Optimal Dimension-Free Sampling for Regularized Classification

Meysam Alishahi
University of Utah
USA

Alexander Munteanu
TU Dortmund
Germany

Simon Omlor
TU Dortmund
Germany

Jeff M. Phillips
University of Utah
USA

Abstract

We prove optimal sampling bounds achieving $(1\pm\varepsilon)$-relative error for a broad class of Lipschitz continuous classification loss functions under various regularization terms. This includes important functions such as logistic and sigmoid loss, hinge loss, and ReLU loss, as prominent and popular representative examples. In particular, we prove k^2/ε^2 upper and lower bounds for $\|\cdot\|_2/k$ regularization, and k/ε^2 upper and lower bounds for $\|\cdot\|_1/k$ regularization. For $\|\cdot\|_2^2/k$ regularization, the sampling complexity depends mainly on a bounded derivative property: if $|g'(x)| \leq g(x)$, and $g(0) > 0$, and g is monotonic or convex, then it admits linear in k sampling complexity; otherwise the general bound is k^2/ε^2. However, if $g(0) = 0$, our results indicate that no dimension-free bounds are possible, and even sublinear bounds are ruled out. All upper bounds are complemented by matching lower bounds up to polylogarithmic terms. Moreover, our work relies conceptually and algorithmically on simple uniform or (squared) norm sampling and hereby improves over recent cubic k^3/ε^2 sensitivity sampling bounds of (Alishahi and Phillips, ICML'24). This is achieved by refined arguments involving higher moment bounds and empirical process analyses to avoid overcounting that appears in the de-facto standard VC-dimension and sensitivity framework.

1 Introduction

Sample complexity of linear classifiers is one of the most central questions in machine learning. It quantifies how many samples (observations) of data are needed to achieve certain guarantees in accuracy. These bounds assume data $(Z, y) \sim \mathcal{P}$ drawn i.i.d. from an unknown distribution $\mathcal{P}$ with $Z = \{z_1, \ldots, z_m\} \subset \mathbb{R}^d$ and labels $y \in \{-1, +1\}^m$. We simplify to row vectors $a_i = y_i z_i \in \mathbb{R}^d$, and say $a \sim \mathcal{P}$. Our ultimate goal is to approximate a function $f(x) = \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \mathcal{R}(x)/k$, with probability $1-\delta$ within a $(1\pm\varepsilon)$-relative approximation, for $\varepsilon, \delta > 0$. Thereby, $g(\cdot)$ is a monotonic, Lipschitz, and often convex loss function, such as hinge, sigmoid, ReLU, or logistic loss, $\mathcal{R}(x)/k$ is a regularization term with $\mathcal{R}(x) \in \{\|x\|_2^2, \|x\|_2, \|x\|_1\}$, and the factor $1/k$ controls the strength of regularization. To facilitate efficiency in empirical risk minimization, we aim to do this with *as few samples* $\{a_i\}_{i=1}^m$ *as possible* from a properly designed distribution $\mathcal{Q}$, and assigning weights $\{w_i\}_{i=1}^m$ to obtain $f_S(x) = \frac{1}{m}\sum_{i=1}^m w_i g(\langle a_i, x\rangle) + \mathcal{R}(x)/k$. The following guarantee implies that the true risk minimizer is approximately preserved, and our goal is to ensure it holds with probability $1-\delta$:

$$\forall x \in \mathbb{R}^d\colon\ |f(x) - f_S(x)| \leq \varepsilon f(x)\,.$$

A special case is uniform sampling, where $\mathcal{Q} = \mathcal{P}$ and all weights are set to $w_i = 1$. Setting $f(x) = f_0(x) + \mathcal{R}(x)/k$, since the regularization appears in both $f(x)$ and its approximation $f_S(x)$, the guarantee is equivalent to

$$\forall x \in \mathbb{R}^d\colon\ \left| f_0(x) - \left(\frac{1}{m}\sum_{i=1}^m w_i g(\langle a_i, x\rangle)\right)\right| \leq \varepsilon f(x), \tag{1}$$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

which will be used prevalently throughout our paper. We now give an overview of classic and modern related work before getting to our contribution.

Classic related work. Classically, for 0-1 loss (which counts the number of misclassifications), let ξ be the Bayes' optimal rate for linear classifiers; which is, for the best choice of linear classifier, the probability a new data point will be misclassified by it. In this setting, the core bounds were established by Vapnik and Chervonenkis (2015) to depend on ε, the tolerated misclassification rate above ξ; on δ, the probability of exceeding that $\xi + \varepsilon$ rate; and on the VC-dimension, which for linear classifiers is $d + 1$ (Kearns and Vazirani, 1994). The tight bound, shown by Li et al. (2001) requires

$$m = \Theta\left(\frac{1}{\varepsilon^2}\Big(d + \log(1/\delta)\Big)\right) \text{ samples.}$$

However, in practice one solves for a linear classifier not in this 0-1 setting, but by a proxy monotonic, often convex, loss function $g(\langle a, x\rangle)$, such as hinge, sigmoid, ReLU, or logistic loss. Partially towards this question Bartlett and Mendelson (2002) developed the use of Rademacher complexity, and then Bartlett et al. (2005) established a bound that instead achieves an *additive ε-approximation of the true loss function* itself (on which gradient descent is run). Their bound also depends on a maximum radius $\|a\|_2 \leq D$, control on the Lipschitz constant L of the loss function g, an upper bound of $\|x\|_2 \leq R$ on the scale of the normal direction used in the loss function (x are the learned parameters). This sort of bound uses

$$m = \Theta\left(\frac{1}{\varepsilon^2}\Big(L^2 D^2 R^2 + \log(1/\delta)\Big)\right) \text{ samples.}$$

The L-Lipschitz property is standard (and basically disallows loss functions from growing too quickly) and a scale parameter D is necessary when sampling uniformly from an unknown distribution, otherwise a single point can affect the results significantly. While $\|x\|_2 \leq R$ leads to a bounded loss function, it is not the way practitioners most commonly formulate and solve linear classification. Moreover, while these bounds are independent of the dimension d, they only provide an additive error. Thus it is not scale-free, and means as gradient descent approaches the minimum, the relative error induced by this analysis increases.

Contemporary related work. More recently, through the perspective of coresets, this problem was revisited by Munteanu et al. (2018); and they achieved a *relative* $(1 \pm \varepsilon)$*-approximation* bound. It also removed the need to bound the radius of data D and scale of parameters with R, but instead introduced a data dependent parameter $\mu(A)$. This $\mu(A)$ roughly captures how separable the data is. They introduced a sampling bound, using VC-dimension arguments, of $m = \tilde{O}(\mu(A)^3 d^3/\varepsilon^4)$, which was improved to quadratic $m = \tilde{O}(\mu(A)^2 d/\varepsilon^2)$ by Mai et al. (2021), and $m = \tilde{O}(\mu(A) d^3/\varepsilon^2)$ by Munteanu et al. (2022), and finally to linear $m = \tilde{O}(\mu(A) d/\varepsilon^2)$ in (Munteanu and Omlor, 2024a). Note that these coreset approaches do not sample uniformly from the unknown distribution, but rather assume there is already a large enough representative sample, that is further reduced to the target size. Moreover, they also do not subsample uniformly from that larger sample. Instead they sample using importance sampling via square-root leverage scores, ℓ_1 Lewis weights, or a mixture of ℓ_1 and ℓ_2 leverage scores, and re-weight the sampled points for their use in the approximation. Also, this reintroduces the dependence on the dimension d, which while natural from VC-dimension theory, may be seen as unusual in the context of modern billion parameter models (Shoeybi et al., 2019; Brown et al., 2020; Narayanan et al., 2021), or whole genome data analysis (Teschke et al., 2024).

Additional work from the coreset perspective, Samadian et al. (2020) and Tolochinsky et al. (2022) both consider a variant that adds a regularization term $\mathcal{R}(x)/k$ for $\mathcal{R}(x) \in \{\|x\|_2^2, \|x\|_2, \|x\|_1\}$. As a result, no bound is required on the norm of x. Instead, it is constrained implicitly as part of the loss function. While Tolochinsky et al. (2022) drew a sample proportional to the leverage scores, Samadian et al. (2020) only need a uniform sample; using a VC-dimension-based argument they require $O(dk \log k/\varepsilon^2)$ samples (assuming $D = 1$). After this, Alishahi and Phillips (2024) revived the Rademacher complexity approach, and this led to dimension-free bounds of $m = O(k^3/\varepsilon^2)$ (for $D = 1$); so the dependence on k has increased, but there is no more dependence on the dimension d. Additional, more loosely related work are deferred to Appendix A.1 due to page limitations.

In this paper we resolve the sample complexity of regularized linear classifiers (up to $\mathrm{polylog}(k)$ factors), i.e. the number of samples that are required to achieve the guarantee given in Equation (1).

To this end, we leverage the weighted sampling scheme of coresets. Our weights are simple, corresponding to a balanced mixture of uniform and norm-sampling, essentially proportional to $1 +$

$\|a\|_2$, when $\int \|a\|_2 \, d\mathcal{P}(a) = B < \infty$. The intuition is that points of high norm tend to dominate the approximation error. The addition of norm-sampling allows us to control their contribution. Further if in addition $\|a\|_2 \leq D$ then uniform i.i.d. sampling from $\mathcal{P}$ provides the same asymptotic upper bounds when the average norm B is replaced by the maximum norm D in the sample complexity.

Our paper will demonstrate the sample complexity story is more intricate than has been revealed up to this point. The tight answer depends on a combination of both the loss function and the regularization term. Our first results apply widely to loss functions g that are only required to be L-Lipschitz and obtain k^2/ε^2 upper bounds. We then improve to linear bounds of k/ε^2 in settings that do not have quadratic lower bounds. For a few combinations of loss and regularizer, we need to restrict further to loss functions with bounded derivative g' (specifically $|g'(r)| < g(r)$ for all $r \in \mathbb{R}$), in cases where this is necessary to improve quadratic upper bounds to k/ε^2. As applications of these bounds and for proving lower bounds, we choose more specific examples of highly popular loss functions for regression and classification tasks that are also widely studied in the aforementioned previous work:

- logistic loss: $g(r) = \ln(1 + \exp(-r))$,
- sigmoid loss: $g(r) = 1/(1 + \exp(r))$,
- hinge (or SVM) loss: $g(r) = \max\{0, 1 - r\}$,
- ReLU loss: $g(r) = \max\{0, -r\}$.

We note that all these functions are monotonic and all except sigmoid loss are convex.

Our sampling complexity bounds (which we state formally in Section 3) will also depend on the choice of regularization, and we consider $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_2^2, \|\cdot\|_1\}$, which correspond to the arguably most popular norm-, ridge-, and LASSO-regularization. In each case except for ReLU with $\mathcal{R}(x) = \|x\|_2^2$, we show the bound must be either $m = \tilde{\Theta}(k^2/\varepsilon^2)$ or $m = \tilde{\Theta}(k/\varepsilon^2)$. Ultimately our bounds will be tight up to poly-logarithmic factors in k for each combination. For ReLU with $\mathcal{R}(x) = \|x\|_2^2$, we show that surprisingly $\Omega(n \log n)$ samples are required, i.e., no dimension-free, nor sublinear sample size are possible.

2 Sampling, algorithms, and applications

The algorithm is a general blueprint of a coreset algorithm (Munteanu and Schwiegelshohn, 2018):

1. compute scores s(a) for every data point $a \in P$.
2. sample m points each with probability $p(a) = \frac{s(a)}{\sum_{a \in P} s(a)} = \frac{s(a)}{S}$.
3. return the sampled points with corresponding weights $w(a) = \frac{1}{m \cdot p(a)}$.

Sampling distribution. The sampling distribution we analyze is proportional to a score function $s(a)$, with $S = \int s(a) \, d\mathcal{P}$. We then adjust the results via importance or rejection sampling. Specifically, we sample in our paper from a mixture of this distribution with an equal proportion of uniform sampling. More formally, our technical derivations focus on the importance sampling version where we set $d\mathcal{Q} = (\frac{s}{2S} + \frac{1}{2}) \, d\mathcal{P} = (\frac{s+S}{2S}) \, d\mathcal{P}$ and the weights become $w_i = \frac{2S}{s(a_i)+S}$. The function $s(a)$ will be chosen to be the (squared) Euclidean norm which augments the uniform part and will allow us to control data points with high Euclidean norm, as these tend to dominate the approximation error. We comment on how to realize this distribution in the paragraphs below.

PAC learning with relative error. Under the assumption that $\Pr_{a \sim \mathcal{P}}(\|a\|_2 > D) = 0$, which (as discussed before) is common and arguably required in all uniform i.i.d. sample complexity literature, our techniques also work for this classic PAC learning setting where we can only draw samples uniformly from $\mathcal{P}$; we extend it by giving *multiplicative* rather than the standard additive guarantees. Our claim is formalized and corroborated in Corollary 3.4.

Sampling algorithms. Returning to sampling from probability distribution $\mathcal{Q}$, the simplest way to realize this, is through rejection sampling. I.e. we take m uniform samples $a_i \sim \mathcal{P}$, and reject each with probability $\frac{1}{2} - \frac{s(a_i)}{2\hat{S}}$. However, implementing this online requires prior knowledge of $\hat{S} = \sum_{j=1}^{m} s(a_j)$ (or a $(1 \pm \varepsilon)$-approximation), which is not always available.

Alternatives have been developed in data stream algorithms: in reservoir sampling one only needs to feed the current $s(a_i)$ to the sampler, which in our case is simply $s(a_i) = \|a_i\|_2 + 1$ (or similar), and the algorithm samples proportionally to $s(a_i) / \sum_{j=1}^{i} s(a_j)$. At the end of the sampling process

Table 1: **Overview of our results.** All bounds have ε^{-2} dependence, and $\widetilde{k} \coloneqq k(\log k)^3 \log\log k$. The best and only previous dimension-free result was $\tilde{O}(k^3/\varepsilon^2)$ by Alishahi and Phillips (2024).

$\mathcal{R}(\cdot)$	LOGISTIC		SIGMOID		HINGE		RELU	
	UPPER	LOWER	UPPER	LOWER	UPPER	LOWER	UPPER	LOWER
$\|\cdot\|_2^2$	$\widetilde{k}$	$k/\log k$	$\widetilde{k}$	$k/(\log k)^3$	k^2	k^2	-	$n \log n$
$\|\cdot\|_1$	$\tilde{k}$	$k \log k$	$\tilde{k}$	$k/\log k$	$\tilde{k}$	$k \log k$	$\tilde{k}$	$k \log k$
$\|\cdot\|_2$	k^2	k^2	k^2	k^2	k^2	k^2	k^2	k^2

we have m samples with distribution $s(a_i)/\hat{S}$, where $\hat{S} = \sum_{j=1}^{m} s(a_j)$ Chao (1982); Efraimidis (2015), and $\hat{S}/m = (1 \pm \varepsilon)S$ for sufficiently large m (see Lemmas B.2 and B.3). Another alternative (that even admits distributions over additive single-coordinate updates rather than insertion of entire points) is turnstile row-norm sampling (Mahabadi et al., 2020; Munteanu and Omlor, 2024b). Using either of the above streaming sampling algorithm as black box, we can realize $\mathcal{Q}$ by taking a uniform sample $a \sim \mathcal{P}$, and then tossing a coin. If it turns heads, we keep it right away. If it turns tails, we feed it to the sampling algorithm. The final sample is the union of the direct samples and the ones that are accepted by the sampling black box.

Finite data or coresets. In the *coreset* setting, see (Phillips, 2017; Munteanu, 2023), we are given a finite amount of n data points. Let $\mathcal{P}$ be the uniform measure supported on these points. Then $\int g(\langle a, x\rangle) d\mathcal{P}(a) = \frac{1}{n}\sum_{i=1}^{n} g(\langle a_i, x\rangle)$; thus approximating the integral to within $(1 \pm \varepsilon)$ is equivalent to constructing a coreset for the dataset. Also, we can simply calculate S (or estimate it using streaming algorithms Woodruff (2014)). Then our theorems apply directly and thus yield optimal coresets for a variety of regularized classification loss functions improving over previous works (Samadian et al., 2020; Tolochinsky et al., 2022; Alishahi and Phillips, 2024).

Distributional setting. In the distributional setting we only have access to samples from the unknown distribution $\mathcal{P}$, and we require additionally $P(\|a\|_2 > D) = 0$ for some term $D < \infty$. As discussed before, this is not needed for sampling from $\mathcal{Q}$, but for setting the weights, it is required to estimate S. Since the final estimate of S will be a $(1\pm\varepsilon)$-approximation, we actually feed $2s(a)$ to the sampler. The total sample size suffers an *additive* $O(D\log(1/\delta)/\varepsilon^2)$ compared to our main theorems Theorems 3.1 to 3.3, see Lemmas B.2 and B.3. We note that $S = B + 1$. Then if $D = O(B)$, this is a lower order term in sample size. D may also be much larger than B, but it is *only* necessary for estimating S; the final sample size is again independent of D.

Kernels and bounded spaces. Most data distributions naturally have some bound B, or D on the (average) norm of data points. If for no other reason, then values are standardized or normalized before learning. Layer-normalization Ba et al. (2016) is, for instance, a very common element of neural networks which leads to $B = D = 1$. This is also made explicit in kernel methods Schölkopf and Smola (2001) where a kernel support-vector-machine (SVM) is mathematically equivalent to linear classification operating in a (perhaps infinite-dimensional) kernel Hilbert space $\mathcal{H}_K$. Similarly, support vector data descriptor (SVDD) (Tax and Duin, 2004) is equivalent to the smallest enclosing ball problem in kernel Hilbert space (Krivosija and Munteanu, 2019). For a wide variety of common kernels $K(a, x)$ (e.g., normalized angular $K(a, x) = \langle a, x\rangle/(\|a\|\|x\|)$; Gaussian $K(a, x) = \exp(-\|a - x\|^2/\sigma^2)$; Laplace $K(a, x) = \exp(-\|a - x\|/\sigma)$), then again $\|a\|_{\mathcal{H}_K} = K(a, a) = 1$, and hence we have $B = D = 1$. When reproducing kernels are used, then this space $\mathcal{H}_K$ has no bound on the dimension, and a dimension-free bound (such as ours) is necessary for a finite sample complexity bound. In all of these cases, our analysis directly applies with uniform random sampling.

3 Our results

We refer to Table 1 for an overview of our contributions when applied to popular example functions. As our results build on and refine the prior work of Alishahi and Phillips (2024), we would like to stress that the model and all assumptions in our paper are either the same or even less strict.

3.1 Upper bounds

Quadratic upper bounds. Our first bound can be viewed as a refinement (and in our view, simplification) of that by Alishahi and Phillips (2024); it is allowed by the choice of sampling proportional to $\|a\|_2 + 1$, and the employment of more powerful bounds. We first symmetrize the approximation error, leading to an upper bound in terms of a canonical Rademacher process. Then we leverage Lipschitz contraction bounds of Ledoux and Talagrand (1991) and a special variant of Khintchine's inequality due to Kahane (1964) to derive a higher moment bound on the Rademacher process. This leads to our approximation result by employing the higher moment method (Alon and Spencer, 2008), i.e., an application of Markov's inequality to a higher power of the random variable.

Theorem 3.1. *Let $f(x) \coloneqq \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \mathcal{R}(x)/k$ such that g is L-Lipschitz. Let $s(a) \geq \|a\|_2 + 1$ for all $a \in \mathbb{R}^d$. Let $T \subseteq \mathbb{R}^d$, and $OPT_T = \inf_{x\in T} f(x)$. Set $C = (SL)^2$. There exists*

$$m_0 = \begin{cases} O(C\varepsilon^{-2} k \ln(\delta^{-1})/OPT_T), & \text{for } \mathcal{R}(\cdot) = \|\cdot\|_2^2 \\ O(C\varepsilon^{-2} k^2 \ln(\delta^{-1})), & \text{for } \mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\}, \end{cases}$$

such that if we sample $m \geq m_0$ points according to $\mathcal{Q}$, then with probability at least $1 - \delta$ we have

$$\forall x \in T\colon \left| f_0(x) - \frac{1}{m}\sum_{i=1}^{m} w_i g(\langle a_i, x\rangle) \right| \leq \varepsilon f(x)\,.$$

We note that in full generality it holds that $OPT \geq \Omega(1/k)$ (see Lemma B.4) if $g(0) > 0$, which implies that both bounds of Theorem 3.1 are quadratic in k in the worst case. Surprisingly, this k^2 upper bound is already optimal in several cases, see Table 1 and Section 3.2. This improves over the k^3 sensitivity sampling bounds of Alishahi and Phillips (2024). Also note that in the case $\mathcal{R}(\cdot) = \|\cdot\|_2^2$, we can handle subsets $T \subseteq \mathbb{R}^d$ with linear k dependence whenever $f(x) \geq OPT_T$ is bounded below by an absolute constant for all $x \in T$.

Moreover, in contrast to sensitivity sampling that was used in previous literature to obtain dimension-free bounds Alishahi and Phillips (2024), our theorem allows to set $s(a) = \|a\|_2 + 1$. This corresponds to row-norm sampling, which is almost as simple to implement as uniform sampling, e.g., using a weighted reservoir sampler Chao (1982), and does not require estimating sensitivities. Further, unlike sensitivity sampling, setting $s(a) = \|a\|_2 + 1$ suffices to ensure that $S = B + 1$ is independent of k for our row norm sampling.

Linear upper bounds. We identify two settings for which we can conduct improved analyses to obtain linear k dependence:

Setting 1: $\|\cdot\|_2^2$-regularization when g is either convex or concave or monotonic and more crucially has a bounded derivative, i.e., $|g'| \leq g$. This applies to our examples of logistic and sigmoid loss, which decay sufficiently slowly towards zero.

Theorem 3.2. *Let $f(x) \coloneqq \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \|x\|_2^2/k$ such that g is L-Lipschitz, and either convex or concave or monotonic, and for all $r \in \mathbb{R}$ it holds that $|g'(r)| \leq g(r)$. Let $s(a) \geq \|a\|_2^2 + 2$ for all $a \in \mathbb{R}^d$. Set $C = SBL^2/g(0)$. There exists $m_0 = O(C\varepsilon^{-2} k \ln(C\varepsilon^{-1} k \ln(\delta^{-1}))^3 \ln(\delta^{-1} \ln(SLk/g(0)))) = \tilde{O}(C\varepsilon^{-2} k \ln(\delta^{-1}))$, such that if we sample $m \geq m_0$ points according to $\mathcal{Q}$, then with probability at least $1 - \delta$ we have that*

$$\forall x \in \mathbb{R}^d\colon \left| f_0(x) - \left(\frac{1}{m}\sum_{i=1}^{m} w_i g(\langle a_i, x\rangle) \right) \right| \leq \varepsilon f(x).$$

If additionally $B = O(\log(k))$ holds, then choosing $C = SL/g(0)$ suffices for the same guarantee.

As argued before, it often holds that $B = O(1)$ in typical applications, which is in favor of the improved constants. Another common case is that $k = n^c$, for some $c \in (0, 1)$. Samadian et al. (2020); Tolochinsky et al. (2022); Mai et al. (2021), in which $B = O(\log(k)) = O(\log(n))$ is an arguable assumption.

Setting 2: The second setting with linear k dependence is for general Lipschitz functions with $\|\cdot\|_1$-regularization.

Theorem 3.3. *Let $f(x) \coloneqq \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \|x\|_1/k$ such that g is L-Lipschitz. Let $s(a) \geq \|a\|_2 + 1$ for all $a \in \mathbb{R}^d$. Set $C = SL$. There exists $m_0 =$*

$O(C\varepsilon^{-2}k\ln(C\varepsilon^{-1}k\ln(\delta^{-1}))^3\ln(\delta^{-1}\ln(\varepsilon^{-1}BLk))) = \tilde{O}(C\varepsilon^{-2}k\ln(\delta^{-1}))$*, such that if we sample* $m \geq m_0$ *points according to* $\mathcal{Q}$*, then with probability at least* $1-\delta$ *it holds for all* $x \in \{x \in \mathbb{R}^d \mid f(x) \leq g(0)/\varepsilon\}$ *that*

$$\left| f_0(x) - \left(\frac{1}{m} \sum_{i=1}^{m} w_i g(\langle a_i, x \rangle) \right) \right| \leq \varepsilon f(x) \,.$$

If g *is homogeneous, then it is sufficient to set* $m_0 = O(C\varepsilon^{-2}k\ln(\delta^{-1})\ln(C\varepsilon^{-1}k\ln(\delta^{-1}))^3)$ *to get the same result for all* $x \in \mathbb{R}^d$ *simultaneously.*

We comment on the restriction to $x \in \{x \in \mathbb{R}^d \mid f(x) \leq g(0)/\varepsilon\}$ in the case of a non-homogeneous function g. First, this set contains and thus allows to approximate the optimal value, since $OPT \leq f(0) = g(0)$. Second, all four example functions can be split into $g(r) = h(r) + b(r)$, where $h(r)$ is homogeneous, and $b(r)$ is bounded by $g(0)$. Whenever such a decomposition exists, then the guarantee of Theorem 3.3 extends to all $x \in \mathbb{R}^d$ (see Lemma D.25).

Both the above improvements are achieved via refined empirical process analyses. The aforementioned symmetrization argument relates the relative error up to an additional small constant to a canonical Gaussian process. The Gaussian process induces a pseudo-metric, which can be bounded in terms of the *metric entropy* $\mathcal{E}$ and the *metric diameter* $\mathcal{D}$ by Sudakov's tail bound, also known as Dudley's bound (Dudley, 2016; Ledoux and Talagrand, 1991). Recently, Woodruff and Yasuda (2023) transformed this classic result into a higher moment bound involving the ratio of the two quantities. The remaining challenge is to bound $\mathcal{E}, \mathcal{D}$ appropriately and use the higher moment method again to obtain the final approximation via Markov's inequality.

More details on the general outline of empirical process analyses and on our specific techniques to handle the two different settings are provided in the technical overview, and the full details can be found in Appendix D, D.3 resp. D.4.

Relative error PAC learning with uniform sampling. Our upper bound theorems apply to the classic problem under plain uniform sampling $a \sim \mathcal{P}$ as special case, resolving the sample complexity, and matching our lower bounds.

Corollary 3.4. *Let* $f(x) \coloneqq \int g(\langle a, x \rangle)\, d\mathcal{P}(a) + \mathcal{R}(x)/k$ *such that* g *is* L*-Lipschitz. Assume that* $\Pr_{a\sim P}(\|a\|_2 > D) = 0$*. Let* $T \subseteq \mathbb{R}^d$*. If we sample a number of* m *points according to* $\mathcal{P}$*, then with probability at least* $1-\delta$

$$\forall x \in T\colon \left| f_0(x) - \left(\frac{1}{m} \sum_{i=1}^{m} g(\langle a_i, x \rangle) \right) \right| \leq \varepsilon f(x) \,,$$

holds under the same restrictions on T*, conditions on* g*, and sizes of* m*, as stated in Theorems 3.1 and 3.3, with every occurrence of* S *replaced by* D*. In Theorem 3.2,* S *must be replaced by* D^2*.*

Proof. The result follows directly by setting $s(a) = D+1 \geq \|a\|_2+1$, resp. $s(a) = D^2+2 \geq \|a\|_2^2+2$ and calling Theorems 3.1 to 3.3. For this choice of $s(a)$ it holds that $S = \int s(a)\, d\mathcal{P}(a) = D+1$, resp. $S = D^2+2$, $\mathcal{Q}(a) = (\frac{s(a)}{2S} + \frac{1}{2})\mathcal{P}(a) = \mathcal{P}(a)$, and $w_i = \frac{2S}{s(a_i)+S} = 1$, which corresponds exactly to uniform sampling of $a \sim \mathcal{P}$. □

Parameter limitations. We note that all above theorems crucially rely on the boundedness of all parameters such as $D, S, B, L, k, 1/\varepsilon, 1/\delta$. For example unnormalized Gaussians would reintroduce a dependence on d through these parameters. In such a situation however, obtaining dimension-free bounds is impossible, cf. (Alishahi and Phillips, 2024).

