Agnostic Active Learning of Single Index Models with Linear Sample Complexity

Aarshvi Gajjar   AARSHVI@NYU.EDU
New York University

Wai Ming Tai   WAIMING.TAI@NTU.EDU.SG
Nanyang Technological University

Xingyu Xu   XINGYUXU@ANDREW.CMU.EDU
Carnegie Mellon University

Chinmay Hegde   CHINMAY.H@NYU.EDU
New York University

Christopher Musco   CMUSCO@NYU.EDU
New York University

Yi Li   YILI@NTU.EDU.SG
Nanyang Technological University

Abstract

We study active learning methods for single index models of the form $F(\boldsymbol{x}) = f(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$, where $f : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^d$. In addition to their theoretical interest as simple examples of non-linear neural networks, single index models have received significant recent attention due to applications in scientific machine learning like surrogate modeling for partial differential equations (PDEs). Such applications require sample-efficient active learning methods that are robust to adversarial noise. I.e., that work even in the challenging agnostic learning setting.

We provide two main results on agnostic active learning of single index models. First, when f is known and Lipschitz, we show that $O(d)$ samples collected via statistical leverage score sampling are sufficient to learn a near-optimal single index model. Leverage score sampling is simple to implement, efficient, and already widely used for actively learning linear models. Our result requires no assumptions on the data distribution, is optimal up to log factors, and improves quadratically on a recent $O(d^2)$ bound of Gajjar et al. (2023). Second, we show that $\tilde{O}(d)$ samples suffice even in the more difficult setting when f is *unknown*. Our results leverage tools from high dimensional probability, including Dudley's inequality and dual Sudakov minoration, as well as a novel, distribution-aware discretization of the class of Lipschitz functions.

Keywords: active learning, leverage scores, single index models, high dimensional probability

1. Introduction

The goal in active learning is to effectively fit a model based on a small amount of labeled data selected from a vast pool of unlabeled data. Active learning finds applications in settings where labeling data is expensive. For example, a common task in scientific machine learning is to learn parameter-to-solution maps for a parametric partial differential equation (PDE) (Cohen and DeVore, 2015). Obtaining each label involves the high-cost numerical solution of the PDE, so learning based on few labels is necessary for efficiency. In other settings, such as experimental science or sensor placement, each label requires running a physical experiment or placing a physical device, so is even more expensive (Alexanderian et al., 2016).

In this paper, we study active learning algorithms that address the challenging *agnostic setting*, aka the adversarial noise setting. Focusing on least squares error for concreteness, given a set of examples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, query access to an arbitrary target vector $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, and some function class $\mathcal{H}$, our goal is to use a small number of queries to find $\tilde{h} \in \mathcal{H}$ such that:

$$\sum_{i=1}^{n} (\tilde{h}(\boldsymbol{x}_i) - y_i)^2 \leq \min_{h \in \mathcal{H}} \sum_{i=1}^{n} (h(\boldsymbol{x}_i) - y_i)^2 + \Delta,$$

for some error parameter Δ. This setting is agnostic because we make *no assumptions* on how y_i is related to $\boldsymbol{x}_i$. E.g., in contrast to most work on statistical experimental design (Pukelsheim, 2006), contextual bandits (Xu et al., 2018), and other formulations of active learning (Balcan et al., 2006; Kääriäinen, 2006), we do not assume a "realizable" setting where $y_i = h^*(\boldsymbol{x}_i)$ or $y_i = h^*(\boldsymbol{x}_i) + \eta_i$ for some ground truth $h^* \in \mathcal{H}$ and mean-zero random noise η_i. We want to compete with the best approximation to y in $\mathcal{H}$, *even if that approximation is poor.* The agnostic setting is important in scientific applications where model misspecification is expected: a simple and efficient machine learning model is being used to approximate a complex physical process or function.

Algorithms for agnostic active learning have received significant recent attention. Even for the simplest possible setting of relative-error linear regression, where $\mathcal{H}$ contains linear functions of the form $h(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{w} \rangle$ and $\Delta = \varepsilon \cdot \min_{h \in \mathcal{H}} \sum_{i=1}^{n} (h(\boldsymbol{x}_i) - y_i)^2$ for $\varepsilon \in (0, 1)$, the optimal active sample complexity was only recently settled to be $\Theta(d/\varepsilon)$ in the agnostic setting (Chen and Price, 2019). Other recent work establishes sample complexity bounds for ℓ_p linear regression (Chen and Derezinski, 2021; Musco et al., 2022), logistic regression (Mai et al., 2021; Munteanu et al., 2018), polynomial regression (Shimizu et al., 2024), and kernel learning (Erdélyi et al., 2020).

1.1. Single Index Models

In contrast to the majority of this prior work, and motivated by applications in scientific machine learning, in this paper we study agnostic active learning methods for *non-linear function families*. In particular, we are interested in the class of *single index models* of the form:

$$h(\boldsymbol{x}) = f(\langle \boldsymbol{x}, \boldsymbol{w} \rangle),$$

where $f : \mathbb{R} \to \mathbb{R}$ is either a known non-linearity (ReLU, sigmoid, etc.) or a learnable function. Also known as "ridge functions" or "plane waves", single index models play an important role in many estimation problems and have been extensively studied in statistics (Hristache et al., 2001; Härdle et al., 2004; Dalalyan et al., 2008). They are effective at modeling physical phenomena, so have also been applied e.g. to PDE surrogate modeling (Cohen et al., 2011; Hokanson and Constantine, 2018; Bigoni et al., 2022). As discussed, agnostic learning is important in such applications due to model misspecification, and agnostic active learning is already widely used for fitting simpler functions classes like polynomials (Hampton and Doostan, 2015; Rauhut and Ward, 2012).

Moreover, single index model are a natural first step towards understanding active learning for neural networks more broadly, an important goal given the increasing importance of neural networks in approximating parameter-to-solution maps (Geist et al., 2021; Bhattacharya et al., 2021; Kutyniok et al., 2022) and quantities of interest (Tripathy and Bilionis, 2018; Khoo et al., 2021; Zhang et al., 2019; O'Leary-Roseberry et al., 2022; Cardenas et al., 2023) in scientific ML.

While there has been prior work in the realizable or i.i.d. mean zero noise setting (Cohen et al., 2011; Fornasier et al., 2012; Tyagi and Cevher, 2012) the first result on actively learning single index models under adversarial agnostic noise is due to Gajjar et al. (2023). That work shows:

Theorem 1 (Theorem 1 from Gajjar et al. (2023)) *Let f be a fixed L-Lipschitz function, let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be a data matrix, and let $\boldsymbol{w}^{\star} = \arg\min_{\boldsymbol{w}} \|f(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{y}\|_2^2$. There is an algorithm that, for any $\varepsilon \in (0, 1)$, observes $\tilde{O}\left(d^2 \cdot \frac{L^8}{\varepsilon^4}\right)$ entries of $\boldsymbol{y}$ and returns $\hat{\boldsymbol{w}}$ such that, for a fixed constant $C > 0$,*

$$\|f(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{y}\|_2^2 \leq C\|f(\boldsymbol{X}\boldsymbol{w}^{\star}) - \boldsymbol{y}\|_2^2 + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^{\star}\|_2^2,$$

with high probability. Above, $f(\boldsymbol{X}\boldsymbol{w})$ denotes the entrywise application of f to the vector $\boldsymbol{X}\boldsymbol{w}$.

The above guarantee is strong since it gives relative error with respect to the *best* approximation to the target $\boldsymbol{y}$, plus a small additive term depending on $\|\boldsymbol{X}\boldsymbol{w}^*\|_2^2$. Gajjar et al. (2023) show that the additive term is necessary: pure multiplicative error requires $\Omega(2^d)$ samples in the worst case.

1.2. Comparison to Non-active Supervised Learning

We highlight that, for the strong guarantee of Theorem 1, there is an inherent gap between active and non-active supervised learning.

