{d})$, therefore the previous bound reduces to

$$\log E(B, \|\cdot\|_X, t) \leq O(1)\frac{(\mathbf{E}_{g\sim\mathcal{N}(0,I_d)}\|g\|_X)^2}{t^2}.$$

A very important step in the proof is thus the following bound, which also required to be reproved to account for the weighted norm for $q \geq 2$.

Consider an orthonormal matrix $A' \in \mathbb{R}^{m\times d}$ and $w \in \mathbb{R}^n_{\geq 1}$. Let $\tau \geq \max_{i=1}^m w_i \left\|e_i^T A'\right\|_2^2$ be the largest weighted ℓ_2 leverage score. Then we have that

$$\mathop{\mathbf{E}}_{g\sim\mathcal{N}(0,I_d)}\|A'g\|_{w,q} \leq m^{1/q}\sqrt{q\cdot\tau}.$$

Here is also where the ℓ_2 leverage scores play a role in our bounds, and will later be used together with the ℓ_p leverage scores in order to balance the diameter bound with the entropy bound.

This bound allows us to control the covering numbers for various ℓ_q norms, including the ℓ_∞ norm for the respective weighted balls. In particular, we obtain bounds on the number of weighted ℓ_∞ balls required to cover weighted ℓ_p balls, by first covering B_w^p using B_w^2 balls and then covering each B_w^2 ball again with B_w^∞ balls. Applying a chaining technique using a telescoping sum over varying scales, yields a bound of roughly

$$\log E(B_w^p, d_X, t) = O(1)\frac{\tau\log m}{t^p}. \tag{6}$$

2.1.4. Bounding the entropy

For bounding the entropy, we need to control the following quantity

$$\int_0^\infty \sqrt{\log E(T_S, d_X, t)}\,\mathrm{d}t.$$

To this end, we first derive the following final covering bounds: one for small t with a logarithmic dependence on t^{-1} and a different bound for larger t with a squared dependence on t^{-1} but lower dependence on d.

$$1)\ \log E(T_S, d_X, t) \leq O(d)\log\left(\frac{Gm}{t}\right),$$
$$2)\ \log E(T_S, d_X, t) \leq O(\log m)\frac{G^2\tau}{t^2}.$$

The first item follows by relating to the unweighted case, where a simple net construction suffices, i.e.,

$$\begin{aligned}\log E(B_w^p, B_w^\infty, t) &\leq \log E(B^p, B^\infty, t)\\ &\leq O(d)\log\left(\frac{Gm}{t}\right).\end{aligned}$$

In fact, this is the only place in our proof where the weighted case can simply be reduced to the unweighted case. The second bound follows by applying the previous Equation (6).

We split the entropy integral at an appropriate point λ into

$$\int_0^\lambda \sqrt{\log E(T_S, d_X, t)}\,\mathrm{d}t + \int_\lambda^{\mathcal{D}} \sqrt{\log E(T_S, d_X, t)}\,\mathrm{d}t$$

where the latter can be cut off at our previous diameter bound $\mathcal{D}$ because when the integral exceeds the diameter, it becomes 0.

The two parts of the integral can now be bounded using the covering bounds 1) respectively 2) from above. That is, for small radii less than λ we use the first bound and for radii larger than λ, we use the the second bound.

Choosing the right value for λ so as to keep both terms appropriately small, we obtain the following entropy bound:

$$\int_0^\infty \sqrt{\log E(T_S, d_X, t)}\mathrm{d}t \leq O(G\tau^{1/2})(\log m)^{1/2}\log\frac{d\sigma}{\tau}$$

Note, in particular the dependence on the weighted largest ℓ_2 leverage score τ which will be crucial to balance the diameter with the entropy bound in the main proof.

2.1.5. Outline of the main proof

We have now worked out all pieces that we need in order to prove our main result given in Theorem 1.5. Again, we refer to the appendix for the full technical details. Here, we present a sketch of the final proof:

Let us begin with the sample size m. This is handled in a standard way by defining an indicator random variable that attains $X_i = 1$ if row i is in the sample and otherwise it attains $X_i = 0$. The expected size equals

$$\mathbf{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n p_i = \alpha\left(1 + \mu\sum_{i=1}^n l_i^{(p)} + \sum_{i=1}^n l_i^{(2)}\right).$$

Since $d \geq 1$, $\sum_{i=1}^n l_i^{(p)} \leq d$, and $\sum_{i=1}^n l_i^{(2)} = d$, we can thus bound the expected size by

$$\alpha d \leq \mathbf{E}\left(\sum_{i=1}^n X_i\right) \leq 3\alpha\mu d.$$

An application of Chernoff's bounds yields in particular that $m \leq 6\alpha d\mu$ holds with probability at least $1-\delta$.

Next, note that by our choice of α it holds that $m \geq \tilde{O}(d + \log(1/\delta))$, and each sample is taken with probability $p_i \geq l_i^{(2)}$. This is sufficient to achieve the ℓ_2 norm subspace embedding up to a factor $1/2$ with probability at least $1-\delta$ (Mahoney, 2011).

It allows us to relate τ to the largest weighted ℓ_2 leverage score of the original matrix, rather than the subsample, i.e.,

$$\tau \leq 4 \max_{i\in[n]} w_i l_i^{(2)}.$$

We assume without loss of generality that $0 < p_i < 1$ and thus noting that $\alpha l_i^{(2)} < p_i < 1$, we have that $w_i l_i^{(2)} = l_i^{(2)}/p_i \leq 1/\alpha$ and thus $\tau \leq 4/\alpha$. A very similar argument yields $\sigma \leq 1/\alpha$. Consequently, we have that

$$\sigma \approx \tau \approx 1/\alpha.$$

Plugging this into our diameter bound, we obtain

$$4(G\sigma)^{1/2} \leq 8(G/\alpha)^{1/2} \leq G\frac{\varepsilon}{2\sqrt{l}} := \mathcal{D},$$

where $l = \Theta(\varepsilon^2\alpha)$ for a suitable constant.

Plugging this into the entropy integral bound, we obtain

$$\int_0^\infty \sqrt{\log E(T_S, d_X, t)}\,\mathrm{d}t \leq G\varepsilon/8 := \mathcal{E}$$

Finally, we found bounds for $\mathcal{D}$, and $\mathcal{E}$ which are suitably balanced and allow us to apply the moment bound of Woodruff & Yasuda (2023c), which yields

$$\mathbf{E}[|\Lambda|^l] \leq (C'\mathcal{E})^l(\mathcal{E}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l \leq C^l\varepsilon^l\delta,$$

for suitably large absolute constants $C' \leq C$.

Using this higher moment bound in an application of Markov's inequality, we get that $|\Lambda| \leq C\varepsilon$ holds with probability at least $1-\delta$, since

$$\mathbf{Pr}(|\Lambda| \geq C\varepsilon) = \mathbf{Pr}(|\Lambda|^l \geq C^l\varepsilon^l) \leq \frac{C^l\varepsilon^l\delta}{C^l\varepsilon^l} = \delta.$$

This concludes the proof by taking a union bound over the three probabilistic events, and rescaling ε and δ respectively.

3. Application to logistic regression

Here we provide an outline and some high level intuition behind the proof of our second result given in Theorem 1.6.

The logistic loss function is given by

$$f(Ax) := \sum_{i=1}^n \ln(1+\exp(a_i x)),$$

so in our previous notation we have to deal with individual loss functions $h(r) = \ln(1+\exp(r))$.

Unfortunately, f does not fully satisfy the assumptions of our main theorem. Therefore, we cannot apply Theorem 1.5 directly. Instead, we observe that f can be rewritten in terms of the coordinate-wise ReLU function and the remainder.

We thus split f into two parts $f(Ax) = f_1(Ax) + f_2(Ax)$ as follows:

$$\begin{aligned} f(Ax) &= \sum_{i=1}^n \ln(1+\exp(a_i x)) \\ &= \sum_{i=1}^n \ln(1+\exp(-|a_i x|)) + \sum_{i=1}^n \max\{0, a_i x\} \\ &= \sum_{i=1}^n h_1(a_i x) + \sum_{i=1}^n h_2(a_i x) \\ &= f_1(Ax) + f_2(Ax). \end{aligned}$$

Using this split, we show that taking a sample S with probabilities $p_i = \min\{1, \alpha(\mu l_i^{(1)} + l_i^{(2)} + \mu d/n))\}$ where $\alpha = O(\frac{\log^3(\mu d \log(\delta^{-1})/\varepsilon) + \ln(\delta^{-1})}{\varepsilon^2})$ preserves the logistic loss function for all $x \in \mathbb{R}^d$ up to a relative error of at most ε. Note, that we oversample only by a linear factor μ.

Indeed, this allows the p-ReLU function $h_2(r) = \max\{0, r\}^p$, with $p = 1$ to be handled by a direct application of our main result, which yields

$$\forall x \in \mathbb{R}^d\colon\ |f_2(Ax) - f_{2,w}(SAx)| \leq \varepsilon f_2(Ax).$$

The remaining part $f_1(Ax)$ is a bounded function, and can be handled by the uniform part of our sample. We note that this can be proven by a simple additive concentration bound, and charging the additive error by the optimal cost.

We take a slight detour using the standard sensitivity framework, which allows us to draw from existing previous work and saves a lot of technicalities. To this end, we restrict the function $h(r)$ to the negative domain to obtain $\tilde{h}\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\tilde{h}(r) = \ln(1+\exp(-r))$.

Now, observe that f_1 can be rewritten as

$$f_1(Ax) = f_1^{(1)}(Ax) + f_1^{(2)}(Ax) - 2\frac{n}{\mu},$$

where

$$f_1^{(1)}(Ax) = \frac{n}{\mu} + \sum_{a_i x \geq 0} \tilde{h}(a_i x)$$

and

$$f_1^{(2)}(Ax) = \frac{n}{\mu} + \sum_{a_i x < 0} \tilde{h}(-a_i x).$$

Next, we observe that since $\tilde{h}(r) \leq \tilde{h}(0) = \ln(2) < 1$, and $f_1^{(1)}, f_1^{(2)} \geq \frac{n}{\mu}$, we have that each sensitivity for both of the two functions is bounded by $\varsigma_i \leq \frac{\mu}{n}$, and the total sensitivity is $O(\mu)$.

Using the strict monotonicity of both functions, we can relate the VC dimension of a set system associated with the two functions to the VC dimension of affine hyperplane classifiers. Using a thresholding and rounding trick, this yields a final VC dimension bound of $O(d \log(\mu/\varepsilon))$. These arguments are standard from recent literature, see Munteanu et al. (2018; 2022) for details.

Putting the VC dimension and sensitivity bounds into the standard sensitivity framework and defining the approximate functions

$$f_{1,w}^{(1)}(SAx) = \frac{n}{\mu} + \sum_{a_i x \geq 0} w_i \tilde{h}(a_i x)$$

and similarly

$$f_{1,w}^{(2)}(SAx) = \frac{n}{\mu} + \sum_{a_i x < 0} w_i \tilde{h}(-a_i x).$$

yields for both $i \in \{1, 2\}$ separately that

$$\forall x \in \mathbb{R}^d \colon \left| f_1^{(i)}(Ax) - f_{1,w}^{(i)}(SAx) \right| \leq \varepsilon f_1^{(i)}(Ax),$$

each with probability at least $1 - \delta$.