3.2 Lower bounds

Quadratic lower bounds. As mentioned before when treating the quadratic k^2/ε^2 upper bound, it is already optimal for $\|\cdot\|_2$-regularization, indicating that this regularizer is the weakest among our examples. However, $\|\cdot\|_2^2$-regularization also exposes weaknesses if the function g is not well-behaved (in particular if its derivative is not uniformly bounded by its value $|g'| \leq g$ as in Theorem 3.2). In fact, the combinations with hinge and ReLU loss functions do not have this property that seems required for better than quadratic sampling complexity, as our quadratic lower bounds match our corresponding upper bounds:

Theorem 3.5. *Consider any of the following combinations of loss functions and regularizers:*

- *logistic/sigmoid with* $\mathcal{R}(\cdot) = \|\cdot\|_2$*, or*
- *hinge/ReLU with* $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_2^2\}$.

There exists a distribution $\mathcal{P}$ *over vectors* $a \in \mathbb{R}^d$ *such that* $m = \Omega(k^2/\varepsilon^2)$ *samples are required to satisfy Equation* (1)*.*

Linear lower bounds. We complement our linear sampling complexity by matching lower bounds.

Theorem 3.6. *Suppose we want to satisfy Equation* (1) *for some combination of loss function and regularizer. In each of the following settings there exists a distribution* $\mathcal{P}$ *over vectors* $a \in \mathbb{R}^d$ *such that the stated lower bounds on the number of samples* m *hold:*

1. *For either ReLU/hinge/logistic loss with* $\mathcal{R}(\cdot) = \|\cdot\|_1$ *regularization,* $m = \Omega(\varepsilon^{-2} k \log k)$ *is required,*
2. *For either sigmoid loss with* $\mathcal{R}(\cdot) = \|\cdot\|_1$*, or logistic loss with* $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ *regularization,* $m = \Omega(\varepsilon^{-2} k / \log k)$ *is required,*
3. *For sigmoid loss with* $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ *regularization,* $m = \Omega(\varepsilon^{-2} k / (\log k)^3)$ *is required.*

Impossibility results. Finally, recall that our bound in Theorem 3.1 for handling $\|\cdot\|_2^2$-regularization is sensitive to the optimal value OPT and even Theorem 3.2 is sensitive to the value of $g(0)$, and it is required that these values are bounded away from 0. This, however, is not the case with the ReLU function since $f(0) = \mathrm{ReLU}(0) + \|0\|_2^2/k = g(0) = 0$, which makes the upper bound infinite. We corroborate this phenomenon by two lower bounds. The first is conceptually simple and yields $\Omega(d \log d)$, showing impossibility of dimension-free sampling complexity. The second refines the idea and yields $\Omega(n \log n)$, for a distribution $\mathcal{P}$ with arbitrarily large support size $n \geq d + 1$.

Theorem 3.7. *Consider the ReLU function where* $g(r) = \max\{0, -r\}$*, and the* $\mathcal{R}(x) = \|x\|_2^2$ *regularizer. For any* $n \geq d$ *there exist distributions supported on* n *points in* $\mathbb{R}^d$*, such that with high probability* $m = \Omega(n \log n)$ *samples are required to satisfy Equation* (1)*.*

We note that $O(n)$ is clearly sufficient *if* the support is known, for instance in a finite data setting. However, if we can access data items only through sampling from an unknown distribution it is known that with high probability $\Omega(n \log n)$ draws are necessary to see all n distinct items by the coupon collector's theorem (Erdős and Rényi, 1961).

4 Technical overview

Even though none of our analyses are conducted in this way, we would like to point the interested reader to Lemma B.1, which may be of independent interest. It states that for $\mathcal{R}(x) \in \{\|\cdot\|_1, \|\cdot\|_2\}$, the analysis can be performed for any $k' \in \mathbb{N}$ assuming $B' = L' = 1$. The resulting bounds work with arbitrary values B, L, k just by substituting $k' = LBk$.

All upper bounds start with a standard symmetrization (cf. Cohen and Peng, 2015), that relates the relative error

$$E \coloneqq \sup_{x \in \mathbb{R}^d} \left| f_0(x) - \left(\frac{1}{m} \sum_{i=1}^{m} w_i g(a_i x) \right) \right| \Big/ f(x)$$

we aim to bound, to the supremum of an empirical process. I.e., up to absolute constants, we show for i.i.d. Rademacher $\sigma_i \sim U\{-1, +1\}$ or Gaussian $\sigma_i \sim N(0, 1)$ variables, that

$$\mathbb{E}(E) \leq \mathbb{E} \sup_{x \in \mathbb{R}^d} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i w_i g(a_i x) \right| \Big/ f(x)\,. \tag{2}$$

All remaining efforts aim at bounding the ℓ-th moment of the empirical process by $(\varepsilon/2)^\ell$. Using Markov's inequality and $\ell = O(\ln(1/\delta))$, such a bound allows us to conclude

$$\Pr(E \geq \varepsilon) = \Pr(E^\ell \geq \varepsilon^\ell) \leq \mathbb{E}(E^\ell)/\varepsilon^\ell \leq 2^{-\ell} \leq \delta\,,$$

known as higher moment method Alon and Spencer (2008).

Quadratic upper bounds. Towards Theorem 3.1, we first bound the ℓ-th moment essentially by

$$\mathbb{E}(E^\ell) \leq \left(\frac{SL\sqrt{\ell}}{\sqrt{m}} \cdot \sup_{x \in \mathbb{R}^d} \left(\frac{\|x\|_2}{OPT + \mathcal{R}(x)/k} \right) \right)^\ell . \tag{3}$$

The main steps in this analysis are taken by a careful application of Talagrand's Lipschitz Contraction (Lemma F.5) followed by Cauchy-Schwartz and the Khintchine-Kahane inequalities (see Corollary F.7). This gives essentially

$$\mathbb{E} \left| \sum_{i=1}^m \sigma_i w_i g(a_i x) \right|^\ell \leq (L\|x\|_2)^\ell \ \mathbb{E} \left(\ell \sum_{i=1}^m w_i^2 \, \|a_i\|_2^2 \right)^{\ell/2} .$$

We can see that the norm of a point has a high impact on the bound of their error contribution. Due to our choice of the sampling distribution and corresponding weights, we can control the norm of data points up to a factor S, i.e., $w_i \|a_i\|_2 \leq \frac{S}{\|a_i\|_2} \cdot \|a_i\|_2 = S$. It then remains to bound the supremum term in Equation (3). We get in Lemma C.5 that

$$\sup_{x \in \mathbb{R}^d} \left(\frac{2\|x\|_2}{OPT + \mathcal{R}(x)/k} \right) \leq \begin{cases} \sqrt{k/OPT}, & \text{if } \mathcal{R}(\cdot) = \|\cdot\|_2^2 \\ 2k, & \text{if } \mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\} \end{cases}$$

and note that it leads to tight bounds in the general case when $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_2^2\}$. By plugging this into Equation (3) we get up to absolute constants

$$\mathbb{E}(E^\ell) \leq \left(\frac{(SL)^2 k^{2/p} \ell}{m OPT^{p-1}} \right)^{\ell/2} ,$$

where $p \in \{1, 2\}$ is the exponent of the regularizer. By our choice of m in Theorem 3.1, it yields $\mathbb{E}(E^\ell) \leq (\varepsilon/2)^\ell$ in both cases, which concludes the proof.

Since $OPT \geq \Omega(1/k)$, this yields only quadratic bounds in k for both $p \in \{1, 2\}$. But for $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ it is useful in special cases where $OPT \geq \Omega(1)$, and leaves hope for a refined analysis. For $\mathcal{R}(\cdot) = \|\cdot\|_1$, the bound is an immediate consequence of the inter-norm inequality $\|x\|_1 \geq \|x\|_2$, which leaves sufficient gap for further improvement.

Linear upper bounds.

Setting 1: For $\mathcal{R}(\cdot) = \|\cdot\|_2^2$, Theorem 3.1 is tight in general. In Theorem 3.2, we improve it for monotonic, or convex functions that in particular do not decay too quickly towards zero; formally, we require that $|g'(r)| \leq g(r)$ for all $r \in \mathbb{R}$. We bound the error akin to Equation (2) by a Gaussian process, and use the framework of empirical process analysis. Sudakov's resp. Dudley's tail bound (Dudley, 2016; Ledoux and Talagrand, 1991) was recently reformulated by (Woodruff and Yasuda, 2023) to obtain a higher moment bound

$$\mathbb{E}[\sup_{t \in T} |X_t|^\ell] \leq (2\mathcal{E})^\ell (\mathcal{E}/\mathcal{D}) + O(\sqrt{\ell}\mathcal{D})^\ell \tag{4}$$

in terms of the *metric entropy* $\mathcal{E}$ and the *metric diameter* $\mathcal{D}$ of a pseudo-metric d_X induced by the Gaussian process $(X_t)_{t \in T}$. It thus remains to bound $\mathcal{E}, \mathcal{D}$ appropriately. To this end, the analysis is partitioned into slices $T_\alpha = \{x \in \mathbb{R}^d \mid \alpha \leq f(x) \leq 2\alpha\}$. First fix a specific value, say $\alpha = 1$; we will later union bound over a small number of values for α. The diameter for T_α can essentially be related to the approximation ratio $G = \sup_{x \in T_\alpha} f_S(x)/f(x)$ and the sensitivity of our samples, i.e.,

$$\mathcal{D} = \sup_{s,t \in T_\alpha} d_X(s,t) \leq G\alpha \left(\frac{1}{m} \sup_{x \in T_\alpha, j \in [m]} \frac{w_j g(a_j x)}{f(x)} \right)^{1/2} ;$$

and we thus bound the sensitivities by $SBL^2 k/g(0)$. Given the common case that $B = O(\log(k))$, we also derive another bound that yields $SLk/g(0)$, which is linear and optimal in the problem specific constants, not only in k.

To bound the metric entropy, we first relate the metric d_X to an ℓ_∞ norm $\|\cdot\|_X$ in the subspace of our weighted samples

$$d_X(s,t) \leq \left(G\alpha \max_{j \in [m]} w_j |\langle a_j, s - t \rangle|^2 L/m \right)^{1/2} .$$

This step relies on either monotonicity or convexity, and in particular it requires the additional assumption that $|g'(r)| \leq g(r)$ for all $r \in \mathbb{R}$. So this works for logistic or sigmoid loss but we will see later that our examples without this assertion (hinge, ReLU) have k^2 lower bounds. Next, we bound the Gaussian complexity $\mathbb{E}_{g\sim\mathcal{N}(0,I_d)}\|g\|_X$ of the norm $\|\cdot\|_X$. Using that $\max_{j\in[m]} w_j|\langle a_j, g\rangle| \leq S\max_{j\in[m]}|\langle \frac{a_j}{\|a_j\|_2}, g\rangle|$, it reduces to bounding the largest in a collection of m standard half-Gaussians which is known to be in the order of $O(\sqrt{\log(m)})$. Then, Sudakov's dual-minorization bound (cf. Bourgain et al., 1989) yields an upper bound on the number of $\|\cdot\|_X$-balls (thus d_X-balls) of large radii $t > 1/m$ to cover the current set T_α, i.e.,

$$\log N(T_\alpha, \|\cdot\|_X, t) \leq O(G^2\alpha^2 SLk\ln(m)/m)/t^2\,.$$

Due to the $1/t^2$ dependence, it is better to use a standard t-net construction for small radii $0 < t < 1/m$. This yields an $O(m\log(1/t))$ dependence, i.e.,

$$\log N(T_\alpha, \|\cdot\|_X, t) \leq O(m\ln(t^{-1}G\alpha\sqrt{SLk\ln(m)/m})).$$

This enables us to calculate the entropy integral by splitting at $\lambda = 1/m$, and using the two bounds for $0 < t \leq \lambda$ resp. $\lambda \leq t \leq \mathcal{D}$. We can cut off the calculation at the diameter, because for larger radii the covering number is trivially 1 and its logarithm becomes 0. This argumentation yields up to absolute constants and polylogarithmic terms that

$$\mathcal{E} = \int_0^\infty \sqrt{\log N(T_\alpha, \|\cdot\|_X, t)}dt \leq \tilde{O}(G\alpha\sqrt{SLk/m})\,.$$

Choosing m carefully, as stated in Theorem 3.2, balances both our bounds on $\mathcal{D}$ and $\mathcal{E}$ to be in $O(\varepsilon G\alpha)$, so we can use the higher moment method on Equation (4) to obtain the desired approximation via Markov's inequality. To complete the argument, we union bound over all T_α for $1 \geq \alpha \geq OPT \geq \Omega(1/k)$ in powers of two. For the remaining $x \in T = \{x \in \mathbb{R}^d \mid f(x) \geq 1\}$ we invoke Theorem 3.1. Since $OPT_T = \inf_{x\in T} f(x) \geq 1$, it yields a linear bound in k.

Setting 2: For $\mathcal{R}(\cdot) = \|\cdot\|_1$ we do not need further assumptions to get a linear bound in k. The outline is similar to the case $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ and we discuss only the most significant changes to obtain Theorem 3.3. Using the relationship between the diameter and sensitivity, we need to bound the sensitivities, which are $O(SLk)$ in the new setting and yield

$$\mathcal{D} = \sup_{s,t\in T_\alpha} d_X(s,t) \leq G\alpha\,(SLk/m)^{1/2}\,.$$

Instead of embedding the metric into a norm, we have the square root of the ℓ_∞-norm. Unlike the previous analysis, this requires squaring the radii before applying Sudakov's bounds. For this not to harm the complexities in subsequent steps, we need to counteract the squaring by proving a bound on the covering numbers that is only linear in $\log(m)/t$, i.e.,

$$\log N(B_1, B_\infty(A), t) \leq O(S)\log(m)/t\,,$$

by chaining the calculation into

$$\log N(B_1, B_2(V^T), \sqrt{t}) + \log N(B_2(V^T), B_\infty(A), \sqrt{t}),$$

where $B_p(A) = \{x \in \mathbb{R}^d \mid \|Ax\|_p \leq 1\}$ and $V^T \in \mathbb{R}^{m\times d}$ is an orthonormal basis for the subspace spanned by $A \in \mathbb{R}^{m\times d}$. We apply Sudakov's minorization bound again, but with radius $\sqrt{t}$, to the second term. For the first term, we employ Sudakov's dual minorization. We need to bound the Gaussian complexity $\mathbb{E}_{g\sim\mathcal{N}(0,I_m)}\|Vg\|_X$, which naïvely results in a prohibitive $O(\sqrt{\log(d)})$. To circumvent this, we leverage the fact that we work in an m dimensional subspace and the additive variance property of squared Gaussians, to aggregate the rows v_i resp. their leverage scores $\|v_i\|_2^2$ into a matrix $W \in \mathbb{R}^{2m\times m}$ where each row norm is $\|w_i\|_2^2 \leq 1$. Notably, this is the *opposite* to flattening lemmas (cf. Woodruff and Yasuda, 2023) and yields $\mathbb{E}_{g\sim\mathcal{N}(0,I_m)}\|Vg\|_X \leq \mathbb{E}_g\|Wg\|_X \leq O(\sqrt{\log(m)})$. With the new bounds on covering numbers, and incorporating the additional factors of our embedding of d_X into $\|\cdot\|_X$, the metric entropy integral calculation works exactly as before. Choosing m carefully, as stated in Theorem 3.3, balances both our bounds on $\mathcal{D}$ and $\mathcal{E}$, so we can use the higher moment method on Equation (4) to conclude the error bound via Markov's inequality. We conclude by a union bound over all T_α, for $g(0)/\varepsilon \geq \alpha \geq OPT \geq \Omega(1/k)$ in powers of two. The analysis is so far restricted to $x \in \{x \in \mathbb{R}^d \mid f(x) \leq g(0)/\varepsilon\}$, but we cannot invoke Theorem 3.1

since it would yield k^2 dependence. However, this set contains and thus allows to approximate the optimal value, since $OPT \leq f(0) = g(0)$. Furthermore, consider our four example functions: ReLU is homogeneous, for which analyzing one fixed value, say $\alpha = 1$ suffices to extend to all $x \in \mathbb{R}^d$ simply by scaling. All other three functions can be split into $g(r) = h(r) + b(r)$, where $h(r)$ is homogeneous, and $b(r)$ is bounded by $g(0)$. Whenever such a decomposition applies, we argue in Lemma D.25, that the guarantee of Theorem 3.3 extends to all $x \in \mathbb{R}^d$ since the error can be separated by the triangle inequality. The homogeneous part works for $x \in \mathbb{R}^d$ as explained in the case of ReLU and the bounded part has a maximum error of $O(g(0))$ which can be charged for $\varepsilon f(x)$, since we need this special treatment only for $x \in \mathbb{R}^d$ that obey $f(x) \geq g(0)/\varepsilon$. Sigmoid loss is bounded in $[0, 1]$, so $h(x) = 0$ and $b(x) = g(x)$. Logistic and hinge loss, both allow for $h(x) = ReLU(x)$, and $b(x) = g(x) - ReLU(x) \in [0, g(0)]$ (cf. Mai et al., 2021).

Quadratic lower bounds. Our constructions for logistic and sigmoid are similar up to constants, and hold only against $\mathcal{R}(\cdot) = \|\cdot\|_2$ regularization. The input is a uniform distribution supported on $n = \Theta((k/\varepsilon)^2)$ standard basis vectors e_i. Any subsample of $m \leq n/2$ elements will miss at least half of the support; let them be indexed by $i \in [n/2]$. The adversarial query is $x = \sum_{i=1}^{n/2} e_i$. Then the original loss is bounded by a constant, say $g(0)/2$, and the regularizer contributes $\|x\|_2/k = \Theta(\sqrt{n}/k) \approx 1/\varepsilon$, so $f(x) = \frac{g(0)}{2} + \frac{1}{\varepsilon}$. However, since our sample misses the indexes of the query, regularization is preserved, but the main loss grows to a significantly larger constant $g(0) > (1+\varepsilon)g(0)/2$. We then show that if the approximation of $f_S(0) \approx \frac{1}{m}\sum w_i \in [1-\varepsilon, 1+\varepsilon]$, then the approximation of $f_S(x) > (1+\varepsilon)f(x)$, thus fails on either of $x' \in \{0, x\}$. Our lower bounds against hinge and ReLU loss work for both $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_2^2\}$, however, for ReLU with $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ we will see worse results later. The construction is very similar, but fixes one index to 1, so $a_i = e_n + e_i/\sqrt{2}$, and accordingly $x = e_n + \sum_{i=1}^{(n-1)/2} e_i/\sqrt{2(n-1)}$ are scaled to constant norm. The regularizer is roughly $1/k$ and our construction ensures that the original loss is roughly $2\varepsilon/k$, while the approximated loss drops to 0. But then the approximation error is only the original loss, i.e., $|f_S(x) - f(x)| = |f(x)| = 2\varepsilon/k > \varepsilon(1+2\varepsilon)/k = \varepsilon f(x)$, which concludes Theorem 3.5. In particular, note how this exploits the fact that hinge and ReLU drop to 0 with slope 1, and thus *do not* have the nice property $|g'(r)| \leq g(r)$ as in Theorem 3.2, which allows linear in k upper bounds for logistic and sigmoid loss with $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ regularization.

Linear lower bounds. All linear in k lower-bounds follow the same high level template. We build a distribution supported on $N = \Theta(k)$ carefully chosen points. For each point a_j we build a query $x^{(j)}$ whose loss is highly sensitive to that single point; it tests whether the sample estimated the contribution of the j-th point accurately. Concretely, an i.i.d. importance sampler induces a multinomial count N_j for each point a_j. Since the sampling scores $s(\cdot)$ are normalized, by Markov's inequality, a constant fraction of supported points must have comparatively small score (e.g. $s_j \leq 2$), hence their expected counts satisfy $\mu_j = \mathbb{E}[N_j] = O(m/k)$. The query $x^{(j)}$ is designed so the failure occurs whenever we have a a constant-factor deviation of N_j from its mean. By Feller's inequality (Lemma F.1), each deviation occurs at least with probability roughly $p = \exp\left(-\varepsilon^2\Omega(m/k)\right)$. By negative association of multinomial coordinates, the probability that *none* of these deviations happens over $\Omega(k)$ disjoint indices is at most $(1-p)^{\Omega(k)}$. A failure probability at most δ forces $p \lesssim \delta/k$, which yields the stated $\Omega(\varepsilon^{-2}k\log k)$ sample lower bounds.For ReLU loss, we place mass on the signed basis vectors $\{\pm e_j\}_{j=1}^k$ and test at $x = -\sigma e_j$, which isolates the contribution of a single point and yields the $\Omega(\varepsilon^{-2}k\log k)$ bound for $\|\cdot\|_1/k$ regularization. For hinge and logistic with $\|\cdot\|_1/k$, we transfer this bound via a reduction from the ReLU problem (see Lemma E.5): as the dot product grows, both (scaled) losses converge to ReLU, and with homogeneous regularizers, which are not affected by scaling, this allows us to reduce from the ReLU hard instance (Theorem 3.6 part (1)). For logistic with $\|\cdot\|_2^2/k$ and sigmoid with $\|\cdot\|_1/k$ or $\|\cdot\|_2^2/k$, we use a k-point construction $e_i + e_{k+1} \in \mathbb{R}^{k+1}$ (for $i \in [k]$) together with a family of queries $x^{(j)} = \alpha(-2e_j + e_{k+1})$ that make the dot product of one point $-\alpha$ and the remaining points have $+\alpha$. Choosing $\alpha = \Theta(\log(k))$ makes $g(\alpha) = 1/k$, keeping $f(x^{(j)})$ small while the estimate becomes sensitive to the single count N_j providing a sufficient condition for failure. For logistic, the exact identity $g(-\alpha) - g(\alpha) = \alpha$ amplifies deviations and yields the $\Omega(\varepsilon^{-2}k/\log k)$ bound. For sigmoid, $g(-\alpha) = 1 - g(\alpha)$ gives weaker amplification, leading to $\Omega(\varepsilon^{-2}k/\log k)$ for $\|\cdot\|_1/k$ and $\Omega(\varepsilon^{-2}k/\log^3 k)$ for $\|\cdot\|_2^2/k$ (Theorem 3.6 parts (2)/(3)).

See Appendix A.2 for a technical overview of additional *impossibility results*.

Acknowledgments and Disclosure of Funding

The authors thank the anonymous reviewers for their constructive feedback and discussions. A.M. and S.O. were supported by the German Research Foundation (DFG) – grant MU 4662/2-1 (535889065) and by the TU Dortmund – Center for Data Science & Simulation (DoDaS).

References

Meysam Alishahi and Jeff M. Phillips. No dimensional sampling coresets for classification. In *Forty-first International Conference on Machine Learning, (ICML)*, 2024.

Noga Alon and Joel H. Spencer. *The Probabilistic Method, Third Edition*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2008. ISBN 978-0-470-17020-5.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Nikhil Bansal, Vincent Cohen-Addad, Milind Prabhu, David Saulpic, and Chris Schwiegelshohn. Sensitivity sampling for k-means: Worst case and stability optimal coreset bounds. In *65th IEEE Annual Symposium on Foundations of Computer Science, (FOCS)*, pages 1707–1723, 2024.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html`.

Min-Te Chao. A general purpose unequal probability sampling plan. *Biometrika*, 69(3):653–656, 1982.

Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, (STOC)*, pages 205–214, 2009.

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

Richard M. Dudley. V.N. Sudakov's work on expected suprema of Gaussian processes. In Christian Houdré, David M. Mason, Patricia Reynaud-Bouret, and Jan Rosiński, editors, *High Dimensional Probability VII*, pages 37–43. Springer International Publishing, 2016.

Pavlos S. Efraimidis. Weighted random sampling over data streams. In *Algorithms, Probability, Networks, and Games*, pages 183–195. Springer, 2015.

P. Erdős and A. Rényi. On a classical problem of probability theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 6:215–220, 1961.

William Feller. Generalization of a probability limit theorem of Cramér. *Trans. Am. Math. Soc.*, 54: 361–372, 1943.

F. G. Foster. On the stochastic matrices associated with certain queuing processes. *The Annals of Mathematical Statistics*, 24(3):355–360, 1953. URL `http://www.jstor.org/stable/2236286`.

Uffe Haagerup. The best constants in the Khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981. URL http://eudml.org/doc/218383.

Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying sums of norms. In *64th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1953–1962, 2023.

Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying generalized linear models. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, (STOC)*, pages 1665–1675, 2024.

Kumar Joag-Dev and Frank Proschan. Negative Association of Random Variables with Applications. *The Annals of Statistics*, 11(1):286 – 295, 1983. doi: 10.1214/aos/1176346079. URL https://doi.org/10.1214/aos/1176346079.

J.-P. Kahane. Sur les sommes vectorielles $\sum \pm u_n$. *Comptes Rendus de l'Académie des Sciences (Paris)*, 259:2577–2580, 1964.

Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, 1994.

Apoorva Khare and Bala Rajaratnam. The Khinchin–Kahane and Lévy inequalities for Abelian metric groups, and transfer from normed (Abelian semi)groups to Banach spaces. *Journal of Mathematical Analysis and Applications*, 528(2):127545, 2023.

A. Khintchine. Über dyadische brüche. *Mathematische Zeitschrift*, 18:109–116, 1923.

Amer Krivosija and Alexander Munteanu. Probabilistic smallest enclosing ball in high dimensions via subgradient sampling. In *35th International Symposium on Computational Geometry (SoCG)*, pages 47:1–47:14, 2019. doi: 10.4230/LIPICS.SOCG.2019.47. URL https://doi.org/10.4230/LIPIcs.SoCG.2019.47.

Rafał Latała and Krzysztof Oleszkiewicz. On the best constant in the Khinchin-Kahane inequality. *Studia Mathematica*, 109(1):101–104, 1994. URL http://eudml.org/doc/216056.

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

Yi Li, Philip M Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527, 2001.

Sepideh Mahabadi, Ilya P. Razenshteyn, David P. Woodruff, and Samson Zhou. Non-adaptive adaptive sampling on turnstile streams. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 1251–1264, 2020.

Tung Mai, Cameron Musco, and Anup Rao. Coresets for classification - simplified and strengthened. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 11643–11654, 2021.

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

Alexander Munteanu and Simon Omlor. Optimal bounds for ℓ_p sensitivity sampling via ℓ_2 augmentation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024a.

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

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

Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David Woodruff. On coresets for logistic regression. *Advances in Neural Information Processing Systems*, 31, 2018.

Alexander Munteanu, Simon Omlor, and Christian Peters. 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)*, pages 2073–2100, 2022.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. Efficient large-scale language model training on GPU clusters using megatron-lm. In Bronis R. de Supinski, Mary W. Hall, and Todd Gamblin, editors, *International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2021, St. Louis, Missouri, USA, November 14-19, 2021*, page 58. ACM, 2021. doi: 10.1145/3458817.3476209. URL https://doi.org/10.1145/3458817.3476209.

Jeff M Phillips. Coresets and sketches. In *Handbook of Discrete and Computational Geometry*, pages 1269–1288. Chapman and Hall/CRC, 3rd edition, 2017.

Alireza Samadian, Kirk Pruhs, Benjamin Moseley, Sungjin Im, and Ryan R. Curtin. Unconditional coresets for regularized loss minimization. In *The 23rd International Conference on Artificial Intelligence and Statistics, (AISTATS)*, pages 482–492, 2020.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053, 2019. URL http://arxiv.org/abs/1909.08053.

V. N. Sudakov. Gaussian random processes and measures of solid angles in Hilbert space. *Sov. Math. Dokl.*, 12:412–415, 1971.

David M. J. Tax and Robert P. W. Duin. Support vector data description. *Machine Learning*, 54(1): 45–66, 2004. doi: 10.1023/B:MACH.0000008084.60811.49.

Sven Teschke, Katja Ickstadt, and Alexander Munteanu. Detecting interactions in high-dimensional data using cross leverage scores. *Biometrical Journal*, 66(8):e70014, 2024. doi: https://doi.org/10.1002/bimj.70014. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.70014.