In particular, consider the uniform distribution over the rows of $\boldsymbol{X}$. Theorem 1 ensures that poly(d) actively collected labels are needed to nearly minimize the expected squared error over all single index models. In contrast, consider an extreme case where $\boldsymbol{X}$ has d rows equal to the standard basis, and all other rows are zero. Even when f is the identity function, we *must* observe target values for those d rows if we want expected squared error competitive with $\boldsymbol{w}^*$. For large n, a poly(d) sample algorithm that uniformly samples rows of $\boldsymbol{X}$ will only observe those d labels with arbitrarily small probability, so will fail to obtain a bound as strong as Theorem 1.

As such, while there is a rich line of work on learning d-dimensional single index models in the agnostic setting with poly(d) or even fewer samples (Gollakota et al., 2023; Goel et al., 2017; Diakonikolas et al., 2020; Frei et al., 2020; Diakonikolas et al., 2022), that work inherently requires strong assumptions on $\boldsymbol{X}$ or $\boldsymbol{y}$, or provides a weaker notion of near optimal learning than Theorem 1.[1] Such assumptions are reasonable in many settings. In others, however, experimental and theoretical evidence shows that active learning can lead to significant improvements in sample complexity (Derezinski et al., 2018; Shimizu et al., 2024). For example, even for polynomial regression on one dimensional data drawn uniformly from an interval, where f in the identity and $\boldsymbol{X}$'s columns contain a basis for the degree d polynomials, there is a well-known gap of $O(d^2)$ vs. $O(d)$ samples for non-active vs. active methods (Cohen et al., 2013; Cohen and Migliorati, 2017). Notably, such a basis will have widely varying row norms, as in the hard case discussed above.

Since our goal is to study theoretical guarantees that inspire and motivate better active learning algorithms, in this work we are primarily interested in guarantees like Theorem 1 with no distributional assumptions on $\boldsymbol{X}$ or functional assumptions on $\boldsymbol{y}$.

1.3. Our Results

In this paper, we significantly improve on the results of Gajjar et al. (2023) in two ways. First, when f is a known L-Lipschitz function, we obtain a sample complexity bound with a *linear* dependence on d and a quadratically improved dependence on ε and L^2:

---

1. As a concrete example, Gollakota et al. (2023) assumes $\boldsymbol{y}$ is bounded by 1 and provides additive error ε on the average squared error. So, missing the d basis rows in our hard example is acceptable whenever $n \geq d/\varepsilon$.
2. A paper accepted to ICLR 2024 (**?**) also claims a linear dependence on d. However, there was an unfixable flaw in an earlier version of the paper's proof, which has been communicated to the authors.

Theorem 2 *Let f be a fixed L-Lipschitz function with $f(0) = 0$, let $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ be a data matrix, and let $\boldsymbol{w}^\star = \arg\min_{\boldsymbol{w}} \|f(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{y}\|^2$. There is an algorithm that, for any $\varepsilon \in (0, 1)$, observes $\tilde{O}\left(d \cdot \frac{L^4}{\varepsilon^2}\right)$ entries of $\boldsymbol{y}$ and returns $\hat{\boldsymbol{w}}$ such that, for a fixed constant $C > 0$, with high probability,*

$$\|f(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{y}\|^2 \le C\|f(\boldsymbol{X}\boldsymbol{w}^*) - \boldsymbol{y}\|^2 + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^*\|^2,$$

As noted in Gajjar et al. (2023), assuming that $f(0) = 0$ is without loss of generality, since we can always shift $\boldsymbol{y}$ by a fixed constant before fitting. It is easy to check that our near-linear dependence on d is optimal up to log factors, since $\Omega(d)$ samples are required even when $f(t) = t$. Theorem 2 establishes that, in terms of dimension, there is no gap between the active agnostic single index learning problem and linear regression. An interesting question for future work is if the dependence on ε can be improved to linear, as is possible for linear regression (Chen and Price, 2019).

Leverage score sampling. As in Gajjar et al. (2023), Theorem 2 is based on collecting samples via *statistical leverage score sampling*. A label for the ith row of $\boldsymbol{X}$ is selected with probability proportional to the ith leverage score $\boldsymbol{x}_i^T(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Also known as coherence motivated, Christoffel function, or effective resistance sampling (Hampton and Doostan, 2015; Adcock et al., 2023; Spielman and Srivastava, 2011), leverage score sampling yields near-optimal active learning bounds for least squares regression, polynomial regression, kernel learning, and variety of other "linear" problems (Sarlos, 2006; Cohen and Migliorati, 2017; Avron et al., 2019) . Leverage score sampling has two major advantages. First, it is computationally efficient: all leverage scores can be computed exactly in $O(nd^2)$ time, or approximately using faster randomized methods (Mahoney et al., 2012). Second, sampling is done in a completely non-adaptive way: the choice of which indices to label does not depend on prior labels collected. This allows for fully parallel data collection.

Technical contributions in Theorem 2. Our improved analysis of leverage score sampling in Theorem 2 requires two new contributions:

First, we provide an improved "subspace embedding" result for single index functions (Lemma 6), which shows that leverage score sampling preserves the ℓ_2 distance between any two vectors of the form $f(\boldsymbol{X}\boldsymbol{w}_1)$ and $f(\boldsymbol{X}\boldsymbol{w}_2)$. When f is the identity function, an optimal subspace embedding result for leverage score sampling follows from standard matrix Chernoff bounds (Tropp, 2015). However, when f is non-linearity, an analysis from first principles is required. Gajjar et al. (2023) employs an ε-net argument, which we improve using more powerful tools from high-dimensional probability, including Dudley's inequality and dual Sudakov minoration. This improvement accounts for our linear vs. quadratic dependence on the dimension d.

Second, to improve on the ε and L dependence, we show a more efficient translation from our subspace embedding result to the active learning problem by analyzing a natural regularized loss minimization procedure for finding $\hat{\boldsymbol{w}}$. The full proof is discussed in detail in Section 4.

Our other improvement on Theorem 1 from Gajjar et al. (2023) is that we are able to extend the result to the more challenging setting where f is an *unknown* Lipschitz function, and can be optimized as part of the training procedure. This setting is well-motivated in computational science, where f is typically parameterized as a piecewise constant or polynomial function, and has been studied in the realizable or i.i.d. noise setting (Cohen et al., 2011; Tyagi and Cevher, 2012; Hemant and Cevher, 2012). To the best of our knowledge, we provide the first result in the agnostic setting:

Theorem 3 *Let* Lip_L *denote the set of all L-Lipschitz, real-valued scalar functions f with $f(0) = 0$. Let $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ be a data matrix, and let $(f^\star, \boldsymbol{w}^\star) = \arg\min_{f \in \mathrm{Lip}_L, \; \boldsymbol{w} \in \mathbb{R}^d} \big\| f(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{y} \big\|^2$. There is an algorithm that, for any $\varepsilon \in (0,1)$, observes $\tilde{O}\left(d \cdot \frac{L^4}{\varepsilon^2} \cdot \log^2 n \right)$ entries of $\boldsymbol{y}$ and returns $\hat{\boldsymbol{w}}$ such that, for a fixed constant $C > 0$, with high probability,*

$$\left\| \hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{y} \right\|^2 \le C \big\| f^\star(\boldsymbol{X}\boldsymbol{w}^\star) - \boldsymbol{y} \big\|^2 + \varepsilon \big\| \boldsymbol{X}\boldsymbol{w}^\star \big\|^2 .$$

Theorem 3 nearly matches Theorem 2 for the case when f is known, except for a mild logarithmic dependence on n.

Technical contributions in Theorem 3. Proving the result requires significant additional work beyond Theorem 2. In particular, a natural approach would be to construct an ε-net $\mathcal{N}_\varepsilon$ over all L-Lipschitz functions such that e.g. for all $f \in \mathrm{Lip}_L$, there is some $\tilde{f} \in \mathcal{N}_\varepsilon$ such that $\| f(\boldsymbol{X}\boldsymbol{w}) - \tilde{f}(\boldsymbol{X}\boldsymbol{w})\|_2 \le \varepsilon \|\boldsymbol{X}\boldsymbol{w}\|_2$. The challenge in doing so is that $f(\boldsymbol{X}\boldsymbol{w})$ is a length n vector, our net would inherently requires a discretization of Lip_L with coarseness depending on $1/n$. This would introduce undesirable polynomial dependencies on n into our sample complexity.