Now, by a union bound, and using the triangle inequality we can put all three approximations together, which yields

$$\begin{aligned}
&\left| f(Ax) - f_w(SAx) \right| \\
&= \Bigg| f_1^{(1)}(Ax) + f_1^{(2)}(Ax) - \frac{2n}{\mu} + f_2(Ax) \\
&\qquad - f_{1,w}^{(1)}(SAx) - f_{1,w}^{(2)}(SAx) + \frac{2n}{\mu} - f_{2,w}(SAx) \Bigg| \\
&\leq \left| f_1^{(1)}(Ax) - f_{1,w}^{(1)}(SAx) \right| + \left| f_1^{(2)}(Ax) - f_{1,w}^{(2)}(SAx) \right| \\
&\qquad + \left| f_2(Ax) - f_{2,w}(SAx) \right| \\
&\leq \varepsilon \left(f_1^{(1)}(Ax) + f_1^{(2)}(Ax) + f_2(Ax) \right) \\
&\leq \varepsilon f(Ax) + 2\varepsilon \frac{n}{\mu} \leq 3\varepsilon f(Ax).
\end{aligned}$$

with probability at least $1 - 3\delta$. We conclude the proof by rescaling ε, and δ.

As a final remark, previous work aimed at approximating the coordinate-wise ReLU function $f_2(Ax)$ by bounding the error additively to within an ε fraction of the ℓ_1 norm and then using a rescaled $\varepsilon' = \varepsilon/\mu$ to relate the ℓ_1 norm back to the ReLU function.

Since the dependence on ε is typically quadratic, this approach is prone to a μ^2 factor in the final sample size. Our direct approximation of the ReLU function handled via Theorem 1.5 requires oversampling by only $\mu l_i^{(p)}$, which results in a linear μ dependence for logistic regression as well.

4. Conclusion and open directions

In this paper, we resolve the sample complexity of ℓ_p subspace embedding via ℓ_p sensitivity sampling for all $p \in [1, 2]$.

Specifically, our work establishes new $\tilde{\Omega}(d^{2-p/2})$ lower bounds against pure ℓ_p leverage score sampling, showing that upper bounds implied by previous work of Chen & Derezinski (2021) are tight in the worst case up to polylogarithmic factors.

By generalizing the approach of Woodruff & Yasuda (2023c) to deal with weighted norms and augmenting ℓ_p sampling probabilities with ℓ_2 leverage scores, our work strengthens previous upper bounds of Chen & Derezinski (2021); Woodruff & Yasuda (2023c) for all $p \in [1, 2]$ to linear $\tilde{O}(\varepsilon^{-2}(\mathfrak{S}^{(p)} + d))$ sample complexity, matching known $\tilde{\Omega}(d/\varepsilon^2)$ lower bounds by Li et al. (2021) in the worst case.

In particular, this resolves an open question of Woodruff & Yasuda (2023c) in the affirmative for $p \in [1, 2]$ and brings the conceptually and computationally simple sensitivity subsampling approach into the regime that was previously only known to be possible using Lewis weights (Cohen & Peng, 2015), or other alternatives (Jambulapati et al., 2023).

As an application of our results, we also obtain the first fully linear $\tilde{O}(\varepsilon^{-2}\mu d)$ bound for approximating logistic regression, obtained via a special treatment of the p-generalization of the ReLU function, and improving over a previous $\tilde{O}(\varepsilon^{-2}\mu^2 d)$ bound (Mai et al., 2021) as well as over $\tilde{O}(\varepsilon^{-2}\mu d^c)$ bounds with $c \gg 1$. Our p-generalization of the ReLU function suggests similar improvements for p-generalized probit regression (Munteanu et al., 2022), which we leave as an open problem.

We note that in the case of $p > 2$, our generalization does not yield an improvement over the previous bounds of Woodruff & Yasuda (2023c). Therefore, obtaining $\tilde{O}(\varepsilon^{-2}(\mathfrak{S}^{(p)} + d))$ bounds or any improvement towards that goal remains an important open problem for future research. Further, we hope that our methods will help to improve sensitivity sampling bounds for other, more general loss functions, and distance based loss functions beyond ℓ_p norms.

Acknowledgements

The authors would like to thank the anonymous reviewers for very valuable comments and discussion. This work was supported by the German Research Foundation (DFG), grant MU 4662/2-1 (535889065), and by the Federal Ministry of Education and Research of Germany (BMBF) and the state of North Rhine-Westphalia (MKW.NRW) as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany. Alexander Munteanu was additionally supported by the TU Dortmund - Center for Data Science and Simulation (DoDaS).

Impact Statement

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

References

Bansal, N., Cohen-Addad, V., Prabhu, M., Saulpic, D., and Schwiegelshohn, C. Sensitivity sampling for k-means: Worst case and stability optimal coreset bounds. *CoRR*, abs/2405.01339, 2024.

Bhaskara, A., Mahabadi, S., and Vakilian, A. Tight bounds for volumetric spanners and applications. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

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

Chen, X. and Derezinski, M. Query complexity of least absolute deviation regression via robust uniform convergence. In *Conference on Learning Theory (COLT)*, pp. 1144–1179, 2021.

Clarkson, K. L. and Woodruff, D. P. Sketching for M-estimators: A unified approach to robust regression. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 921–939, 2015a.

Clarkson, K. L. and Woodruff, D. P. Input sparsity and hardness for robust subspace approximation. In Guruswami, V. (ed.), *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 310–329, 2015b.

Clarkson, K. L., Drineas, P., Magdon-Ismail, M., Mahoney, M. W., Meng, X., and Woodruff, D. P. The fast Cauchy transform and faster robust linear regression. *SIAM Journal on Computing*, 45(3):763–810, 2016.

Cohen, M. B. and Peng, R. ℓ_p row sampling by Lewis weights. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC)*, pp. 183–192, 2015.

Cohen-Addad, V., Saulpic, D., and Schwiegelshohn, C. A new coreset framework for clustering. In *53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 169–182, 2021.

Cohen-Addad, V., Larsen, K. G., Saulpic, D., Schwiegelshohn, C., and Sheikh-Omar, O. A. Improved coresets for Euclidean k-means. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

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

Dudley, R. M. V.N. Sudakov's work on expected suprema of Gaussian processes. In Houdré, C., Mason, D. M., Reynaud-Bouret, P., and Rosiński, J. (eds.), *High Dimensional Probability VII*, pp. 37–43. Springer International Publishing, 2016.

Feldman, D. and Langberg, M. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pp. 569–578, 2011.

Feldman, D., Schmidt, M., and Sohler, C. Turning Big Data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. *SIAM J. Comput.*, 49(3): 601–657, 2020.

Frick, S., Krivosija, A., and Munteanu, A. Scalable learning of item response theory models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1234–1242, 2024.

Huang, L., Li, J., and Wu, X. Towards optimal coreset construction for (k,z)-clustering: Breaking the quadratic dependency on k. *CoRR*, abs/2211.11923, 2022.

Jambulapati, A., Lee, J. R., Liu, Y. P., and Sidford, A. Sparsifying sums of norms. In *64th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1953–1962, 2023.

Langberg, M. and Schulman, L. J. Universal ε-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 598–607, 2010.

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

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

Mahabadi, S., Razenshteyn, I. P., Woodruff, D. P., and Zhou, S. Non-adaptive adaptive sampling on turnstile streams. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 1251–1264, 2020.

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

Mai, T., Musco, C., and Rao, A. Coresets for classification - simplified and strengthened. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 11643–11654, 2021.

Molina, A., Munteanu, A., and Kersting, K. Core dependency networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3820–3827, 2018.

Munteanu, A. Coresets and sketches for regression problems on data streams and distributed data. In *Machine Learning under Resource Constraints, Volume 1 - Fundamentals*, pp. 85–98. De Gruyter, Berlin, Boston, 2023.

Munteanu, A. and Omlor, S. Turnstile ℓ_p leverage score sampling with applications. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Munteanu, A. and Schwiegelshohn, C. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intell.*, 32 (1):37–53, 2018.

Munteanu, A., Schwiegelshohn, C., Sohler, C., and Woodruff, D. P. On coresets for logistic regression. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 6562–6571, 2018.

Munteanu, A., Omlor, S., and Woodruff, D. P. Oblivious sketching for logistic regression. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 7861–7871, 2021.

Munteanu, A., Omlor, S., and Peters, C. p-Generalized probit regression and scalable maximum likelihood estimation via sketching and coresets. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2073–2100, 2022.

Munteanu, A., Omlor, S., and Woodruff, D. P. Almost linear constant-factor sketching for ℓ_1 and logistic regression. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Nelson, J. Chaining introduction with some computer science applications. *Bull. EATCS*, 120, 2016.

Padmanabhan, S., Woodruff, D. P., and Zhang, R. Computing approximate ℓ_p sensitivities. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Schwiegelshohn, C. and Sheikh-Omar, O. A. An empirical evaluation of k-means coresets. In *30th Annual European Symposium on Algorithms (ESA)*, pp. 84:1–84:17, 2022.

Spielman, D. A. and Srivastava, N. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011.

Tolochinsky, E., Jubran, I., and Feldman, D. Generic coreset for scalable learning of monotonic kernels: Logistic regression, sigmoid and more. In *International Conference on Machine Learning (ICML)*, pp. 21520–21547, 2022.

Tukan, M., Maalouf, A., and Feldman, D. Coresets for near-convex functions. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Wang, R. and Woodruff, D. P. Tight bounds for ℓ_1 oblivious subspace embeddings. *ACM Trans. Algorithms*, 18(1): 8:1–8:32, 2022.

Woodruff, D. P. and Yasuda, T. Online Lewis weight sampling. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 4622–4666, 2023a.

Woodruff, D. P. and Yasuda, T. New subset selection algorithms for low rank approximation: Offline and online. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 1802–1813, 2023b.

Woodruff, D. P. and Yasuda, T. Sharper bounds for ℓ_p sensitivity sampling. In *International Conference on Machine Learning (ICML)*, pp. 37238–37272, 2023c.

Woodruff, D. P. and Zhang, Q. Subspace embeddings and ℓ_p-regression using exponential random variables. In *The 26th Annual Conference on Learning Theory (COLT)*, pp. 546–567, 2013.

A. Setting

We are given some dataset consisting of n points $a_i \in \mathbb{R}^d$ for $i \in [n]$ where $d \ll n$. We set $A \in \mathbb{R}^{n\times d}$ to be the matrix with rows a_i. Further we are given a possibly weighted function $f_w(v) = \sum_{i=1}^n w_i h(a_i x)$ for some function $h : \mathbb{R} \to \mathbb{R}_{\geq 0}$ that measures the contribution of each point. We drop the subscript if the weights are uniformly $w_i = 1$ for all $i \in [n]$

Specifically, we consider in this paper $h(r) = |r|^p$ and $h(r) = \max\{0, r\}^p$ for $p \in [1, \infty)$. We will also extend the latter case for $p = 1$ to logistic regression $h(r) = \ln(1 + \exp(r))$.

Our goal is to show that when sampling with probability proportional to a mixture of the ℓ_p-leverage scores, the ℓ_2-leverage scores and $1/n$, then with a sample of $m \ll n$ elements we can guarantee with failure probability at most δ that it holds that

$$\sup_{x\in\mathbb{R}^d} |f(Ax) - f_w(SAx)| \leq \varepsilon f(Ax), \tag{7}$$

where $S \in \{0, 1\}^{m\times n}$ is a sampling matrix with exactly one non-zero entry in each row such that SA extracts m rows out of n from the data matrix A, see Definition C.6.

We first prove a lower bound against pure ℓ_p leverage score sampling. The remainder is dedicated to proving our main results, namely optimal upper bounds for ℓ_p sensitivity sampling with ℓ_2 augmentation.