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

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: Festschrift for Alexey Chervonenkis*, pages 11–30. Springer, 2015.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1-2):1–157, 2014.

David P. Woodruff and Taisuke Yasuda. Sharper bounds for ℓ_p sensitivity sampling. In *International Conference on Machine Learning (ICML)*, pages 37238–37272, 2023.

A Omitted material (due to page limitations)

A.1 Further related work.

Our work combines Rademacher complexity Alishahi and Phillips (2024) with empirical processes, by deriving higher moment bounds on a Rademacher process Cohen and Peng (2015); Mai et al. (2021) and using the classic higher moment method Alon and Spencer (2008); Clarkson and Woodruff (2009); Woodruff (2014). Further, our work continues a recent series of works that achieve optimal linear complexities (whenever possible) through empirical process analyses of sensitivity sampling for ℓ_p-subspace embeddings (Woodruff and Yasuda, 2023; Munteanu and Omlor, 2024a), k-means/median clustering (Bansal et al., 2024), and sparsification of sums of semi-norms, and generalizations of linear models (Jambulapati et al., 2023, 2024).

A.2 Technical overview of impossibility results

Impossibility results. A simple $\Omega(d \log d)$ bound shows no dimension-free sampling is possible for the combination of ReLU with $\mathcal{R}(\cdot) = \|\cdot\|_2^2$. Consider the uniform distribution over d standard basis vectors in $\mathbb{R}^d$. Suppose our sample misses one of them, say e_1. If we query $x = \eta e_1$ for a tiny $\eta > 0$, the loss becomes η/d, its approximation is 0 and the regularizer is η^2/k. When we let η be small enough, the regularizer has no effect, so $|f_S(x) - f(x)| = |f(x)| = \eta/d > \varepsilon(\eta/d + \eta^2/k) = \varepsilon f(x)$. Thus, all d points need to be in the sample, and $\Omega(d \log d)$ samples are required to succeed this task by the coupon collector's theorem (Erdős and Rényi, 1961).

Our $\Omega(n \log n)$ lower bound in Theorem 3.7 refines the above idea of a *scaling mismatch* near the origin and highlights a limitation of the sampling model to provide a uniform relative-error guarantee for sublinear sample size. The proof builds a geometric hard instance by placing n points $a_i \in \mathbb{R}^d, i \in [n]$ along the moment curve, so their convex hull is a cyclic polytope. As a result, each point can be separated from all others by a supporting hyperplane. Equivalently, for every point a_j there exists a query direction $x^{(j)}$ such that a_j has negative dot product while all other points have non-negative dot products. For ReLU, this makes the loss at $x^{(j)}$ depend essentially on the contribution of that single point. Ridge regularization does not prevent this isolation, because we can scale the query by a tiny factor $\eta > 0$. The ReLU term scales as $\Theta(\eta)$, while the regularization scales as $\Theta(\eta^2)$. In particular, for the isolating direction $x^{(j)}$ we can choose $\eta \to 0$ so that $\eta^2/g(\eta c_j) \to 0$, and the regularization becomes negligible compared to the (already tiny) ReLU loss. Writing N_j for the number of times a_j appears in the m i.i.d. samples, in the limit $\eta \to 0$, the required relative-error guarantee at x_j essentially reduces to $|N_j - \mathbb{E}N_j| \leq \varepsilon\, \mathbb{E}N_j$. Because scores are normalized, many points receive only $O(1/n)$ sampling probability in each draw. Therefore, ensuring relative error requires a coupon collector number of samples $m = \Omega(\varepsilon^{-2} n \log n)$.

B Preliminaries

B.1 Setting

We study the problem, where we are given i.i.d. sampling access a distribution $\mathcal{P}$ over $\mathbb{R}^d$. Further we fix a loss function $g\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$. We will consider the regularized loss function $f\colon \mathbb{R}^d \to \mathbb{R}_{\geq 0}$

$$f(x) \coloneqq \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \frac{\mathcal{R}(x)}{k}$$

with regularization parameter $k \in \mathbb{N}$ and one of the following regularizers $\mathcal{R}(x) \in \{\|x\|_2^2, \|x\|_2, \|x\|_1\}$. We further set $f_0(x) \coloneqq \int g(\langle a, x\rangle)\, d\mathcal{P}(a)$, and $g_0(r) = g(r) - g(0)$. We note that for many important functions, it holds that $g(0) \leq 1$.

For our sampling algorithms, we will define a suitable function $s\colon \mathbb{R}^d \to [0, \infty)$ such that $S = \int s(a)\, d\mathcal{P}(a) < \infty$. Given approximation and confidence parameters $\varepsilon, \delta \geq 0$, our goal is to draw a small number $m \in \mathbb{N}$ of points $a_1, \ldots, a_m \in \mathbb{R}^d$ via a probability distribution $\mathcal{Q}$ with density $d\mathcal{Q}(a) = (\frac{s(a)}{2S} + \frac{1}{2})\, d\mathcal{P}(a) = \frac{s(a)+S}{2S}\, d\mathcal{P}(a)$ such that if we reweight the contributions of point a_i

by $w_i = w(a_i) = \frac{2S}{s(a_i)+S}$, then it holds with probability at least $1-\delta$ that

$$\forall x \in \mathbb{R}^d \colon \left| f_0(x) - \frac{1}{m}\sum_{i=1}^{m} w_i g(\langle a_i, x\rangle) \right| \leq \varepsilon f(x)\,.$$

We note that $\frac{1}{m}\sum_{i=1}^{m} w_i \leq 2$, since each $w_i = \frac{2S}{s(a_i)+S} \leq \frac{2S}{S} = 2$.

We can draw from $\mathcal{Q}$ via a mixture of uniform and rejection sampling. I.e., we draw a from $\mathcal{P}$ and throw a coin. If the coin turns heads then we keep the sample and we are done. On tails, we decide with probability $s(a)/S$ to keep the sample or otherwise we reject the sample.

B.2 Assumptions and parameters

We make the following assumptions on the distribution $\mathcal{P}$ and the loss function for some parameters $L \geq 1, B \geq 0$.

[A1] g is L-Lipschitz

[A2] $\mathbb{E}_{a\sim\mathcal{P}}(\|a\|_2) = B < \infty$

We note that only in one case of $\mathcal{R}(x) = \|x\|_2^2$, and $g(x)$ with bounded derivatives (this will be made clear in Appendix D.3), we will need the slightly stronger assumption that $\mathbb{E}_{a\sim\mathcal{P}}(\|a\|_2^2) = B$. A similar assumption as Assumption [A2] is then implied by Jensen's inequality since $\mathbb{E}_{a\sim\mathcal{P}}(\|a\|_2)^2 \leq \mathbb{E}_{a\sim\mathcal{P}}(\|a\|_2^2) \leq B$. Thus $\mathbb{E}_{a\sim\mathcal{P}}(\|a\|_2) \leq \sqrt{B}$.

We note that in any case, we will use the upper bound $B \leq S$. This is valid since we will ensure that $s(a) \geq \|a\|_2 + 1$, resp. $s(a) \geq \|a\|_2^2 + 2$. So in the first case we have $S = B + 1 \geq B$, and in the second case, we have that either $B \geq 1$, so both $\sqrt{B} \leq B \leq B + 2 = S$, or if $B < 1$, then $B < \sqrt{B} < 1 < B + 2 = S$.

These fairly standard assumptions ensure that the regularization has an effect. The constants L and B will show up in our bounds. We note that we often have $L = B = 1$. In parts of the analysis it is even possible to simply assume $L = B = 1$, irrespective of the actual value and conduct the analysis with $k' = LBk$. This is shown in Lemma B.1 below to hold if $\mathcal{R}(x)$ is sub-multiplicative. It could also be adapted to work for $\mathcal{R}(x) = \|x\|_2^2$ with $\|x\|_2 \leq 1$, if $f(x) \geq c_0$ for all $x \in \mathbb{R}^d$.

Lemma B.1. *Let $f(x) := \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \mathcal{R}(x)/k$ and assume that g is L-Lipschitz, $\mathbb{E}_{a\sim\mathcal{P}}(\|a\|_2) \leq B$ and $\mathcal{R}(Bx) \leq B\mathcal{R}(x)$. Then there exists a function $g_2 = g/L$ that is 1-Lipschitz, such that if*

$$\forall x \in \mathbb{R}^d \colon \left| \int g_2(\langle a/B, x\rangle)\, d\mathcal{P}(a) - \frac{1}{m}\sum_{i=1}^{m} w_i g_2(\langle a_i/B, x\rangle) \right| \leq \varepsilon \left(\int g_2(\langle a/B, x\rangle)\, d\mathcal{P}(a) + \frac{\mathcal{R}(x)}{LBk} \right)\,,$$

then it also holds that

$$\forall x \in \mathbb{R}^d \colon \left| f_0(x) - \frac{1}{m}\sum_{i=1}^{m} w_i g(a_i x) \right| \leq \varepsilon f(x)\,.$$

Proof. We set $g_2(x) = g(x)/L$ and assume that for all $x \in \mathbb{R}^d$ it holds that

$$\left| \int g_2(\langle a/B, x\rangle)\, d\mathcal{P}(a) - \frac{1}{m}\sum_{i=1}^{m} w_i g_2(\langle a/B, x\rangle) \right| \leq \varepsilon \left(\int g_2(\langle a/B, x\rangle)\, d\mathcal{P}(a) + \frac{\mathcal{R}(x)}{LBk} \right).$$

Now, let $x \in \mathbb{R}^d$ be arbitrary. We set $x' = Bx$ and note that this is a one-to one-mapping. Then we have that

$$\begin{aligned} \left| f(x) - \frac{1}{m}\sum_{i=1}^{m} w_i g(a_i x) \right| &= \left| \int L g_2(\langle a/B, x'\rangle)\, d\mathcal{P}'(a) - \frac{L}{m}\sum_{i=1}^{m} w_i g_2(\langle a_i/B, x'\rangle) \right| \\ &\leq \varepsilon L \left(\int g_2(\langle a/B, x'\rangle)\, d\mathcal{P}(a) + \frac{\mathcal{R}(x')}{LBk} \right) \end{aligned}$$

$$\leq \varepsilon \left(\int g(ax)\, d\mathcal{P}(a) + \frac{\mathcal{R}(x)}{k} \right).$$

□

The following lemma states that we can estimate S up to a $(1 \pm \varepsilon)$ factor based on a small uniform sample. This will not be needed for taking importance samples but it will only be needed for determining the weights using some $\hat{S} = \Theta(S)$ in the distributional setting and requires that the distribution is bounded, i.e., $\Pr(\|a\|_2 > D) = 0$. We note that such assumption is also standard whenever uniform sampling from an unknown distribution is involved. We note that in a finite setting such as coresets, S can be determined exactly simply by summing the (squared) row norms and does not require this lemma.

Lemma B.2. *Assume we have i.i.d. sampling access to an unknown distribution $\mathcal{P}$ with $\Pr(\|a\|_2 > D) = 0$. Set $s(a) = \|a\|_2^p + 1$, for $p \in \{1, 2\}$. If $m \geq \frac{D^p \ln(1/\delta)}{\varepsilon^2}$, then we have for $S = \int s(a_i)\, d\mathcal{P}(a)$ with probability $1 - \delta$ that the empirical estimator satisfies*

$$\left| \frac{1}{m} \sum_{i=1}^{m} s(a_i) - S \right| \leq \varepsilon S\,.$$

Proof. First note that $\mathbb{E}_{a_i \sim \mathcal{P}} \left(\frac{1}{m} \sum_{i=1}^{m} s(a_i) \right) = S$, and recall that $S \geq 1$. We also calculate

$$\mathbb{V}(s(a_i)) = \mathbb{V}(\|a_i\|_2^p) \leq \mathbb{E}(\|a_i\|_2^{2p}) = \int \|a_i\|_2^{2p}\, d\mathcal{P}(a) \leq D^p \int \|a_i\|_2^p\, d\mathcal{P}(a) \leq D^p S.$$

By an application of Bernstein's inequality, we thus get

$$\begin{aligned} \Pr\left(\left| \frac{1}{m} \sum_{i=1}^{m} s(a_i) - \int s(a_i)\, d\mathcal{P}(a) \right| \geq \varepsilon S \right) &\leq \exp\left(-\frac{\varepsilon^2 m^2 S^2}{2mD^p S + \frac{2}{3} D^p m \varepsilon S} \right) \\ &\leq \exp\left(-\frac{\varepsilon^2 m}{3D^p} \right) \leq \delta \end{aligned}$$

□

Lemma B.3. *Let $\varepsilon \in (0, 1/2]$, and let $\hat{S} \in [1-\varepsilon, 1+\varepsilon] \cdot S$. Let $w(a) = \frac{2S}{s(a)+S}$, and $w'(a) = \frac{2\hat{S}}{s(a)+\hat{S}}$. Then it holds that*

$$w'(a) \in [1 - \varepsilon, 1 + \varepsilon] \cdot w(a).$$

Proof. We let $\eta \in [-\varepsilon, \varepsilon]$. By assumption it holds that $|\eta| \leq \varepsilon \leq 1/2$. Now fix $\hat{S} = (1 + \eta)S$.

We have that

$$\begin{aligned} \frac{|w'(a) - w(a)|}{w(a)} &= \left| \frac{w'(a)}{w(a)} - 1 \right| = \left| \frac{s(a) + S}{2S} \cdot \frac{(1+\eta)2S}{s(a) + (1+\eta)S} - 1 \right| \\ &= \left| \frac{s(a) + (1+\eta)S + \eta s(a)}{s(a) + (1+\eta)S} - 1 \right| = \left| \frac{\eta s(a)}{s(a) + (1+\eta)S} \right| = |\eta| \cdot \left| \frac{s(a)}{s(a) + (1+\eta)S} \right| \leq \varepsilon\,, \end{aligned}$$

where the last step follows since $2S \geq (1 + \eta)S \geq S/2 > 0$. □

B.3 Bounds on the optimal value

Parts of our analysis will depend on the optimal value $OPT_T = \inf_{x \in T} f(x)$. In particular this will be the case when the regularizer is $\mathcal{R}(x) = \|\cdot\|_2^2$, but we will also require a lower bound for $\mathcal{R}(x) = \|\cdot\|_1$. To handle these cases, we will need the following lemmas.

Lemma B.4. *For any $T \subseteq \mathbb{R}^d$, let $OPT_T = \inf_{x \in T} f(x)$. It holds that*

$$\forall x \in T\colon\ f(x) \geq \frac{OPT_T + \mathcal{R}(x)/k}{2}.$$

Proof. We have for every $x \in \mathbb{R}^d$ that $f(x) = f_0(x) + \mathcal{R}(x)/k \geq \mathcal{R}(x)/k$. Moreover we have by definition that $\forall x \in T\colon f(x) \geq OPT_T$. It follows for all $x \in T$ that $(OPT_T + \mathcal{R}(x)/k)/2 \leq 2f(x)/2 = f(x)$. □

We will denote the global optimum for $T = \mathbb{R}^d$ by $OPT \coloneqq OPT_{\mathbb{R}^d}$, and note that $OPT_T \geq OPT$ holds for any $T \subseteq \mathbb{R}^d$. We can derive a general bound on OPT based on $g(0)$, since in this case the regularization term is $R(0)/k = 0$.

Lemma B.5. *If $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ then $OPT \geq g(0)^2/(4(LB)^2k)$. If $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\}$ then $OPT \geq g(0)/(LBk)$.*

Proof. We first bound $f_0(x)$ using a Lipschitz expansion of g around 0.

$$\begin{aligned} f_0(x) &= \int g(\langle a, x\rangle)\, d\mathcal{P}(a) \\ &\geq \int \max\{0, (g(0) - L\|a\|_2\|x\|_2)\}\, d\mathcal{P}(a) && \text{Assumption [A1]; Cauchy-Schwarz} \\ &\geq \max\{0, g(0) - L\|x\|_2 \int \|a\|_2\, d\mathcal{P}(a)\} \\ &\geq \max\{0, g(0) - LB\|x\|_2\}. && \text{Assumption [A2]} \end{aligned}$$

For $\mathcal{R}(x) = \|x\|_2^2$, we let $r = \|x\|_2$. Then setting $h(r) = \max\{0, g(0) - LBr\} + r^2/k$, we have that

$$OPT = \inf_{x\in\mathbb{R}^d} f(x) = \inf_{x\in\mathbb{R}^d} f_0(x) + \|x\|_2^2/k \geq \inf_{x\in\mathbb{R}^d} h(\|x\|_2) \geq \inf_{r\in\mathbb{R}} h(r).$$

Using the derivative of $h(r)$, we observe two options for the minimizing r, depending on how the $\max$ within h evaluates. If $r \leq g(0)/(LB)$ then h is minimized at $r = LBk/2$ with $h(LBk/2) = g(0) - (LB)^2k/4 \geq (LB)^2k/4$. Otherwise if $r \geq g(0)/(LB)$, then h is minimized on the boundary where $r = g(0)/(LB)$, and $h(g(0)/LB) = \frac{g(0)^2}{(LB)^2k}$. The final claimed bound of OPT takes the minimum of these two options, and under the previous parameter assumptions, we get that

$$OPT \geq \min\{(LB)^2k/4, g(0)^2/((LB)^2k)\} \geq g(0)^2/(4(LB)^2k).$$

For the remaining cases $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\}$, we note that $\|x\|_1 \geq \|x\|_2$. It thus suffices to prove a bound for $\mathcal{R}(\cdot) = \|\cdot\|_2$. A similar analysis as above but with $h(r) = \max\{0, g(0) - LBr\} + r/k$ yields $OPT \geq g(0)/(LBk)$ in that case. □

C A quadratic upper bound depending on OPT

In this section, we prove a higher moment bound that will turn out to be quadratic in k in the worst case. However, we will see that in some settings considered in this paper, this bound is already tight and improves over previous cubic bounds.

On the other hand, for $\mathcal{R}(\cdot) = \|\cdot\|_2^2$, if we get the promise that OPT is bounded below by a constant, then it yields a linear k-dependence, which will be helpful to handle marginal cases in our improved upper bound analysis Appendix D.3.

C.1 Analysis

The proof of Theorem 3.1 is divided into multiple parts for the sake of a clear presentation. Our first lemma states that the expected value of our subsampled loss function is the same as its original value. The proof is straightforward but we include it for completeness and since it is an important piece of the puzzle.

Lemma C.1. *For any $x \in \mathbb{R}^d$ it holds that*

$$\mathbb{E}_{a\sim\mathcal{Q}^m}\left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i x) + \mathcal{R}(x)/k\right) = f(x)\,.$$

Proof. First note that

$$\begin{aligned}\mathbb{E}_{A\sim\mathcal{Q}^m}\left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i x)\right) &= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{a_i\sim\mathcal{Q}}(w_i g(a_i x)) \\ &= \mathbb{E}_{a\sim\mathcal{Q}}(w(a)g(ax)) \\ &= \int g(ax)w(a)\,d\mathcal{Q}(a) \\ &= \int g(ax)\frac{2S}{s(a)+S}\cdot\frac{s(a)+S}{2S}\,d\mathcal{P}(a) = f_0(x)\,.\end{aligned}$$

Thus, we have that

$$\begin{aligned}\mathbb{E}_{A\sim\mathcal{Q}^m}\left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i x)+\mathcal{R}(x)/k\right) &= \mathbb{E}_{A\sim\mathcal{Q}^m}\left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i x)\right)+\mathcal{R}(x)/k \\ &= f_0(x)+\mathcal{R}(x)/k = f(x)\,.\end{aligned}$$

□

We proceed by considering the relative approximation error

$$E \coloneqq \sup_{x\in\mathbb{R}^d}\left|f(x)-\left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i x)+\mathcal{R}(x)/k\right)\right| \Big/ f(x)\,.$$

We first show that for a suitable choice of ℓ, bounding the ℓ-th moment of the error term suffices to conclude the theorem.

Lemma C.2. *Assume that $\mathbb{E}(E^\ell) \leq (\varepsilon/2)^\ell$ for some even integer $\ell \geq 2\ln(\delta^{-1})$. Then with probability at least $1-\delta$ we have that $E < \varepsilon$*

Proof. Using Markov's inequality we have that

$$\Pr(E\geq\varepsilon) = \Pr(E^\ell\geq\varepsilon^\ell) \leq \frac{\mathbb{E}(E^\ell)}{\varepsilon^\ell} \leq 2^{-\ell} \leq \exp(-2\ln(2)\ln(\delta^{-1})) < \delta\,.$$

□

It remains to bound the ℓ-th moment $\mathbb{E}(E^\ell)$. Our first attempt is a Rademacher complexity bound that follows from a standard symmetrization argument involving Rademacher random variables.

Lemma C.3. *Let $\ell \in \mathbb{N}$ be an even integer. Then we have that*

$$\mathbb{E}(E^\ell) \leq 2^\ell \cdot \mathbb{E}_{a\sim\mathcal{Q}^m,\sigma\sim\mathcal{U}\{-1,1\}^m}\left(\sup_{x\in\mathbb{R}^d}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i x)\right|^\ell \Big/ f(x)^\ell\right).$$

Proof. We first show that $\mathbb{E}(E^\ell)$ is bounded by the expected difference between two independently drawn samples $a, a' \sim \mathcal{Q}^m$ with their corresponding weights $w, w' \in \mathbb{R}^m$. From Lemma C.1, we know that $\mathbb{E}_{a\sim\mathcal{Q}^m}(f_0(x) - \frac{1}{m}\sum_{i=1}^m w_i g(a_i x)) = 0$. We thus get

$$\begin{aligned}\mathbb{E}(E^\ell) &= \mathbb{E}_{a\sim\mathcal{Q}^m}\left(\sup_{x\in\mathbb{R}^d}\left|f(x)-\left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i x)+\mathcal{R}(x)/k\right)\right|^\ell \Big/ f(x)^\ell\right) \\ &= \mathbb{E}_{a\sim\mathcal{Q}^m}\left(\sup_{x\in\mathbb{R}^d}\left|f_0(x)-\frac{1}{m}\sum_{i=1}^m w_i g(a_i x)\right|^\ell \Big/ f(x)^\ell\right) \\ &= \mathbb{E}_{a\sim\mathcal{Q}^m}\left(\sup_{x\in\mathbb{R}^d}\left|f_0(x)-\frac{1}{m}\sum_{i=1}^m w_i g(a_i x)-\mathbb{E}_{a'\sim\mathcal{Q}^m}\left(f_0(x)-\frac{1}{m}\sum_{i=1}^m w'_i g(a'_i x)\right)\right|^\ell \Big/ f(x)^\ell\right)\end{aligned}$$

$$= \mathbb{E}_{a\sim\mathcal{Q}^m}\left(\sup_{x\in\mathbb{R}^d}\left|\mathbb{E}_{a'\sim\mathcal{Q}^m}\left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i x) - \frac{1}{m}\sum_{i=1}^m w_i' g(a_i' x)\right)\right|^\ell \Big/ f(x)^\ell\right)$$

$$\leq \mathbb{E}_{a,a'\sim\mathcal{Q}^m}\left(\sup_{x\in\mathbb{R}^d}\left|\frac{1}{m}\sum_{i=1}^m w_i g(a_i x) - \frac{1}{m}\sum_{i=1}^m w_i' g(a_i' x))\right|^\ell \Big/ f(x)^\ell\right).$$

The last inequality follows by Jensen's inequality since $\sup$ is a convex function.

Next, by a standard symmetrization argument (Ledoux and Talagrand, 1991; Cohen and Peng, 2015), we bound the last term back in terms of a single sample but with random Rademacher sign variables. Recall that ℓ is an even number. We have that

$$\mathbb{E}(E^\ell) \leq \mathbb{E}_{a,a'\sim\mathcal{Q}^m}\left(\sup_{x\in\mathbb{R}^d}\left|\frac{1}{m}\sum_{i=1}^m w_i g(a_i x) - \frac{1}{m}\sum_{i=1}^m w_i' g(a_i' x))\right|^\ell \Big/ f(x)^\ell\right)$$

$$= \mathbb{E}_{a,a'\sim\mathcal{Q}^m,\sigma\sim\mathcal{U}\{-1,1\}^m}\left(\sup_{x\in\mathbb{R}^d}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i(w_i g(a_i x) - w_i' g(a_i' x))\right|^\ell \Big/ f(x)^\ell\right)$$

$$\leq \mathbb{E}_{a,a'\sim\mathcal{Q}^m,\sigma\sim\mathcal{U}\{-1,1\}^m}\left(\sup_{x\in\mathbb{R}^d}\left(\left|\frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i x)\right| + \left|\frac{1}{m}\sum_{i=1}^m \sigma_i w_i' g(a_i' x))\right|\right)^\ell \Big/ f(x)^\ell\right)$$

$$\leq 2^\ell \cdot \mathbb{E}_{a\sim\mathcal{Q}^m,\sigma\sim\mathcal{U}\{-1,1\}^m}\left(\sup_{x\in\mathbb{R}^d}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i x)\right|^\ell \Big/ f(x)^\ell\right).$$

□

As a final step, we need to control the ℓ-th moment of our Rademacher bound. To this end, we set

$$\Lambda \coloneqq \sup_{x\in\mathbb{R}^d}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i x)\right| / f(x).$$

We need to upper bound $\mathbb{E}(\Lambda^\ell)$. To this end, we will use a Lipschitz contraction bound (Lemma F.5) proven in (Alishahi and Phillips, 2024) and the Khinchine-Kahane inequality (Proposition F.6 and Corollary F.7), which extends Khintchine's classic inequality (Khintchine, 1923) from sequences of complex numbers to elements of Banach spaces.

Lemma C.4. *If $s(a) \geq \|a\|_2$ holds for all $a \in \mathbb{R}^d$ then we have that*

$$\mathbb{E}(\Lambda^\ell) \leq \left(\frac{8\sqrt{\ell}}{\sqrt{m}}\right)^\ell \left(\frac{g(0)^\ell}{OPT^\ell} + (SL)^\ell \sup_{x\in\mathbb{R}^d}\left(\frac{2\|x\|_2}{OPT + \mathcal{R}(x)/k}\right)^\ell\right).$$

Proof. Recall that ℓ is an even number, and $s(a) \geq \|a\|_2 \geq 0$. Then we have $w_i = 2S/(s(a_i)+S) \leq \min\{2S/\|a\|_2, 2\}$ and thus

$$\mathbb{E}(\Lambda^\ell) = \mathbb{E}\sup_{x\in\mathbb{R}^d}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i x)\right|^\ell \Big/ f(x)^\ell$$

$$= \left(\frac{1}{m}\right)^\ell \mathbb{E}\sup_{x\in\mathbb{R}^d}\left|\sum_{i=1}^m \sigma_i w_i (g(0) + g_0(a_i x))\right|^\ell \Big/ f(x)^\ell$$

$$\overset{Jensen}{\leq} \left(\frac{2}{m}\right)^\ell \mathbb{E}\sup_{x\in\mathbb{R}^d}\left(\left|\sum_{i=1}^m \sigma_i w_i g(0)\right|^\ell + \left|\sum_{i=1}^m \sigma_i w_i g_0(a_i x)\right|^\ell\right) \Big/ f(x)^\ell$$

$$\overset{Lemma\ F.5}{\leq} \left(\frac{2}{m}\right)^\ell \mathbb{E}\sup_{x\in\mathbb{R}^d}\left((2g(0))^\ell \left|\sum_{i=1}^m \sigma_i\right|^\ell + (2L)^\ell \left|\sum_{i=1}^m \sigma_i w_i a_i x\right|^\ell\right) \Big/ f(x)^\ell$$

$$
\begin{aligned}
&\overset{CSI,\,Lemma\ B.4}{\leq} \left(\frac{4}{m}\right)^\ell \left(\frac{g(0)^\ell\, \mathbb{E}\left|\sum_{i=1}^m \sigma_i\right|^\ell}{OPT^\ell} + L^\ell\, \mathbb{E} \sup_{x\in\mathbb{R}^d} \left(\frac{2\|x\|_2}{OPT+\mathcal{R}(x)/k}\right)^\ell \left\|\sum_{i=1}^m \sigma_i w_i a_i\right\|_2^\ell \right) \\
&\overset{Corollary\ F.7}{\leq} \left(\frac{4}{m}\right)^\ell \ell^{\ell/2} \left(\frac{g(0)^\ell\, m^{\ell/2}}{OPT^\ell} + L^\ell \sup_{x\in\mathbb{R}^d} \left(\frac{2\|x\|_2}{OPT+\mathcal{R}(x)/k}\right)^\ell \mathbb{E}\left(\sum_{i=1}^m w_i^2 \left\|a_i\right\|_2^2\right)^{\ell/2} \right) \\
&\leq \left(\frac{4\sqrt{\ell}}{m}\right)^\ell \left(\frac{g(0)^\ell m^{\ell/2}}{OPT^\ell} + L^\ell \sup_{x\in\mathbb{R}^d} \left(\frac{2\|x\|_2}{OPT+\mathcal{R}(x)/k}\right)^\ell (2S)^\ell m^{\ell/2} \right) \\
&\leq \left(\frac{8\sqrt{\ell}}{\sqrt{m}}\right)^\ell \left(\frac{g(0)^\ell}{OPT^\ell} + (SL)^\ell \sup_{x\in\mathbb{R}^d} \left(\frac{2\|x\|_2}{OPT+\mathcal{R}(x)/k}\right)^\ell \right)
\end{aligned}
$$

Here the first inequality follows by Jensen's inequality. The next inequality uses $w_i \leq 2$ and further the application of Lemma F.5 uses $\phi_i(r) = w_i g_0(r/w_i)$, $t_i = w_i a_i x$ and $T = \{Ax \mid x \in \mathbb{R}^d\}$. Note that T is not bounded but since $w_i g_0(a_i x)/f(x)^\ell$ goes to 0 for $\|x\|_2 \to \infty$ we can apply the lemma to any bounded subset of T to get the unbounded version. The subsequent (in)equalities follow in elementary steps by linearity of expectation, the Cauchy-Schwarz inequality, Lemma B.4, followed by the Khintchine-Kahane inequality (see Corollary F.7) on both expected values and $w_i \leq 2S/\|a_i\|_2$. □

It now remains to bound the supremum term of Lemma C.4 depending on the choice of the regularizer. We have the following lemma.