We avoid this issue by building a sampling-aware discretization, in such a way that $|f(\langle \boldsymbol{x}_j, \boldsymbol{w}\rangle) - \tilde{f}(\langle \boldsymbol{x}_j, \boldsymbol{w}\rangle)| \le \varepsilon |\langle \boldsymbol{x}_j, \boldsymbol{w}\rangle|$ for most of the indices j sampled in $\boldsymbol{y}$. Roughly speaking, our discretization ensures a finer approximation to f for inputs close to 0, and is coarser for values further from 0. It takes significant effort to bound the size of discretization (and even to figure out what is the proper measure of "size"), which requires leveraging a diverse set of techniques including generic chaining for Bernoulli processes, dyadic decompositions, and a construction of an embedding to simpler spaces to control the size of our discretization

1.4. Additional Discussion of Related Work

As discussed, actively learning single index models has been studied in prior literature, although not in the agnostic setting (Cohen et al., 2011; Tyagi and Cevher, 2012). Prior work also considers more challenging "multi-index" models of the form $\sum_{i=1}^{k} f_i(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)$ (Fornasier et al., 2012; Hemant and Cevher, 2012). Understanding the multi-index problem in the agnostic setting is an interesting direction for future work. In prior work, various assumptions on the non-linearity f have been considered, including that f is a low-degree polynomial (Chen et al., 2020) and that f has bounded derivatives of high order (Cohen et al., 2011; Hemant and Cevher, 2012). In line with recent work (Gajjar et al., 2023; Gollakota et al., 2023), we make a Lipschitz assumption because it is simple, yet captures most fixed non-linearities commonly used in neural networks, like ReLUs and sigmoids.

A different but related line of work considers the setting where $\boldsymbol{y}$ is assumed to be generated by a single index model, and the goal is to approximate the target with few (non-active) samples, and in a computationally efficient way, by fitting it with a shallow neural network (Bietti et al., 2022; Kakade et al., 2011; Mousavi-Hosseini et al., 2023; Abbe et al., 2022; Dudeja and Hsu, 2018; Damian et al., 2022, 2023). Such results are motivated by an effort to understand the representation capability and optimization properties of neural networks. One recent result in this setting considers an agnostic guarantee, where the goal is to find an approximation to $\boldsymbol{y}$ competitive with the best single neuron approximation (Gollakota et al., 2023). However, as detailed in Section 1.2, that work inherently requires stronger assumptions than ours, since strong, distribution-independent guarantees of the form provided by Theorems 1 to 3 are not possible without active learning.

Remark on computational efficiency. In contrast to some of the work above, we *do not* analyze computational efficiency, only sample complexity. However, our active sampling algorithm is computationally efficient since it simply requires computing leverage scores. Moreover, in the fixed f setting, after collecting samples, we prove that it suffice to simply fit a single index model to those samples using a regularized ℓ_2 loss. As confirmed experimentally in Gajjar et al. (2023), this can be done easily and efficiently in practice using gradient descent methods. Nevertheless, while beyond the scope of our work, we think formally analyzing the computation efficiency of single index learning methods in the active, agnostic setting is a nice direction for future work.

Beyond research on single index models, we also note that there has been a significant recent work on minimising objectives that can be expressed as $\min_{\boldsymbol{w}} \big\| f(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) \big\|_2$, where f is a non linearity (Jambulapati et al., 2023; Mai et al., 2021; Munteanu et al., 2018; Musco et al., 2022). While this problem differs in important ways from ours, given that $\boldsymbol{y}$ is *inside* the non-linearity, methods like leverage score sampling have also proven valuable in this setting. For example, Jambulapati et al. (2023) employs tools similar to those used in the proof of Theorem 2 to obtain bounds for functions f with a natural "auto-Lipschitz" property. However, we note that this property is incomparable to our L-Lipschitz assumption on f; for instance the 1-Lipschitz ReLU function is not L-auto-Lipschitz for any finite L.

2. Notation

For a natural number n, we let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For a vector $\boldsymbol{x}$ in $\mathbb{R}^d$ with entries $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d, \|\boldsymbol{x}\| = (\sum_{i=1}^d \boldsymbol{x}_i^2)^{1/2}$ denotes its ℓ_2 norm. We use Lip_L to represent the class of L-Lipschitz functions on $\mathbb{R}$ vanishing at 0, i.e., $\mathrm{Lip}_L = \{f \in \mathcal{C}(\mathbb{R}) : f(0) = 0, \big|f(x_1) - f(x_2)\big| \leq L|x_1 - x_2|, \forall x_1, x_2 \in \mathbb{R}\}$. We extend the notation of $f(\cdot)$ to n-dimensional vectors: for $\boldsymbol{z} \in \mathbb{R}^n$, denote $f(\boldsymbol{z}) \in \mathbb{R}^n$ as the entrywise application of f to $\boldsymbol{z}$, i.e. $f(\boldsymbol{z}) = (f(\boldsymbol{z}_1), f(\boldsymbol{z}_2), \ldots, f(\boldsymbol{z}_n))$. We denote the i-th standard basis vector as $\mathbf{e}_i$. The Euclidean ball of radius R centered at $\boldsymbol{x} \in \mathbb{R}^d$ is denoted by $B_{\boldsymbol{x}}(R)$. In the case where the ball is centered at the origin, we simply use $B(R)$.

Throughout the paper, c and C will denote positive universal constants that may vary upon each occurrence. The notation $\tilde{O}(m)$ denotes $O(m \log^c m)$ for a fixed constant c. Moreover $a \lesssim b$ means that there exists a positive constant $C > 0$ such that $a \leq Cb$.

3. Preliminaries

Our goal is to find a single index model that best fits a given set of data points, $(\boldsymbol{x}_j, \boldsymbol{y}_j)$ for $j \in [n]$. In particular, for some class of functions $\mathcal{F}$, we wish to solve the problem

$$\min_{f \in \mathcal{F}, \boldsymbol{w} \in \mathbb{R}^d} \mathcal{L}(f, \boldsymbol{w}), \quad \text{where } \mathcal{L}(f, \boldsymbol{w}) \coloneqq \sum_{j=1}^n \big|f(\langle \boldsymbol{w}, \boldsymbol{x}_j \rangle) - y_j\big|^2 = \|f(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{y}\|^2. \quad (1)$$

In this work, we consider the case when $\mathcal{F} = \{f\}$, containting just a single known function f, as well as the case when it is the set of all Lipschitz functions with $f(0) = 0$.

Suppose that $(f^\star, \boldsymbol{w}^\star)$ minimize the loss function $\mathcal{L}$ over $f \in \mathcal{F}, \boldsymbol{w} \in \mathbb{R}^d$. We define

$$\mathrm{OPT}(\mathcal{F}) \coloneqq \min_{f \in \mathcal{F}, \boldsymbol{w} \in \mathbb{R}^d} \mathcal{L}(f, \boldsymbol{w}) = \mathcal{L}(f^\star, \boldsymbol{w}^\star).$$

We measure the accuracy of an approximate solution to (1) by quantifying the difference between its loss and the optimal loss. Specifically, for a given solution $(f, \boldsymbol{w})$, we define accuracy as follows:

Definition 4 (ε-accurate solution) *Fix some sufficiently large constant $C > 0$. Given $\varepsilon > 0$, a pair $(f, \boldsymbol{w})$ with $f \in \mathcal{F}$, $\boldsymbol{w} \in \mathbb{R}^d$ is said to be an ε-accurate solution to the problem (1), if*

$$\mathcal{L}(f, \boldsymbol{w}) \leq C \cdot \mathsf{OPT}(\mathcal{F}) + \varepsilon \|\boldsymbol{X}\boldsymbol{w}^\star\|^2.$$

This notion of accuracy was also used in Gajjar et al. (2023). Similar notions also appear in work on agnostic active learning in other contexts (Avron et al., 2019).