B. Lower bound against pure ℓ_p leverage score sampling

Theorem B.1. *There exists a matrix $A \in \mathbb{R}^{m\times 2d}$, for sufficiently large $m \gg 2d$, such that if we sample each row $i \in [n]$ with probability $p_i := \min\{1, kl_i^{(p)}\}$ for some $k \in \mathbb{N}$, then with high probability, the ℓ_p subspace embedding guarantee (see Equation (5)) does not hold unless $k = \Omega(d^{2-p/2}/(\log d)^{p/2})$.*

Proof. Theorem 1.6 of (Woodruff & Yasuda, 2023c) implies the existence of an $n \times d$ matrix A_1 of full-rank d, where the dimensions $n \geq d$ are sufficiently large and A_1 has total ℓ_p sensitivity $O((d \log d)^{p/2})$.

Let A_2 be an $n' \times d$ matrix, where $n' \geq n$ is divisible by d, that consists of the identity matrix stacked n'/d times. Note, that it has total ℓ_p sensitivity d by construction. We let $m = n + n'$.

We combine the two matrices to get $A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$. Note that by construction, $A \in \mathbb{R}^{m\times 2d}$ has full rank $2d$ and its total ℓ_p sensitivity is $O((d \log d)^{p/2}) + d \geq d$.

Obtaining an ℓ_p subspace embedding for A requires to preserve the rank, which requires at least d rows from each matrix, A_1 and A_2, to be sampled.

Sampling via pure ℓ_p scores as defined in the theorem, the probability to hit a row from A_1 is bounded above by $\kappa(d \log d)^{p/2}/d$ for some absolute constant κ.

Thus, if we sample less than $\frac{d^{2-p/2}}{4\kappa(\log d)^{p/2}}$ rows of A, the expected number of rows that we collect from A_1 is bounded by

$$\frac{d^{2-p/2}}{4\kappa(\log d)^{p/2}} \cdot \frac{\kappa(d \log d)^{p/2}}{d} = d/4.$$

By a standard application of Chernoff's bound, our sample will thus comprise less than $d/2 < d$ rows from A_1 with high probability, implying that the subsample is rank deficient.

This proves an $\Omega(d^{2-p/2}/(\log d)^{p/2})$ lower bound against ℓ_p subspace embeddings via pure ℓ_p leverage score sampling. □

C. Preliminaries

C.1. Covering numbers

Covering numbers are the minimum numbers to cover one shape or body using multiple (possibly overlapping) copies of another shape or body. A prominent example is the minimum number of Euclidean balls of radius $1/2$ to cover the unit radius Euclidean ball.

More generally, if the second body is an ε-ball with respect to a certain metric then the covering number corresponds to the minimum size of an ε-net.

Given two convex bodies $T, K \subseteq \mathbb{R}^d$, we define the covering number $E(T, K)$ by

$$E(T, K) = \inf_{k \in \mathbb{N}} \{\exists X \subseteq \mathbb{R}^d, T \subseteq \bigcup_{x \in X} x + K\}$$

where $x + K = \{y = x + z \mid z \in K\}$ denotes the Minkowski sum.

Further, given a metric d_X and $t \geq 0$, we define the ball of radius t, denoted $B_X(t)$, by

$$B_X(t) = \{x \in \mathbb{R}^d \mid d_X(0, x) \leq t\}.$$

Further we define $E(T, d_X, t)$ to be the minimum cardinality of any t-net of T with respect to d_X, formally

$$E(T, d_X, t) = \inf\{|N| \mid N \subset T, \text{ for any } x \in T \text{ there exists } x' \in N \text{ with } d_X(x, x') \leq t\}$$

If for any $x', x \in T$ it holds that $d_X(x, x') \leq d_{X'}(0, x - x')$ then it holds that $E(T, d_X, t) \leq E(T, B_{X'}(t))$.

Given $\varepsilon > 0$ we say that $N \subset \mathbb{R}^d$ is an ε-net of T with respect to d_X if for any point $x \in T$ there exists a point $y \in N$ such that $d_X(x, y) \leq \varepsilon$. In particular, it holds for any ε-net N that $E(T, d_X, \varepsilon) \leq |N|$.

Dual Sudakov Minoration The following result will help us bounding covering numbers for the case where T is the Euclidean ball.

Definition C.1 (Lévy mean)**.** Let $\|\cdot\|_X$ be a norm. Then, the *Lévy mean of* $\|\cdot\|_X$ is defined as

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

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

Theorem C.2 (Dual Sudakov minoration, Proposition 4.2 of Bourgain et al., 1989)**.** *Let $\|\cdot\|_X$ be a norm, and let $B \subseteq \mathbb{R}^d$ denote the Euclidean unit ball in d dimensions. Then,*

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

C.2. Dudleys theorem

The following moment bound follows from the so-called Dudley's theorem and is one of the main tools we use in our analysis:

Lemma C.3 (Woodruff & Yasuda, 2023c, slightly modified)**.** *Let $(X_t)_{t \in T}$ be a Gaussian process with pseudo-metric $d_X(s, t) := \|X_s - X_t\|_2 = \sqrt{\mathbf{E}(X_s - X_t)^2}$. Further let $\operatorname{diam}(T) := \sup\{d_X(s, t) : s, t \in T\} \leq \mathcal{D}$ be a bound on the diameter of d_X on T and $\int_0^\infty \sqrt{\log E(T, d_X, u)}\, \mathrm{d}u \leq \mathcal{E}$ be a bound on the entropy. Then, there is a constant $C = O(1)$ such that for*

$$\Lambda := \mathbf{E} \sup_{t \in T} X_t$$

it holds that

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

The only modification is that we consider a more general Λ than Woodruff & Yasuda (2023c). However, we stress that their proof remains unchanged.

The moment bound can be used for a suitably large choice of l to obtain low failure probability bounds using Markov's inequality. The moment parameter l has two functions: first, it allows us to remove the dependence on the term $(\mathcal{E}/\mathcal{D})$, more precisely to replace it by 2^l which holds whenever $l \geq \log_2(\mathcal{E}/\mathcal{D})$. Second, by choosing $l \geq \log(\delta^{-1})$ we can bound the failure probability by δ while having only an additive $\log(\delta^{-1})$ in the sample size. The task then reduces to obtaining best possible bounds for the diameter $\mathcal{D}$ and the entropy $\mathcal{E}$ to allow a small sample size.

C.3. Definitions

Let $p \in [1, \infty]$ and $h : \mathbb{R} \setminus \{0\} \to \mathbb{R}_{\geq 0}$ be either the function with $h(y) = |r|^p$ or $h(y) = \max\{r, 0\}^p$. Further given a data set $A \in \mathbb{R}^{n \times d}$ consisting of rows $a_1, \ldots, a_n \in \mathbb{R}^d$ and a weight vector $w \in \mathbb{R}^n_{\geq 1}$ and $x \in \mathbb{R}^d$ we set

$$f_w(Ax) = \sum_{i=1}^{n} w_i h(a_i x).$$

Definition C.4 (μ-complex,Munteanu et al., 2018; 2022, slightly modified)**.** Given a data set $A \in \mathbb{R}^{n \times d}$ we define the parameter μ as

$$\mu(A) = \sup_{\|Ax\|_p = 1} \frac{\|Ax\|_p^p}{\|(Ax)^-\|_p^p} = \sup_{\|Ax\|_p = 1} \frac{\|Ax\|_p^p}{\|(Ax)^+\|_p^p},$$

where $(Ax)^+$, and $(Ax)^-$ are the vectors comprising only the positive resp. negative entries of Ax and all others set to 0.

Definition C.5 (ℓ_p-leverage scores)**.** Given a data set $A \in \mathbb{R}^{n \times d}$ we define the i-th ℓ_p leverage score of A by

$$l_i^{(p)} = \sup_{x \in \mathbb{R} \setminus \{0\}} \frac{|a_i x|^p}{\|Ax\|_p^p}.$$

Definition C.6 (sampling)**.** For $i \in [n]$ let $p_i \in (0, 1]$. Then sampling with probability p_i is the following concept: We sample point a_i with probability p_i and set its weight to $w_i = 1/p_i$. This sampling process can also be described using a matrix $S \in \{0, 1\}^{m \times n}$ where m is the number of sampled elements and $S_{ij} = 1$ if and only if $j = j(i) \in [n]$ is the i-th sample. The weight vector $w \in \mathbb{R}^n_{\geq 1}$ is then defined by $w_i = w_{j(i)}$. We slightly overload the notation and identify with S either the sampling matrix or the set of sampled indices.

Note that, different to most previous work, we do not put the weights into the matrix S.

The following definition extends norms to work with (auxiliary) weights.

Definition C.7 (weighted norms)**.** Given a vector $v \in \mathbb{R}^n$, $p \in [1, \infty)$ and weights $w \in \mathbb{R}^n_{\geq 1}$, we let

$$\|v\|_{w,p} := \left(\sum_{i=1}^{n} w_i |v_i|^p \right)^{1/p}.$$

For the ℓ_∞ norm, we also let

$$\|v\|_{w,\infty} := \lim_{p \to \infty} \|v\|_{w,p} = \max_{i \in [n]} |v_i| \qquad \text{and} \qquad \|v\|_{w,\infty,p} := \max_{i \in [n]} |w_i^{1/p} v_i|.$$

D. Outline of the analysis

The proof consists of several main steps, where the goal is to apply Lemma C.3 to bound the deviation of the weighted subsample from the original function:

1) First, we show that the deviation of the weighted subsample from the original function can be bounded by a Gaussian process.

2) We give a bound on the diameter of the Gaussian process of step 1).

3) To bound the entropy, we first generalize some of the theory presented in (Woodruff & Yasuda, 2023c) to cope with auxiliary weights.

4) Using the results of 2) and 3), we are able to bound the entropy of the Gaussian process of step 1).

5) We then put everything together and proceed by proving the main theorem.

E. Bounding by a Gaussian process

We will analyze the following term:

$$\mathop{\mathbf{E}}_{S} \sup_{f(Ax)=1} |f_w(SAx) - f(Ax)|^\ell \tag{8}$$

for some integer $\ell \geq 1$. Since both functions $h(r) = |r|^p$ and $h(r) = \max\{0, r\}^p$ are absolutely homogeneous, and we are interested in a relative error approximation, it suffices to consider points $x \in \mathbb{R}^d$ with $f(Ax) = 1$. Towards applying Lemma C.3, we first bound above term by a Gaussian process:

Lemma E.1. *For $\ell \geq 1$ it holds that*

$$\mathop{\mathbf{E}}_{S} \sup_{f(Ax)=1} |f_w(SAx) - f(Ax)|^\ell \leq (2\pi)^{\ell/2} \mathop{\mathbf{E}}_{S,g} \sup_{f(Ax)=1} \left| \sum_{i \in S} g_i w_i h(a_i x) \right|^\ell$$

where g_i are independent standard Gaussians.

Proof. We first note that $|c + \cdot|^\ell$ is a convex function for any $c \in \mathbb{R}$ and the sup over convex functions is again convex. Thus by applying Jensen's inequality twice, we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{S} \sup_{f_w(Ax)=1} |f_w(SAx) - f(Ax)|^\ell &= \mathop{\mathbf{E}}_{S} \sup_{f(Ax)=1} |f_w(SAx) - f(Ax) + 0|^\ell \\
&= \mathop{\mathbf{E}}_{S} \sup_{f(Ax)=1} |f_w(SAx) - f(Ax) + \mathop{\mathbf{E}}_{S'}(f(Ax) - f_{w'}(S'Ax))|^\ell \\
&\leq \mathop{\mathbf{E}}_{S} \sup_{f(Ax)=1} \mathop{\mathbf{E}}_{S'} |f_w(SAx) - f(Ax) + (f(Ax) - f_{w'}(S'Ax))|^\ell \\
&\leq \mathop{\mathbf{E}}_{S,S'} \sup_{f(Ax)=1} |f_w(SAx) - f_{w'}(S'Ax)|^\ell,
\end{aligned}$$

where S', w' are a second realization of our sampling matrix, and the corresponding weights.