Lemma C.5. *It holds that*

$$
\sup_{x\in\mathbb{R}^d} \left(\frac{2\|x\|_2}{OPT+\mathcal{R}(x)/k}\right) \leq \begin{cases} \sqrt{k/OPT} & ,\ \text{if } \mathcal{R}(\cdot) = \|\cdot\|_2^2 \\ 2k & ,\ \text{if } \mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\} \end{cases}
$$

Proof. For the first item, we need a bound on $\sup_{x\in\mathbb{R}^d} \frac{2\|x\|_2}{OPT+\|x\|_2^2/k}$. To this end, we consider the function $h(r) = \frac{r}{OPT+r^2/k} = \frac{1}{OPT/r+r/k}$ for $r \geq 0$. Note, that it is maximized if $OPT/r + r/k$ is minimized, which is the case when its derivative satisfies $1/k - OPT/r^2 = 0$. Rearranging leads to $r = \sqrt{k\,OPT}$.

This implies that

$$
\sup_{x\in\mathbb{R}^d} \frac{2\|x\|_2}{OPT+\|x\|_2^2/k} \leq \frac{2\sqrt{k\,OPT}}{OPT+OPT} = \frac{\sqrt{k}}{\sqrt{OPT}}\,. \tag{5}
$$

For the second item, recall that $\|x\|_1 \geq \|x\|_2$ and $OPT > 0$. Thus

$$
\sup_{x\in\mathbb{R}^d} \left(\frac{2\|x\|_2}{OPT+\|x\|_1/k}\right) \leq \sup_{x\in\mathbb{R}^d} \left(\frac{2\|x\|_2}{OPT+\|x\|_2/k}\right) \leq \sup_{x\in\mathbb{R}^d} \left(\frac{2\|x\|_2}{\|x\|_2/k}\right) = 2k\,.
$$

□

C.2 Proof of the main sampling result

Theorem 3.1. *Let $f(x) \coloneqq \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \mathcal{R}(x)/k$ such that g is L-Lipschitz. Let $s(a) \geq \|a\|_2 + 1$ for all $a \in \mathbb{R}^d$. Let $T \subseteq \mathbb{R}^d$, and $OPT_T = \inf_{x\in T} f(x)$. Set $C = (SL)^2$. There exists*

$$
m_0 = \begin{cases} O(C\varepsilon^{-2} k \ln(\delta^{-1})/OPT_T), & \text{for } \mathcal{R}(\cdot) = \|\cdot\|_2^2 \\ O(C\varepsilon^{-2} k^2 \ln(\delta^{-1})), & \text{for } \mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\}, \end{cases}
$$

such that if we sample $m \geq m_0$ points according to $\mathcal{Q}$, then with probability at least $1-\delta$ we have

$$
\forall x \in T\colon \left| f_0(x) - \frac{1}{m}\sum_{i=1}^m w_i g(\langle a_i, x\rangle) \right| \leq \varepsilon f(x)\,.
$$

Proof. First assume that $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\}$. By chaining the arguments of Lemmas C.3, C.4 and C.5, and using that $OPT \geq g(0)/(LBk)$ by Lemma B.5, we get for sufficiently large absolute constants $C'' > C' > 1$ that

$$\mathbb{E}(E^\ell) \leq \left(\frac{C'\sqrt{\ell}}{\sqrt{m}}\right)^\ell \left(\frac{g(0)^\ell}{OPT^\ell} + (SLk)^\ell\right) \leq \left(\frac{C'\sqrt{\ell}}{\sqrt{m}}\right)^\ell \left((BLk)^\ell + (SLk)^\ell\right) \leq \left(\frac{C''SLk\sqrt{\ell}}{\sqrt{m}}\right)^\ell .$$

As a consequence, setting $\ell = 2\lceil \ln(\delta^{-1})\rceil$ and $m = (C'')^2 S^2 L^2 k^2 \ell/\varepsilon^2 = O(S^2L^2\varepsilon^{-2}k^2\log(\delta^{-1}))$, the final moment bound becomes $\mathbb{E}(E^\ell) \leq (\varepsilon/2)^\ell$ and we can conclude by applying Lemma C.2.

For the case $\mathcal{R}(\cdot) = \|\cdot\|_2^2$, we can use the appropriate bound of Lemma C.5 and the lower bound $OPT \geq g(0)^2/(4(BL)^2k)$ of Lemma B.5. We thus get for sufficiently large absolute constants $C'' > C' > 1$ that

$$\begin{aligned}\mathbb{E}(E^\ell) &\leq \left(\frac{C'\sqrt{\ell}}{\sqrt{m}}\right)^\ell \left(\frac{g(0)^\ell}{OPT^\ell} + SL^\ell \frac{k^{\ell/2}}{OPT^{\ell/2}}\right) = \left(\frac{C'\sqrt{\ell}}{\sqrt{m}}\right)^\ell \left(\frac{g(0)^\ell}{OPT^{\ell/2}} \cdot \frac{1}{OPT^{\ell/2}} + (SL)^\ell \frac{k^{\ell/2}}{OPT^{\ell/2}}\right) \\ &\leq \left(\frac{C'\sqrt{\ell}}{\sqrt{m}}\right)^\ell \left((2BL)^\ell \cdot \frac{k^{\ell/2}}{OPT^{\ell/2}} + (SL)^\ell \frac{k^{\ell/2}}{OPT^{\ell/2}}\right) \leq \left(\frac{C''SL\sqrt{k\ell}}{\sqrt{mOPT}}\right)^\ell .\end{aligned}$$

Setting $\ell = 2\lceil \ln(\delta^{-1})\rceil$ and $m = (C'')^2 S^2 L^2 k\ell/(\varepsilon^2 OPT) = O(S^2L^2\varepsilon^{-2}k\log(\delta^{-1})/OPT)$, the final moment bound becomes $\mathbb{E}(E^\ell) \leq (\varepsilon/2)^\ell$ and we can conclude by applying Lemma C.2. □

D Improving the upper bound to linear

In this section, we present an improved analysis that yields linear dependence on k. This is achieved in two different settings:

1. $\|\cdot\|_2^2$-regularization when g is either monotonic or convex/concave and has a bounded derivative, i.e., $|g'| \leq g$,
2. general functions with $\|\cdot\|_1$-regularization.

We start wit a few preliminaries on Gaussian processes and their analysis.

D.1 Moment bound for Gaussian processes

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

Lemma D.1 (Woodruff and Yasuda, 2023, slightly modified)**.** *Let $(X_t)_{t\in T}$ be a Gaussian process with pseudo-metric $d_X(s,t) := \|X_s - X_t\|_2 = \sqrt{\mathbb{E}(X_s - X_t)^2}$. Further, we 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 $\int_0^\infty \sqrt{\log N(T, d_X, u)}\, du \leq \mathcal{E}$ be a bound on the metric entropy. Then*

$$\mathbb{E}[\sup_{t\in T}|X_t|^\ell] \leq (2\mathcal{E})^\ell(\mathcal{E}/\mathcal{D}) + O(\sqrt{\ell}\mathcal{D})^\ell.$$

The only modification is that we consider a more general Λ than Woodruff and Yasuda (2023). 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. Our task reduces to obtaining best possible bounds for the diameter $\mathcal{D}$ and the entropy $\mathcal{E}$ to allow a small sample size.

D.2 Sudakov's bounds on the covering numbers of ℓ_2 balls

In this section we give some basic definitions and results regarding covering numbers.

Definition D.2. Let $(V, \|\cdot\|)$ be a normed space and $W \subseteq V$. Then $C \subseteq W$ is a t-covering of W of size $n = |C|$ if $\forall w \in W\ \exists c \in C\colon \|w - c\| \leq t$. The covering number of covering W with balls of radius t with respect to $\|\cdot\|$ is defined as $N(W, \|\cdot\|, t) = \min\{n \in \mathbb{N} \mid \exists t\text{-covering of } W \text{of size } n\}$.

Definition D.3. Let $\|\cdot\|_X$ be a norm. Then, we define

$$M_X \coloneqq \mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,I_d)} \|g\|_X \,.$$

The following results will help us bounding covering numbers for the cases where T are balls with respect to the Euclidean or the ℓ_1 norm. Bounds on M_X imply bounds for covering the Euclidean ball by $\|\cdot\|_X$-balls using the following result:

Proposition D.4 (Sudakov minoration, Sudakov, 1971, cf. Proposition 4.1 of Bourgain et al., 1989)**.** *Let $\|\cdot\|_X$ be a norm, and let $\|\cdot\|_{X^*}$ be the dual norm with respect to the Euclidean inner product. Let $B_2 \subseteq \mathbb{R}^d$ denote the Euclidean unit ball in d dimensions. Then*

$$\log N(\|\cdot\|_X, B_2, t) \leq O(1)\frac{M_{X^*}^2}{t^2} \,.$$

Proposition D.5 (Dual Sudakov minoration, Proposition 4.2 of Bourgain et al., 1989)**.** *Let $\|\cdot\|_X$ be a norm, and let $B_2 \subseteq \mathbb{R}^d$ denote the Euclidean unit ball in d dimensions. Then*

$$\log N(B_2, \|\cdot\|_X, t) \leq O(1)\frac{M_X^2}{t^2} \,.$$

We will also need a standard covering argument:

Lemma D.6.

$$\log N(B_2, \|\cdot\|_2, t) \leq d\log(1+2/t) \,.$$

Proof. Consider a maximal set $C = \{x_1, \ldots, x_2\} \subseteq B_2$ such that $\min_{i\neq j} \|x_i - x_j\|_2 > t/2$. Then C is a t-covering of B_2, otherwise there exists $x \in B_2 \setminus C$ such that $\forall c \in C\colon \|x-c\|_2 \geq t$ contradicting the maximality of C. We have that $\bigcup(x_i + t/2)B_2 \subseteq (1+t/2)B_2$. Thus, we have by a standard volume argument $N(B_2, \|\cdot\|_2, t) \leq \frac{\mathrm{vol}((1+t/2)B_2)}{\mathrm{vol}\,((t/2)B_2)} \leq (1+2/t)^d$. The Lemma follows by taking logs on both sides. □

D.3 Sampling bound for monotonic, convex or concave functions with bounded derivative and ridge regularization

In this subsection we study the setting where

$$f(x) = \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \frac{\|x\|_2^2}{k}$$

for some $k \in \mathbb{N}$. Set $f_0(x) = \int g(\langle a, x\rangle)\, d\mathcal{P}(a)$. Let $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be convex or concave or monotonic and satisfy $|g'(r)| \leq g(r)$ for all $r \in \mathbb{R}$.

Further, we assume that g is L-Lipschitz (Assumption [A1]) and that $\mathbb{E}_{a\sim\mathcal{P}}(\|a\|_2^2) = B < \infty$ for some $L \geq 1, B \geq 0$, which is slightly stronger as Assumption [A2] before as discussed in Appendix B.2. We will use $s(a) \geq \|a\|_2^2 + 2$, which is squared norm sampling and can be handled analogously to the previous norm sampling. Indeed, it implies $s(a) \geq \|a\|_2 + 1$ as before and under the strengthened assumption, if we set $s(a) = \|a\|_2^2 + 2$, we have that $B \leq S = B + 2$, is in particular independent of k.

Fix some subset $T \subseteq \mathbb{R}^d$. Given a sample S we further set $G_S = \sup_{x\in T} 1 + |f_S(x) - f_0(x)|/f(x)$ where $f_S(x) = \frac{1}{m}\sum_{i=1}^m w_i g(a_i x)$. Note that $\sqrt{G_S} \leq G_S$ since $G_S \geq 1$.

We will frequently use the following bound.

Lemma D.7. *It holds for any $T \subseteq \mathbb{R}^d$ and all $x \in T$ that $\frac{f_S(x)}{f(x)} \leq G_S$.*

Proof. It always holds that

$$\frac{f_S(x)}{f(x)} - \frac{f_0(x)}{f(x)} \leq \left|\frac{f_S(x)}{f(x)} - \frac{f_0(x)}{f(x)}\right| \,.$$

Thus,

$$\frac{f_S(x)}{f(x)} \leq 1 + \left|\frac{f_S(x)}{f(x)} - \frac{f_0(x)}{f(x)}\right| \leq \sup_{x\in T} 1 + \frac{|f_S(x) - f_0(x)|}{f(x)} = G_S \,.$$

□

A final (very weak) bound will be required later to claim that $G_S \leq \mathrm{poly}(m)$.

Lemma D.8. *It holds that*

$$G_S \leq \begin{cases} 18(LS)^3 k/g(0)^2, & \text{for } \mathcal{R}(\cdot) = \|\cdot\|_2^2 \\ 6LSk, & \text{for } \mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\}. \end{cases}$$

Proof. We will see in the course of the analysis, that since $\|x\|_1 \geq \|x\|_2$, it suffices to show the claim for $\|x\|_2^p$, for $p \in \{1, 2\}$. We start with a common analysis and then make a case distinction depending on p in the final derivations.

Recall that $w_i = \frac{2S}{s(a_i)+S} \leq \min\{2, 2S/\|a_i\|_2\}$. The two bounds of Lemma B.5 combined yield $f(x) \geq \frac{g(0)^p}{4^{p-1}(BL)^p k}$.

It thus holds that

$$\begin{aligned} G_S &= 1 + \sup_{x \in T} \frac{|f_S(x) - f_0(x)|}{f(x)} \\ &\leq 1 + \sup_{x \in T} \frac{|f_0(x)|}{f(x)} + \sup_{x \in T} \frac{|f_S(x)|}{f(x)} \\ &\leq 2 + \sup_{x \in T} \frac{|f_S(x)|}{f(x)} \\ &\leq 2 + \sup_{x \in T} \frac{1}{m} \sum_{i=1}^m w_i \frac{g(0) + g_0(a_i x)}{f(x)} \\ &\leq 2 + \frac{2g(0)}{f(x)} + \sup_{x \in T} \frac{L}{m} \sum_{i=1}^m w_i \frac{\|a_i\|_2 \|x\|_2}{f(x)} \\ &\leq 2 + \frac{2 \cdot 4^{p-1}(LB)^p k}{g(0)^{p-1}} + 2 \sup_{x \in T} \frac{LS}{m} \sum_{i=1}^m \frac{\|x\|_2}{f(x)} \,. \end{aligned}$$

For $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\}$, we have that $f(x) \geq \frac{\|x\|_1}{k} \geq \frac{\|x\|_2}{k}$. The above derivation with $p = 1$ yields

$$G_S \leq 2 + 2LBk + 2LSk \leq 6LSk \,.$$

For $\mathcal{R}(\cdot) = \|\cdot\|_2^2$, we consider two sub-cases. If $\|x\|_2 \geq 1$, then $\|x\|_2^2 \geq \|x\|_2$ and we can simply apply the same continuation with $p = 1$ to get $G_S \leq 6LSk$.

In the remaining case $\|x\|_2 \leq 1$, we need to set $p = 2$. Our derivation yields

$$G_S \leq 2 + \frac{8(LB)^2 k}{g(0)} + \frac{2LS}{f(x)} \leq 2 + \frac{8(LB)^2 k}{g(0)} + \frac{8LS(LB)^2 k}{g(0)^2} \leq \frac{18(LS)^3 k}{g(0)^2} \,.$$

□

A standard symmetrization argument yields the following bound that relates higher moments of the error $E = G_S - 1$ to a Gaussian process defined over T.

Lemma D.9. *It holds that* $\mathbb{E}(G_S - 1)^\ell \leq (2\pi)^{\ell/2} \mathbb{E}(\Lambda^\ell)$ *where* $\Lambda = \sup_{t \in T} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i w_i g(a_i t) \right| / f(t)$*, and* σ_i *are i.i.d standard Gaussian random variables.*

Proof. Note that $G_S - 1 = E$ as used in Lemma C.3, which gives us

$$\mathbb{E}((G_S - 1)^\ell) \leq 2^\ell \cdot \mathbb{E}_{a \sim \mathcal{Q}^m, \sigma \sim \mathcal{U}\{-1,1\}^m} \left(\sup_{x \in \mathbb{R}^d} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i w_i g(a_i x) \right|^\ell \Big/ f(x)^\ell \right).$$

We then convert the σ_i from Rademacher to Gaussian random variables, which takes an additional factor $(\frac{\pi}{2})^{\ell/2}$, see Eq. (4.8) of (Ledoux and Talagrand, 1991). Overall this changes the constant factor from 2^ℓ to $2^\ell \cdot (\frac{\pi}{2})^{\ell/2} = (2\pi)^{\ell/2}$. □

We will analyze the error in slices of similar function values. To this end, we set $T_\alpha = \{x \in \mathbb{R}^d \mid f(x) \in [\alpha, 2\alpha)\}$. Our analysis is first restricted to such a slice, but we will later union bound over a suitably small number of different values of α to extend to the entire range of function values. We note that for any $x \in T_\alpha$ it holds that $\|x\|_2^2 \leq 2\alpha k$ and thus $T_\alpha \subseteq (2\alpha k)^{1/2} B_2$. For now, we consider the canonical Gaussian process $X_s = \frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i x)$ for $x \in T_\alpha$ for some fixed $\alpha \in \mathbb{R}$, where σ_i denote i.i.d. standard Gaussian random variables.

This process induces a metric d_X over T_α that is given by

$$\begin{aligned} d_X(x,x') &= \left(\mathop{\mathbb{E}}_{\sigma} \left(\frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i x) - \frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i x') \right)^2 \right)^{1/2} \\ &= \left(\frac{1}{m^2}\sum_{i=1}^m w_i^2 \left(g(a_i x) - g(a_i x')\right)^2 \right)^{1/2}, \end{aligned}$$

where we used the fact that σ_i are independent zero-mean and unit-variance random variables. This implies for $i \neq j$ that $\mathbb{E}(\sigma_i \sigma_j) = \mathbb{E}(\sigma_i)\,\mathbb{E}(\sigma_j) = 0$. Thus only the squared terms survive and we can use $\mathbb{E}(\sigma_i^2) = 1$ to conclude.

D.3.1 Bounding the diameter

We first relate the diameter to the sensitivity. This will be reused in the case of $\|\cdot\|_1$ regularization.

Lemma D.10. *It holds that* $\sup_{s,t\in T_\alpha} d_X(s,t) \leq 2G_S\alpha \left(\frac{2}{m}\sup_{x\in T_\alpha, j\in[m]} \frac{w_j g(a_j x)}{f(x)}\right)^{1/2}$.

Proof. Note that for all $x \in \mathbb{R}^d$ we have $g(a_i s) \geq 0$. Then

$$\begin{aligned} \sup_{s,t\in T_\alpha} d_X(s,t) &= \sup_{s,t\in T_\alpha} \left(\frac{1}{m^2}\sum_{i=1}^m w_i^2 \left(g(a_i s) - g(a_i t)\right)^2 \right)^{1/2} \\ &= \sup_{s,t\in T_\alpha} \left(\frac{1}{m^2}\sum_{i=1}^m w_i^2 \left(g(a_i s)^2 - 2\, g(a_i s) g(a_i t) + g(a_i t)^2\right) \right)^{1/2} \\ &\leq \sup_{s,t\in T_\alpha} \left(\frac{1}{m^2}\sum_{i=1}^m w_i^2 \left(g(a_i s)^2 + g(a_i t)^2\right) \right)^{1/2} \\ &= \sup_{x\in T_\alpha} \left(\frac{2}{m^2}\sum_{i=1}^m w_i^2 g(a_i x)^2 \right)^{1/2} \\ &\leq \sup_{x\in T_\alpha} \left(\frac{2}{m} \max_{j\in[m]} w_j g(a_j x) \frac{1}{m}\sum_{i=1}^m w_i g(a_i x) \right)^{1/2} \\ &\leq \sup_{x\in T_\alpha} \left(\frac{2}{m} \max_{j\in[m]} w_j g(a_j x) \cdot G_S f(x) \right)^{1/2} \\ &= \sup_{x\in T_\alpha} \left(\frac{2}{m} \max_{j\in[m]} \frac{w_j g(a_j x)}{f(x)} \cdot G_S f(x)^2 \right)^{1/2} \\ &\leq 2G_S\alpha \left(\frac{2}{m} \sup_{x\in T_\alpha, j\in[m]} \frac{w_j g(a_j x)}{f(x)} \right)^{1/2}. \end{aligned}$$

□

It remains to bound the sensitivity.

Lemma D.11. *Let $w_i = \frac{2S}{s(a_i)+S}$. Then for any $i \in [m]$ and any $x \in T_\alpha$ it holds that*

$$\frac{w_i g(a_i x)}{f(x)} \leq \frac{16SBL^2k}{g(0)}.$$

Proof. Recall that $s(a_i) \geq \|a_i\|_2^2 + 1 \geq \max\{\|a_i\|_2, 1\}$. Also recall that $\|x\|_2 \leq \sqrt{2\alpha k}$ and $f(x) \geq \alpha \geq \frac{g(0)^2}{4(BL)^2 k}$. Thus, it holds that

$$\begin{aligned}
\frac{w_i g(a_i x)}{f(x)} &= \frac{w_i(g(0) + g_0(a_i x))}{f(x)} \leq \frac{2S}{s(a_i)+S} \cdot \frac{g(0) + L|a_i x|}{f(x)} \\
&\overset{CSI}{\leq} \frac{2S}{S}\frac{g(0)}{f(x)} + \frac{2S}{\|a_i\|_2}\frac{L\|a_i\|_2\|x\|_2}{f(x)} \leq 2\left(\frac{4(BL)^2k}{g(0)} + \frac{SL\|x\|_2}{f(x)}\right) \\
&\leq 2\left(\frac{4(BL)^2k}{g(0)} + \frac{SL\sqrt{2k}}{\sqrt{\alpha}}\right) \leq 2\left(\frac{4(BL)^2k}{g(0)} + \frac{4SBL^2k}{g(0)}\right) \leq \frac{16SBL^2k}{g(0)}. \qquad \square
\end{aligned}$$

The following lemma shows that the sensitivity bound (and thus using Lemma D.10 also the diameter bound) can be improved to be linear when we have a small $B = O(\log k)$.

Lemma D.12. *Let $w_i = \frac{2S}{s(a_i)+S}$ and assume that $g(r) \geq |g'(r)|$ for all $r \in \mathbb{R}$. Then for any $i \in [m]$ and any $x \in T_\alpha$ it holds that*

$$\frac{w_i g(a_i x)}{f(x)} \leq 4SL \max\left\{k, \frac{\exp(B)}{g(0)}\right\}$$

Proof. Recall that $s(a_i) \geq \|a_i\|_2^2 + 2 \geq \|a_i\|^2 + 1 \geq \max\{\|a_i\|_2, 1\}$. First, note that our assumption on the derivative implies that $g(r) \geq g(0)\exp(-r)$.

To see this, assume $g(0) > 0$, otherwise the statement is trivial. Start with $g(r) \geq |g'(r)| \geq -g'(r)$, thus $\frac{g'(r)}{g(r)} \geq -1$. It follows that

$$\ln(g(r)) - \ln(g(0)) = \int_0^r \frac{g'(x)}{g(x)}\,dx \geq -\int_0^r 1 = -r\,.$$

Taking the exponential function on both sides and rearranging, gives

$$\begin{aligned}
&\frac{g(r)}{g(0)} \geq \exp(-r) \\
\iff \quad &g(r) \geq g(0)\exp(-r)\,. \qquad (6)
\end{aligned}$$

Now, for $\|x\|_2 \leq 1$ we have that

$$\begin{aligned}
f(x) &= \int g(\langle a, x\rangle)\,d\mathcal{P}(a) + \frac{\|x\|_2^2}{k} \\
&\overset{Equation\ (6)}{\geq} \int \exp(-\langle a, x\rangle) g(0)\,d\mathcal{P}(a) \\
&\overset{Jensen}{\geq} g(0)\exp\left(-\int |\langle a, x\rangle|\,d\mathcal{P}(a)\right) \\
&\overset{CSI}{\geq} g(0)\exp\left(-\int \|a\|_2\|x\|_2\,d\mathcal{P}(a)\right) \geq g(0)\exp(-B).
\end{aligned}$$

and thus

$$\begin{aligned}
\frac{w_i g(a_i x)}{f(x)} &= \frac{w_i(g(0) + g_0(a_i x))}{f(x)} \leq \frac{2S}{s(a_i)+S} \cdot \frac{g(0) + L|\langle a_i, x\rangle|}{f(x)} \\
&\overset{CSI}{\leq} \frac{2g(0) + 2SL\|x\|_2}{g(0)\exp(-B)} \leq 2\left(\exp(B) + \frac{SL}{g(0)\exp(-B)}\right) \leq \frac{2SL\exp(B)}{g(0)}\,.
\end{aligned}$$

For any x with $\|x\|_2 \geq 1$ we have that $\|x\|_2^2 \geq \|x\|_2$. Therefore $f(x) = f_0(x) + \|x\|_2^2/k \geq f_0(x) + \|x\|_2/k$. Thus $f(x) \geq g(0)/(LBk)$ by Lemma B.5. Then,

$$\frac{w_i g(a_i x)}{f(x)} = \frac{w_i(g(0) + g_0(a_i x))}{f(x)} \leq \frac{2S}{s(a_i) + S} \cdot \frac{g(0) + L|a_i x|}{f(x)}$$
$$\overset{CSI}{\leq} \frac{2S}{S}\frac{g(0)}{f(x)} + \frac{2S}{\|a_i\|_2}\frac{L\|a_i\|_2\|x\|_2}{f(x)} \leq 2BLk + 2SLk \leq 4SLk\,.$$

□

We conclude the following two bounds on the diameter.

Corollary D.13. *It holds that* $\sup_{s,t \in T_\alpha} d_X(s,t) \leq 12 G_S \alpha \left(\frac{SBL^2k}{mg(0)}\right)^{1/2}$.