3.1. Subsampled regression

As in Gajjar et al. (2023), we collect data in an active way via *importance sampling*. Every row of $\boldsymbol{X}$ is assigned a score, and rows are sampled with probability proportional to those scores. Target values in $\boldsymbol{y}$ only need to be observed for rows that get sampled. Concretely, we use statistical leverage scores, which have been widely applied for active linear regression (see e.g. Chen and Price (2019); Mahoney et al. (2011). Intuitively, leverage scores measure how influential a row is in forming the column space of the data matrix, $\boldsymbol{X}$. Formally, they are definied as follows:

Definition 5 (Statistical leverage score) *The leverage score of the j-th row $\boldsymbol{x}_j$ of a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is defined as $\tau_j(\boldsymbol{X}) \coloneqq \boldsymbol{x}_j^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_j = \sup_{\boldsymbol{w} \in \mathbb{R}^d} \frac{\langle \boldsymbol{x}_j, \boldsymbol{w}\rangle^2}{\|\boldsymbol{X}\boldsymbol{w}\|_2^2}$.*

Sampling process. We consider a sample-with-replacement variant of leverage score sampling. Let p denote a probability distribution over $[n]$ where $j \in [n]$ is assigned probability $p_j = \frac{\tau_j(\boldsymbol{X})}{\sum_{i=1}^n \tau_i(\boldsymbol{X})}$. We generate m i.i.d. random indices $j_1, \ldots, j_m \sim p$, each taking values in $[n]$.

We next define a sampling-and-reweighting matrix $\boldsymbol{S}$ to succinctly represent a subsampled regression problem, which will be used for active learning. Given indices $j_1, j_2, \ldots, j_m$ sampled from p, we construct $\boldsymbol{S} \in \mathbb{R}^{m \times n}$, by setting the i-th row of $\boldsymbol{S}$ to be $\frac{1}{\sqrt{mp_{j_i}}} \boldsymbol{e}_{j_i}$. We let $\hat{\mathcal{L}}(f, \boldsymbol{w})$ denote:

$$\hat{\mathcal{L}}(f, \boldsymbol{w}) \coloneqq \|\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{S}\boldsymbol{y}\|^2 = \frac{1}{m} \sum_{i=1}^m \frac{1}{p_{j_i}} \big| f(\langle \boldsymbol{w}, \boldsymbol{x}_{j_i}\rangle) - y_{j_i} \big|^2. \tag{2}$$

One can verify that $\mathbb{E}[\boldsymbol{S}^\top \boldsymbol{S}] = \boldsymbol{I}$ and hence $\hat{\mathcal{L}}(f, \boldsymbol{w})$ is equal to $\mathcal{L}(f, \boldsymbol{w})$ in expectation. Importantly, however, this subsampled loss can be evaluated using only those target values that appear in $\boldsymbol{S}\boldsymbol{y}$, i.e. using at most m entries from $\boldsymbol{y}$. So, our approach is to minimize $\hat{\mathcal{L}}$ as a surrogate for $\mathcal{L}$.

4. Main results

While it is tempting to try to obtain an ε-accurate solution by returning any minimizer for the subsampled loss $\hat{\mathcal{L}}$, and doing so works in the linear setting, it can be seen that such an approach will fail when f is non-linear. In particular, consider the case when the rows of $\boldsymbol{X}$ contain all 2^d binary vectors of length d, f is a ReLU non-linearity, and $\boldsymbol{y}$ contains a 1 in a single vector in $\{0,1\}^d$. Any sampling strategy that takes $o(2^d)$ samples will only observe labels equal to 0. Thus, since $f(x) = \max(x, 0)$, *any vector $\hat{\boldsymbol{w}}$ with non-positive entries is a valid minimizer for $\hat{\mathcal{L}}$.* However, by choose $\hat{\boldsymbol{w}}$ to have very large negative entries, we can make $\mathcal{L}(f, \hat{\boldsymbol{w}})$ arbitrarily bad.

Gajjar et al. (2023) deal with this issue by imposing a hard constraint on the norm of $\boldsymbol{w}$. We instead consider minimizing a regularized version of $\hat{\mathcal{L}}$ that penalizes $\boldsymbol{w}$ with high norm. Ultimately,

our regularization approach allows us to improve the $O(\varepsilon^{-4}L^8)$ dependence in sample complexity in Gajjar et al. (2023) to $O(\varepsilon^{-2}L^4)$. Concretely, we define the regularized loss as:

$$\hat{\mathcal{L}}_{\mathsf{reg}}(f, \boldsymbol{w}) \coloneqq \hat{\mathcal{L}}(f, \boldsymbol{w}) + \varepsilon\|\boldsymbol{X}\boldsymbol{w}\|^2 = \big\|\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{S}\boldsymbol{y}\big\|^2 + \varepsilon\|\boldsymbol{X}\boldsymbol{w}\|^2. \tag{3}$$

Our first main result is that, for a fixed f, the weight $\hat{\boldsymbol{w}}$ that minimizes (3), is an ε-accurate solution for $\mathcal{L}(f, \boldsymbol{w})$ with high probability.

Theorem 2 *Let $\hat{\boldsymbol{w}}$ solve the subsampled regularized least squares problem:*

$$\hat{\boldsymbol{w}} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \hat{\mathcal{L}}_{\mathsf{reg}}(f, \boldsymbol{w}). \tag{4}$$

Then, for some universal constant $c > 0$, as long as $m \geq cL^4\varepsilon^{-2}d\log^3 d$, $(f, \hat{\boldsymbol{w}})$ is an ε-accurate solution to of (1) with $\mathcal{F} = \{f\}$ with probability at least $99/100$.

The main ingredient for proving this theorem is to show that $\|\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}^\star)\|^2$ is close to $\|f(\boldsymbol{X}\boldsymbol{w}) - f(\boldsymbol{X}\boldsymbol{w}^\star)\|^2$ for all $\boldsymbol{w}$ with sufficiently bounded norm, which is formally characterized by the following nonlinear subspace embedding lemma.

Lemma 6 (Non-linear subspace embedding with fixed non-linearity) *Let $\boldsymbol{S}$ be a leverage score subsampling matrix with m rows as defined in Section 3. Assume $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has orthonormal columns.[3] There is some universal constant $C > 0$ such that, for any $f \in \mathrm{Lip}_L$, and any $R > 0$, as long as $m \geq CL^4\varepsilon^{-2}d\log^3(d)\cdot\log(1/\delta)$, the following holds with probability $\geq 1 - \delta$.*

$$\left| \big\|\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}_1) - \boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}_2)\big\|^2 - \big\|f(\boldsymbol{X}\boldsymbol{w}_1) - f(\boldsymbol{X}\boldsymbol{w}_2)\big\|^2 \right| \leq \varepsilon R^2, \quad \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R).$$

This name "nonlinear subspace embedding" is a nod to the now standard subspace embedding result for leverage score sampling (Sarlos, 2006; Drineas et al., 2006). We will use this result as well, which is a special (and stronger) case of Lemma 6 when f is the identity function.

Lemma 7 (Subspace embedding) *For any $\varepsilon, \delta \in (0, 1)$, as long as $m \gtrsim \frac{d\log(d/\delta)}{\varepsilon^2}$, then with probability at least $1 - \delta$, for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$,*

$$(1-\varepsilon)\|\boldsymbol{X}\boldsymbol{w}_1 - \boldsymbol{X}\boldsymbol{w}_2\|^2 \leq \|\boldsymbol{S}\boldsymbol{X}\boldsymbol{w}_1 - \boldsymbol{S}\boldsymbol{X}\boldsymbol{w}_2\|^2 \leq (1+\varepsilon)\|\boldsymbol{X}\boldsymbol{w}_1 - \boldsymbol{X}\boldsymbol{w}_2\|^2. \tag{5}$$

Lemma 7 can be proven using a matrix Chernoff bound (see, e.g., Woodruff (2014)). Our Lemma 6 requires more work due to the potential nonlinearity of f. Its proof is discussed in Section 5 and, along with a generalization to unknown f in Lemma 8, is the major technical contribution of our work. Before getting into the details, we show how Lemma 6 can be used to prove Theorem 2.