The last term is bounded by

$$\begin{aligned}
\mathop{\mathbf{E}}_{S,S'} \sup_{f(Ax)=1} |f_w(SAx) - f_{w'}(S'Ax)|^\ell &= \mathop{\mathbf{E}}_{S,S'} \sup_{f(Ax)=1} \left| \sum_{i \in S} w_i h(a_i x) - \sum_{i \in S'} w_i h(a_i x) \right|^\ell \\
&\leq \mathop{\mathbf{E}}_{S,S',\sigma} \sup_{f(Ax)=1} \left| \sum_{i \in S \cup S'} \sigma_i w_i h(a_i x) \right|^\ell
\end{aligned}$$

where $\sigma_i \in \{-1, 1\}$ are uniform random signs indicating whether i are sampled by S or S'. Terms that are sampled by both cancel and can thus only decrease the expected value. We may thus assume that no index is sampled in both copies S, S'. Further note that two copies of the same process can at most increase the expected value by a factor of two. Thus, we get that

$$\mathop{\mathbf{E}}_{S} \sup_{f(Ax)=1} |f_w(SAx) - f(Ax)|^\ell \leq \mathop{\mathbf{E}}_{S,\sigma} \sup_{f(Ax)=1} \left| 2 \sum_{i \in S} \sigma_i w_i h(a_i x) \right|^\ell = 2^\ell \mathop{\mathbf{E}}_{S,\sigma} \sup_{f(Ax)=1} \left| \sum_{i \in S} \sigma_i w_i h(a_i x) \right|^\ell.$$

Finally, note that

$$2^\ell \mathop{\mathbf{E}}_{S,\sigma} \sup_{f(Ax)=1} \left| \sum_{i \in S} \sigma_i w_i h(a_i x) \right|^\ell \leq (2\pi)^{\ell/2} \mathop{\mathbf{E}}_{S,g} \sup_{f(Ax)=1} \left| \sum_{i \in S} g_i w_i h(a_i x) \right|^\ell,$$

where g_i are independent standard Gaussians, using a comparison between Rademacher and Gaussian variables (see, e.g., Lemma 4.5, and Equation 4.8 of Ledoux & Talagrand, 1991). □

Given a realization S we set $G_S := 1 + \sup_{f(Ax)=1} |f_w(SAx) - f(Ax)|$. Then, we have that $\sup_{f(Ax)=1} |f_w(SAx) - f(Ax)| = G_S - 1$. We thus get the following corollary

Corollary E.2. *For $\ell \geq 1$ it holds that*

$$\mathop{\mathbf{E}}_{S}(G_S - 1)^\ell \leq (2\pi)^{\ell/2} \mathop{\mathbf{E}}_{S,g} \sup_{f(Ax)=1} \left| \sum_{i \in S} g_i w_i h(a_i x) \right|^\ell .$$

Our goal is now to bound the right hand side of Corollary E.2 using Lemma C.3. To this end, we will dedicate the following sections to bounding the diameter and the entropy of the Gaussian process $(2\pi)^{\ell/2} \mathbf{E}_{S,g} \sup_{f(Ax)=1} |\sum_{i \in S} g_i w_i h(a_i x)|^\ell$.

F. Analyzing the Gaussian pseudo metric

In the following sections we consider a fixed realization S. We set

$$G := G_S = 1 + \sup_{f(Ax)=1} |f_w(SAx) - f(Ax)| \geq \sup_{f(Ax)=1} f_w(SAx) \tag{9}$$

and for $y, y' \in \mathbb{R}^m$ we define

$$d_X(y, y') = \left(\mathop{\mathbf{E}}_{g \sim \mathcal{N}(0, I_m)} \left(\sum_{i \in S} g_i w_i h(y_i) - \sum_{i \in S} g_i w_i h(y'_i)) \right)^2 \right)^{1/2} .$$

$$T_A = \{x \in \mathbb{R}^d \mid f(Ax) = 1\} \subseteq \{x \in \mathbb{R}^d \mid \|Ax\|_p^p \leq \mu\}$$

and $\sigma = \max_i w_i l_i^{(p)}$. Further we set $T_S = SA(T_A) = \{y \in \mathbb{R}^m \mid y = SAx, x \in T_A\}$

F.1. Bounding the diameter

Before bounding the diameter we prove the following lemma:

Lemma F.1. *If $p \in (0, 2]$ for any $r, r' \in \mathbb{R}$ holds that*

$$||r|^{p/2} - |r'|^{p/2}| \leq |r - r'|^{p/2}.$$

and

$$|\max\{0, r\}^{p/2} - \max\{0, r'\}^{p/2}| \leq |\max\{0, r\} - \max\{0, r'\}|^{p/2}.$$

Proof. Let $0 \leq a \leq b$. For $p \in (0, 2]$ note that $p/2 - 1 \leq 0$, and in the following calculation $t \geq t - a$. We thus have that

$$b^{p/2} - a^{p/2} = \int_a^b (p/2) t^{p/2-1} \, dt \leq \int_a^b (p/2)(t-a)^{p/2-1} \, dt = \int_0^{b-a} (p/2) t^{p/2-1} \, dt = (b-a)^{p/2}.$$

Assuming without loss of generality that $|r| \geq |r'|$, we can apply this followed by a triangle inequality to get

$$|r|^{p/2} - |r'|^{p/2} \leq (|r| - |r'|)^{p/2} \leq |r - r'|^{p/2}.$$

Similarly, if $r, r' \geq 0$, we have that

$$|\max\{0, r\}^{p/2} - \max\{0, r'\}^{p/2}| = |r|^{p/2} - |r'|^{p/2} \leq |r - r'|^{p/2} = |\max\{0, r\} - \max\{0, r'\}|^{p/2}$$

or, if $r < 0$ or $r' < 0$ holds, then

$$|\max\{0, r\}^{p/2} - \max\{0, r'\}^{p/2}| = \max\{0, r, r'\}^{p/2} \leq |\max\{0, r\} - \max\{0, r'\}|^{p/2}.$$

□

The following lemma bounds d_X by the weighted infinity norm for $h(r) = |r|^p$. It allows us to deduce a bound on the diameter and we will later need it to bound the entropy as well.

Lemma F.2. *Let $p \in [1,2]$ and let $h(r) = |r|^p$. For any $y, y' \in \mathbb{R}^m$ it holds that*

$$d_X(y,y') \leq (2(f_w(y) + f_w(y'))\|y - y'\|^p_{w,\infty,p})^{1/2}. \tag{10}$$

Further we have that

$$d_X(y,0) \leq (f_w(y)\|y\|^p_{w,\infty,p})^{1/2}. \tag{11}$$

Proof. First note that since $\mathbf{E}(g_i g_j) = \mathbf{1}_{i=j}$ we have that

$$d_X(y,y')^2 = \sum_{i \in S} w_i^2 (h(y_i) - h(y_i'))^2.$$

Thus we have that

$$\begin{aligned}
d_X(y,y')^2 &= \sum_{i=1}^m w_i^2 (h(y_i) - h(y_i'))^2 \\
&= \sum_{i=1}^m w_i^2 (h(y_i)^{1/2} - h(y_i')^{1/2})^2 (h(y_i)^{1/2} + h(y_i')^{1/2})^2 \\
&\overset{\text{Lemma F.1}}{\leq} \sum_{i=1}^m w_i^2 (|y_i - y_i'|^{p/2})^2 (h(y_i)^{1/2} + h(y_i')^{1/2})^2 \\
&\leq \|y - y'\|^p_{w,\infty,1/p} \cdot \sum_{i=1}^m w_i 2(h(y_i) + h(y_i'))) \\
&\leq 2\|y - y'\|^p_{w,\infty,1/p} (f_w(y) + f_w(y')).
\end{aligned}$$

For the second part of the lemma note that

$$\begin{aligned}
d_X(y,0)^2 &= \sum_{i=1}^m w_i^2 (h(y_i) - h(0))^2 \\
&= \sum_{i=1}^m w_i^2 h(y_i)^2 \leq \sum_{i=1}^m w_i |y_i|^p w_i h(y_j) \leq f_w(y)\|y\|^p_{w,\infty,p}.
\end{aligned}$$

□

We will now move our attention to $h(r) = \max\{0, r\}^p$.

Lemma F.3. *Let $p \in [1,2]$ and let $h(r) = \max\{0,r\}^p$. For any $y, y' \in \mathbb{R}^m$, let $(y)^+ = (\max\{0, y_i\})_{i=1}^m$ denote the vector that contains only the non-negative entries of y and all others are set to 0. Then, it holds that*

$$d_X(y,y') \leq (2(f_w(y) + f_w(y'))\|(y)^+ - (y')^+\|^p_{w,\infty,p})^{1/2}. \tag{12}$$

Further we have that

$$d_X(y,0) \leq (f_w(y)\|(y)^+\|^p_{w,\infty,p})^{1/2}. \tag{13}$$

Proof. First note that since $\mathbf{E}(g_i g_j) = \mathbf{1}_{i=j}$ we have that

$$d_X(y,y')^2 = \sum_{i \in S} w_i^2 (h(y_i) - h(y_i'))^2.$$

Thus we have that

$$d_X(y,y')^2 = \sum_{i=1}^m w_i^2 (h(y_i) - h(y_i'))^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{m} w_i^2 (h(y_i)^{1/2} - h(y_i')^{1/2})^2 (h(y_i)^{1/2} + h(y_i')1/2)^2 \\
&\overset{\text{Lemma F.1}}{\leq} \sum_{i=1}^{m} w_i^2 (|\max\{y_i, 0\} - \max\{y_i', 0\}|^{p/2})^2 (h(y_i)^{1/2} + h(y_i')^{1/2})^2 \\
&\leq \|(y)^+ - (y')^+\|_{w,\infty,1/p}^p \cdot \sum_{i=1}^{m} w_i 2(h(y_i) + h(y_i'))) \\
&\leq 2\|(y)^+ - (y')^+\|_{w,\infty,1/p}^p (f_w(y) + f_w(y')).
\end{aligned}
$$

For the second part of the lemma note that

$$
\begin{aligned}
d_X(y,0)^2 &= \sum_{i=1}^{m} w_i^2 (h(y_i) - h(0))^2 \\
&= \sum_{i=1}^{m} w_i^2 h(y_i)^2 \leq \sum_{i=1}^{m} w_i \max\{0, y_i\}^p w_i h(y_i) \leq f_w(y) \|(y)^+\|_{w,\infty,p}^p.
\end{aligned}
$$

□

We thus conclude the following bound on the diameter.

Lemma F.4. *Let $p \in [1,2]$ It holds that $\operatorname{diam}(T_S) \leq 4(G\sigma)^{1/2}$ where $\sigma = \sup_{f(Ax)=1} \|SAx\|_{w,\infty,p}^p$.*

Proof. Let $y, y' \in T_S$. First note that there exist $x, x' \in T_A$ with $y = SAx$ and $y' = SAx'$. We thus have by Equation (9) that $f_w(y) \leq G$.