Proof. The bound follows directly from using the sensitivity bound of Lemma D.11 with Lemma D.10 □

Corollary D.14. *It holds that* $\sup_{s,t \in T_\alpha} d_X(s,t) \leq 6 G_S \alpha \left(\frac{SL(k+\exp(B))}{mg(0)}\right)^{1/2}$.

Proof. The bound follows directly from using the sensitivity bound of Lemma D.12 with Lemma D.10 □

D.3.2 Relating the metric to a norm

To bound the entropy, we first bound the metric d_X by a norm $\|\cdot\|_X$ that will be used in subsequent steps.

Lemma D.15. *Assume that g is either convex or concave or monotonic and for all r we have that $|g'(r)| \leq g(r)$. Then for any $s, t \in T_\alpha$ it holds that*

$$d_X(s,t) \leq \left(4 G_S \alpha \max_{j \in [m]} w_j |\langle a_j, s - t\rangle|^2 L/m\right)^{1/2}.$$

Proof. We have

$$\begin{aligned}
d_X(s,t)^2 &= \frac{1}{m^2}\sum_{i=1}^m w_i^2 \left(g(a_i s) - g(a_i t)\right)^2 \\
&= \frac{1}{m^2}\sum_{i=1}^m w_i^2 \left|g(a_i s) - g(a_i t)\right| \cdot \left|g(a_i s) - g(a_i t)\right| \\
&= \frac{1}{m^2}\sum_{i=1}^m w_i^2 \left|\int_{a_i t}^{a_i s} g'(x)\,dx\right| \cdot \left|g(a_i s) - g(a_i t)\right| \\
&\leq \frac{1}{m^2}\sum_{i=1}^m w_i^2 |a_i s - a_i t| \cdot \left(\max_{x \in [a_i t, a_i s]} |g'(x)|\right) \cdot \left|g(a_i s) - g(a_i t)\right| \qquad (*)
\end{aligned}$$

If the function is convex or concave, then the derivative is monotonic and the maximum is attained on the boundary of the interval. Thus,

$$(*) \leq \frac{1}{m^2}\sum_{i=1}^m w_i^2 |a_i s - a_i t| \left(|g'(a_i s)| + |g'(a_i t)|\right) \cdot \left|g(a_i s) - g(a_i t)\right| \qquad (**)$$

Alternatively, if the function itself is monotonic we first use the derivative bound and then the maximum is again attained on the boundary of the interval. Thus,

$$(*) \leq \frac{1}{m^2}\sum_{i=1}^m w_i^2 |a_i s - a_i t| \cdot \left(\max_{x \in [a_i t, a_i s]} |g(x)|\right) \cdot \left|g(a_i s) - g(a_i t)\right|$$

$$\leq \frac{1}{m^2}\sum_{i=1}^{m} w_i^2 |a_i s - a_i t| \left(|g(a_i s)| + |g(a_i t)|\right) \cdot |g(a_i s) - g(a_i t)| \qquad (**)$$

Finally we can finish both derivations as follows.

$$\begin{aligned}
(**) &\leq \frac{L}{m^2}\sum_{i=1}^{m} w_i^2 |a_i s - a_i t|^2 \left(|g(a_i s)| + |g(a_i t)|\right) \\
&\leq \frac{L}{m} \max_{j\in[m]} w_j |\langle a_j, s-t\rangle|^2 \cdot \left(\frac{1}{m}\sum_{i=1}^{m} w_i g(a_i s) + \frac{1}{m}\sum_{i=1}^{m} w_i g(a_i t)\right) \\
&\leq \frac{L}{m} \max_{j\in[m]} w_j |\langle a_j, s-t\rangle|^2 \cdot G_S(f(s)+f(t)) \\
&\leq 4 G_S \alpha \max_{j\in[m]} w_j |\langle a_j, s-t\rangle|^2 L/m.
\end{aligned}$$

□

D.3.3 Bounding the covering numbers

Clearly $\|\cdot\|_X := \left(4G_S\alpha \max_{j\in[m]} w_j \left|\langle a_j, \cdot\rangle\right|^2 L/m\right)^{1/2}$ defines a norm. We will use the following lemma to bound the covering numbers.

Lemma D.16. *Assume that g is either convex or concave or monotonic and for all r we have that $|g'(r)| \leq g(r)$. Then for any radius $t > 0$ it holds that*

$$\log N(B_2, \|\cdot\|_X, t) \leq O(G_S^2 \alpha S L \ln(m)/m)/t^2$$

and

$$\log N(T_\alpha, \|\cdot\|_X, t) \leq O(G_S^2 \alpha^2 S L k \ln(m)/m)/t^2\,.$$

Proof. Since $s(a_j) \geq \|a_j\|_2^2 + 2 \geq \|a_j\|_2^2$, and thus $w_j = \frac{2S}{s(a_j)+S} \leq 2S/\|a_j\|_2^2$, it follows that

$$\begin{aligned}
\mathop{\mathbb{E}}_{y\sim\mathcal{N}(0,I_d)} \|y\|_X &= \mathop{\mathbb{E}}_{y\sim\mathcal{N}(0,I_d)} \left(4G_S\alpha \max_{j\in[m]} w_j |\langle a_j, y\rangle|^2 L/m\right)^{1/2} \\
&\leq 2G_S\sqrt{\alpha L/m} \mathop{\mathbb{E}}_{y\sim\mathcal{N}(0,I_d)} \left(\max_{j\in[m]} w_j |\langle a_j, y\rangle|^2\right)^{1/2} \\
&= 2G_S\sqrt{\alpha L/m} \mathop{\mathbb{E}}_{y\sim\mathcal{N}(0,I_d)} \max_{j\in[m]} \sqrt{w_j} |\langle a_j, y\rangle| \\
&\leq 2G_S\sqrt{2\alpha S L/m} \mathop{\mathbb{E}}_{y\sim\mathcal{N}(0,I_d)} \max_{j\in[m]} \left|\left\langle \frac{a_j}{\|a_j\|_2}, y\right\rangle\right| \\
&\leq O(G_S\sqrt{\alpha S L \ln(m)/m})\,, \qquad (7)
\end{aligned}$$

where the last inequality follows from Lemma F.2 and the fact that each $Z_j = \left\langle \frac{a_j}{\|a_j\|_2}, y\right\rangle$ is identically distributed as a zero mean Gaussian with unit variance $\left\|\frac{a_j}{\|a_j\|_2}\right\|_2^2 = 1$ by self-stability of the Gaussian distribution.

Finally, we can apply Proposition D.5 using Equation (7) as follows.

$$\log N(B_2, \|\cdot\|_X, t) \leq O(G_S^2 \alpha S L \ln(m)/m)/t^2$$

The second bound follows using that $T_\alpha \subseteq (2\alpha k)^{1/2} B_2$. Then

$$\begin{aligned}
\log N(T_\alpha, \|\cdot\|_X, t) &\leq \log N((2\alpha k)^{1/2} B_2, \|\cdot\|_X, t) \\
&= \log N(B_2, \|\cdot\|_X, t/(2\alpha k)^{1/2}) \leq O(G_S^2 \alpha^2 S L k \ln(m)/m)/t^2\,.
\end{aligned}$$

□

For very small t the following bound gives better dependencies.

Lemma D.17. *It holds that*

$$\log N(B_2, \|\cdot\|_X, t) \leq O(m \ln(t^{-1} G_S \sqrt{\alpha SL \ln(m)/m}))$$

and

$$\log N(T_\alpha, \|\cdot\|_X, t) \leq O(m \ln(t^{-1} G_S \alpha \sqrt{SLk \ln(m)/m})).$$

Proof. Let $\{v_1, \ldots, v_m\} \subset \mathbb{R}^d$ be an orthonormal basis of the subspace V spanned by $a_1, \ldots, a_m$. We show that the following set is a covering of B_2 of size $|N| \leq \exp(O(m \ln(t^{-1} G_S \sqrt{\alpha SL \ln(m)/m})))$ with respect to $\|\cdot\|_X$.

$$N = \left\{ \sum_{i=1}^m z_i v_i t / (G_S \sqrt{\alpha SL \ln(m)/m}) \;\middle|\; z \in \mathbb{Z}^m, |z_i| \leq t^{-1} G_S \sqrt{\alpha SL \ln(m)/m} \right\}.$$

We fix an arbitrary element $x \in B_2$ and let $x' = \sum_{i=1}^m x_i v_i$ be the orthogonal projection of x onto V. Further, we set $x_i'' = \lfloor x_i' \cdot t^{-1} G_S \sqrt{\alpha SL \ln(m)/m} \rfloor \cdot t / (G_S \sqrt{\alpha SL \ln(m)/m})$. Note that $x_i'' \in N$ and $\|x' - x''\|_X \leq t$ hold by construction. Finally, since $x - x'$ is orthogonal to V, the triangle inequality yields that

$$\|x - x''\|_X = \|x - x' - (x'' - x')\|_X \leq \|x - x'\|_X + \|x' - x''\|_X = \|x' - x''\|_X \leq t.$$

The second bound follows from rescaling t by a factor $(2\alpha k)^{-1/2}$ as in the previous lemma. □

D.3.4 Bounding the metric entropy

We are now ready to bound the entropy integral.

Lemma D.18. *Assume that g is either convex or concave or monotonic and for all r we have that $|g'(r)| \leq g(r)$. Let $m \geq C \cdot \min\{SBL^2 k/g(0), SL(k + \exp(B))/g(0)\}$, for a sufficiently large absolute constant $C > 1$, and assume that $\alpha \leq 1$. Then it holds that $\int_0^\infty \sqrt{\log N(T_\alpha, \|\cdot\|_X, t)}\, dt \leq O(\ln(m)^{3/2} G_S \alpha \sqrt{SLk/m})$*

Proof. We split the entropy integral at some small $\lambda \ll 1$ into two parts. We use Lemma D.17 and the technical Lemma F.4 to bound the first part and Lemma D.16 for the second. It is sufficient to integrate up to the diameter, since for larger radii, the covering number is simply 1 and its logarithm becomes 0. By Corollaries D.13 and D.14 and Lemma D.8 we have that $\mathcal{D} \leq G_S \alpha \leq \mathrm{poly}(m)$ for our choice of m and assumption that $\alpha \leq 1$.

Then there exist constants C_1, C_2 such that

$$\begin{aligned}
&\int_0^\infty \sqrt{\log N(T_\alpha, \|\cdot\|_X, t)}\, dt \\
&= \int_0^\lambda \sqrt{\log N(T_\alpha, \|\cdot\|_X, t)}\, dt \;+\; \int_\lambda^{\mathcal{D}} \sqrt{\log N(T_\alpha, \|\cdot\|_X, t)}\, dt \\
&\leq \int_0^\lambda C_1 (m \ln(t^{-1} G_S \alpha \sqrt{SLk \ln(m)/m}))^{1/2}\, dt + \int_\lambda^{\mathcal{D}} C_2 t^{-1} G_S \alpha \sqrt{SLk \ln(m)/m}\, dt \\
&\leq O(\lambda) \sqrt{m \ln(\lambda^{-1})} + O(\lambda) \sqrt{m \ln\left(G_S \alpha \sqrt{SLk \ln(m)/m}\right)} + O(\ln(\mathcal{D}) + \ln(\lambda^{-1}))\, G_S \alpha \sqrt{SLk \ln(m)/m}
\end{aligned}$$

Now setting $\lambda = \frac{1}{m}$, we can assert that the last term dominates and thus we get that

$$\int_0^\infty \sqrt{\log N(T_\alpha, \|\cdot\|_X, t)}\, dt \leq O(G_S \alpha \ln(m)^{3/2} \sqrt{SLk/m})$$

□

D.3.5 Proof of the main sampling result

We are now ready to prove the main theorem of this section.

Theorem 3.2. *Let $f(x) \coloneqq \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \|x\|_2^2/k$ such that g is L-Lipschitz, and either convex or concave or monotonic, and for all $r \in \mathbb{R}$ it holds that $|g'(r)| \leq g(r)$. Let $s(a) \geq \|a\|_2^2 + 2$ for all $a \in \mathbb{R}^d$. Set $C = SBL^2/g(0)$. There exists $m_0 = O(C\varepsilon^{-2}k\ln(C\varepsilon^{-1}k\ln(\delta^{-1}))^3\ln(\delta^{-1}\ln(SLk/g(0)))) = \tilde{O}(C\varepsilon^{-2}k\ln(\delta^{-1}))$, such that if we sample $m \geq m_0$ points according to $\mathcal{Q}$, then with probability at least $1-\delta$ we have that*

$$\forall x \in \mathbb{R}^d\colon \left| f_0(x) - \left(\frac{1}{m}\sum_{i=1}^{m} w_i g(\langle a_i, x\rangle)\right)\right| \leq \varepsilon f(x).$$

If additionally $B = O(\log(k))$ holds, then choosing $C = SL/g(0)$ suffices for the same guarantee.

Proof. We set $\ell = 2\lceil\ln(\delta^{-1})\rceil$, and $m = O(C\varepsilon^{-2}k\ell\ln(m)^3)$. Let $\alpha = C_0\beta/k$ for a fixed $\beta \in \{2^0, 2, \ldots, 2^{\lfloor\log(k/C_0)\rfloor}\}$, and C_0 to be chosen later. By Lemma D.9 it holds that $\mathbb{E}(G_S - 1)^\ell \leq (2\pi)^{\ell/2}\,\mathbb{E}(\Lambda^\ell)$ where $\Lambda = \sup_{t\in T_\alpha}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i t)\right|/f(t)$. Note that $X_t = \frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i t)$ defines a Gaussian process. We have that $\Lambda = \sup_{t\in T_\alpha} X_t/f(t) \leq \sup_{t\in T_\alpha} X_t/\alpha$. Using Corollary D.13 we get that the diameter is bounded by

$$\mathcal{D} = 12 G_S \alpha \left(\frac{SBL^2 k}{m g(0)}\right)^{1/2} = \Theta(\varepsilon G_S \alpha/\sqrt{\ell})$$

and using Lemma D.18 we get that the entropy is bounded by

$$\mathcal{E} = O(G_S \alpha \ln(m)^{3/2}\sqrt{SLk/m}) = O(\varepsilon G_S \alpha).$$

Lemma D.9 relates the error term to the Gaussian process, and we continue with Lemma D.1 to bound its ℓ-th moments

$$\begin{aligned}
\mathop{\mathbb{E}}_S (G_S - 1)^\ell &\overset{Lem\ D.9}{\leq} \mathop{\mathbb{E}}_S (2\pi)^{\ell/2}\mathop{\mathbb{E}}_\sigma |\Lambda|^\ell \\
&\leq \mathop{\mathbb{E}}_S (2\pi)^{\ell/2}\alpha^{-\ell}\mathop{\mathbb{E}}_\sigma \sup_{t\in T_\alpha}|X_t|^\ell \\
&\overset{Lem\ D.1}{\leq} \mathop{\mathbb{E}}_S (2\pi)^{\ell/2}\frac{(2\mathcal{E})^\ell(\mathcal{E}/\mathcal{D}) + O(\sqrt{\ell}\mathcal{D})^\ell}{\alpha^\ell} \\
&\leq \mathop{\mathbb{E}}_S ((2e)^{-1}\varepsilon G_S)^\ell \leq \frac{\delta\varepsilon^\ell}{2\cdot 2^\ell}\mathop{\mathbb{E}}_S((G_S - 1) + 1)^\ell \\
&\leq \frac{\delta\varepsilon^\ell}{2\cdot 2^\ell}\mathop{\mathbb{E}}_S 2^\ell((G_S - 1)^\ell + 1) \leq \varepsilon^\ell \mathop{\mathbb{E}}_S (G_S - 1)^\ell + \frac{\delta}{2}\varepsilon^\ell.
\end{aligned}$$

By rearranging the terms, this implies that

$$\mathop{\mathbb{E}}_S (G_S - 1)^\ell \leq \frac{\delta\varepsilon^\ell}{2(1-\varepsilon^\ell)} \leq \delta\varepsilon^\ell\,.$$

Using Markov's inequality we get that

$$\Pr(G_S - 1 \geq \varepsilon) = \Pr((G_S - 1)^\ell \geq \varepsilon^\ell) \leq \frac{\delta\varepsilon^\ell}{\varepsilon^\ell} = \delta\,.$$

Thus, it holds with probability $1-\delta$ that

$$\forall x \in T_\alpha\colon \left|\frac{\int g(ax)\, d\mathcal{P}(a) - \frac{1}{m}\sum_{i=1}^m w_i g(a_i x)}{f(x)}\right| \leq G_S - 1 \leq \varepsilon\,.$$

Recall that $f(x) \geq g(0)^2/(4(LB)^2 k)$ by Lemma B.5. We set $C_0 = g(0)^2/(4(LB)^2)$ accordingly. Substituting δ by $\delta/\log((SL)^2 k/C_0)$, we can use the union bound to apply the result to all $\alpha = C_0\beta/k$ for $\beta \in \{2^0, 2^1, \ldots, 2^{\lfloor\log((SL)^2 k/C_0)\rfloor}\}$ simultaneously.

For larger values of α, we have $f(x) \geq \alpha \geq C_0 2^{\lceil \log((SL)^2 k/C_0)\rceil}/k \geq (SL)^2$. Note that we have $s(a) \geq \|a\|_2^2 + 2 \geq \|a\|_2 + 1$. We can thus apply Theorem 3.1 to $T = \{x \in \mathbb{R}^d \mid f(x) \geq (SL)^2\}$ and thus with $OPT_T \geq (SL)^2$, which yields $O((SL)^2 \varepsilon^{-2} k \ln(\delta^{-1})/OPT_T) = O(\varepsilon^{-2} k \ln(\delta^{-1})) \leq O(m)$ sample size.

The improved constant $C = SLk/g(0)$ follows verbatim by replacing Corollary D.13 with Corollary D.14, which gives $\mathcal{D} = \Theta(SL(k + \exp(B))/g(0))$ together with the additional assumption that $B = O(\log(k))$. □

D.4 Sampling bound for ℓ_1 regularization

In this subsection we study the setting where

$$f(x) = \int g(\langle a, x\rangle)\, d\mathcal{P}(a) + \frac{\|x\|_1}{k}.$$

We work again with our standard L-Lipschitz and B-bounded distribution assumptions [A1] and [A2], and set $s(a) \geq \|a\|_2 + g(0)$. We note that the analysis will be similar to the one in Appendix D.3. Therefore, several details and definitions are borrowed from the previous sections and will not be explained again in all details.

Fix some subset $T \subseteq \mathbb{R}^d$. Given a sample S recall that we set $G_S = \sup_{x \in T} 1 + |f_S(x) - f_0(x)|/f(x)$ where $f_S(x) = \frac{1}{m}\sum_{i=1}^m w_i g(a_i x)$. Note that $\sqrt{G_S} \leq G_S$ since $G_S \geq 1$. We also remind of the bound $f_S(x)/f(x) \leq G_S$, see Lemma D.7.

A standard symmetrization argument relates the error $E = G_S - 1$ to a Gaussian process defined over T, see Lemma D.9. We then consider the Gaussian process $X_s = \left|\frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i s)\right|$ for $s \in T_\alpha = \{x \in \mathbb{R}^d \mid f(x) \in [\alpha, 2\alpha)\}$ for some $\alpha \in \mathbb{R}$. We note that for any $x \in T_\alpha$ it holds that $\|x\|_1 \leq 2\alpha k$ and thus $T_\alpha \subseteq 2\alpha k B_1$. The metric d_X is again given by

$$d_X(x, x') = \left(\frac{1}{m^2}\sum_{i=1}^m w_i^2 (g(a_i x) - g(a_i x'))^2\right)^{1/2}.$$

and thus, the diameter bound is similar to the one in Appendix D.3.1.

D.4.1 Bounding the diameter

Lemma D.19. *Let $w_i = \frac{2S}{s(a_i)+S}$ and $s(a_i) \geq \|a_i\|_2$. Then for any $i \in [m]$ and any $x \in T_\alpha$ it holds that*

$$\frac{w_i g(a_i x)}{f(x)} \leq 4SLk\,.$$

Proof. Recall that $s(a_i) \geq \|a_i\|_2^2 + 1 \geq \max\{\|a_i\|_2, 1\}$. Also recall that $f(x) \geq \frac{g(0)}{BLk}$, and $f(x) \geq \frac{\|x\|_1}{k}$. Thus, it holds that

$$\begin{aligned}\frac{w_i g(a_i x)}{f(x)} = \frac{w_i(g(0) + g_0(a_i x))}{f(x)} &\leq \frac{2S}{s(a_i)+S}\cdot\frac{g(0) + L|a_i x|}{f(x)} \\ &\overset{CSI}{\leq} \frac{2S}{S}\frac{g(0)}{f(x)} + \frac{2S}{\|a_i\|_2}\frac{L\|a_i\|_2\|x\|_2}{f(x)} \leq 2\left(BLk + \frac{SL\|x\|_2}{f(x)}\right) \leq 4SLk\,.\end{aligned}$$

□

We have the following bound on the diameter.

Lemma D.20. *It holds that* $\sup_{s,t \in T_\alpha} d_X(s,t) \leq 6G_S\alpha\left(\frac{SLk}{m}\right)^{1/2}$.

Proof. By Lemma D.10 we have that $\sup_{s,t\in T_\alpha} d_X(s,t) \leq 2G_S\alpha\left(\frac{2}{m}\sup_{x\in T_\alpha, j\in[m]}\frac{w_j g(a_j x)}{f(x)}\right)^{1/2}$. Using Lemma D.19, yields that

$$\sup_{s,t\in T_\alpha} d_X(s,t) \leq 6G_S\alpha\left(\frac{SLk}{m}\right)^{1/2}$$

□

D.4.2 Relating the metric to a norm

This time we do not need a proper norm, the square root of an ℓ_∞-norm will suffice for our purposes.

Lemma D.21. *Let A be the matrix with rows $w_i a_i$. Then for any $s, t \in T_\alpha$ it holds that*

$$d_X(s,t) \leq \left(\frac{4G_S \alpha L \|A(s-t)\|_\infty}{m}\right)^{1/2}.$$

Proof. We have

$$\begin{aligned}
d_X(s,t)^2 &= \frac{1}{m^2}\sum_{i=1}^m w_i^2 \left(g(a_i s) - g(a_i t)\right)^2 \\
&\leq \frac{1}{m^2}\sum_{i=1}^m w_i^2 \left|g(a_i s) - g(a_i t)\right| \cdot \left(g(a_i s) + g(a_i t)\right) \\
&\leq \frac{1}{m^2}\sum_{i=1}^m w_i^2 L|a_i s - a_i t| \cdot \left(g(a_i s) + g(a_i t)\right) \\
&\leq \frac{L}{m^2}\sum_{i=1}^m \max_{j\in[m]} w_j |a_j(s-t)| \cdot w_i\left(g(a_i s) + g(a_i t)\right) \\
&\leq \frac{L}{m}\max_{j\in[m]} w_j |a_j(s-t)| \cdot \left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i s) + \frac{1}{m}\sum_{i=1}^m w_i g(a_i t)\right) \\
&\leq \frac{L}{m}\max_{j\in[m]} w_j |a_j(s-t)| \cdot G_S\left(f(s) + f(t)\right) \\
&\leq \frac{4G_S\alpha L}{m}\max_{j\in[m]} \left|(w_j a_j)(s-t)\right|.
\end{aligned}$$

□

D.4.3 Bounding the covering numbers

We set $A \in \mathbb{R}^{m\times d}$ to be the matrix with rows $w_i a_i$ and for $p \in \mathbb{R}_{\geq 0}$ we set $B_p(A) = \{x \in \mathbb{R}^d \mid \|Ax\|_p \leq 1\}$, and $B_p = B_p(I_d)$. We first bound the covering numbers for covering B_1 with $B_\infty(A)$ balls.

Lemma D.22. *For any radius $t > 0$ it holds that*

$$\log N(B_1, B_\infty(A), t) \leq O(S)\frac{\ln(m)}{t}.$$

Proof. Assume the rank of $A \in \mathbb{R}^{m\times d}$, is bounded by $m < d$. Let $A = U\Sigma V^T$ be the singular value decomposition. The matrix $V \in \mathbb{R}^{d\times m}$ is a matrix with orthonormal columns that span the same subspace as A. We first cover B_1 with $B_2(V^T)$ and then we cover $B_2(V^T)$ with $B_\infty(A)$. Note that $B_2(V^T)$ is isometric to the Euclidean ball in dimension m.

For the first part, we aim to use Proposition D.4 with $\|\cdot\|_X = \|\cdot\|_1$. We note that the dual norm is $\|\cdot\|_{X^*} = \|\cdot\|_\infty$. We thus need to bound $\mathbb{E}_{g\sim\mathcal{N}(0,I_m)}\|Vg\|_\infty$. However, we cannot use Lemma F.2 directly since this is a d-dimensional Gaussian. Since it is projected to an m-dimensional subspace, there is hope that $O(\sqrt{\log(m)})$ will suffice:

Each row corresponds to a Gaussian as $v_i g = \sum_{j=1}^m V_{ij} g_j \sim \|v_i\|_2 \cdot g'$, where $g' \sim N(0,1)$ In particular $|v_i g|^2 \sim \|v_i\|_2^2 \cdot g'^2$ follows a squared Gaussian distribution. It is also known that since V is an orthonormal basis, its squared row norms are the statistical leverage scores, for which it is known (Foster, 1953) that they are bounded by $\|v_i\|_2^2 \leq 1$ and sum to the rank, so $\sum_{i=1}^d \|v_i\|_2^2 \leq m < d$.

Flattening lemmas take high leverage points and replicate them several times to replace one heavy coordinate by several ones with uniformly small leverage Woodruff and Yasuda (2023). Reversely to these result, we aim to combine rows in such a way that we end up with $\Theta(m)$ rows, each with $\Theta(1)$ leverage.

To this end, consider Lemma F.3. Using this lemma, whenever, we have a group G of rows, i.e., Gaussians $v_i g, i \in G$, we can replace them by a row vector $v \in \mathbb{R}^d$, i.e., by a Gaussian vg that has variance $\|v\|_2^2 = \sum_{i \in G} \|v_i\|_2^2$ to aggregate the norm of the entire group into one single row. The lemma ensures that the expected maximum can only grow.

Due to the aforementioned properties of leverage scores, we now put the rows of V into at most $2m$ groups G_i such that $1/2 \leq \sum_{j \in G_i} \|v_j\|_2^2 \leq 1$ (e.g. by greedy bin-packing). We use Lemma F.3 with $\sigma_{\max} \leq 1$ to aggregate all vectors within G_i to only one vector w_i with $\|w_i\|_2^2 = \sum_{j \in G_i} \|v_j\|_2^2 \leq 1$ and arrange them into a new matrix $W \in \mathbb{R}^{2m \times m}$.

This yields

$$\begin{aligned}\left(\mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, I_m)} \|Vg\|_\infty\right)^2 &\overset{Jensen}{\leq} \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, I_m)} \max_{i \in [d]} |\langle v_i, g \rangle|^2 \\ &\leq \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, I_m)} \max_{i \in [2m]} |\langle w_i, g \rangle|^2 = O(\ln(m))\,,\end{aligned}$$

where we used Jensen's inequality followed by the above argument, and then followed by Lemma F.2.

We thus get from Proposition D.4 that

$$\log N(B_1, B_2(V^T), t) \leq O(1) \frac{\ln(m)}{t^2}\,.$$

Similarly, we have that $A_i = w_i a_i \leq 2S a_i / \|a_i\|_2$ and thus

$$\mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, I_d)} \|Ag\|_\infty \leq 2S \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, I_d)} \max_{i \in m} \left| \left\langle \frac{a_i}{\|a_i\|_2}, g \right\rangle \right| = O(S\sqrt{\ln(m)})\,.$$

where the last inequality follows from Lemma F.2 and the fact that each $Z_i = \left\langle \frac{a_i}{\|a_i\|_2}, g \right\rangle$ is identically distributed as a zero mean Gaussian with unit variance $\left\| \frac{a_j}{\|a_j\|_2} \right\|_2^2 = 1$ by self-stability of the Gaussian distribution.