Proof of Theorem 2 We simplify the discussion by assuming that, without loss of generality, $\boldsymbol{X}$ has orthonormal columns. This can be done because leverage score are invariant under column transformations. If $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{Q} \in \mathbb{R}^{n\times d}$ has orthonormal columns and $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ is invertible, denoting by $\boldsymbol{q}_j$ the j-th row of $\boldsymbol{Q}$, one has $\boldsymbol{q}_j = \boldsymbol{R}^\top\boldsymbol{x}_j$ and may verify

$$\tau_j(\boldsymbol{Q}) = \boldsymbol{q}_j^\top(\boldsymbol{Q}^\top\boldsymbol{Q})^{-1}\boldsymbol{q}_j = \boldsymbol{x}_j^\top(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{x}_j = \tau_j(\boldsymbol{X}).$$

---

3. As we will discuss later, $\boldsymbol{X}$ can be assumed to be orthonormal without loss of generality.

The conclusion then follows from observing that all the involved statements are not affected if we substitute $\boldsymbol{X}$ with $\boldsymbol{Q}$ and $\boldsymbol{w}$ with $\boldsymbol{Rw}$. The same approach was used in Gajjar et al. (2023).

When $\boldsymbol{X}$ is orthonormal, the leverage scores have a particularly simple form.

$$\tau_j(\boldsymbol{X}) = \|\boldsymbol{x}_j\|^2, \quad \sum_{j=1}^{n} \tau_j(\boldsymbol{X}) = d. \tag{6}$$

Throughout the proof, we shall condition on the event that

$$\hat{\mathcal{L}}(f, \boldsymbol{w}^\star) \leq 1000 \cdot \mathcal{L}(f, \boldsymbol{w}^\star) = 1000 \cdot \mathsf{OPT}, \tag{7}$$

which happens with probability at least 0.999 by Markov's inequality (since $\mathbb{E}\hat{\mathcal{L}} = \mathcal{L}$). Now, since

$$f(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{y} = \big(f(\boldsymbol{X}\hat{\boldsymbol{w}}) - f(\boldsymbol{X}\boldsymbol{w}^\star)\big) + \big(f(\boldsymbol{X}\boldsymbol{w}^\star) - \boldsymbol{y}\big),$$

from triangle inequality and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, we get the following

$$\mathcal{L}(f, \hat{\boldsymbol{w}}) = \big\|f(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{y}\big\|^2 \leq 2\big\|f(\boldsymbol{X}\hat{\boldsymbol{w}}) - f(\boldsymbol{X}\boldsymbol{w}^\star)\big\|^2 + 2\mathsf{OPT}. \tag{8}$$

Next, by optimality of $\hat{\boldsymbol{w}}$, we have $\hat{\mathcal{L}}_{\mathsf{reg}}(f, \hat{\boldsymbol{w}}) \leq \hat{\mathcal{L}}_{\mathsf{reg}}(f, \boldsymbol{w}^\star)$, i.e.,

$$\|\boldsymbol{S}f(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{S}\boldsymbol{y}\|^2 + \varepsilon\|\boldsymbol{X}\hat{\boldsymbol{w}}\|^2 \leq \hat{\mathcal{L}}(f, \boldsymbol{w}^\star) + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^\star\|^2. \tag{9}$$

In particular, we have (recalling that $\boldsymbol{X}$ is assumed to have orthonormal columns)

$$\|\hat{\boldsymbol{w}}\|^2 = \|\boldsymbol{X}\hat{\boldsymbol{w}}\|^2 \leq \frac{1}{\varepsilon}\hat{\mathcal{L}}(f, \boldsymbol{w}^\star) + \|\boldsymbol{X}\boldsymbol{w}^\star\|^2 \leq \frac{1000}{\varepsilon}\mathsf{OPT} + \|\boldsymbol{X}\boldsymbol{w}^\star\|^2.$$

This implies that $\hat{\boldsymbol{w}} \in B(R)$ where $R = \sqrt{\frac{1000}{\varepsilon}\mathsf{OPT} + \|\boldsymbol{X}\boldsymbol{w}^\star\|^2}$. Then, from Lemma 6, the following holds with probability at least 0.999 given $m \geq CL^4\varepsilon^{-2}d\log^3 d$:

$$\begin{aligned}
\|f(\boldsymbol{X}\hat{\boldsymbol{w}}) - f(\boldsymbol{X}\boldsymbol{w}^\star)\|^2 &\leq \|\boldsymbol{S}f(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}^\star)\|^2 + \varepsilon R^2 \\
&\lesssim \|\boldsymbol{S}f(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{S}\boldsymbol{y}\|^2 + \|\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}^\star) - \boldsymbol{S}\boldsymbol{y}\|^2 + \varepsilon R^2 \\
&\lesssim \left(\hat{\mathcal{L}}(f, \boldsymbol{w}^\star) + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^\star\|^2\right) + \hat{\mathcal{L}}(f, \boldsymbol{w}^\star) + \varepsilon\left(\frac{1}{\varepsilon}\mathsf{OPT} + \|\boldsymbol{X}\boldsymbol{w}^\star\|^2\right) \\
&\lesssim \mathsf{OPT} + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^\star\|^2,
\end{aligned}$$

where the second line is triangle inequality, the third line used (9) and the definition of R, and the last line used (7). Plugging the above inequality into (8) readily gives

$$\mathcal{L}(f, \hat{\boldsymbol{w}}) \lesssim \mathsf{OPT} + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^\star\|^2,$$

as desired (after replacing ε by ε/C for some sufficiently large constant $C > 0$). ∎

Our second main result is for the setting where f is an unknown L-Lipschitz function vanishing at 0. Again, we show that an ε-accurate solution to the problem of $\min_{f,\boldsymbol{w}} \mathcal{L}(f, \boldsymbol{w})$ can be obtained by minimizing the regularized subsampled loss, but this time over both f and $\boldsymbol{w}$. Formally, we have:

Theorem 3 *Let $\hat{f}$ and $\hat{\boldsymbol{w}}$ be the solution to the subsampled regularized least square problem:*

$$(\hat{f}, \hat{\boldsymbol{w}}) = \arg\min_{\substack{f \in \mathrm{Lip}_L \\ \boldsymbol{w} \in \mathbb{R}^d}} \hat{\mathcal{L}}_{\mathsf{reg}}(f, \boldsymbol{w}). \tag{10}$$

Then, for some universal constant $C > 0$, as long as $m \geq CL^4 \varepsilon^{-2} d(\log^2 n + \log^3 d)$, $(\hat{f}, \hat{\boldsymbol{w}})$ is an ε-accurate solution to (1) with $\mathcal{F} = \mathrm{Lip}_L$ probability at least $99/100$.

Similar to Theorem 2, this result crucially depends on the following nonlinear concentration result, which extends Lemma 6 to the case when $\boldsymbol{X}\boldsymbol{w}_1$ and $\boldsymbol{X}\boldsymbol{w}_2$ are transformed by two potentially different L-Lipschitz functions, f_1 and f_2.