Next note that

$$\|(y)^+\|_{w,\infty,p}^p \leq \|y\|_{w,\infty,p}^p \leq \sup_{f(Ax)=1} \|SAx\|_{w,\infty,p}^p = \sigma.$$

Similarly, we have that $f_w(y') \leq G$ and $\|(y')^+\|_{w,\infty,p}^p \leq \sigma$. By Lemma F.2 and using the triangle inequality we have that

$$d_X(y, y') \leq (2(f_w(y) + f_w(y')) \|y - y'\|_{w,\infty,p}^p)^{1/2} \leq (8G\sigma)^{1/2} \leq 4(G\sigma)^{1/2}.$$

□

F.2. Bounds on covering numbers

Next, we aim to bound the entropy. To this end, we first need to bound the log of the covering numbers $\log E(T, d_X, t)$. We will use two bounds, one for small t with a small dependence on t^{-1} and a different bound for larger t with larger dependence on t^{-1} but smaller dependence on d. For bounding the covering numbers we will use an approach similar to Woodruff & Yasuda (2023c). Given a Matrix $A' \in \mathbb{R}^{m \times d}$, a weight vector $w \in \mathbb{R}^n$ and $q > 1$, we set $B_w^q := B_w^q(A') = \{x \in \mathbb{R}^d \mid \|A'x\|_{w,q} \leq 1\}$.

We will start with the following simple lemma that helps to gain a better understanding of covering numbers:

Lemma F.5. *Let $X \subseteq \mathbb{R}^n$ with $0 \in X$ and $r \in \mathbb{R}_{\geq 0}$ and let $rX = \{rx \mid x \in X\}$. Further let $\theta \geq 1$ and d_X be a metric. Then it holds that*

$$\log E(rX, d_X^\theta, t) = \log E(X, r^\theta d_X^\theta, t) = \log E(X, d_X, (t/r)^{1/\theta}).$$

Further, let d_X be a metric with $d_X'(y,0) \geq d_X(y,0)$ for all $y \in \mathbb{R}^d$. Then it holds that

$$\log E(X, d_X, t) \leq \log E(X, d_X', t).$$

Proof. The first equality follows by homogeneity. For the second, notice that

$$\{x \in \mathbb{R}^n \mid r^\theta d_X(0,x)^\theta \leq t\} = \{x \in \mathbb{R}^n \mid d_X(0,x) \leq (t/r)^{1/\theta}\}.$$

For the inequality, observe that

$$\{x \in \mathbb{R}^n \mid d_X(0,x) \le t\} \supseteq \{x \in \mathbb{R}^n \mid d'_X(0,x) \le t\}.$$

Thus, we need more t-balls with respect to norm d'_X sets to cover X than we need t-balls with respect to d_X and consequently $\log E(X, d_X, t) \le \log E(X, d'_X, t)$. □

We will now bound the number of q-balls we need to cover the Euclidean ball for $q \ge 2$.

We note that since $\mathbf{E}_{g\sim\mathcal{N}(0,I_d)}\|g\|_2 = \Theta(\sqrt{d})$, Theorem C.2 implies

$$\log E(B, \|\cdot\|_X, t) \le O(1)\frac{(\mathbf{E}_{g\sim\mathcal{N}(0,I_d)}\|g\|_X)^2}{t^2}.$$

We thus proceed with a bound on the enumerator:

Lemma F.6. *Let $q \ge 2$ and let $A' \in \mathbb{R}^{m\times d}$ and $w \in \mathbb{R}^m_{\ge 1}$. Let $\tau \ge \max_{i=1}^m w_i\left\|e_i^T A'\right\|_2^2$. Then,*

$$\mathop{\mathbf{E}}_{g\sim\mathcal{N}(0,I_d)}\|A'g\|_{w,q} \le m^{1/q}\sqrt{q\cdot\tau}.$$

Proof. We have for each row $a'_i, i \in [n]$ that $a'_i g$ is distributed as a Gaussian random variable with zero mean and standard deviation $\|a'_i\|_2$. By a known bound for their q-th absolute moment, and applying the known upper bound on Stirling's approximation $\Gamma(x+1) \le \sqrt{e\pi x}\left(\frac{x}{e}\right)^x$, we obtain

$$\begin{aligned}
\mathop{\mathbf{E}}_{g\sim\mathcal{N}(0,I_d)} w_i|a'_i g|^q &= w_i \cdot \frac{2^{q/2}\Gamma(\frac{q+1}{2})}{\sqrt{\pi}}\|a'_i\|_2^q \\
&\le w_i \cdot \frac{2^{q/2}}{\sqrt{\pi}} \cdot \sqrt{e\pi\frac{q-1}{2}}\left(\frac{q-1}{2e}\right)^{\frac{q-1}{2}} \cdot \|a'_i\|_2^q \\
&\le w_i \cdot 2^{q/2} \cdot e^{-\frac{q-1}{2}+\frac{1}{2}}\left(\frac{q-1}{2}\right)^{\frac{q-1}{2}+\frac{1}{2}} \cdot \|a'_i\|_2^q \\
&\le w_i \cdot e^{-\frac{q-2}{2}}(q-1)^{\frac{q}{2}} \cdot \|a'_i\|_2^q \\
&\le w_i \cdot q^{q/2} \cdot \|a'_i\|_2^q .
\end{aligned}$$

Then by Jensen's inequality and linearity of expectation, we get

$$\begin{aligned}
\mathop{\mathbf{E}}_{g\sim\mathcal{N}(0,I_d)}\|A'g\|_{w,q} &\le \left(\mathop{\mathbf{E}}_{g\sim\mathcal{N}(0,I_d)}\|A'g\|_{w,q}^q\right)^{1/q} \\
&= \left(\sum_{i=1}^m w_i q^{q/2} \cdot \|a'_i\|_2^q\right)^{1/q} \\
&\le \left(m \cdot q^{q/2} \cdot \max_{i=1}^m w_i\|a'_i\|_2^q\right)^{1/q} \\
&\le m^{1/q} \cdot q^{1/2} \cdot \max_{i=1}^m w_i^{1/q}\|a'_i\|_2 \\
&= m^{1/q}\sqrt{q\cdot\tau} .
\end{aligned}$$

□

By combining the above calculation with Theorem C.2, we obtain the following bound.

Corollary F.7. *Let $2 \le q < \infty$ and let $A' \in \mathbb{R}^{m\times d}$ be orthonormal with respect to the weighted ℓ_2 norm. Let $\tau \ge \max_{i=1}^m w_i\left\|e_i^T A'\right\|_2^2$. Then,*

$$\log E(B_w^2(A'), \|.\|_{w,q}, t) \le O(1)\frac{m^{2/q}q\cdot\tau}{t^2}$$

Proof. Since A' is orthonormal with respect to the weighted ℓ_2 norm, $B_w^2(A') = B_w^2$ is isometric to the Euclidean ball in d dimensions, and thus Theorem C.2 applies. □

We also get a similar result for $q = \infty$. To this end it suffices to apply Corollary F.7 with $q = O(\log n)$.

Corollary F.8. *Let $A' \in \mathbb{R}^{m\times d}$ be orthonormal with respect to the weighted ℓ_2 norm. Let $\tau \geq \max_{i=1}^m w_i \left\|e_i^T A'\right\|_2^2$. Then,*

$$\log E(B_w^2, \|.\|_{w,\infty}, t) \leq O(1)\frac{(\log m)\cdot\tau}{t^2}.$$

Proof. We take a B_w^q cover, for $q = 2\log m$, of $B_w^2(A')$ whose size is bounded in Corollary F.7 by at most

$$C\frac{m^{2/q}q\cdot\tau}{t^2} = C\frac{2e\log(m)\cdot\tau}{t^2} \leq O(1)\frac{\log(m)\cdot\tau}{t^2}$$

for an absolute constant C. Now we replace every q-ball with an ∞-ball with the same center and radius t. Note that for any $y \in \mathbb{R}^m$ we have that $\|y\|_\infty \leq \|y\|_q$. This implies that $B_w^q \subseteq B_w^\infty$ for any fixed radius t. By this subset relation, the set of ∞-balls is a B_w^∞ cover of $B_w^2(A')$ of the same size. □

By interpolation, we can improve the bound in Corollary F.7:

Lemma F.9. *Let $2 < r < \infty$ and let $A' \in \mathbb{R}^{m\times d}$ be orthonormal with respect to the weighted ℓ_2 norm, for weights $w \in \mathbb{R}_{\geq 1}^m$. Let $\tau \geq \max_{i=1}^n w_i \left\|e_i^T A'\right\|_2^2$. Let $t^{-1} \leq \mathrm{poly}(m)$. Then,*

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

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

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

We define a measure $\nu : \mathcal{P}([n]) \to \mathbb{R}_{\geq 0}$ by $\nu(T) = \sum_{i\in T} w_i$ for $T \subseteq [n]$. Then by Hölder's inequality, we have for any $y \in \mathbb{R}^m$ that

$$\begin{aligned}
\|y\|_{w,r} &= \left(\sum_{i=1}^m w_i |y(i)|^{r(1-\theta)}|y(i)|^{r\theta}\right)^{1/r} \\
&= \left(\int_{[n]} |y(i)^{2r}|^{(1-\theta)/2}|y(i)^{rq}|^{\theta/q}\, d\nu(i)\right)^{1/r} \\
&\leq \left(\int_{[n]} |y(i)|^2\, d\nu(i)\right)^{(1-\theta)/2}\left(\int_{[n]} |y(i)|^q\, d\nu(i)\right)^{\theta/q} \\
&= \left(\sum_{i=1}^m w_i|y(i)|^2\right)^{(1-\theta)/2}\left(\sum_{i=1}^m w_i|y(i)|^q\right)^{\theta/q} = \|y\|_{w,2}^{1-\theta}\|y\|_{w,q}^{\theta}
\end{aligned}$$

For any $y \in B_w^2$ we have that $\|y\|_{w,2} \leq 1$ and thus

$$\|y\|_{w,r} \leq \|y\|_{w,2}^{1-\theta}\|y\|_{w,q}^{\theta} \leq \|y\|_{w,q}^{\theta} = \|y\|_{w,q}^{\theta}$$

so by Lemma F.5 and Corollary F.7

$$\log E(B_w^2, B_w^r, t) \leq \log E(B_w^2, \|.\|_{w,q}^{\theta}, t) \leq \log E(B_w^2, B_w^q, t^{1/\theta}) \leq O(1)\frac{m^{2/q}q\cdot\tau}{t^{2/\theta}}.$$

Now, we have

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

so by taking $q = \frac{2r}{r-2}\log m$, we have that $m^{2/q} = \exp(\frac{2\ln(m)}{q}) = O(1)$ and

$$t^{2/\theta} = t^{(\frac{2r}{r-2}\log m - 2)/\log m} = t^{2r/(r-2)}/t^{2/\log(m)} = \Theta(1)(t/2)^{2r/(r-2)}$$

since $t^{-2/\log(m)} = \mathrm{poly}(m)^{2/\log(m)} = \Theta(1)$, so we conclude our claim. □

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

Lemma F.10. *Let $1 \le p < 2$ and let $A' \in \mathbb{R}^{m\times d}$ be orthonormal with respect to the weighted ℓ_2 norm. Let $\tau \ge \max_{i=1}^m w_i \left\|e_i^T A'\right\|_2^2$ and $t \ge 1/\mathrm{poly}(d)$. Then,*