Using the dual Sudakov bound stated in Proposition D.5, we get that

$$\log N(B_2(V^T), B_\infty(A), t) \leq O(S) \frac{\ln(m)}{t^2}$$

Combining both results implies that

$$\begin{aligned}\log N(B_1, B_\infty(A), t) &\leq \log N(B_1, B_2(V^T), \sqrt{t}) + \log N(\sqrt{t} B_2(V^T), B_\infty(A), t) \\ &= \log N(B_1, B_2(V^T), \sqrt{t}) + \log N(B_2(V^T), B_\infty(A), \sqrt{t}) \\ &\leq O(1) \frac{\ln(m)}{t} + O(S) \frac{\ln(m)}{t} = O(S) \frac{\ln(m)}{t}\end{aligned}$$

□

Using this result, we bound in the following lemma the covering numbers for covering T_α with balls according to our metric.

Lemma D.23. *For any radius $t > 0$ it holds that*

$$\log N(T_\alpha, d_X, t) \leq O(1) \frac{G_S \alpha^2 SLk}{mt^2} \log(m)\,.$$

and

$$\log N(T_\alpha, d_X, t) \leq O(1) m \log\left(\frac{G_S \alpha^2 SLk}{mt^2}\right)\,.$$

Proof. Recall that $T_\alpha \subseteq 2\alpha k B_1$. Using Lemmas D.21 and D.22, we get that

$$\begin{aligned}\log N(T_\alpha, d_X, t) &\le \log N\left(2\alpha k B_1, \left(\frac{4G_S\alpha L\|A\cdot\|_\infty}{m}\right)^{1/2}, t\right)\\ &= \log N\left(2\alpha k B_1, B_\infty(A), \frac{mt^2}{4G_S\alpha L}\right)\\ &= \log N\left(B_1, B_\infty(A), \frac{mt^2}{8G_S\alpha^2 Lk}\right)\\ &\le O(1)\frac{G_S\alpha^2 SLk}{mt^2}\log(m)\,.\end{aligned}$$

For the second part, assume the rank of $A \in \mathbb{R}^{m\times d}$, is bounded by $m < d$. Let $A = U\Sigma V^T$ be the singular value decomposition. The matrix $V \in \mathbb{R}^{d\times m}$ is a matrix with orthonormal columns that span the same subspace as A. We have $B_1 \subseteq B_2 \subseteq B_2(V^T)$ as the projection to the subspace can make the norm only smaller, i.e., $\|V^T x\|_2 \le \|x\|_2 \le \|x\|_1 \le 1$.

Recall that $w_i = \frac{2S}{s(a_i)+S} \le \frac{2S}{\|a_i\|_2}$, and let A' be the matrix whose rows equal $a'_i = a_i/\|a_i\|_2$.

For $x \in B_2(V^T)/(2S)$ it holds by definition that $\|V^T x\|_2 \le 1/(2S)$. Using $V^T V = I$, this implies

$$\begin{aligned}\|Ax\|_\infty &= \|U\Sigma V^T x\|_\infty = \|U\Sigma V^T V V^T x\|_\infty\\ &= \|AVV^T x\|_\infty \le 2S\|A'VV^T x\|_\infty\\ &= 2S\max_{i\in[m]} x^T VV^T a'_i \le 2S\sup_{a\in B_2(V^T)} x^T VV^T a'_i\\ &= 2S\cdot x^T VV^T x/\|V^T x\|_2 = 2S\|V^T x\|_2 \le 1\,.\end{aligned}$$

where the second inequality follows since all $a'_i \in B_2(V^T)$. Thus, we have that $B_2(V^T)/(2S) \subseteq B_\infty(A)$.

Consequently, it holds that

$$\log N(B_1, B_\infty(A), t) \le \log N(B_2(V^T), B_2(V^T), t/(2S)) \le O(1)m\log(S/t),$$

where the last inequality follows from the standard volume argument, see Lemma D.6. Similar to our previous bound, we arrive at

$$\begin{aligned}\log N(T_\alpha, d_X, t) &\le \log N\left(B_1, B_\infty(A), \frac{mt^2}{8G_S\alpha^2 Lk}\right)\\ &\le O(1)m\log\left(\frac{G_S\alpha^2 SLk}{mt^2}\right)\,.\end{aligned}$$

□

D.4.4 Bounding the metric entropy

It remains to bound the entropy integral.

Lemma D.24. *Let $m \ge \frac{CSLk}{\varepsilon^2}$, for a sufficiently large absolute constant $C > 1$, and assume that $\alpha \le 1/\varepsilon$. Then it holds that $\int_0^\infty \sqrt{\log N(T_\alpha, \|\cdot\|_X, t)}\,dt \le O(\ln(m)^{3/2} G_S\alpha\sqrt{SLk/m})$*

Proof. Again we split the entropy integral at some small $\lambda \ll 1$ into two parts. We use Lemma D.23 and the technical Lemma F.4. It is sufficient to integrate up to the diameter, since for larger radii, the covering number is simply 1 and its logarithm becomes 0. By Lemma D.20 we have that $\mathcal{D} \le \varepsilon G_S\alpha \le G_S \le \mathrm{poly}(m)$ for our choice of m using Lemma D.8 and the assumption that $\alpha \le 1/\varepsilon$.

Using Lemma D.23 and the fact that $\sqrt{G_S} \le G_S$, we get that

$$\int_0^\infty \sqrt{\log N(T_\alpha, d_X, t)}\,dt$$

$$\leq \int_0^{\lambda} O(1)\sqrt{m\log\left(\frac{G_S\alpha^2 SLk}{mt^2}\right)}\,dt \;+\; \int_{\lambda}^{\mathcal{D}} O(1)\frac{G_S\alpha}{t}\sqrt{\frac{SLk}{m}\log(m)}\,dt$$

$$\leq \int_0^{\lambda} O(1)\sqrt{m\log\left(\frac{1}{t}\right)}\,dt + \int_0^{\lambda} O(1)\sqrt{m\log\left(\frac{G_S\alpha SLk}{m}\right)}\,dt + \int_{\lambda}^{\mathcal{D}} O(1)\frac{G_S\alpha}{t}\sqrt{\frac{SLk}{m}\log(m)}\,dt\,.$$

$$\leq O(\lambda)\sqrt{m\log\left(\lambda^{-1}\right)} + O(\lambda)\sqrt{m\log\left(\frac{G_S\alpha SLk}{m}\right)} + O(\log(\mathcal{D})+\log(\lambda^{-1}))G_S\alpha\sqrt{\frac{SLk}{m}\log(m)}\,dt\,.$$

Now setting $\lambda = \frac{1}{m}$, we can assert that the last term dominates and thus we get that

$$\int_0^{\infty} \sqrt{\log N(T_\alpha, \|\cdot\|_X, t)}\,dt \leq O(G_S\alpha\ln(m)^{3/2}\sqrt{SLk/m})$$

□

D.4.5 Proof of the main sampling result

We are now ready to prove our general bound for ℓ_1 regularized functions.

Theorem 3.3. *Let $f(x) := \int g(\langle a,x\rangle)\,d\mathcal{P}(a) + \|x\|_1/k$ such that g is L-Lipschitz. Let $s(a) \geq \|a\|_2 + 1$ for all $a \in \mathbb{R}^d$. Set $C = SL$. There exists $m_0 = O(C\varepsilon^{-2}k\ln(C\varepsilon^{-1}k\ln(\delta^{-1}))^3\ln(\delta^{-1}\ln(\varepsilon^{-1}BLk))) = \tilde{O}(C\varepsilon^{-2}k\ln(\delta^{-1}))$, such that if we sample $m \geq m_0$ points according to $\mathcal{Q}$, then with probability at least $1-\delta$ it holds for all $x \in \{x \in \mathbb{R}^d \mid f(x) \leq g(0)/\varepsilon\}$ that*

$$\left| f_0(x) - \left(\frac{1}{m}\sum_{i=1}^{m} w_i g(\langle a_i, x\rangle)\right)\right| \leq \varepsilon f(x)\,.$$

If g is homogeneous, then it is sufficient to set $m_0 = O(C\varepsilon^{-2}k\ln(\delta^{-1})\ln(C\varepsilon^{-1}k\ln(\delta^{-1}))^3)$ to get the same result for all $x \in \mathbb{R}^d$ simultaneously.

Proof. We set $\ell = 2\lceil\ln(\delta^{-1})\rceil$, and $m = C\cdot\frac{k\ell\ln(m)^3}{\varepsilon^2}$. Let $\alpha = C_0\beta/k$ for some fixed $\beta \in \{2^0, 2, \ldots, 2^{\lfloor\log(k/\varepsilon C_0)\rfloor}\}$, and C_0 to be chosen later. By Lemma D.9 it holds that $\mathbb{E}(G_S-1)^\ell \leq (2\pi)^{\ell/2}\,\mathbb{E}(\Lambda^\ell)$ where $\Lambda = \sup_{t\in T_\alpha}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i t)\right|/f(t)$. Note that $X_t = \frac{1}{m}\sum_{i=1}^m \sigma_i w_i g(a_i t)$ defines a Gaussian process. We have that $\Lambda = \sup_{t\in T_\alpha} X_t/f(t) \leq \sup_{t\in T_\alpha} X_t/\alpha$. Using Lemma D.20 we get that the diameter is bounded by

$$\mathcal{D} = 6G_S\alpha\left(\frac{SLk}{m}\right)^{1/2} = \Theta(\varepsilon G_S\alpha/\sqrt{\ell})$$

and using Lemma D.24 we get that the entropy is bounded by

$$\mathcal{E} = O(G_S\alpha\ln(m)^{3/2}\sqrt{SLk/m}) = O(\varepsilon G_S\alpha).$$

Lemma D.9 relates the error term to the Gaussian process, and we continue with Lemma D.1 to bound its ℓ-th moments

$$\begin{aligned}
\mathop{\mathbb{E}}_S(G_S-1)^\ell &\overset{Lem\ D.9}{\leq} \mathop{\mathbb{E}}_S(2\pi)^{\ell/2}\mathop{\mathbb{E}}_\sigma|\Lambda|^\ell \\
&\leq \mathop{\mathbb{E}}_S(2\pi)^{\ell/2}\alpha^{-\ell}\mathop{\mathbb{E}}_\sigma\sup_{t\in T_\alpha}|X_t|^\ell \\
&\overset{Lem\ D.1}{\leq} \mathop{\mathbb{E}}_S(2\pi)^{\ell/2}\frac{(2\mathcal{E})^\ell(\mathcal{E}/\mathcal{D}) + O(\sqrt{\ell}\mathcal{D})^\ell}{\alpha^\ell} \\
&\leq \mathop{\mathbb{E}}_S((2e)^{-1}\varepsilon G_S)^\ell \leq \frac{\delta\varepsilon^\ell}{2\cdot 2^\ell}\mathop{\mathbb{E}}_S((G_S-1)+1)^\ell \\
&\leq \frac{\delta\varepsilon^\ell}{2\cdot 2^\ell}\mathop{\mathbb{E}}_S 2^\ell((G_S-1)^\ell+1) \leq \varepsilon^\ell\mathop{\mathbb{E}}_S(G_S-1)^\ell + \frac{\delta}{2}\varepsilon^\ell.
\end{aligned}$$

By rearranging the terms, this implies that

$$\mathop{\mathbb{E}}_{S}(G_S - 1)^\ell \leq \frac{\delta\varepsilon^\ell}{2(1-\varepsilon^\ell)} \leq \delta\varepsilon^\ell\,.$$

Using Markov's inequality we get that

$$\Pr(G_S - 1 \geq \varepsilon) = \Pr((G_S-1)^\ell \geq \varepsilon^\ell) \leq \frac{\delta\varepsilon^\ell}{\varepsilon^\ell} = \delta\,.$$

Thus, it holds with probability $1-\delta$ that

$$\forall x \in T_\alpha\colon \left|\frac{\int g(ax)\,d\mathcal{P}(a) - \frac{1}{m}\sum_{i=1}^m w_i g(a_i x)}{f(x)}\right| \leq G_S - 1 \leq \varepsilon\,.$$

This concludes the claim for $x \in T_\alpha$. If g is homogeneous, then it suffices to apply the above for a fixed $\alpha = 1$. The result extends to all $x \in \mathbb{R}^d$ simply by scaling.

Otherwise, recall that $f(x) \geq g(0)/(LBk)$ by Lemma B.5. We set $C_0 = 1/(LB)$. Substituting δ by $\delta/\log(k/(\varepsilon C_0))$, we can use the union bound to apply the result to all $\alpha = g(0)C_0\beta/k$ for $\beta \in \{2^0, 2^1, \ldots, 2^{\lfloor \log(k/(\varepsilon C_0))\rfloor}\}$ simultaneously. This handles all $x \in \{x \in \mathbb{R}^d \mid f(x) \leq g(0)/\varepsilon\}$ with the stated sample size. □

For larger values of α, not handled by Theorem 3.3, we have $f(x) \geq \alpha = g(0)C_0 2^{\lceil \log(k/(\varepsilon C_0))\rceil}/k \geq g(0)/\varepsilon$. This enables a different argument, if $g(\cdot)$ can be decomposed into a homogeneous part and a bounded part (as all our example functions).

Lemma D.25. *Assume that $g(x) = h(x)+b(x)$ such that $h(x)$ is homogeneous, i.e., for any $\lambda \in \mathbb{R}_{\geq 0}$ it holds that $h(\lambda x) = \lambda h(x)$. Further assume that $b(x)$ is bounded by $g(0)$, i.e., it holds for all $r \in \mathbb{R}$ that $b(r) \leq g(0)$. If*

$$\left|f_0(x) - \left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i x)\right)\right| \leq \varepsilon f(x).$$

holds for all $x \in \mathbb{R}^d$ such that $f(x) \leq g(0)/\varepsilon$, then the inequality holds for all $x \in \mathbb{R}^d$ with $\varepsilon' = 4\varepsilon$.

Proof. We need to prove the claim for all $x \in \mathbb{R}^d$ for which it holds that $f(x) \geq g(0)/\varepsilon$. Note that for the homogeneous part, if we have a relative error guarantee e.g. for $h(x) = 1$, then it is also a relative error guarantee for $\lambda h(x) = h(\lambda x)$ for any scaling $\lambda \geq 0$. Thus the guarantee extends to all $x \in \mathbb{R}^d$. Recall that $w_i = \frac{2S}{s(a_i)+S} \leq 2$. By our assumption and the triangle inequality, we have

$$\begin{aligned}
&\left|f_0(x) - \left(\frac{1}{m}\sum_{i=1}^m w_i g(a_i x)\right)\right| \\
&= \left|\int h(ax) + b(ax)\,d\mathcal{P}(a) - \left(\frac{1}{m}\sum_{i=1}^m w_i(h(a_i x) + b(a_i x))\right)\right| \\
&\leq \left|\int h(ax)\,d\mathcal{P}(a) - \left(\frac{1}{m}\sum_{i=1}^m w_i h(a_i x)\right)\right| + \left|\int b(ax)\,d\mathcal{P}(a) - \left(\frac{1}{m}\sum_{i=1}^m w_i b(a_i x)\right)\right| \\
&\leq \varepsilon h(x) + \left|\int b(ax)\,d\mathcal{P}(a)\right| + \left|\frac{1}{m}\sum_{i=1}^m w_i b(a_i x)\right| \\
&\leq \varepsilon h(x) + g(0) + 2g(0) \leq \varepsilon f(x) + 3\varepsilon f(x) = 4\varepsilon f(x).
\end{aligned}$$

□

E Lower bounds

In this section, we prove various lower bounds.

E.1 Quadratic lower bounds

We prove a series of quadratic lower bounds summarized in the following theorem:

Theorem 3.5. *Consider any of the following combinations of loss functions and regularizers:*

- *logistic/sigmoid with* $\mathcal{R}(\cdot) = \|\cdot\|_2$*, or*
- *hinge/ReLU with* $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_2^2\}$.

There exists a distribution $\mathcal{P}$ *over vectors* $a \in \mathbb{R}^d$ *such that* $m = \Omega(k^2/\varepsilon^2)$ *samples are required to satisfy Equation* (1)*.*

Proof. The proof is treated in Lemma E.1 for logistic loss and Corollary E.2 for sigmoid loss, both with $\mathcal{R}(\cdot) = \|\cdot\|_2$. The remaining combinations are treated in Lemma E.3 for hinge loss, and in Lemma E.4 for ReLU loss. Both results work with either of the regularizers $\mathcal{R}(\cdot) \in \{\|\cdot\|_2, \|\cdot\|_1\}$. □

We get an $\Omega((k/\varepsilon)^2)$ LB against logistic regression with $\|\cdot\|_2$.

Lemma E.1. *Consider the distribution* $\mathcal{P}$ *with* $P(X = e_j) = 1/d$*, where* e_j *is the* j*-th unit vector. Let* $\varepsilon \in (0, 1/10)$*. Consider any sample* $(a_i, w_i)_{i\in[m]}$ *such that each* (a_i, w_i) *is set to* $a_i = e_j$ *and* $w_i = 1/(dp_j)$ *with some probability* $p_j \in [0, 1]$*. Further let* $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ *be the logistic loss function with*

$$g(r) = \ln(1 + \exp(-r)) .$$

If $\mathcal{R}(x) = \|x\|_2$*,* $d = 2(k \ln(2)/40\varepsilon)^2$ *and* $m \leq d/2$ *then there exists* $x \in \mathbb{R}^d$ *such that*

$$\left| \frac{1}{d} \sum_{j=1}^{d} g(xe_j) - \frac{1}{m} \sum_{i=1}^{m} w_i g(xa_i) \right| > \varepsilon \left(\frac{1}{d} \sum_{j=1}^{d} g(xe_j) + \frac{\mathcal{R}(x)}{k} \right) .$$

Proof. We consider the case $\mathcal{R}(x) = \|x\|_2$, $d = 2(k \ln(2)/40\varepsilon)^2$. For notational convenience, we scale the loss and regularizer by a factor $1/\ln(2)$. This obviously preserves the multiplicative error guarantee. For the proof consider $g(r) = \ln(1 + \exp(-r))/\ln(2)$.

Let $m \leq d/2$. Without loss of generality, we assume that there are no samples $a_i = e_j$ for any $j \leq d/2$. Let $x = \sum_{j=1}^{d/2} e_j$. Note that for half of the unit vectors, we have $g(e_j x) = g(0) = \ln(2)/\ln(2) = 1$, and for the other half it holds that $g(e_j x) = g(1) = \ln(1 + \exp(-1))/\ln(2)$. We define $c = \ln(1 + \exp(-1))/(2\ln(2))$ and observe that $c \approx 0.225971 < \frac{1}{4}$.

The (rescaled) loss of x thus equals

$$\frac{1}{d} \sum_{j=1}^{d} g(e_j x) = \frac{1}{2} \left(1 + \frac{\ln(1 + \exp(-1))}{\ln(2)} \right) = \frac{1}{2} + c .$$

For the (rescaled) regularization term, we have

$$\frac{\mathcal{R}(x)}{k \ln(2)} = \frac{\sqrt{d/2}}{k \ln(2)} = \frac{\sqrt{(k \ln(2)/40\varepsilon)^2}}{k \ln(2)} = \frac{1}{40\varepsilon} .$$

Overall, we get that

$$\frac{1}{d} \sum_{j=1}^{d} g(e_j x) + \frac{\mathcal{R}(x)}{k \ln(2)} = \frac{1}{2} + c + \frac{1}{40\varepsilon}$$

and the loss of taking $x = 0$ instead equals $\frac{1}{d} \sum_{j=1}^{d} g(0) = 1$.

By construction of x it always holds that $a_i x = 0$ for the subsampled version and thus

$$\frac{1}{m} \sum_{i=1}^{m} w_i g(a_i x) = \frac{1}{m} \sum_{i=1}^{m} w_i g(0) .$$

Now assume $\frac{1}{m}\sum_{i=1}^{m} w_i g(0) \in [1-\varepsilon, 1+\varepsilon]$. Then using $\varepsilon \leq 1/10$ and $c < 1/4$ we have that

$$\frac{1}{2} - c - \varepsilon > \frac{1}{2} - \frac{1}{4} - \varepsilon > \frac{1}{4} - \frac{1}{10} > \frac{1}{10} = \frac{1}{20} + \frac{1}{40} + \frac{1}{40} > \frac{\varepsilon}{2} + \varepsilon c + \frac{1}{40}\,. \tag{8}$$

It follows that

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m} w_i g(a_i x) + \frac{\mathcal{R}(x)}{k\ln(2)} \quad &\geq \quad 1 - \varepsilon + \frac{1}{40\varepsilon} = \frac{1}{2} + c + \frac{1}{40\varepsilon} + \frac{1}{2} - c - \varepsilon \\ &\overset{Equation\ (8)}{>} \frac{1}{2} + c + \frac{1}{40\varepsilon} + \frac{\varepsilon}{2} + \varepsilon c + \frac{1}{40} = (1+\varepsilon)\left(\frac{1}{2} + c + \frac{1}{40\varepsilon}\right).\end{aligned}$$

We conclude that for at least one choice of $x' \in \{x, 0\}$ it holds for the original loss and regularizer that

$$\left|\frac{1}{d}\sum_{j=1}^{d} g(e_j x') - \frac{1}{m}\sum_{i=1}^{m} w_i g(a_i x')\right| > \varepsilon\left(\frac{1}{d}\sum_{j=1}^{d} g(e_j x') + \frac{\mathcal{R}(x')}{k}\right).$$

□

We get the same bound against sigmoid loss with $\|\cdot\|_2$ in a very similar way; only some constants change.

Corollary E.2. *Consider the distribution $\mathcal{P}$ with $P(X = e_j) = 1/d$, where e_j is the j-th unit vector. Let $\varepsilon \in (0, 1/10)$. Consider any sample $(a_i, w_i)_{i\in[m]}$ such that each (a_i, w_i) is set to $a_i = e_j$ and $w_i = 1/(dp_j)$ with some probability $p_j \in [0,1]$. Further let $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be the sigmoid loss function with*

$$g(r) = \frac{1}{1+\exp(r)}\,.$$

If $\mathcal{R}(x) = \|x\|_2$, $d = 2((k/2)/50\varepsilon)^2$ and $m \leq d/2$ then there exists $x \in \mathbb{R}^d$ such that

$$\left|\frac{1}{d}\sum_{j=1}^{d} g(xe_j) - \frac{1}{m}\sum_{i=1}^{m} w_i g(xa_i)\right| > \varepsilon\left(\frac{1}{d}\sum_{j=1}^{d} g(xe_j) + \frac{\mathcal{R}(x)}{k}\right).$$

Proof. The proof is exactly as the proof of Lemma E.1. Only some constants change. In particular the scaling factor becomes $1/g(0) = 2$ instead of $1/\ln(2)$, and the value of c changes to $c = 1/(1+\exp(1)) < 0.3$ instead of $c < 1/4$. Using $d = 2((k/2)/50\varepsilon)^2$ instead of $d = 2(k\ln(2)/40\varepsilon)^2$ allow the remaining calculations to work verbatim. □

The lemma below gives $\Omega((k/\varepsilon)^2)$ against hinge with both ℓ_2 regularizers.

Lemma E.3. *Consider the distribution $\mathcal{P}$ with $P(X = v_j) = 1/(d-1)$ where $v_j = e_d + e_j/\sqrt{2}$. Consider any sample $(a_i, w_i)_{i\in[m]}$ such that each (a_i, w_i) is set to $a_i = v_j$ and $w_i = 1/((d-1)p_j)$ with some probability $p_j \in [0,1]$ and let $\varepsilon \in (0, 1/4)$. Further let $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be the function with*

$$g(r) = \begin{cases} \textit{arbitrary} & \textit{if } r < 0 \\ 1 - r & \textit{if } r \in [0,1] \\ 0 & \textit{otherwise.}\end{cases}$$

If $\mathcal{R}(x) \in \{\|x\|_2, \|x\|_2^2\}$, $d = (k/6\varepsilon)^2 + 1$ and $m \leq (d-1)/2$ then there exists an $x \in \mathbb{R}^d$ such that

$$\left|\frac{1}{d-1}\sum_{j=1}^{d-1} g(v_j x) - \frac{1}{m}\sum_{i=1}^{m} w_i g(a_i x)\right| > \varepsilon\left(\frac{1}{d-1}\sum_{j=1}^{d-1} g(v_j x) + \frac{\mathcal{R}(x)}{k}\right).$$

Proof. Without loss of generality we assume that there are no samples $a_i = v_j$ for any $j \leq (d-1)/2$. Let $x = e_d - \sum_{i=1}^{(d-1)/2} e_i/\sqrt{(d-1)/2}$. We make the following observations:

1. We have for both choices that $\mathcal{R}(x) \leq 2$ since

(a) $\mathcal{R}(x) = \|x\|_2^2 = 1^2 + \sum_{i=1}^{(d-1)/2} (1/\sqrt{(d-1)/2})^2 = 2\,,$
(b) $\mathcal{R}(x) = \|x\|_2 = \sqrt{\|x\|_2^2} = \sqrt{2} < 2\,.$

2. Also note for $j > (d-1)/2$ that $v_j^T x = 1$ and thus $g(v_j^T x) = 1 - 1 = 0$,
by the assumption it also holds for all samples that $a_i x = 1$ resp. $g(a_i x) = g(1) = 0$.

3. Finally, for $j \leq (d-1)/2$ we have $v_j^T x = 1 - 1/\sqrt{2} \cdot 1/\sqrt{(d-1)/2} = 1 - 1/\sqrt{(d-1)}$
and
thus $g(v_j^T x) = 1 - 1 + 1/\sqrt{(d-1)} = 1/\sqrt{(d-1)}$.

These observations together with the choice of $d - 1 = (k/6\varepsilon)^2$ yield that

$$\left(\frac{1}{d-1} \sum_{j=1}^{d-1} g(v_j x) + \frac{\mathcal{R}(x)}{k} \right) = \frac{1}{d-1} \frac{d-1}{2} \frac{1}{\sqrt{(d-1)}} + \frac{\mathcal{R}(x)}{k} \leq \frac{6\varepsilon}{2k} + \frac{2}{k} = \frac{4 + 6\varepsilon}{2k}\,.$$

Further we have that

$$\frac{1}{m} \sum_{i=1}^{m} w_i g(a_i x) = \frac{1}{m} \sum_{i=1}^{m} w_i g(1) = 0\,.$$

Using $\varepsilon < 1/4$ we conclude that

$$\left| \frac{1}{m} \sum_{i=1}^{m} w_i g(a_i x) - \frac{1}{d-1} \sum_{j=1}^{d-1} g(v_j x) \right| = \left| \frac{1}{d-1} \sum_{j=1}^{d-1} g(v_j x) \right| = \frac{6\varepsilon}{2k} > \varepsilon \cdot \left(\frac{4 + 6\varepsilon}{2k} \right).$$

□

The lemma below gives $\Omega((k/\varepsilon)^2)$ against ReLU with both ℓ_2 regularizers.

Lemma E.4. *Consider the distribution $\mathcal{P}$ with $P(X = e_j) = 1/d$. Consider any sample $(a_i, w_i)_{i \in [m]}$ such that each (a_i, w_i) is set to $a_i = v_j$ and $w_i = 1/(dp_j)$ with some probability $p_j \in [0, 1]$ and let $\varepsilon \in (0, 1/4)$. Further let $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be the function with*

$$g(r) = \begin{cases} -r & \text{if } r < 0 \\ 0 & \text{otherwise.} \end{cases}$$

If $\mathcal{R}(x) \in \{\|x\|_2, \|x\|_2^2\}$, $d = (k/6\varepsilon)^2$ and $m \leq d/2$ then there exists an $x \in \mathbb{R}^d$ such that

$$\left| \frac{1}{d} \sum_{j=1}^{d} g(v_j x) - \frac{1}{m} \sum_{i=1}^{m} w_i g(a_i x) \right| > \varepsilon \left(\frac{1}{d} \sum_{j=1}^{d} g(v_j x) + \frac{\mathcal{R}(x)}{k} \right).$$

Proof. Without loss of generality we assume that there are no samples $a_i = e_j$ for any $j \leq d/2$. Let $x = -\sum_{i=1}^{d/2} e_i/\sqrt{d}$. We make the following observations:

1. We have for both choices that $\mathcal{R}(x) = 1$ since

(a) $\mathcal{R}(x) = \|x\|_2^2 = \sum_{i=1}^{d} (1/\sqrt{d})^2 = 1\,,$
(b) $\mathcal{R}(x) = \|x\|_2 = \sqrt{\|x\|_2^2} = 1\,.$

2. Also note for $j > d/2$ that $v_j^T x = 0$ and thus $g(v_j^T x) = g(0) = 0$,
by the assumption it also holds for all samples that $a_i x = 0$ resp. $g(a_i x) = g(0) = 0$.