Lemma 8 (Non-linear subspace embedding with unknown non-linearity) *Let $\boldsymbol{S}$ be a leverage score subsampling matrix with m rows as defined in Section 3. Assume $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ has orthonormal columns. As long as $m \geq CL^4 \varepsilon^{-2} d(\log^2 n + \log^3(d/\delta))$ for some fixed constant $C > 0$, the following holds with probability $\geq 1 - \delta$.*

$$\left| \left\| \boldsymbol{S}f_1(\boldsymbol{X}\boldsymbol{w}_1) - \boldsymbol{S}f_2(\boldsymbol{X}\boldsymbol{w}_2) \right\|^2 - \left\| f_1(\boldsymbol{X}\boldsymbol{w}_1) - f_2(\boldsymbol{X}\boldsymbol{w}_2) \right\|^2 \right| \leq \varepsilon R^2,$$

for all $f_1, f_2 \in \mathrm{Lip}_L$ and for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)$.

The proof of Theorem 3 is deferred to Appendix A and follows the same steps as that of Theorem 2, with the distinction that we must compare $\mathcal{L}(\hat{f}, \hat{\boldsymbol{w}})$ with the optimal loss $\mathcal{L}(f^\star, \boldsymbol{w}^\star)$.

5. Proof of the non-linear concentrations

The proofs of Lemma 6 and Lemma 8, especially the latter one, are the major challenges in this work. We sketch the key ideas below.

Proof roadmap for Lemma 6 Our proof (given in Appendix C) is based on ideas from Rudelson (1999). However, we employ a simplified version, using dual Sudakov minoration, deviating from of the construction of chaining functionals therein. This simplification significantly streamlines the explanation, although at the cost of losing a $\log^2 d$ factor. Similar approach also appeared in a recent work Jambulapati et al. (2023) for a different problem.

Step 1: Symmetrization. We want to obtain a tail bound for the supremum of our random process. For this, we apply a standard symmetrization technique (Lemma 18), where we introduce i.i.d. Rademacher random variables $\xi_1, \xi_2, \ldots, \xi_m$, where each ξ_i takes values in -1 and 1 with probabilities $1/2$ each. This reduces the problem to finding tail bounds of the following Bernoulli process:

$$Z(\boldsymbol{w}_1, \boldsymbol{w}_2) := \sum_{i=1}^{m} \xi_i \frac{\left(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle) \right)^2}{p_{j_i}}, \qquad (\boldsymbol{w}_1, \boldsymbol{w}_2) \in B(R) \times B(R).$$

In particular, this allows us to prove Lemma 6 by showing $\frac{1}{m} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| \lesssim \varepsilon$ with overwhelming probability.

Step 2: Applying sub-Gaussianity of symmetrized process. It turns out that, conditioned on the samples $\{j_1, j_2, \ldots, j_m\}$ (fixing $\boldsymbol{S}$), the above random process is sub-Gaussian w.r.t. the index $(\boldsymbol{w}_1, \boldsymbol{w}_2)$, endowed with the following metric on $B(R) \times B(R)$ (Appendix C.5).

$$\rho_f\big((\boldsymbol{w}_1, \boldsymbol{w}_2), (\boldsymbol{w}_1', \boldsymbol{w}_2')\big)$$
$$:= \left(\sum_{i=1}^m \frac{1}{p_{j_i}^2} \big(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle)\big)^2 - \big(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1' \rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2' \rangle)\big)^2 \right)^{1/2}.$$

Step 3: Bounding the sub-Gaussian norm by the dual norm of a polytope. The Lipschitz continuity of f allows us to obtain an upper bound (Lemma 19) of ρ_f in terms of the dual norm of the polytope $\mathcal{P}$ formed by vertices $\{\pm \boldsymbol{x}_{j_i} / \sqrt{p_{j_i}}\}_{i \in [m]}$:

$$\rho_f\big((\boldsymbol{w}_1, \boldsymbol{w}_2), (\boldsymbol{w}_1', \boldsymbol{w}_2')\big) \lesssim L^2 R \sqrt{m} \left(\|\boldsymbol{w}_1 - \boldsymbol{w}_1'\|_{\mathcal{P}^\circ} + \|\boldsymbol{w}_2 - \boldsymbol{w}_2'\|_{\mathcal{P}^\circ} \right), \forall \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_1', \boldsymbol{w}_2' \in B(R)$$

with high probability, where $\mathcal{P}^\circ$ is the polar of $\mathcal{P}$ and $\|\cdot\|_{\mathcal{P}^\circ}$ is its Minkowski norm. This bound gets rid of any dependence on the non-linearity f and prepares us for the application of dual Sudakov minoration (Lemma 15).

Step 4: Bound in expectation via Dudley's integral and dual Sudakov minoration. The upper bound of ρ_f, and the sub-Gaussianity of $Z(\boldsymbol{w}_1, \boldsymbol{w}_2)$ allows us to obtain upper bound for the expectation of the random process using Dudley's integral with respect to $\|\cdot\|_{\mathcal{P}^\circ}$ (Appendix C.3). This, in turn, is controlled using dual Sudakov minoration, leading to

$$\mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| \lesssim L^2 R^2 \sqrt{m d \log^2 d \cdot \log m}.$$

When $m \gtrsim \varepsilon^{-2} d \log^3 d$, this implies $\frac{1}{m} \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| \lesssim \varepsilon$, which is already close to what we desire.

Step 5: Tail bound via concentration of measure. Finally, we deduce the desired tail bound from the above bound in expectation using a concentration of measure argument in Appendix C.4.

Proof roadmap for Lemma 8. As discussed, a primary challenge in dealing with the unknown f is that, to prove the bound, we must construct a discretization for the infinite class of functions Lip_L. We require an efficient discretization of Lip_L that avoids any polynomial dependencies on n. We outline our approach below, and the complete proofs are provided in Appendix D.

Step 1: Symmetrization. Similar to Lemma 6, the problem can be reduced to obtaining tail bounds for the supremum of the symmetrized process, but now considering two different functions f_1 and f_2:

$$Z_{f_1, f_2}(\boldsymbol{w}_1, \boldsymbol{w}_2) := \sum_{i=1}^m \xi_i \frac{\big(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle)\big)^2}{p_{j_i}},$$

where $f_1, f_2 \in \mathrm{Lip}_L$, $(\boldsymbol{w}_1, \boldsymbol{w}_2) \in B(R) \times B(R)$.

11

Step 2: Applying generic chaining to find guidelines for discretization. The next step involves discretizing Lip_L for a "good" approximation, while maintaining a small "size". Using generic chaining for Bernoulli process, we construct (in Appendix D.1) two "sampling-aware" metrics D_∞ and D_2 (depending on $\boldsymbol{S}$) on $\mathrm{Lip}_L \times \mathrm{Lip}_L$ measuring respectively the "goodness" of approximation and the size of the discretization. We show (in Lemma 21) that if $\mathcal{N}_\Delta \subset \mathrm{Lip}_L$ is such that $\mathcal{N}_\Delta \times \mathcal{N}_\Delta$ is a Δ-net of $\mathrm{Lip}_L \times \mathrm{Lip}_L$ in D_∞ distance, then[4]

$$\mathbb{E} \sup_{\substack{f_1,f_2 \in \mathrm{Lip}_L \\ \boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)}} \left| Z_{f_1,f_2}(\boldsymbol{w}_1, \boldsymbol{w}_2) \right| \lesssim L^2 R^2 \sqrt{md\log^2 d \cdot \log m} + \mathbb{E}\Delta + \mathbb{E}\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{N}_\Delta \times \mathcal{N}_\Delta, D_2, \varepsilon)} \, \mathrm{d}\varepsilon.$$

Step 3: Net construction. The construction of the desired net, carried out in Appendix D.3, is the major technical challenge in the proof. Our approach is guided by the following upper bound on D_∞, proved in Lemma 22:

$$D_\infty\big((f_1, f_2), (f_1', f_2')\big) \lesssim \rho_\infty(f_1, f_1') + \rho_\infty(f_2, f_2'),$$

where ρ_∞ is some metric on Lip_L obeying (with high probability)

$$\rho_\infty(f_1, f_2) \lesssim LR\sqrt{m} \left(\sum_{i=1}^m \frac{1}{\|\boldsymbol{x}_{j_i}\|^2} \|f_1 - f_2\|_{L^\infty([-R\|\boldsymbol{x}_{j_i}\|, R\|\boldsymbol{x}_{j_i}\|])}^2 \right)^{1/2}.$$

We then construct a net $\mathcal{N}_\Delta$ (details are in Appendix D.3.1) by imposing that for any $f \in \mathrm{Lip}_L$, there is some $\tilde{f} \in \mathcal{N}_\Delta$, such that $|\tilde{f}(x) - f(x)| \lesssim \varepsilon|x|$ for larger values of $|x|$, while $|\tilde{f}(x) - f(x)| \lesssim \varepsilon$ for smaller values of $|x|$. The former restriction ensures $\frac{1}{\|\boldsymbol{x}_{j_i}\|^2} \|f_1 - f_2\|_{L^\infty([-R\|\boldsymbol{x}_{j_i}\|, R\|\boldsymbol{x}_{j_i}\|])}^2 \lesssim \varepsilon^2$ if $\|\boldsymbol{x}_{j_i}\|$ is large, but will require a net of infinite cardinality if demanded to hold for all x. The latter restriction is a careful relaxation of the former, which, as we shall show in the proof, does not hurt the quality of approximation by the net (up to logarithmic factors).