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

Further if $2 - p \le \frac{1}{\ln(d)}$ then we have that

$$\log E(B_w^p, \|.\|_{w,\infty}, t) = O(1)\frac{(\log m)\cdot\tau}{t^p}$$

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

We will first bound $E(B_w^p, B_w^2, t)$ using Lemma F.9. For each $k \ge 0$, let $\mathcal{E}_k \subseteq B_w^p$ be a maximal subset of B_w^p of minimum size such that for every pair of distinct $y, y' \in \mathcal{E}_k$, $\|y - y'\|_{w,2} > 8^k t$, and for $8^k t \ge 1$ we define $\mathcal{E}_k := \{0\}$. Note that

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

Since for any point in B_w^p and thus in particular for any point in $\mathcal{E}_{k+1}$ there exists a point in $\mathcal{E}_k$ by averaging, for each k, there exists $y^{(k)} \in \mathcal{E}_k$ such that if

$$\mathcal{F}_k := \left\{y \in \mathcal{E}_k : \|y - y^{(k)}\|_{w,2} \le 8^{k+1}t\right\},$$

then

$$|\mathcal{F}_k| \ge \frac{|\mathcal{E}_k|}{|\mathcal{E}_{k+1}|} = \frac{E(B_w^p, B_w^2, 8^k t)}{E(B_w^p, B_w^2, 8^{k+1}t)}$$

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

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

Then, $\mathcal{G}_k \subseteq B_w^2$ and since for $y, y' \in \mathcal{F}_k$ it holds that $\|y - y^{(k)}\|_{w,p} \le \|y\|_{w,p} + \|y^{(k)}\|_{w,p} \le 2$ we also have that $\mathcal{G}_k \subseteq B_w^p \cdot 2/8^{k+1}t$. Further since $F_k \subseteq \mathcal{E}_k$ it holds that $\|y - y'\|_{w,2} > 1/8$ for every distinct $y, y' \in \mathcal{G}_k$. Then by Hölder's inequality,

$$\frac{1}{8^2} \le \|y - y'\|_{w,2}^2 \le \|y - y'\|_{w,p}\|y - y'\|_{w,p'} \le (\|y\|_{w,p} + \|y'\|_{w,p})\|y - y'\|_{w,p'} \le \frac{4}{8^{k+1}t}\|y - y'\|_{w,p'}$$

so $\|y - y'\|_{w,p'} \ge 2\cdot 8^{k-2}t$ which implies that the sets $S_y = \{x \in B_2 \mid \|x - y\|_{w,p'} \le 8^{k-2}t\}$ and $S_{y'} = \{x \in B_2 \mid \|x - y'\|_{w,p'} \le 8^{k-2}t\}$ are disjoint for any different $y, y' \in \mathcal{G}_k$. Thus any maximal subset $S \subseteq \mathbb{R}^d$ such that for each distinct $y, y' \in \mathcal{E}_k$, $\|y - y'\|_{w,p'} > 8^{k-2}t$ must have at least one point in S_y for any $y \in \mathcal{G}_k$. Consequently

$$\log E(B_w^2, B_w^{p'}, 8^{k-2}t) \ge \log|\mathcal{G}_k| = \log|\mathcal{F}_k| \ge \log E(B_w^p, B_w^2, 8^k t) - \log E(B_w^p, B_w^2, 8^{k+1}t). \tag{14}$$

Using that

$$\frac{p'}{p'-2} = \frac{(1-1/p)^{-1}}{(1-1/p)^{-1} - 2} = \frac{(1-1/p)^{-1}}{(1-1/p)^{-1}(1 - 2(1-1/p))} = \frac{1}{2/p - 1} = \frac{p}{2-p}$$

and summing over k gives

$$\begin{aligned}
\log E(B_w^p, B_w^2, t) &= \sum_{k\geq 0} \log E(B_w^p, B_w^2, 8^k t) - \log E(B_w^p, B_w^2, 8^{k+1} t) \\
&\leq \sum_{k\geq 0} \log E(B_w^2, B_w^{p'}, 8^{k-2} t) && (14) \\
&\leq O(1) \frac{1}{(t/2)^{2p'/(p'-2)}} \cdot \left(\frac{p'}{p'-2} \log d + \log m \right) \tau && \text{Lemma F.9} \\
&= O(1) \frac{1}{(t/2)^{2p/(2-p)}} \cdot \left(\frac{p}{2-p} \log d + \log m \right) \tau
\end{aligned}$$

where we take $p'/(p'-2) = 1$ for $p' = \infty$. Combining this with Lemma F.5 and Corollary F.8, we now bound

$$\begin{aligned}
\log E(B_w^p, B_w^\infty, t) &\leq \log E(B_w^p, B_w^2, t') + \log E(B_w^2(t'), B_w^\infty, t) \\
&\leq \log E(B_w^p, B_w^2, t') + \log E(B_w^2, B_w^\infty, t/t') \\
&\leq O(1) \frac{1}{(t')^{2p/(2-p)}} \cdot \left(\frac{p}{2-p} \log d + \log m \right) \tau + O(1) \frac{(\log m) \cdot \tau}{(t/t')^2}
\end{aligned}$$

for any $t'^{-1} \leq \mathrm{poly}(m)$. We choose t' satisfying

$$\frac{1}{(t')^{2p/(2-p)}} = \frac{(t')^2}{t^2},$$

which gives

$$t' = t^{2/(2+2p/(2-p))} = t^{2(2-p)/4} = t^{1-p/2} \geq t^{1/2}$$

which implies that $t'^{-1} \leq \mathrm{poly}(m)$. Further we get that

$$(t')^{2p/(2-p)} = \left(t^2\right)^{\frac{2p/(2-p)}{2+2p/(2-p)}} = t^p$$

so we obtain a bound of

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

For the final part of the lemma, first note that $B_p \subseteq B_2$ and thus by Corollary F.8 and using that $t^{-1} = O(d^c)$ we have that

$$\begin{aligned}
\log E(B_w^p, \|.\|_{w,\infty}, t) &\leq \log E(B_w^2, \|.\|_{w,\infty}, t) \\
&\leq O(1) \frac{(\log m) \cdot \tau}{t^2} \\
&= O(1) \frac{(\log m) \cdot \tau}{t^p \cdot t^{2-p}} \\
&= O(\exp(\log(d^c)/(2-p)) \frac{(\log m) \cdot \tau}{t^p}
\end{aligned}$$

Thus if $1/(2-p) = \Omega(\ln(d))$ then we have that $\log E(B_w^p, \|.\|_{w,\infty}, t) = O(1) \frac{(\log m) \cdot \tau}{t^p}$. □

To deal with very small t we will need another lemma. We set $B_1^q(A) = \{x \in \mathbb{R}^d \mid \|Ax\|_q \leq 1\}$, i.e., the case where the weights are uniformly 1.

Lemma F.11. *For any $1 \leq r \leq q$, any weight vector $w \in \mathbb{R}^n_{\geq 1}$ and any $t \in \mathbb{R}_{\geq 1}$ it holds that*

$$E(B_1^r(A), B_1^q(A), t) \geq E(B_w^r(A), B_w^q(A), t).$$

Proof. Assume that for $N \subseteq \mathbb{R}^n$ it holds that for any point in $y \in B_1^r(A)$ there exists a point $x \in N$ such that $\|x-y\|_q \leq t$. Given $x \in \mathbb{R}^n$ we define $x^{(q)} \in \mathbb{R}^n$ by $x_i^{(q)} = \frac{x_i}{w_i^{1/q}}$ and we set $N_w = \{x^{(q)} \mid x \in N\}$. Now let $y' \in B_w^r(A)$. We define $y \in \mathbb{R}^n$ by $y_i = w_i^{1/q} \cdot y_i'$. Recall that $w_i = 1/p_i \geq 1$, since any probabilities satisfy $p_i \leq 1$. We thus have that

$$\|y\|_r^r = \sum_{i=1}^n |y_i|^r = \sum_{i=1}^n w_i^{r/q}|y_i'|^r \leq \sum_{i=1}^n w_i|y_i'|^r \leq 1$$

since $y' \in B_w^r(A)$. Thus $y \in B_1^r(A)$ and there exists $x \in N$ such that $\|x-y\|_q \leq t$. Notice that

$$|y_i' - x_i^{(q)}| = \left|\frac{y_i}{w_i^{1/q}} - \frac{x_i}{w_i^{1/q}}\right| = \frac{|y_i - x_i|}{w_i^{1/q}}.$$

Then we have that

$$\|y' - x^{(q)}\|_{w,q}^q = \sum_{i=1}^n w_i|y_i' - x_i^{(q)}|^q = \sum_{i=1}^n |y_i - x_i|^q \leq t^q$$

and thus it holds that $\|y' - x^{(q)}\|_{w,q} \leq t$ and N_w is a suitable net proving that

$$E(B_1^r(A), B_1^q(A), t) \geq E(B_w^r(A), B_w^q(A), t).$$

□

F.3. Bounding the entropy

Recall the original setting where $A \in \mathbb{R}^{n\times d}$, $T_A = \{x \in \mathbb{R}^d \mid f(Ax) = 1\}$ and $T_S = SA(T_A)$, and $d_X(y,y')^2 = \sum_{i\in S} w_i^2(h(y_i) - h(y_i'))^2$. Further let m be the number of rows of S and thus also of $A' = SA$.

Using the results from previous section we can deduce following bounds for the covering numbers of T_S with respect to d_X:

Lemma F.12. *Let $1 \leq p \leq 2$. Then for any $t \in (0,1]$ it holds that*

$$1)\ \log E(T_S, d_X, t) \leq O(d)\log\left(\frac{Gm}{t}\right),$$

$$2)\ \log E(T_S, d_X, t) \leq O(\log(m)) \cdot \frac{\gamma G^2 \tau}{t^2}.$$

where τ is the maximum weighted ℓ_2-leverage score of (SA, w) and $\gamma = \frac{1}{2-p}$ for $2-p \geq 1/\ln(d)$ and $\gamma = 1$ for $2-p \leq 1/\ln(d)$.

Proof. By Lemma F.2 and Lemma F.3 it holds for all $y, y' \in T_S$ that $d_X(y,y') \leq (2(f_w(y)+f_w(y'))\|y-y'\|_{w,\infty,p}^p)^{1/2}$ for $h(r) = |r|^p$ and $d_X(y,y') \leq (2(f_w(y)+f_w(y'))\|(y)^+ - (y')^+\|_{w,\infty,p}^p)^{1/2}$ for $h(r) = \max\{r,0\}^p$. For any $y, y' \in B_w^p(A')$ we thus have that $d_X(y,y') \leq 2\|y-y'\|_{w,\infty,p}^{p/2}$. We set $S_{w,p}$ to be the matrix we get by replacing each 1 entry at column i with $w_i^{1/p}$. For $h(r) = |r|^p$ it holds that $T_S = B_w^p(A', G) \subseteq B^p(S_{w,2}A, G)$ as

$$B_w^p(A', G) = \left\{y \in \text{range}(A') \mid \sum_i w_i y_i^p \leq 1\right\} \subseteq \left\{y \in \text{range}(A') \mid \sum_i w_i^{1/2} y_i^p \leq 1\right\} = B^p(S_{w,2}A, G)$$

since $w_i \geq 1$. For $h(r) = \max\{r,0\}$ it holds that $\{(y)^+ \mid y \in T_S\} \subseteq B_w^p(A', G) \subseteq B^p(S_{w,2}A, G)$ and by Lemma F.3 it suffices to restrict to $(y)^+$ in this case. Thus, rather than just covering the 1-ball of A', we need to cover the G-ball which is the same as covering the 1-ball with an adjusted $\frac{t}{G}$ instead of t. Thus, we have by Lemma F.5 that

$$\begin{aligned}\log E(T_S, d_X, t) &\leq \log E(B^p(S_{w,2}A), 2\|\cdot\|_{w,\infty,p}^{p/2}, t/G)\\ &= \log E(B^p(S_{w,2}A), 2\|S_{w,2}\cdot\|_{w,\infty}^{p/2}, t/G)\end{aligned}$$

$$= \log E\left(B^p(S_{w,2}A), B^\infty(S_{w,2}A), \left(\frac{t}{2G}\right)^{2/p}\right).$$

$$= \log E\left(B^p(S_{w,2}A), B^\infty(S_{w,2}A), \left(\frac{t}{2G}\right)^{2/p}\right).$$

Next, to prove the claimed inequalities, we combine these bounds with bounds for the covering number $E(B^p(S_{w,2}A), B_w^\infty(S_{w,2}A), t)$.