3. Finally, for $j \leq d/2$ we have $v_j^T x = -1/\sqrt{d}$ and
thus $g(v_j^T x) = 1/\sqrt{d}$.

These observations together with the choice of $d = (k/6\varepsilon)^2$ yield that

$$\left(\frac{1}{d}\sum_{j=1}^{d} g(v_j x) + \frac{\mathcal{R}(x)}{k}\right) = \frac{1}{d}\frac{d}{2}\frac{1}{\sqrt{d}} + \frac{\mathcal{R}(x)}{k} \le \frac{6\varepsilon}{2k} + \frac{1}{k} = \frac{2+6\varepsilon}{2k}\,.$$

Further we have that

$$\frac{1}{m}\sum_{i=1}^{m} w_i g(a_i x) = \frac{1}{m}\sum_{i=1}^{m} w_i g(0) = 0\,.$$

Using $\varepsilon < 1/4$ we conclude that

$$\left|\frac{1}{m}\sum_{i=1}^{m} w_i g(a_i x) - \frac{1}{d}\sum_{j=1}^{d} g(v_j x)\right| = \left|\frac{1}{d}\sum_{j=1}^{d} g(v_j x)\right| = \frac{6\varepsilon}{2k} > \varepsilon \cdot \left(\frac{2+6\varepsilon}{2k}\right).$$

□

E.2 Linear lower bounds

Let $\mathcal{P}$ be a distribution on $\mathbb{R}^d$, $g\colon \mathbb{R} \to [0,\infty)$ a non–negative loss function, and $\mathcal{R}\colon \mathbb{R}^d \longleftarrow \mathbb{R}_{\ge 0}$ a regularizer function. For $k \in \mathbb{N}$ define

$$f(x) = \underbrace{\int g(\langle a, x\rangle)\, d\mathcal{P}(a)}_{=:f_0(x)} + \frac{\mathcal{R}(x)}{k}, \qquad x \in \mathbb{R}^d.$$

Given a non–negative sampling score $s(a)$ with $S = \int s(a)\, d\mathcal{P}(a) < \infty$, we draw m points $a_1, \ldots, a_m \sim \mathcal{Q}$ where $d\mathcal{Q}(a) = \frac{s(a)}{S}\, d\mathcal{P}(a)$, and attach the importance weights $w_i = w(a_i) = \frac{S}{s(a_i)}$. A *relative–error approximation* of sample size m must satisfy, for given $\varepsilon, \delta \in (0,1)$,

$$\Pr\Big[\forall x \in \mathbb{R}^d\colon \big|f_0(x) - \hat{f}_0(x)\big| \le \varepsilon f(x)\Big] \ge 1-\delta, \quad \hat{f}_0(x) = \frac{1}{m}\sum_{i=1}^{m} w_i g\big(\langle a_i, x\rangle\big). \tag{9}$$

Theorem 3.6. *Suppose we want to satisfy Equation* (1) *for some combination of loss function and regularizer. In each of the following settings there exists a distribution $\mathcal{P}$ over vectors $a \in \mathbb{R}^d$ such that the stated lower bounds on the number of samples m hold:*

1. *For either ReLU/hinge/logistic loss with $\mathcal{R}(\cdot) = \|\cdot\|_1$ regularization, $m = \Omega(\varepsilon^{-2} k \log k)$ is required,*
2. *For either sigmoid loss with $\mathcal{R}(\cdot) = \|\cdot\|_1$, or logistic loss with $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ regularization, $m = \Omega(\varepsilon^{-2} k/\log k)$ is required,*
3. *For sigmoid loss with $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ regularization, $m = \Omega(\varepsilon^{-2} k/(\log k)^3)$ is required.*

Proof. 1. For ReLU loss with $\mathcal{R}(\cdot) = \|\cdot\|_1$ regularization, $m = \Omega(\varepsilon^{-2} k \log k)$ is proven in Lemma E.6. For hinge/logistic with the same regularizer, the bound follows by reduction from the case of ReLU, treated in Lemma E.5. 2. For sigmoid loss with $\mathcal{R}(\cdot) = \|\cdot\|_1$, $m = \Omega(\varepsilon^{-2} k/\log k)$ is proven in Lemma E.9. For logistic loss with $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ regularization, we have the same bound proven in Lemma E.8 3. For sigmoid loss with $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ regularization, $m = \Omega(\varepsilon^{-2} k/(\log k)^3)$ is proven in Lemma E.9. □

Lemma E.5. *Assume that $a_1, \ldots, a_m$ satisfy Eq. 9 with a regularizer $\mathcal{R}$ such that $\mathcal{R}(rx) = r\,\mathcal{R}(x)$ for all $r \ge 0$, and that*

$$\lim_{\beta\to\infty} \frac{g(\beta t)}{\beta} = \hat{g}(t).$$

Then $a_1, \ldots, a_m$ also satisfy Eq. 9 when g is replaced by $\hat{g}$.

Proof. Since $a_1, \ldots, a_m$ satisfy Eq. 9, we have

$$\Pr\Big[\forall x \in \mathbb{R}^d\colon \big|f_0(x) - \hat{f}_0(x)\big| \le \varepsilon f(x)\Big] \ge 1-\delta,$$

where

$$f(x) = \underbrace{\int g(\langle a, x\rangle)\, d\mathcal{P}(a)}_{=:f_0(x)} + \frac{\mathcal{R}(x)}{k}, \qquad x \in \mathbb{R}^d$$

and $\hat{f}_0(x) = \frac{1}{m}\sum_{i=1}^m w_i g\big(\langle a_i, x\rangle\big)$. Note that $\forall x \in \mathbb{R}^d\colon \big|f_0(x) - \hat{f}_0(x)\big| \le \varepsilon f(x)$ is equivalent to $\forall x \in \mathbb{R}^d\colon \big|\frac{1}{\beta} f_0(\beta x) - \frac{1}{\beta}\hat{f}_0(\beta x)\big| \le \varepsilon \frac{1}{\beta} f(\beta x)$. Now, sending $\beta \to \infty$ concludes the desired assertion. □

Due to our reduction, we start with a lower bound against ReLU with $\mathcal{R}(\cdot) \in \{\|\cdot\|_1, \|\cdot\|_2\}$ regularization.

Lemma E.6. *Fix $0 < \varepsilon \le 1/4$ and $0 < \delta \le 1/4$. Let $\mathcal{R} \in \{\|\cdot\|_1, \|\cdot\|_2\}$ and define the ReLU loss $g(t) = [-t]_+ = \max\{0, -t\}$. Then there exists a distribution $\mathcal{P}$ on $\mathbb{R}^d$ such that the following holds: if for an importance sampling we have*

$$\Pr\Big[\forall x \in \mathbb{R}^d :\ \big|\hat{f}_0(x) - f_0(x)\big| \ \le\ \varepsilon\, f(x)\Big] \ \ge\ 1 - \delta\,,$$

then the sample size must satisfy

$$m \ \ge\ \Omega\left(\frac{k\log(k/\delta)}{\varepsilon^2}\right).$$

Proof. For $d = k$ place equal mass on the $2k$ signed canonical basis vectors:

$$\mathcal{P}\{+e_j\} = \mathcal{P}\{-e_j\} = \frac{1}{2k}\,, \qquad j = 1, \dots, k\,.$$

Write $S = 1$ without loss of generality (rescaling s leaves $\mathcal{Q}$ and w unchanged). As we assumed $S = 1$, we have

$$\frac{1}{k}\sum_{j=1}^k s(+e_j) + \frac{1}{k}\sum_{j=1}^k s(-e_j) = 2\,,$$

hence one of those two sums is at most 1. Set $\sigma \in \{+1, -1\}$ such that

$$\frac{1}{k}\sum_{j=1}^k s(\sigma e_j) \le 1\,.$$

Define

$$J = \{j \in [k]\colon s(\sigma e_j) \le 2\}$$

and note that we must have $|J| \ge k/2$. Let us check the approximation inequality at $x = -\sigma e_j$, for some $j \in J$:

$$g(\langle \sigma e_j, x\rangle) = g(-1) = 1\,, \qquad\qquad g(\langle -\sigma e_j, x\rangle) = g(1) = 0\,,$$

and $g(\langle \pm e_i, x\rangle) = 0$ for all $i \ne j$. Put these all together, for $\mathcal{R} \in \{\|\cdot\|_1, \|\cdot\|_2\}$, we have

$$f_0(x) = \frac{1}{2k}, \quad \frac{\mathcal{R}(x)}{k} = \frac{1}{k}, \quad f(x) = \frac{3}{2k}\,.$$

Let N_j be the number of occurrence of σe_j in the m samples. Since $\mathcal{Q}(\sigma e_j) = q_j = \frac{s(\sigma e_j)}{2k}$, we obtain

$$N_j \sim \mathrm{Bionomial}(m, q_j)\,.$$

Set $\mu_j = mq_j$. For selected x, we have

$$\hat{f}_0(x) = \frac{1}{m}.\frac{1}{s(\sigma e_j)} N_j\,.$$

This indicates that the approximation property at x is violated if and only if

$$|\hat{f}_0(x) - f_0(x)| = \left|\frac{N_j}{m s(\sigma e_j)} - \frac{1}{2k}\right| > \varepsilon f(x) = \varepsilon\frac{3}{2k}\,.$$

Using that $ms(\sigma e_j) = s(\sigma e_j)\mu_j/q_j = 2k\mu_j$ we have the condition $\left|\frac{N_j}{ms(\sigma e_j)} - \frac{1}{2k}\right| > \frac{3\varepsilon}{2k}$, multiplied by $2k\mu_j$, is equivalent to

$$|N_j - \mu_j| > 3\varepsilon\mu_j.$$

Let us call this event F_j which can be interpreted as the event that our desired approximation is violated at $x = -\sigma e_j$. Also, define F'_j as the event that $N_j - \mu_j > 3\varepsilon\mu_j$ and notice $F'_j \subseteq F_j$, so F'_j implies that the approximation fails.

Note that for each $j \in J$, we have $s(\sigma e_j) \leq 2$ which concludes $q_j \leq \frac{1}{k} \leq \frac{1}{2}$ and thus $\frac{q_j}{1-q_j} \leq \frac{2}{k}$.

Note that $\mathbb{V}(N_j) = mq_j(1-q_j)$. Using Lemma F.1 with $t = 3\varepsilon mq_j$ with $\varepsilon \leq 1/600$, we obtain

$$\Pr(F_j) \geq \Pr(F'_j) \geq c\exp\left(-\frac{9\varepsilon^2 m^2 q_j^2}{3mq_j(1-q_j)}\right) = c\exp\left(-3\varepsilon^2\frac{mq_j}{(1-q_j)}\right) \geq c\exp\left(-6\varepsilon^2\frac{m}{k}\right) = p$$

which is independent of j.

Define, for each $j \in J$,

$$F'_j \;=\; \{\, N_j > z_j \,\}, \qquad z_j = (1+3\varepsilon)\,\mu_j\,, \qquad \mu_j = \mathbb{E}[N_j] = mq_j\,.$$

Then

$$\begin{aligned}
\Pr\Big(\bigcup_{j\in J} F'_j\Big) &= 1 - \Pr\Big(\bigcap_{j\in J}(F'_j)^c\Big) \\
&\geq 1 - \prod_{j\in J}\Pr\big((F'_j)^c\big) \qquad \textit{(negative association of multinomial coordinates)} \\
&= 1 - \prod_{j\in J}\big(1 - \Pr(F'_j)\big) \\
&\geq 1 - (1-p)^{|J|} \\
&\geq 1 - (1-p)^{k/2}\,.
\end{aligned}$$

The vector of counts $(N_j)_{j\in J}$ is a subcollection of a multinomial count vector, which is known to be *negatively associated*. Here $(F'_j)^c = \{N_j \leq z_j\}$ is decreasing in N_j, so $\Pr\left(\bigcap_{j\in J}(F'_j)^c\right) \leq \prod_{j\in J}\Pr\left((F'_j)^c\right)$ (see Joag-Dev and Proschan (1983)), which yields the second line above. On the other hand $1-(1-p)^{\frac{k}{2}} \leq \Pr(\text{approximation failure}) \leq \delta$ which concludes

$$1 - \delta \leq (1-p)^{\frac{k}{2}} \leq \exp\left(-\frac{kp}{2}\right)$$

and thus $kp \leq -2\ln(1-\delta)$. Plugging the value of p, we have

$$\exp\left(-6\varepsilon^2\frac{m}{k}\right) \leq \frac{-2\ln(1-\delta)}{ck} \Longrightarrow m \geq \frac{k}{6\varepsilon^2}\ln\left(\frac{ck}{-2\ln(1-\delta)}\right) = \Theta\left(\frac{k\log(k/\delta)}{\varepsilon^2}\right).$$

□

Using our previous reduction, the bound holds for logistic and hinge loss as well.

Corollary E.7. *Fix $0 < \varepsilon \leq \frac{1}{4}$ and $\delta \leq \frac{1}{4}$. For either the logistic loss $g_{\mathrm{L}}(t) = \ln(1+e^{-t})$ or hinge-loss $g_{\mathrm{H}}(t) = \max\{0, 1-t\}$ and regularizer term $\mathcal{R}(x) \in \{\|x\|_1, \|x\|_2\}$, there exists a distribution $\mathcal{P}$ on $\mathbb{R}^d$ such that* any *importance-sampled approximation that attains*

$$\left|\hat{f}_0(x) - f_0(x)\right| \leq \varepsilon f(x), \quad \forall x \in \mathbb{R}^d,$$

with probability at least $1-\delta$ must have size

$$m = \Omega\Big(\frac{k\log(k/\delta)}{\varepsilon^2}\Big).$$

Proof. Note that $\lim_{\beta\mapsto\infty}\frac{g_L(\beta t)}{\beta} = \lim_{\beta\mapsto\infty}\frac{g_H(\beta t)}{\beta} = \max\{0,-t\}$. Now, using Lemma E.6 and Lemma E.5 we have completed the proof. □

We still need to cover the case of logistic regression with $\mathcal{R}(\cdot) = \|\cdot\|_2^2$ regularization. We note that our general proof also covers other regularization variants, and that the proof outline can be adapted to sigmoid loss as well.

Lemma E.8. *Fix $0 < \varepsilon \le \frac{1}{4}$ and $\delta \le \frac{1}{4}$. Consider the logistic loss $g(r) = \ln(1 + e^{-r})$. For $p \in \{1, 2\}$ and $q \in \{1, 2\}$, set*

$$\mathcal{R}_q^p(x) \coloneqq \frac{\|x\|_q^p}{k}, \qquad f_0(x) \coloneqq \mathbb{E}_{a\sim\mathcal{P}}\big[g(\langle a, x\rangle)\big], \qquad f(x) \coloneqq f_0(x) + \mathcal{R}_q^p(x).$$

Then there exists a distribution $\mathcal{P}$ on $\mathbb{R}^d$ such that the following holds. Any i.i.d. importance–weighted approximation $\hat{f}$ of size m that satisfies

$$\Pr\Big[\sup_{x\in\mathbb{R}^d} \big|\hat{f}(x) - f(x)\big| \le \varepsilon\, f(x)\Big] \ge 1 - \delta$$

for every $k \ge 4$, accuracy parameter $\varepsilon \in (0, 1)$, and failure probability $\delta \in (0, 1/2]$, must obey

$$m \ge \begin{cases} \Omega\left(\dfrac{k}{\varepsilon^2}\,\log\left(\dfrac{k}{\delta}\right)\right) & \text{if } p = 1,\ q \in \{1, 2\}, \\ \Omega\left(\dfrac{k}{\varepsilon^2 \log(k)}\,\log\left(\dfrac{1}{\delta}\right)\right) & \text{if } p = 2,\ q \in \{1, 2\}. \end{cases}$$

Proof. First note that since $\|\cdot\|_1 \ge \|\cdot\|_2$, we only need to prove the statement for $q = 1$. Work in $d = k + 1$ and let $\mathcal{P}$ be uniform on the k atoms

$$p_i = (e_i, 1) = e_i + e_{k+1} \in \mathbb{R}^{k+1} \qquad (i = 1, \ldots, k).$$

An i.i.d. importance sampler uses nonnegative scores $s_i = s(p_i)$ normalized by $\frac{1}{k}\sum_{i=1}^k s_i = 1$ and draws $a_1, \ldots, a_m \sim \mathcal{Q}$ with $\mathcal{Q}\{p_i\} = q_i = \frac{s_i}{k}$. The unbiased importance–weighted estimator is

$$\hat{f}_0(x) = \frac{1}{m}\sum_{t=1}^m \frac{1}{s(a_t)}\, g(\langle a_t, x\rangle), \qquad \mathop{\mathbb{E}}_{x\sim\mathcal{Q}}[\hat{f}_0(x)] = f_0(x).$$

Let $Z_i := |\{t : \ a_t = p_i\}|$ so $(Z_1, \ldots, Z_k) \sim \mathrm{Mult}(m; q_1, \ldots, q_k)$ and $Z_i \sim \mathrm{Bin}(m, q_i)$ with expectation $\mu_i = \mathbb{E}(Z_i) = mq_i$ and variance $\sigma_i^2 := \mathbb{V}(Z_i) = mq_i(1 - q_i)$. Set $H := \{i \in [k] : \ s_i \le 2\}$ and note that Markov's inequality and (as rescaling s leaves $\mathcal{Q}$ and w unchanged, we may assume that) $S = \frac{1}{k}\sum_i s_i = 1$ give $|H| \ge \frac{k}{2}$. For $j \in H$, we have $q_j = s_j/k \le 2/k \le 1/2$ and hence

$$\sigma_j^2 = mq_j(1 - q_j) \ \ge\ \frac{1}{2}\,\frac{ms_j}{k}. \tag{10}$$

Fix $\alpha \ge 1$ to be chosen shortly. For each $j \in [k]$, define the adversarial query

$$x^{(j)} = \alpha(-2e_j + e_{k+1}).$$

Then $\langle p_j, x^{(j)}\rangle = -\alpha$ and $\langle p_i, x^{(j)}\rangle = \alpha$ for $i \ne j$. The logistic identity

$$g(-r) - g(r) = r \qquad \forall r \in \mathbb{R} \tag{11}$$

yields

$$f_0(x^{(j)}) = \frac{1}{k}\, g(-\alpha) + \frac{k-1}{k}\, g(\alpha) = g(\alpha) + \frac{\alpha}{k}. \tag{12}$$

Moreover $3\alpha = \|x^{(j)}\|_1 \le \|x^{(j)}\|_1^2 = 9\alpha^2$, hence for $p \in \{1, 2\}$

$$\mathcal{R}_1^p(x^{(j)}) = \frac{\|x^{(j)}\|_1^p}{k} \ \le\ \frac{9\alpha^p}{k}. \tag{13}$$

Let $\bar{S}_m = \frac{1}{m}\sum_{t=1}^m \frac{1}{s(a_t)} \ge 0$. We may assume that $\bar{S}_m \in [1-\varepsilon, 1+\varepsilon]$, otherwise the approximation fails on $x = 0$ and we are done. Using (11) and $Z_i = \sum_t \mathbf{1}\{a_t = p_i\}$,

$$\hat{f}_0(x^{(j)}) = \frac{Z_j}{ms_j}\, g(-\alpha) + \sum_{i\ne j}\frac{Z_i}{ms_i}\, g(\alpha) \ge \alpha\,\frac{Z_j}{ms_j} + \bar{S}_m\, g(\alpha) \ge \alpha\,\frac{Z_j}{ms_j} + (1 - \varepsilon)\, g(\alpha)\,,$$

which concludes

$$\hat{f}_0(x^{(j)}) - f_0(x^{(j)}) \;\geq\; \alpha\Big(\frac{Z_j}{ms_j} - \frac{1}{k}\Big) \;-\; \varepsilon g(\alpha). \tag{14}$$

In view of (12) and (13), we have

$$f(x^{(j)}) = f_0(x^{(j)}) + \mathcal{R}_1^p(x^{(j)}) \;\leq\; g(\alpha) + \frac{\alpha}{k} + \frac{9\alpha^p}{k} \leq g(\alpha) + \frac{10\alpha^p}{k} \tag{15}$$

Thus, we see that a *sufficient* condition for

$$\hat{f}(x^{(j)}) - f(x^{(j)}) > \varepsilon f(x^{(j)})$$

is

$$\alpha\Big(\frac{Z_j}{ms_j} - \frac{1}{k}\Big) - \varepsilon g(\alpha) \;\geq\; \varepsilon\Big(g(\alpha) + \frac{10\alpha^p}{k}\Big),$$

which is satisfied if $Z_j \geq mq_j + t_j$ for

$$t_j = \varepsilon m q_j \Big(2g(\alpha)k + 10\alpha^{p-1}\Big).$$

Choose $\alpha = \ln\left(\frac{1}{e^{1/k}-1}\right)$ and note that $g(\alpha)k = 1$ and $\alpha \in \left[\frac{1}{2}\log k, \log k\right] = \Theta(\log k)$. This implies

$$t_j = \varepsilon m q_j \Big(2 + 10\alpha^{p-1}\Big) \leq 12\varepsilon m q_j \alpha^{p-1}.$$

Define the (increasing) event

$$E_j = \{Z_j \geq mq_j + t_j\}.$$

If E_j occurs, the approximation fails at the single query $x^{(j)}$. Note $Z_j \sim \mathrm{Bin}(m, q_j)$ and $\sigma_j^2 = mq_j(1-q_j)$. Using $q_j = s_j/k$ and $1 - q_j \geq \frac{1}{2}$, we have $\sigma_j^2 \geq \frac{ms_j}{2k} = \frac{mq_j}{2}$ and then (using $\alpha > 1$)

$$\frac{t_j}{\sigma_j^2} \leq 24\varepsilon\alpha^{p-1} \leq 200,$$

provided that $\varepsilon \leq \frac{25}{3\alpha^{p-1}}$. Also, note that

$$\begin{aligned}\frac{t_j^2}{\sigma_j^2} &\leq 288\varepsilon^2 \frac{ms_j}{k}\alpha^{2p-2} \\ &\leq 576\varepsilon^2 \frac{m}{k}\alpha^{2p-2}.\end{aligned}$$

Hence, using Lemma F.1, for each $j \in H$,

$$\Pr(E_j) \geq p^* = c\exp\Big(-576\varepsilon^2\alpha^{2p-2}\frac{m}{k}\Big). \tag{16}$$

The multinomial vector $(Z_1, \ldots, Z_k)$ is *negatively associated* (see Joag-Dev and Proschan (1983)), in particular, for increasing or decreasing events depending on disjoint coordinates we have

$$\Pr\Big(\bigcap_{j\in H} A_j\Big) \leq \prod_{j\in H} \Pr(A_j).$$

Since each E_j^c depends only on Z_j and is decreasing, we obtain

$$\Pr\Big(\bigcap_{j\in H} E_j^c\Big) \leq \prod_{j\in H} \Pr(E_j^c) \leq \Big(1-p^*\Big)^{|H|} \leq \Big(1-p^*\Big)^{k/2}.$$

Accordingly,

$$\Pr\Big(\bigcup_{j\in H} E_j\Big) = 1 - \Pr\Big(\bigcap_{j\in H} E_j^c\Big) \geq\; 1 - \Big(1-p^*\Big)^{k/2}.$$

For the approximation to succeed with probability at least $1-\delta$ we must have

$$1 - \Big(1-p^*\Big)^{k/2} \leq \delta,$$

which implies

$$p^* \le 1 - (1-\delta)^{2/k} = 1 - e^{-\frac{2}{k}\ln\frac{1}{1-\delta}} \le \frac{2}{k}\ln\frac{1}{1-\delta} \le \frac{2}{k}\frac{\delta}{1-\delta} \le \frac{4\delta}{3k}$$

where the first two inequalities follow from $e^x \ge 1 + x$ for $x \in \mathbb{R}$ and the third from the assumption $\delta \le 1/2$. Solving

$$p^* \le \frac{4\delta}{3k}$$

implies

$$m \ge \Omega\left(\frac{k}{\varepsilon^2\alpha^{2p-2}}\ln\frac{k}{\delta}\right).$$

□

As indicated above, we adapt the proof to sigmoid loss as well.

Lemma E.9. *Fix $0 < \varepsilon \le \frac{1}{4}$ and $\delta \le \frac{1}{4}$. Consider the sigmoid loss $g(r) = 1/(1+e^r)$. For $p \in \{1,2\}$ and $q \in \{1,2\}$, set*

$$\mathcal{R}_q^p(x) \coloneqq \frac{\|x\|_q^p}{k}, \qquad f_0(x) \coloneqq \mathbb{E}_{a\sim\mathcal{P}}\big[g(\langle a, x\rangle)\big], \qquad f(x) \coloneqq f_0(x) + \mathcal{R}_q^p(x).$$

Then there exists a distribution $\mathcal{P}$ on $\mathbb{R}^d$ such that the following holds. Any i.i.d. importance–weighted approximation $\hat{f}$ of size m that satisfies

$$\Pr\Big[\sup_{x\in\mathbb{R}^d}\big|\hat{f}(x) - f(x)\big| \le \varepsilon\, f(x)\Big] \ge 1 - \delta$$

for every $k \ge 4$, accuracy parameter $\varepsilon \in (0,1)$, and failure probability $\delta \in (0, 1/2]$, must obey

$$m \ge \begin{cases} \Omega\left(\dfrac{k}{\varepsilon^2\log(k)}\,\log\left(\dfrac{1}{\delta}\right)\right) & \text{if } p = 1,\ q \in \{1,2\}, \\ \Omega\left(\dfrac{k}{\varepsilon^2\log^3(k)}\,\log\left(\dfrac{1}{\delta}\right)\right) & \text{if } p = 2,\ q \in \{1,2\}. \end{cases}$$

Proof. First note that since $\|\cdot\|_1 \ge \|\cdot\|_2$, we only need to prove the statement for $q = 1$. Work in $d = k + 1$ and let $\mathcal{P}$ be uniform on the k atoms

$$p_i = (e_i, 1) = e_i + e_{k+1} \in \mathbb{R}^{k+1} \qquad (i = 1, \dots, k).$$

An i.i.d. importance sampler uses nonnegative scores $s_i = s(p_i)$ normalized by $\frac{1}{k}\sum_{i=1}^k s_i = 1$ and draws $a_1, \dots, a_m \sim \mathcal{Q}$ with $\mathcal{Q}\{p_i\} = q_i = \frac{s_i}{k}$. The unbiased importance–weighted estimator is

$$\hat{f}_0(x) = \frac{1}{m}\sum_{t=1}^m \frac{1}{s(a_t)}\, g(\langle a_t, x\rangle), \qquad \mathop{\mathbb{E}}_{x\sim\mathcal{Q}}[\hat{f}_0(x)] = f_0(x).$$

Let $Z_i := |\{t : \ a_t = p_i\}|$ so $(Z_1, \dots, Z_k) \sim \mathrm{Mult}(m; q_1, \dots, q_k)$ and $Z_i \sim \mathrm{Bin}(m, q_i)$ with expectation $\mu_i = \mathbb{E}(Z_i) = mq_i$ and variance $\sigma_i^2 := \mathbb{V}(Z_i) = mq_i(1-q_i)$. Set $H := \{i \in [k] : \ s_i \le 2\}$ and note that Markov's inequality and (as rescaling s leaves $\mathcal{Q}$ and w unchanged, we may assume that) $S = \frac{1}{k}\sum_i s_i = 1$ give $|H| \ge \frac{k}{2}$. For $j \in H$, we have $q_j = s_j/k \le 2/k \le 1/2$ and hence

$$\sigma_j^2 = mq_j(1-q_j) \ \ge \ \frac{1}{2}\,\frac{ms_j}{k}. \tag{17}$$

Fix $\alpha \ge 1$ to be chosen shortly. For each $j \in [k]$, define the adversarial query

$$x^{(j)} = \alpha(-2e_j + e_{k+1}).$$

Then $\langle p_j, x^{(j)}\rangle = -\alpha$ and $\langle p_i, x^{(j)}\rangle = \alpha$ for $i \ne j$.

Choose $\alpha = \ln(k-1) > 1$ and note that $g(\alpha)k = 1$ and $\alpha = \Theta(\log k)$. The sigmoid identity

$$g(-r) = 1 - g(r) \qquad \forall r \in \mathbb{R} \tag{18}$$

yields that

$$h(r) = g(-r) - g(r) = 1 - 2g(r) \qquad \forall r \in \mathbb{R} \tag{19}$$

and thus

$$h(\alpha) = 1 - 2g(\alpha) = 1 - \frac{2}{k} \in [0.5, 1] \ . \tag{20}$$

Therefore we have that

$$f_0(x^{(j)}) = \frac{1}{k}\, g(-\alpha) + \frac{k-1}{k}\, g(\alpha) = \frac{1}{k}\big(g(-\alpha) - g(\alpha)\big) + g(\alpha) = \frac{h(\alpha)}{k} + g(\alpha). \tag{21}$$

Moreover $3\alpha = \|x^{(j)}\|_1 \le \|x^{(j)}\|_1^2 = 9\alpha^2$, hence for $p \in \{1, 2\}$

$$\mathcal{R}_1^p(x^{(j)}) = \frac{\|x^{(j)}\|_1^p}{k} \ \le \ \frac{9\alpha^p}{k}. \tag{22}$$

Let $\bar{S}_m = \frac{1}{m}\sum_{t=1}^m \frac{1}{s(a_t)} \ge 0$. We may assume that $\bar{S}_m \in [1-\varepsilon, 1+\varepsilon]$, otherwise the approximation fails on $x = 0$ and we are done. Using (19) and $Z_i = \sum_t \mathbf{1}\{a_t = p_i\}$,

$$\begin{aligned}\hat{f}_0(x^{(j)}) = \frac{Z_j}{ms_j}\, g(-\alpha) + \sum_{i \ne j} \frac{Z_i}{ms_i}\, g(\alpha) &= \frac{Z_j}{ms_j}\big(g(-\alpha) - g(\alpha)\big) + \bar{S}_m\, g(\alpha) \\ &\ge \frac{Z_j}{ms_j} h(\alpha) + (1-\varepsilon)\, g(\alpha)\,,\end{aligned}$$

which concludes by (20) that

$$\hat{f}_0(x^{(j)}) - f_0(x^{(j)}) \ \ge \ \Big(\frac{Z_j}{ms_j} - \frac{1}{k}\Big) h(\alpha) - \varepsilon g(\alpha) \ge \Big(\frac{Z_j}{2ms_j} - \frac{1}{2k}\Big) - \varepsilon g(\alpha). \tag{23}$$

In view of (12) and (13), and using (20), we have

$$f(x^{(j)}) = f_0(x^{(j)}) + \mathcal{R}_1^p(x^{(j)}) \ \le \ g(\alpha) + \frac{h(\alpha)}{k} + \frac{9\alpha^p}{k} \le g(\alpha) + \frac{10\alpha^p}{k} \tag{24}$$

Thus, we see that a *sufficient* condition for

$$\hat{f}(x^{(j)}) - f(x^{(j)}) > \varepsilon f(x^{(j)})$$

is

$$\Big(\frac{Z_j}{2ms_j} - \frac{1}{2k}\Big) - \varepsilon g(\alpha) \ \ge \ \varepsilon\Big(g(\alpha) + \frac{10\alpha^p}{k}\Big),$$

which is satisfied if $Z_j \ge mq_j + t_j$ for

$$t_j = \varepsilon m q_j \Big(4g(\alpha)k + 20\alpha^p\Big).$$

This implies

$$t_j = \varepsilon m q_j \Big(4 + 20\alpha^p\Big) \le 24\varepsilon m q_j \alpha^p.$$

Define the (increasing) event

$$E_j = \{Z_j \ge mq_j + t_j\}.$$

If E_j occurs, the approximation fails at the single query $x^{(j)}$. Note $Z_j \sim \mathrm{Bin}(m, q_j)$ and $\sigma_j^2 = mq_j(1-q_j)$. Using $q_j = s_j/k$ and $1 - q_j \ge \frac{1}{2}$, we have $\sigma_j^2 \ge \frac{ms_j}{2k} = \frac{mq_j}{2}$ and then (using $\alpha > 1$)

$$\frac{t_j}{\sigma_j^2} \le 48\varepsilon\alpha^p \le 200,$$

provided that $\varepsilon \le \frac{4}{\alpha^p}$. Also, note that

$$\begin{aligned}\frac{t_j^2}{\sigma_j^2} &\le 1152\varepsilon^2 \frac{ms_j}{k}\alpha^{2p} \\ &\le 2304\varepsilon^2 \frac{m}{k}\alpha^{2p}.\end{aligned}$$

Hence, using Lemma F.1, for each $j \in H$,

$$\Pr(E_j) \geq p_* = c \exp\Big(- 2304 \varepsilon^2 \alpha^{2p} \frac{m}{k} \Big). \tag{25}$$

The multinomial vector $(Z_1, \ldots, Z_k)$ is *negatively associated* (see Joag-Dev and Proschan (1983)), in particular, for increasing or decreasing events depending on disjoint coordinates we have

$$\Pr\Big(\bigcap_{j \in H} A_j \Big) \leq \prod_{j \in H} \Pr(A_j).$$

Since each E_j^c depends only on Z_j and is decreasing, we obtain

$$\Pr\Big(\bigcap_{j \in H} E_j^c \Big) \leq \prod_{j \in H} \Pr(E_j^c) \leq \Big(1 - p_*\Big)^{|H|} \leq \Big(1 - p_*\Big)^{k/2}.$$

Accordingly,

$$\Pr\Big(\bigcup_{j \in H} E_j \Big) = 1 - \Pr\Big(\bigcap_{j \in H} E_j^c \Big) \geq \; 1 - \Big(1 - p_*\Big)^{k/2}.$$

For the approximation to succeed with probability at least $1 - \delta$ we must have

$$1 - \Big(1 - p_*\Big)^{k/2} \leq \delta,$$

which implies

$$p^* \leq 1 - (1 - \delta)^{2/k} = 1 - e^{-\frac{2}{k} \ln \frac{1}{1-\delta}} \leq \frac{2}{k} \ln \frac{1}{1-\delta} \leq \frac{2}{k} \frac{\delta}{1-\delta} \leq \frac{4\delta}{3k}$$

where the first two inequalities follow from $e^x \geq 1 + x$ for $x \in \mathbb{R}$ and the third from the assumption $\delta \leq 1/2$. Solving

$$p^* \leq \frac{4\delta}{3k}$$

implies

$$m \geq \Omega\left(\frac{k}{\varepsilon^2 \alpha^{2p}} \ln \frac{k}{\delta} \right).$$

□

E.3 Impossibility of dimension-free or sublinear sampling for ReLU with $\|\cdot\|_2^2$ regularization

We first show a simple $\Omega(d \log d)$ lower bound against the combination of ReLU and ridge regularization $\mathcal{R}(x) = \|x\|_2^2$. Then we establish an $\Omega(N \log N)$ lower bound with more sophisticated techniques. We summarize both results in the following theorem.

Theorem 3.7. *Consider the ReLU function where $g(r) = \max\{0, -r\}$, and the $\mathcal{R}(x) = \|x\|_2^2$ regularizer. For any $n \geq d$ there exist distributions supported on n points in $\mathbb{R}^d$, such that with high probability $m = \Omega(n \log n)$ samples are required to satisfy Equation* (1)*.*

Proof. The theorem follows directly from Lemma E.10 for $n = d$, and Lemma E.11 for $n \geq d + 1$. □

Lemma E.10. *Consider the ReLU function where $g(r) = \max\{0, -r\}$, and the $\mathcal{R}(x) = \|x\|_2^2$ regularizer. Then there exists a distribution such that with high probability $m = \Omega(d \log d)$ uniform or row-norm samples are required to satisfy*

$$\forall x \in \mathbb{R}^d \colon \; \left| f_0(x) - \left(\frac{1}{m} \sum_{i=1}^{m} w_i g(a_i x) \right) \right| \leq \varepsilon f(x)\,.$$

Proof. Consider the uniform distribution over the d standard unit vectors, so $P(a = e_j) = 1/d$ for each $j \in [d]$. Set $x = \alpha e_i$ for some small scalar $\alpha > 0$ and index $i \in [d]$ yet to be determined.

Irrespective of the choice of parameters α and i, the original loss and regularization for x equals

$$\frac{1}{d}\sum_{j=1}^{d} g(e_j x) + \frac{\mathcal{R}(x)}{k} = \frac{\alpha}{d} + \frac{\alpha^2}{k}\,.$$

However, by the classic coupon collector's theorem Erdős and Rényi (1961), if we take less than $m \le \Omega(d \log d)$ samples, then with high probability some vector e_{i^*} will not be present, and we set $i = i^*$ in the definition of x. Then every sample satisfies $g(a_i x) = 0$ and the loss degrades to $0 + \alpha^2/k$.

$$\frac{1}{m}\sum_{j=1}^{m} g(e_j x) + \frac{\mathcal{R}(x)}{k} = 0 + \frac{\alpha^2}{k}\,.$$

Now choosing $\alpha \coloneqq 2k/3d$ and using $\varepsilon < 1/2$, we get that

$$\varepsilon\left(\frac{\alpha}{d} + \frac{\alpha^2}{k}\right) < \frac{k}{3d^2} + \frac{2k}{9d^2} = \frac{5k}{9d^2} < \frac{2k}{3d^2} = \frac{\alpha}{d}\,.$$

Thus the error on x is

$$\left|\frac{\alpha}{d} + \frac{\alpha^2}{k} - \left(0 + \frac{\alpha^2}{k}\right)\right| = \frac{\alpha}{d} > \varepsilon\left(\frac{\alpha}{d} + \frac{\alpha^2}{k}\right)\,.$$

□

Lemma E.11. *Consider the ReLU function where $g(r) = \max\{0, -r\}$, and the $\mathcal{R}(x) = \|x\|_2^2$ regularizer. For any fixed d and arbitrary $N \ge d + 1$, there exists a distribution supported on N points with uniform mass for which with high probability $m = \Omega(\varepsilon^{-2} N \log N)$ samples are required to satisfy*

$$\forall x \in \mathbb{R}^d\colon\ \left|f_0(x) - \left(\frac{1}{m}\sum_{i=1}^{m} w_i g(a_i x)\right)\right| \le \varepsilon f(x)\,.$$

Proof. Consider N points $a_1, \ldots, a_N$ on the moment curve $\gamma(t) = (1, t, t^2, \ldots, t^d)$ with distinct t's. Place equal mass on these points:

$$\mathcal{P}\{a_j\} = \frac{1}{N}\,, \qquad j = 1, \ldots, N.$$

Write $S = 1$ without loss of generality (rescaling s leaves $\mathcal{Q}$ and w unchanged). Define

$$J = \{j \in [N]\colon s(a_j) \le 2\}$$

and note that we must have $|J| \ge N/2$.

We know that every a_j is a vertex of the cyclic polytope $\operatorname{conv}\{a_1, \ldots, a_N\}$. So, there must be a supporting hyperplane with a normal x_j for which we have

$$\langle a_j, x_j\rangle < 0, \qquad \langle a_i, x_j\rangle \ge 0 \quad (\forall i \ne j).$$

For ReLU $g(t) = \max\{0, -t\}$, this gives

$$g(\langle a_j, x_j\rangle) = c_j > 0, \qquad \text{and} \qquad g(\langle a_i, x_j\rangle) = 0 \qquad \text{for all } i \ne j.$$

Let N_j be the number of occurrence of a_j in the m samples.

We can now consider scaling the input set $\{x_1, \ldots, x_N\}$ uniformly down by a factor $\eta \in (0, 1]$. We will ultimately drive $\eta \mapsto 0$; while we do this, the ReLu part in f_0 scales linearly with η, but because the residual has a squared term it decreases quadratically, and can be factored out of the analysis. As expected, without a residual (and no other conditions), there can be no useful sampling bound.

In particular, observe that

$$f_0(\eta x_j) = \frac{\eta c_j}{N}, \qquad \hat{f}_0(\eta x_j) = \frac{\eta c_j}{m s(a_j)} N_j, \qquad \frac{\|\eta x_j\|_2^2}{k} = \frac{\eta^2 \|x_j\|_2^2}{k}.$$

Since $\mathcal{Q}(a_j) = q_j = \frac{s(\sigma e_j)}{N}$, we obtain

$$N_j \sim \text{Bionomial}(m, q_j)\,.$$

So, if we write out the approximation inequality, we get

$$\begin{aligned}
\left|\frac{\eta c_j}{ms(a_j)}N_j - \frac{\eta c_j}{N}\right| \le \varepsilon\left(\frac{\eta c_j}{N} + \frac{\eta^2\|x_j\|_2^2}{k}\right) \qquad &\xrightarrow{\times\frac{1}{\eta c_j}} \qquad \left|\frac{N_j}{ms(a_j)} - \frac{1}{N}\right| \le \varepsilon\left(\frac{1}{N} + \frac{\eta\|x_j\|_2^2}{kc_j}\right) \\
&\xrightarrow{\eta\mapsto 0} \qquad \left|\frac{N_j}{ms(a_j)} - \frac{1}{N}\right| \le \varepsilon\frac{1}{N} \\
&\xrightarrow{\times ms(a_j)} \qquad |N_j - \mu_j| \le \varepsilon\mu_j.
\end{aligned}$$

Define F_j as the event $|N_j - \mu_j| > \varepsilon\mu_j$. Note that F_j implies that the approximation inequality fails. Also, define F_j' as the event that $N_j - \mu_j > \varepsilon\mu_j$ and notice $F_j' \subseteq F_j$, so F_j' implies failure of obtaining the desired approximation.

Note that for each $j \in J$, we have $s(a_j) \le 2$ which concludes $q_j \le \frac{2}{N} \le \frac{1}{2}$ and thus $\frac{q_j}{1-q_j} \le \frac{4}{N}$.

Note that $\mathbb{V}(N_j) = mq_j(1-q_j)$. Using Lemma F.1 with $t = \varepsilon mq_j$ with $\varepsilon \le 1/600$, we obtain

$$\Pr(F_j) \ge \Pr(F_j') \ge c\exp\left(-\frac{\varepsilon^2 m^2 q_j^2}{3mq_j(1-q_j)}\right) = c\exp\left(-\varepsilon^2\frac{mq_j}{3(1-q_j)}\right) \ge c\exp\left(-\varepsilon^2\frac{4}{3}\frac{m}{N}\right) = p$$

which is independent of j. Define, for each $j \in J$,

$$F_j' \;=\; \{\, N_j > z_j \,\}\,, \qquad z_j = (1+\varepsilon)\,\mu_j\,, \qquad \mu_j = \mathbb{E}[N_j] = mq_j\,.$$

Then

$$\begin{aligned}
\Pr\Big(\bigcup_{j\in J} F_j'\Big) &= 1 - \Pr\Big(\bigcap_{j\in J}(F_j')^c\Big) \\
&\ge 1 - \prod_{j\in J}\Pr\big((F_j')^c\big) \qquad \textit{(negative association of multinomial coordinates)} \\
&= 1 - \prod_{j\in J}\big(1 - \Pr(F_j')\big) \\
&\ge 1 - (1-p)^{|J|} \\
&\ge 1 - (1-p)^{N/2}\,.
\end{aligned}$$

The vector of counts $(N_j)_{j\in J}$ is a subcollection of a multinomial count vector, which is known to be *negatively associated*. Here $(F_j')^c = \{N_j \le z_j\}$ is decreasing in N_j, so $\Pr\big(\bigcap_{j\in J}(F_j')^c\big) \le \prod_{j\in J}\Pr\big((F_j')^c\big)$ (see Joag-Dev and Proschan (1983)), which yields the second line above. On the other hand $1-(1-p)^{\frac{N}{2}} \le \Pr(\text{approximation failure}) \le \delta$ which concludes

$$1 - \delta \le (1-p)^{\frac{N}{2}} \le \exp\left(-\frac{Np}{2}\right)$$

and thus $Np \le -2\ln(1-\delta)$. Plugging the value of p, we have

$$\exp\left(-\frac{4}{3}\varepsilon^2\frac{m}{N}\right) \le \frac{-2\ln(1-\delta)}{cN} \Longrightarrow m \ge \Theta\left(\frac{N\log N}{\varepsilon^2}\right).$$

□

F Technical lemmas

The following classic anti-concentration result is due to Feller and will help in proving lower bounds.

Lemma F.1 (Feller, 1943)**.** *Let Z be a sum of independent random variables, each attaining values in $[0,1]$, and let $\sigma = \sqrt{\mathbb{V}(Z)} \geq 200$. Then for all $t \in [0, \frac{\sigma^2}{100}]$ we have*

$$\Pr[Z \geq \mathbb{E}[Z] + t] \geq c \cdot \exp(-t^2/(3\sigma^2)),$$

where $c > 0$ is some fixed constant.

Lemma F.2. *Let $Z_1, \ldots, Z_m \sim \mathcal{N}(0, \sigma_i)$ be m Gaussians with $\sigma_i \leq \sigma_{\max}$. Then $\mathbb{E}(\max_{i\in[m]} |Z_i|) = O(\sigma_{\max}\sqrt{\log m})$, and $\mathbb{E}(\max_{i\in[m]} |Z_i|^2) = O(\sigma^2_{\max}\log m)$.*

Proof. For the first claim, let $\lambda \coloneqq \frac{\sqrt{\ln m}}{\sigma_{\max}}$. We have by Jensen's inequality and the well-known moment generating function of Gaussians $\mathbb{E}\left(\exp\left(\lambda Z_j\right)\right) = \exp\left(\sigma_j^2\lambda^2/2\right)$, that

$$\begin{aligned}
\exp\left(\lambda\,\mathbb{E}\left(\max_{j\in[m]} Z_j\right)\right) &\leq \mathbb{E}\left(\exp\left(\lambda \max_{j\in[m]} Z_j\right)\right) = \mathbb{E}\left(\max_{j\in[m]} \exp\left(\lambda Z_j\right)\right) \\
&\leq \mathbb{E}\left(\sum_{j\in[m]} \exp(\lambda Z_j)\right) = \sum_{j\in[m]} \mathbb{E}\left(\exp\left(\lambda Z_j\right)\right) \\
&= \sum_{j\in[m]} \exp\left(\sigma_j^2\lambda^2/2\right) \leq m \cdot \exp\left(\sigma^2_{\max}\lambda^2/2\right).
\end{aligned}$$

Now, taking the natural logarithm on both sides and rearranging, we get

$$\mathbb{E}\left(\max_{j\in[m]} Z_j\right) \leq \frac{\ln m}{\lambda} + \frac{\sigma^2_{\max}\lambda}{2} = \sigma_{\max}\sqrt{\ln m} + \frac{\sigma_{\max}\sqrt{\ln m}}{2} \leq 2\sigma_{\max}\sqrt{\ln m}\,.$$

By symmetry, we have the same bound for the collection of random variables $-Z_j, j \in [m]$. Thus the bound follows for $\max_{j\in[m]} |Z_j| = \max_{j\in[m]} \max\{Z_j, -Z_j\} = \max\{\max_{j\in[m]} Z_j, \max_{j\in[m]} -Z_j\} \leq \max_{j\in[m]} Z_j + \max_{j\in[m]} -Z_j \leq 2\max_{j\in[m]} Z_j$ up to another factor 2 as well.

For the second claim, set $\lambda \coloneqq \frac{1}{4\sigma^2_{\max}}$. Note that $|Z_j|^2 = Z_j^2$ follow a squared Gaussian distribution. We repeat the above argument using the appropriate moment generating function $\mathbb{E}\left(\exp\left(\lambda Z_j\right)\right) = \left(1 - 2\sigma_i^2\lambda\right)^{-1/2} \leq \left(1 - 2\sigma^2_{\max}\lambda\right)^{-1/2}$ to obtain

$$\exp\left(\lambda\,\mathbb{E}\left(\max_{j\in[m]} Z_j\right)\right) \leq m \cdot \left(1 - 2\sigma^2_{\max}\lambda\right)^{-1/2}.$$

Taking the natural logarithm on both sides and rearranging, we get

$$\mathbb{E}\left(\max_{j\in[m]} Z_j\right) \leq \frac{\ln m}{\lambda} + \frac{1}{2\lambda}\ln\left(\frac{1}{1 - 2\sigma^2_{\max}\lambda}\right) = 4\sigma^2_{\max}\ln m + 2\sigma^2_{\max}\ln(2) \leq 6\sigma^2_{\max}\ln m\,.$$

□

Lemma F.3. *For $i \in [m]$, let $X_i \sim N(0, \sigma_i^2), Z \sim N(0, \sum_{i=1}^{\ell} \sigma_i^2)$. Then $\mathbb{E}(\max_{i\in[m]} |X_i|^2) \leq \mathbb{E}(|Z|^2)$.*

Proof. We have that

$$\mathbb{E}\left(\max_{i\in[m]} |X_i|^2\right) = \mathbb{E}\left(\max_{i\in[m]} X_i^2\right) \leq \mathbb{E}\left(\sum_{i\in[m]} X_i^2\right) = \sum_{i\in[m]} \mathbb{E}(X_i^2) = \sum_{i\in[m]} \sigma_i^2 = \mathbb{E}(Z^2) = \mathbb{E}(|Z|^2).$$

□

The following technical lemma will be helpful for bounding the entropy integral.

Lemma F.4 (Woodruff and Yasuda, 2023)**.** *Let $0 < \lambda \leq 1$. Then,*

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

in particular for $0 < \lambda \leq 1/3$, we have that $\int_0^{\lambda} \sqrt{\log\frac{1}{t}}\,dt \leq 2\lambda\left(\sqrt{\log(1/\lambda)}\right)$

Our proofs use the following Lipschitz contraction bound from (Alishahi and Phillips, 2024), which generalizes Theorem 4.12 of Ledoux and Talagrand (1991).

Lemma F.5 (Ledoux and Talagrand, 1991; Alishahi and Phillips, 2024)**.** *Let $T \subset \mathbb{R}^m$ be a bounded set. For any L-Lipschitz functions $\phi_i \colon \mathbb{R} \to \mathbb{R}$ with $\phi_i(0) = 0$ for each $i \in [m]$, we have*

$$\mathop{\mathbb{E}}_{\sigma \sim \mathcal{U}\{-1,1\}^m} \sup_{t \in T} \left| \sum_{i=1}^m \sigma_i \phi_i(t_i) \right| \leq 2L \mathop{\mathbb{E}}_{\sigma \sim \mathcal{U}\{-1,1\}^m} \sup_{t \in T} \left| \sum_{i=1}^m \sigma_i t_i \right| .$$

One can find the following proposition in Khare and Rajaratnam (2023).

Proposition F.6 (Kahane, 1964; Latała and Oleszkiewicz, 1994)**.** *For all $p, q \in [1, \infty)$, there exists a universal constant $C_{p,q} > 0$ depending only on p, q, such that for all choices of Banach spaces $\mathbb{B}$ equipped with norm $\|\cdot\|$, finite sets of vectors $x_1, \ldots, x_n \in \mathbb{B}$, and independent Rademacher variables $r_1, \ldots, r_n$,*

$$\left[\mathbb{E} \left\| \sum_{k=1}^n r_k x_k \right\|^q \right]^{1/q} \leq C_{p,q} \left[\mathbb{E} \left\| \sum_{k=1}^n r_k x_k \right\|^p \right]^{1/p} .$$

If moreover $1 = p \leq q \leq 2$, then the constant $C_{1,q} = 2^{1-1/q}$ is optimal.

The following corollary will be used in our derivations.

Corollary F.7. *Let $x_1, \ldots, x_n \in \mathbb{R}^d$, and let $\sigma_1, \ldots, \sigma_n$ be i.i.d. Rademacher random variables. Let $\ell > 2$ be an even integer. Then it holds that*

$$\mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^\ell \leq \ell^{\ell/2} \left[\sum_{i=1}^n \|x_i\|_2^2 \right]^{\ell/2} . \tag{26}$$

Proof. We would like to apply Proposition F.6. We are interested in the case where $\mathbb{B} = (\mathbb{R}^d, \|\cdot\|_2)$ and $p = 2 < q = \ell$, where $\ell = 2\lceil \log(\delta^{-1}) \rceil$ is an even integer to be used in the exponent of our moment bound. It is known that in this case $C_{2,\ell} = \sqrt{2}(\Gamma((\ell+1)/2)/\sqrt{\pi})^{1/\ell}$ Haagerup (1981).

We first show that $C_{2,\ell}^\ell \leq \ell^{\ell/2}$. To this end, we use the fact that $\Gamma(1/2) = \sqrt{\pi}$, and for any real number x it holds that $\Gamma(x+1) = x\Gamma(x)$. Since ℓ is an even non-negative integer, we have that

$$\begin{aligned} C_{2,\ell}^\ell = \frac{2^{\ell/2}}{\sqrt{\pi}} \Gamma\left(\frac{\ell+1}{2}\right) &= \frac{2^{\ell/2}}{\sqrt{\pi}} \prod_{i=1}^{\ell/2} \left(\frac{\ell}{2} - i + \frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right) = 2^{\ell/2} \prod_{i=1}^{\ell/2} \left(\frac{\ell}{2} - i + \frac{1}{2}\right) \\ &= 2^{-\ell/2} \prod_{i=1}^{\ell/2} (\ell - (2i-1)) = 2^{-\ell/2} \frac{(\ell-1)!}{\prod_{i=1}^{\ell/2-1} (\ell - 2i)} \\ &= 2^{-\ell/2} \frac{\ell \cdot (\ell-1)!}{2^{\ell/2} \frac{\ell}{2} \cdot \prod_{i=1}^{\ell/2-1} \left(\frac{\ell}{2} - i\right)} = 2^{-\ell} \frac{\ell!}{\frac{\ell}{2}!} \leq \ell^{\ell/2} . \end{aligned}$$

Using Proposition F.6, we conclude that

$$\mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i \right\|^\ell \leq \ell^{\ell/2} \left(\mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2 \right)^{\ell/2} = \ell^{\ell/2} \left(\sum_{i=1}^n \|x_i\|_2^2 \right)^{\ell/2} . \tag{27}$$

To see that the equality holds, note that σ_i are independent zero-mean and unit-variance random variables. This implies for $i \neq j$ that $\mathbb{E}(\sigma_i \sigma_j) = \mathbb{E}(\sigma_i)\,\mathbb{E}(\sigma_j) = 0$. Thus only squared terms survive and we can use $\mathbb{E}(\sigma_i^2) = 1$ on these squared terms. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Lower bounds and impossibility results are proven in the paper specifying and corroborating limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All our results are proven in the appendix with all details and the main body contains high level descriptions and intuition. Only the technical overview of our lower bounds had to be deferred to Appendix A due to page limitations, which is clearly stated in the main part. The omitted parts will be put back into the main body, given sufficient space, e.g. in a cam-ready version or in a final 'full version' on arxiv.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: the paper does not include experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not include experiments requiring data or code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our purely theoretical/mathematical research conforms in every respect with the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper presents theoretical/mathematical work whose goal is to advance the field of Machine Learning. If there are potential societal consequences of our work, there are none we feel must be specifically highlighted here.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: this research does not involve LLMs at all.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.