For the first part of the lemma, it holds that $\log E(B_1^p, B_1^\infty, t) \leq O(d \log \frac{m}{t})$. To see this, take an orthonormal basis U of $S_{w,2}A'$. Then $N = \{\frac{t}{m^{1/p}} U(n_1, \dots n_d)^T \mid \forall i \in [d] \colon n_i \in \mathbb{N}, n_i \leq m\}$ is a t-net of B_1^p with respect to the ℓ_∞ norm. By Lemma F.11 it holds that $\log E(B_w^p, B_w^\infty, t) \leq \log E(B_1^p, B_1^\infty, t)$. Consequently we have that $\log E(T_S, d_X, t) \leq O(d) \log\left(\frac{Gm}{t}\right)$.

The second bound follows immediately by combining above argumentation with Lemma F.10 as $\tau(S_{w,2}A) = \tau(SA, w)$. □

For bounding the entropy, we will slightly adapt the proof of Woodruff & Yasuda (2023c) and use their following lemma.

Lemma F.13 (Woodruff & Yasuda, 2023c)**.** *Let $0 < \lambda \leq 1$. Then,*

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

Finally we are ready to bound the entropy:

Lemma F.14. *Let $1 \leq p < 2$ and let $SA' \in \mathbb{R}^{m \times d}$ be orthonormal with respect to the weighted ℓ_2 norm. Let $\tau \geq \max_{i=1}^n \left\|e_i^T SA'\right\|_2^2$ and let $\sigma = \sup_{i \in S, y \in T_S} w_i |y_i|^p$. Then if $\tau \in \Omega(\mathrm{poly}(1/d))$,*

$$\int_0^\infty \sqrt{\log E(T_S, d_X, t)}\, \mathrm{d}t \leq O(\gamma^{1/2} G \tau^{1/2})(\log m)^{1/2} \log \frac{d\sigma}{\tau}$$

where $\gamma = \frac{1}{2-p}$ for $2 - p \geq 1/\ln(d)$ and $\gamma = 1$ for $2 - p \leq 1/\ln(d)$.

Proof. Note that it suffices to integrate the entropy integral to $\mathrm{diam}(T_S)$ rather than ∞, since $\log E(T_S, d_X, t) = 0$ for $t \geq \mathrm{diam}(T_S)$ and recall that the diameter is at most $4(G\sigma)^{1/2}$ by Lemma F.4, and since $p \geq 1$.

For small radii less than λ for a parameter λ to be chosen, we use the first bound of Lemma F.12, i.e.

$$\log E(T_S, d_X, t) \leq O(d) \log\left(\frac{Gm}{t}\right)$$

so by using Lemma F.13 we get that

$$\begin{aligned}
\int_0^\lambda \sqrt{\log E(T_S, d_X, t)}\, \mathrm{d}t = \int_0^\lambda \sqrt{O(d) \log\left(\frac{Gm}{t}\right)}\, \mathrm{d}t &\leq \lambda\sqrt{O(d)\log(Gm)} + \sqrt{O(d)} \int_0^\lambda \sqrt{\log \frac{1}{t}}\, \mathrm{d}t \\
&\leq \lambda\sqrt{O(d)\log(Gm)} + \sqrt{d}\left(\lambda\sqrt{\log\frac{1}{\lambda}} + \frac{\sqrt{\pi}}{2}\lambda\right) \\
&\leq O(\lambda)\sqrt{d \log \frac{Gm}{\lambda}}
\end{aligned}$$

On the other hand, for radii larger than λ, we use the the second bound of Lemma F.12, which gives

$$\log E(T_S, d_X, t) \leq O(\log m) \cdot \frac{\gamma G^2 \tau}{t^2}$$

so the entropy integral gives a bound of

$$O(1)\left[(\log m)\gamma G^2 \tau\right]^{1/2} \int_\lambda^{4(G\sigma)^{1/2}} \frac{1}{t}\, \mathrm{d}t = O(1)\left[(\log m)\gamma G^2 \tau\right]^{1/2} \log \frac{4(G\sigma)^{1/2}}{\lambda}.$$

We choose $\lambda = \sqrt{G^2\tau/d} = \Omega(\mathrm{poly}(1/d))$, which yields the claimed conclusion. □

G. Proof of the main theorem

Theorem G.1. *Let $3/10 > \varepsilon, \delta > 0$, $p \in [1,2]$ and let $f(Ax) = \sum_{i=1}^n h(a_i x)$ where $h(r) = |r|^p$ or $h(r) = \max\{r, 0\}^p$. If $\alpha = O(\frac{\gamma \log(d\mu \log(\delta^{-1})/\varepsilon) \ln^2(d) + \ln(\delta^{-1})}{\varepsilon^2})$ and for all $i \in [n]$ it holds that $p_i \geq \min\{1, \alpha(\mu l_i^{(p)} + l_i^{(2)} + \frac{1}{n})\}$.*

Then with failure probability at most δ it holds that

$$\forall x \in \mathbb{R}^d \colon |f_w(SAx) - f(Ax)| \leq \varepsilon f(Ax)$$

and the number m of samples is bounded by

$$m = O(d\mu\alpha) = O\left(\frac{d\mu}{\varepsilon^2}\left(\gamma \log(d\mu \log(\delta^{-1})/\varepsilon) \ln^2(d) + \ln(\delta^{-1})\right)\right)$$

where S and w are constructed as in Definition C.6 (this corresponds to sampling point i with probability p_i and setting $w_i = 1/p_i$), $\mu = 1$ if $h(r) = |r|^p$ and $\mu = \mu(A)$ if $h(r) = \max\{r, 0\}^p$ and $\gamma = \frac{1}{2-p}$ for $2 - p \geq 1/\ln(d)$ and $\gamma = 1$ for $2 - p \leq 1/\ln(d)$.

Proof. First, without loss of generality we assume that for any $i \in [n]$ we have $p_i < 1$.

Second, note that since $p_i > 0$ for all i we have for any $x \in \mathbb{R}^d$ that $\mathbf{E}(f_w(SAx)) = f(Ax)$.

Third, without loss of generality we assume that $p_i = \min\{1, \alpha(\mu l_i^{(p)} + l_i^{(2)} + \frac{1}{n})\}$ since increasing p_i can only reduce the failure probability for obtaining the same approximation bound.

By definition we have that $\alpha = O(\frac{l}{\varepsilon^2})$ where $l = O(\gamma \log(d\mu \log(\delta^{-1})/\varepsilon) \ln^2(d) + \ln(\delta^{-1}))$. Assume the constants are large enough.

We want to bound the number of samples m. To this end, we define the random variable $X_i = 1$ if $i \in S$ and $X_i = 0$ otherwise. Note that X_i is a Bernoulli random variable and using that the sum of sensitivities is bounded by d for all $p < 2$ and equal to d for $p = 2$ (cf. Woodruff & Yasuda, 2023c), we get that

$$\mathbf{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n p_i = \alpha\left(1 + \sum_{i=1}^n \mu l_i^{(p)} + \sum_{i=1}^n l_i^{(2)}\right) = \alpha\left(1 + \sum_{i=1}^n \mu l_i^{(p)} + d\right) \leq 3\alpha\mu d,$$

and similarly, it also holds that $\mathbf{E}\left(\sum_{i=1}^n X_i\right) \geq \alpha d$.

An application of Chernoff bounds yields

$$m = \sum_{i=1}^n X_i \leq 2 \cdot \mathbf{E}\left(\sum_{i=1}^n X_i\right) \leq 6\alpha\mu d$$

with failure probability at most $2\exp\left(-\mathbf{E}(\sum_{i=1}^n X_i)/3\right) \leq 2\exp(-\alpha d/3) \leq \delta$.

We proceed with proving the correctness of our claim: by Corollary E.2 we have that

$$\mathop{\mathbf{E}}_{S} \sup_{f(Ax)=1} |f_w(SAx) - f(Ax)|^\ell := \mathop{\mathbf{E}}_{S}(G_S - 1)^\ell \leq (2\pi)^{\ell/2} \mathop{\mathbf{E}}_{S,g} \sup_{f(Ax)=1} \left|\sum_{i \in S} g_i w_i h(a_i x)\right|^\ell.$$

For fixed S we set $\Lambda := \sup_{f(Ax)=1} \left|\sum_{i \in S} g_i w_i h(a_i x)\right|$. We bound this quantity by Lemma C.3 to get

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

In the following we use the results of the previous sections to bound the entropy $\mathcal{E}$ and the diameter $\mathcal{D}$. To this end, we determine the parameters $\tau = \sup_{\|SAx\|_{w,2}=1, i \in S} w_i |a_i x|^2$ and $\sigma = \sup_{f(Ax)=1, i \in [n]} w_i |a_i x|^p$.

Taking $m \geq \tilde{O}(d + \log(1/\delta))$ samples with probability $p_i \geq l_i^{(2)}$ preserves the ℓ_2 norm up to a factor $1/2$ with failure probability at most δ (Mahoney, 2011). We thus have that

$$\tau = \sup_{\|SAx\|_{w,2}=1, i \in S} w_i |a_i x|^2$$

$$
\begin{aligned}
&= \sup_{\|SAx\|_{w,2}=1, i\in S} \frac{w_i|a_i x|^2}{\|SAx\|_{w,2}^2} \\
&\leq \sup_{\|SAx\|_{w,2}=1, i\in S} \frac{4w_i|a_i x|^2}{\|Ax\|_2^2} \\
&\leq \sup_{\|Ax\|_2=1, i\in[n]} \frac{4w_i|a_i x|^2}{\|Ax\|_2^2} = 4\max_{i\in[n]} w_i l_i^{(2)}.
\end{aligned}
$$

Now since $\alpha l_i^{(2)} < p_i < 1$ we have that $w_i l_i^{(2)} = l_i^{(2)}/p_i \leq 1/\alpha$ and thus $\tau \leq 4/\alpha$. Similarly, since $p_i < 1$ we also have that

$$
\sigma = \max_{i\in[n]} \sup_{f(Ax)=1} w_i|a_i x|^p \leq \max_{i\in[n]} \frac{\mu l_i^{(p)}}{p_i} \leq \max_{i\in[n]} \frac{\mu l_i^{(p)}}{\alpha\mu l_i^{(p)}} \leq 1/\alpha.
$$

Thus, by Lemma F.14, using that $\sigma/\tau \leq 1$ and choosing the constants for α large enough, we have that the entropy is bounded by

$$
O(\gamma^{1/2} G\tau^{1/2}(\log m)^{1/2}\log(d)) \leq O(\gamma^{1/2} G(1/\alpha)^{1/2}(\log m)^{1/2}\log(d)) \leq G\varepsilon/8 := \mathcal{E}
$$

Thus, we get by Lemma F.4 a bound on the diameter of

$$
4(G\sigma)^{1/2} \leq 8(G/\alpha)^{1/2} \leq G\frac{\varepsilon}{2\sqrt{l}} := \mathcal{D}.
$$

Consequently, we get that

$$
\begin{aligned}
\mathop{\mathbf{E}}_{g\sim\mathcal{N}(0,I_m)}[|\Lambda|^l] &\leq (2\mathcal{E})^l(\mathcal{E}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l \\
&\leq (G\varepsilon/4)^l(\sqrt{l}/4) + O(G\varepsilon/2)^l \\
&\leq (G\varepsilon/4)^l 2^l + O(G\varepsilon/2)^l \\
&\leq G^l\varepsilon^l\delta
\end{aligned}
$$

Recall that $F_S = \sup_{f(Ax)=1} |f_w(SAx) - f(Ax)|$. Thus, we get in expectation over the sample S that

$$
F_S^l = (G_S - 1)^l \leq 3^l G_S^l \varepsilon^l \delta = 3^l(1 + F_S)^l\varepsilon^l\delta \leq 3^l\varepsilon^l\delta + 3^l F_S^l \varepsilon^l
$$

Rearranging the terms we get that

$$
\mathop{\mathbf{E}}_{S} F_S^l - 3^l\varepsilon^l F_S^l\varepsilon^l \leq 3^l\delta.
$$

Dividing both sides of the inequality by $(1 - 3^l\varepsilon^l)$ and using that $\varepsilon < 3/10$ yields

$$
\mathop{\mathbf{E}}_{S} F_S^l \leq \frac{3^l\varepsilon^l\delta}{(1 - 3^l\varepsilon^l)} \leq 30^l\varepsilon^l\delta.
$$

Using Markov's inequality we get that $F_S^l \leq 30^l\varepsilon^l$ with failure probability at most δ.

$$
\mathbf{Pr}(F_S \geq 30\varepsilon) = \mathbf{Pr}(F_S^l \geq 30^l\varepsilon^l) \leq \frac{30^l\varepsilon^l\delta}{30^l\varepsilon^l} = \delta.
$$

To finish the proof, we recall that we need three events to hold: preserving the ℓ_2 norm, the number of samples is $m = O(\alpha d)$ and $F_S \leq 30\varepsilon$. The total failure probability for these events to hold is at most $\delta + \delta + \delta = 3\delta$ by applying the union bound. Rescaling δ and ε completes the proof. □

H. Application to logistic regression

H.1. Sensitivity framework

We use the standard sensitivity framework (Langberg & Schulman, 2010) to handle a uniform sample. This requires first some terminology for the VC-dimension.

Definition H.1. The range space for a set $\mathcal{F}$ is a pair $\mathfrak{R} = (\mathcal{F}, \text{ranges})$ where ranges is a family of subsets of $\mathcal{F}$. The VC-dimension $\Delta(\mathfrak{R})$ of $\mathfrak{R}$ is the size $|G|$ of the largest subset $G \subseteq \mathcal{F}$ such that G is shattered by ranges, i.e., $|\{G \cap R \mid R \in \text{ranges}\}| = 2^{|G|}$.

Definition H.2. Let $\mathcal{F}$ be a finite set of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}_{\geq 0}$. For every $x \in \mathbb{R}^d$ and $r \in \mathbb{R}_{\geq 0}$, let $\text{range}_{\mathcal{F}}(x, r) = \{f \in \mathcal{F} \mid f(x) \geq r\}$, and $\text{ranges}(\mathcal{F}) = \{\text{range}_{\mathcal{F}}(x, r) \mid x \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\}$, and $\mathfrak{R}_{\mathcal{F}} = (\mathcal{F}, \text{ranges}(\mathcal{F}))$ be the range space induced by $\mathcal{F}$.

Proposition H.3. *(Feldman et al., 2020) Consider a family of functions $\mathcal{F} = \{f_1, \ldots, f_n\}$ mapping from $\mathbb{R}^d$ to $[0, \infty)$ and a vector of weights $u \in \mathbb{R}^n_{>0}$. Let $\varepsilon, \delta \in (0, 1/2)$. Let $s_i \geq \zeta_i$. Let $S = \sum_{i=1}^n s_i \geq \sum_{i=1}^n \zeta_i = Z$. Given s_i one can compute in time $O(|\mathcal{F}|)$ a set $R \subset \mathcal{F}$ of*

$$O\left(\frac{S}{\varepsilon^2}\left(\Delta \ln S + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

weighted functions such that with probability $1 - \delta$, we have for all $x \in \mathbb{R}^d$ simultaneously

$$\left|\sum_{f_i \in \mathcal{F}} u_i f_i(x) - \sum_{f_i \in R} w_i f_i(x)\right| \leq \varepsilon \sum_{f_i \in \mathcal{F}} u_i f_i(x),$$

where each element of R is sampled i.i.d. with probability $p_j = \frac{s_j}{S}$ from $\mathcal{F}$, $w_i = \frac{S u_j}{s_j |R|}$ denotes the weight of a function $f_i \in R$ that corresponds to $f_j \in \mathcal{F}$, and where Δ is an upper bound on the VC-dimension of the range space $\mathfrak{R}_{\mathcal{F}^*}$ induced by $\mathcal{F}^*$ obtained by defining $\mathcal{F}^*$ to be the set of functions $f_j \in \mathcal{F}$, where each function is scaled by $\frac{S u_j}{s_j |R|}$.

H.2. Sensitivity sampling for logistic regression

The logistic loss function is given by

$$f(Ax) = \sum_{i=1}^n h(a_i x)$$

where $h(r) = \ln(1 + \exp(r))$. Since f does not satisfy our assumptions, we cannot apply our main theorem directly. Instead we split f into two parts:

$$\begin{aligned} f(Ax) = \sum_{i=1}^n \ln(1 + \exp(a_i x)) &= \sum_{a_i x < 0} \ln(1 + \exp(a_i x)) \; + \sum_{a_i x \geq 0} \ln(1 + \exp(a_i x)) \\ &= \sum_{a_i x < 0} \ln(1 + \exp(a_i x)) \; + \sum_{a_i x \geq 0} \ln(\exp(a_i x)(\exp(-a_i x) + 1)) \\ &= \sum_{a_i x < 0} \ln(1 + \exp(a_i x)) \; + \sum_{a_i x \geq 0} \ln(\exp(-a_i x) + 1) + a_i x \\ &= \sum_{i=1}^n \ln(1 + \exp(-|a_i x|)) + \sum_{a_i x \geq 0} a_i x. \end{aligned} \tag{15}$$

Using this split we show that sampling with probabilities $p_i = \min\{1, \alpha(\mu l_i^{(1)} + l_i^{(2)} + \mu d/n))\}$ where $\alpha = O(\frac{\log^3(\mu d \log(\delta^{-1})/\varepsilon) + \ln(\delta^{-1})}{\varepsilon^2})$ preserves the logistic loss function for all $x \in \mathbb{R}^d$ up to a relative error of at most ε.

Theorem H.4. *Let $A \in \mathbb{R}^{n \times d}$ be μ-complex some $\infty > \mu \geq 1$ and let $\varepsilon, \delta \in (0, 1/2)$. Further assume that we sample with probabilities $p_i \geq \min\{1, \alpha(\mu l_i^{(1)} + l_i^{(2)} + \frac{\mu d}{n})\}$ for $\alpha \geq O((\log^3(d\mu \log(\delta^{-1})/\varepsilon) + \ln(\delta^{-1}))/\varepsilon^2)$, where the number of samples is*

$$m = O\left(\frac{d\mu}{\varepsilon^2}\left(\log^3(d\mu \log(\delta^{-1})/\varepsilon) + \log(\delta^{-1})\right)\right)$$

Then we have that

$$\forall x \in \mathbb{R}^d \colon |f_w(SAx) - f(Ax)| \leq \varepsilon f(Ax)$$

with failure probability at most δ.

Proof. Note that by applying our main result given in Theorem G.1 for $h(t) = \max\{0, t\}$, we get with failure probability at most $\delta/3$ that

$$\forall x \in \mathbb{R}^d \colon \left| \sum_{a_i x \geq 0, i \in [n]} a_i x - \sum_{a_i x \geq 0, i \in S} w_i a_i x \right| \leq \varepsilon \sum_{a_i x \geq 0, i \in [n]} a_i x.$$

This handles the second term in our split of Equation (15) within our framework.

Further consider for the first sum of Equation (15), the functions $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $h(r) = \ln(1 + \exp(-r))$, $f_1(Ax) = \frac{n}{\mu} + \sum_{a_i x \geq 0} h(a_i x)$ and $f_2(Ax) = \frac{n}{\mu} + \sum_{a_i x < 0} h(-a_i x)$.

Our goal is to apply Proposition H.3 to f_1 and f_2.

We first note that $h(r) \leq h(0) = \ln(2) \leq 1$ and $f_1(Ax), f_2(Ax) \geq \frac{n}{\mu}$ and thus the sensitivities are bounded for both functions by $\zeta_i \leq \frac{\mu}{n}$. The total sensitivity is consequently bounded by $O(\mu)$.

Next observe that the VC dimension of the range spaces of $\mathcal{F}_1 = \{h_z : x \mapsto h(zx)\}$ and $\mathcal{F}_1 = \{h_z : x \mapsto h(-zx)\}$ are bounded by $d + 1$ since h is an increasing function which allows to relate to the VC dimension of affine hyperplane classifiers by a standard argument, (cf. Munteanu et al., 2018; 2021).

Further, applying a thresholding and rounding trick to the sensitivities (Munteanu et al., 2022), the VC dimension of the weighted range spaces are bounded by $O(d \log(\mu/\varepsilon))$.

Since $p_i \geq \frac{\mu}{n}$ and $m \geq \Omega(\frac{\mu}{\varepsilon^2} \cdot (d \ln(\mu/\varepsilon) \ln(\mu) + \ln(1/\delta))$ we have by Proposition H.3 with failure probability at most $\delta/3$

$$|f_{w,1}(SAx) - f_1(Ax)| \leq \varepsilon f_1(Ax).$$

Similarly with failure probability at most $\delta/3$ it holds that

$$|f_{w,2}(SAx) - f_2(Ax)| \leq \varepsilon f_2(Ax).$$

Now combining everything by triangle inequality, and using the union bound, we get with failure probability at most δ that

$$\begin{aligned}
&|f(Ax) - f_w(SAx)| \\
&= \left| \sum_{a_i x \geq 0, i \in [n]} a_i x + f_1(Ax) - \frac{n}{\mu} + f_2(Ax) - \frac{n}{\mu} - \sum_{a_i x \geq 0, i \in S} w_i a_i x - f_{w,1}(SAx) + \frac{n}{\mu} - f_{w,2}(SAx) + \frac{n}{\mu} \right| \\
&\leq \left| \sum_{a_i x \geq 0, i \in [n]} a_i x - \sum_{a_i x \geq 0, i \in S} w_i a_i x \right| + |f_1(Ax) - f_{w,1}(SAx)| + |f_2(Ax) - f_{w,2}(SAx)| \\
&\leq \varepsilon \left(\sum_{a_i x \geq 0, i \in [n]} a_i x + f_1(Ax) + f_2(Ax) \right) \\
&\leq \varepsilon f(Ax) + 2\varepsilon \frac{n}{\mu} \leq 3\varepsilon f(Ax).
\end{aligned}$$

The last inequality follows from the lower bound $f(Ax) \geq \frac{n}{\mu}$ of Munteanu et al. (2021). Rescaling ε concludes the proof. □