Step 4: Bounding the size of the discretization via embedding to simple spaces. Subsequently, to bound the Dudley's integral of $\mathcal{N}_\Delta \times \mathcal{N}_\Delta$ in D_2 distance (see Appendix D.3.3), we introduce another technique derived from the proof of Lemma 22 and the explicit construction of $\mathcal{N}_\Delta$. In essence, we construct an embedding (a combination of Lemma 22, Lemma 25 and Lemma 26) of $(\mathcal{N}_\Delta \times \mathcal{N}_\Delta, D_2)$ into $(\mathrm{Lip}_1, L^\infty([-1, 1]))^2$ (within an additive error which is well controlled). Since the Dudley's integral of $(\mathrm{Lip}_1, L^\infty([-1, 1])$ has a well-known upper bound (Lemma 28), this embedding would imply our desired bound easily.

6. Conclusion

This work provides the first sample complexity results for actively learning single index models in the agnostic setting with a nearly linear dependence on the dimension d, and with no strong distributional assumptions. We believe the results suggest a number of avenues for future exploration. For example, while we obtain near optimal sample complexity results, we do not consider computational efficiency, which has been considered in work on fitting single index functions in related

---

4. Here Δ can be a random variable depending on $\boldsymbol{S}$, as D_∞ is already a random distance depending on $\boldsymbol{S}$. This accounts for the expectation in $\mathbb{E}\Delta$.

models. Additionally, to the best of our knowledge, the related (and harder) multi-index model has not been addressed in the same setting that we consider. Finally, although our Theorems 2 and 3 provide a constant-factor multiplicative approximation, we believe it should be possible to obtain a $(1 + \varepsilon)$ approximation using techniques from Musco et al. (2022).

References

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Proceedings of the 35th Annual Conference on Computational Learning Theory (COLT)*, pages 4782–4887, 2022.

Ben Adcock, Juan M Cardenas, and Nick Dexter. CS4ML: A general framework for active learning with arbitrary data based on Christoffel functions. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas. A fast and scalable method for a-optimal design of experiments for infinite-dimensional bayesian nonlinear inverse problems. *SIAM Journal on Scientific Computing*, 38(1):A243–A272, 2016.

Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. A universal sampling method for reconstructing signals with simple fourier transforms. In *Proceedings of the 51st Annual ACM Symposium on Theory of Computing (STOC)*, pages 1051–1063, 2019.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 65–72, 2006.

Kaushik Bhattacharya, Bamdad Hosseini, Nikola B Kovachki, and Andrew M Stuart. Model reduction and neural networks for parametric PDEs. *The SMAI Journal of Computational Mathematics*, 7:121–157, 2021.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 9768–9783, 2022.

Daniele Bigoni, Youssef Marzouk, Clémentine Prieur, and Olivier Zahm. Nonlinear dimension reduction for surrogate modeling using gradient information. *Information and Inference: A Journal of the IMA*, 11(4):1597–1639, 2022.

Juan M Cardenas, Ben Adcock, and Nick Dexter. CS4ML: A general framework for active learning with arbitrary data based on christoffel functions. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Minshuo Chen, Yu Bai, Jason D Lee, Tuo Zhao, Huan Wang, Caiming Xiong, and Richard Socher. Towards understanding hierarchical learning: Benefits of neural representations. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Xue Chen and Michal Derezinski. Query complexity of least absolute deviation regression via robust uniform convergence. In *Proceedings of the 34th Annual Conference on Computational Learning Theory (COLT)*, pages 1144–1179, 2021.

Xue Chen and Eric Price. Active regression via linear-sample sparsification. In *Proceedings of the 32nd Annual Conference on Computational Learning Theory (COLT)*, pages 663–695, 2019.

Albert Cohen and Ronald DeVore. Approximation of high-dimensional parametric pdes. *Acta Numerica*, 24:1–159, 2015.

Albert Cohen and Giovanni Migliorati. Optimal weighted least-squares methods. *SMAI Journal of Computational Mathematics*, 3:181–203, 2017.

Albert Cohen, Ingrid Daubechies, Ronald DeVore, Gerard Kerkyacharian, and Dominique Picard. Capturing ridge functions in high dimensions from point queries. *Constructive Approximation*, 35(2):225–243, 2011.

Albert Cohen, Mark A. Davenport, and Dany Leviatan. On the stability and accuracy of least squares approximations. *Foundations of Computational Mathematics*, 13(5):819–834, 2013.

Arnak S Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *The Journal of Machine Learning Research*, 9:1647–1678, 2008.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of the 35th Annual Conference on Computational Learning Theory (COLT)*, pages 5413–5452, 2022.

Michal Derezinski, Manfred K. K Warmuth, and Daniel J Hsu. Leveraged volume sampling for linear regression. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

Ilias Diakonikolas, Surbhi Goel, Sushrut Karmalkar, Adam R. Klivans, and Mahdi Soltanolkotabi. Approximation schemes for relu regression. In *Proceedings of the 33rd Annual Conference on Computational Learning Theory (COLT)*, 2020.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning a single neuron with adversarial label noise via gradient descent. In *Proceedings of the 35th Annual Conference on Computational Learning Theory (COLT)*, 2022.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1127–1136, 2006.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Annual Conference on Computational Learning Theory (COLT)*, pages 1887–1930, 2018.

Tamás Erdélyi, Cameron Musco, and Christopher Musco. Fourier sparse leverage scores and approximate kernel learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Massimo Fornasier, Karin Schnass, and Jan Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12:229–262, 2012.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Aarshvi Gajjar, Christopher Musco, and Chinmay Hegde. Active learning for single neuron models with lipschitz non-linearities. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4101–4113, 2023.

Moritz Geist, Philipp Petersen, Mones Raslan, Reinhold Schneider, and Gitta Kutyniok. Numerical solution of the parametric diffusion equation by deep neural networks. *Journal of Scientific Computing*, 88(1):22, 2021.

Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the ReLU in polynomial time. In *Proceedings of the 30th Annual Conference on Computational Learning Theory (COLT)*, 2017.

Aravind Gollakota, Parikshit Gopalan, Adam R. Klivans, and Konstantinos Stavropoulos. Agnostically learning single-index models using omnipredictors. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Jerrad Hampton and Alireza Doostan. Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression. *Computer Methods in Applied Mechanics and Engineering*, 290:73–97, 2015.

Wolfgang Härdle, Marlene Müller, Stefan Sperlich, Axel Werwatz, et al. *Nonparametric and semiparametric models*, volume 1. Springer, 2004.

Tyagi Hemant and Volkan Cevher. Active learning of multi-index function models. In *Advances in Neural Information Processing Systems 25 (NeurIPS)*, 2012.

Jeffrey M. Hokanson and Paul G. Constantine. Data-driven polynomial ridge approximation using variable projection. *SIAM Journal on Scientific Computing*, 40(3):A1566–A1589, 2018.

Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying generalized linear models, 2023.

Matti Kääriäinen. Active learning in the non-realizable case. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory (ALT)*, pages 63–77, 2006.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems 24 (NeurIPS)*, 24, 2011.

Yuehaw Khoo, Jianfend Lu, and Lexing Ying. Solving parametric PDE problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, 2021.

Gitta Kutyniok, Philipp Petersen, Mones Raslan, and Reinhold Schneider. A theoretical analysis of deep neural networks and parametric PDEs. *Constructive Approximation*, 55(1):73–125, 2022.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

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

Michael W. Mahoney, Petros Drineas, Malik Magdon-Ismail, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13: 3475–3506, 2012.

Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

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

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 6562–6571, 2018.

Cameron Musco, Christopher Musco, David P. Woodruff, and Taisuke Yasuda. Active linear regression for ℓ_p norms and beyond norms and beyond. In *Proceedings of the 63rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 744–753, 2022.

Thomas O'Leary-Roseberry, Umberto Villa, Peng Chen, and Omar Ghattas. Derivative-informed projected neural networks for high-dimensional parametric maps governed by PDEs. *Computer Methods in Applied Mechanics and Engineering*, 388:114199, 2022.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

Holger Rauhut and Rachel Ward. Sparse Legendre expansions via $\ell 1$-minimization. *Journal of Approximation Theory*, 164(5):517 – 533, 2012.

Mark Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1): 60–72, 1999.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21–es, 2007.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

Atsushi Shimizu, Xiaoou Cheng, and Jonathan Weare Christopher Musco. Improved active learning via dependent leverage score sampling. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011. Preliminary version in the 40th Annual ACM Symposium on Theory of Computing (STOC).

Michel Talagrand. *Upper and lower bounds for stochastic processes: decomposition theorems*. A Series of Modern Surveys in Mathematics. Springer Cham, 2021.

Rohit K. Tripathy and Ilias Bilionis. Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *Journal of Computational Physics*, 375:565–588, 2018.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

Hemant Tyagi and Volkan Cevher. Learning ridge functions with randomized sampling in high dimensions. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2025–2028, 2012.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 843–851, 2018.

Dongkun Zhang, Lu Lu, Ling Guo, and George Em Karniadakis. Quantifying total uncertainty in physics-informed neural networks for solving forward and inverse stochastic problems. *Journal of Computational Physics*, 397, 2019.

Appendix A. Proof of regression guarantee with unknown non-linearity

Proof (Proof of Theorem 3) We simplify the discussion by assuming that, without loss of generality, $\boldsymbol{X}$ has orthonormal columns, just like Theorem 2. Throughout the proof, we shall condition on the event that

$$\hat{\mathcal{L}}(f^\star, \boldsymbol{w}^\star) \leq 1000 \cdot \mathcal{L}(f^\star, \boldsymbol{w}^\star) = 1000 \cdot \mathsf{OPT}, \tag{11}$$

which happens with probability at least 0.999 by Markov's inequality (since $\mathbb{E}\hat{\mathcal{L}} = \mathcal{L}$). Now, since

$$\hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{y} = \left(\hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - f^\star(\boldsymbol{X}\boldsymbol{w}^\star)\right) + \left(f^\star(\boldsymbol{X}\boldsymbol{w}^\star) - \boldsymbol{y}\right),$$

from triangle inequality and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, we get the following

$$\mathcal{L}(\hat{f}, \hat{\boldsymbol{w}}) = \left\|\hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{y}\right\|^2 \leq 2\left\|\hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - f^\star(\boldsymbol{X}\boldsymbol{w}^\star)\right\|^2 + 2\mathsf{OPT}. \tag{12}$$

Next, by optimality of $\hat{f}, \hat{\boldsymbol{w}}$, we have $\hat{\mathcal{L}}_{\mathsf{reg}}(\hat{f}, \hat{\boldsymbol{w}}) \leq \hat{\mathcal{L}}_{\mathsf{reg}}(f^\star, \boldsymbol{w}^\star)$, i.e.,

$$\|\boldsymbol{S}\hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{S}\boldsymbol{y}\|^2 + \varepsilon\|\boldsymbol{X}\hat{\boldsymbol{w}}\|^2 \leq \hat{\mathcal{L}}(f^\star, \boldsymbol{w}^\star) + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^\star\|^2. \tag{13}$$

In particular, we have (recalling that $\boldsymbol{X}$ is assumed to have orthonormal columns)

$$\|\hat{\boldsymbol{w}}\|^2 = \|\boldsymbol{X}\hat{\boldsymbol{w}}\|^2 \leq \frac{1}{\varepsilon}\hat{\mathcal{L}}(f^\star, \boldsymbol{w}^\star) + \|\boldsymbol{X}\boldsymbol{w}^\star\|^2 \leq \frac{1000}{\varepsilon}\mathsf{OPT} + \|\boldsymbol{X}\boldsymbol{w}^\star\|^2.$$

This implies that $\hat{\boldsymbol{w}} \in B(R)$ where $R = \sqrt{\frac{1000}{\varepsilon}\mathsf{OPT} + \|\boldsymbol{X}\boldsymbol{w}^\star\|^2}$. Then, from Lemma 8, the following holds with probability at least 0.999 given $m \geq CL^4\varepsilon^{-2}d\log^3 d$:

$$
\begin{aligned}
\|\hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - f^\star(\boldsymbol{X}\boldsymbol{w}^\star)\|^2 &\leq \|\boldsymbol{S}\hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{S}f^\star(\boldsymbol{X}\boldsymbol{w}^\star)\|^2 + \varepsilon R^2 \\
&\lesssim \|\boldsymbol{S}\hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{S}\boldsymbol{y}\|^2 + \|\boldsymbol{S}f^\star(\boldsymbol{X}\boldsymbol{w}^\star) - \boldsymbol{S}\boldsymbol{y}\|^2 + \varepsilon R^2 \\
&\lesssim \|\boldsymbol{S}\hat{f}(\boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{S}\boldsymbol{y}\|^2 + \|\boldsymbol{S}f^\star(\boldsymbol{X}\boldsymbol{w}^\star) - \boldsymbol{S}\boldsymbol{y}\|^2 + \varepsilon R^2 \\
&\lesssim \left(\hat{\mathcal{L}}(f^\star, \boldsymbol{w}^\star) + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^\star\|^2\right) + \hat{\mathcal{L}}(f^\star, \boldsymbol{w}^\star) + \varepsilon\left(\frac{1}{\varepsilon}\mathsf{OPT} + \|\boldsymbol{X}\boldsymbol{w}^\star\|^2\right) \\
&\lesssim \mathsf{OPT} + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^\star\|^2,
\end{aligned}
$$

where the second line is triangle inequality, the third line used (13) and the definition of R, and the last line used (7). Plugging the above inequality into (12) readily gives

$$\mathcal{L}(\hat{f}, \hat{\boldsymbol{w}}) \lesssim \mathsf{OPT} + \varepsilon\|\boldsymbol{X}\boldsymbol{w}^\star\|^2,$$

as desired (after replacing ε by ε/C for some sufficiently large constant $C > 0$). ∎

Appendix B. Tools from High Dimensional Probability

In this section, we introduce the tools which are used for our main results.

B.1. Covering numbers

We shall need two slightly different definitions of covering numbers, which will coincide in most useful cases.

Covering number with respect to a set. Let K, T be subsets in $\mathbb{R}^n$. The covering number (with respect to set T) $\mathcal{N}(K, T)$ is defined as

$$\mathcal{N}(K, T) := \min\left\{n \in \mathbb{Z}, n \geq 0 : \exists x_1, \ldots, x_n \in K, \text{ such that } K \subset \cup_{i=1}^n (x_i + T)\right\}.$$

The above can also be seen as the minimum number of translations over T required to cover K.

Covering number with respect to a metric. Let $K \subset \mathbb{R}^n$. The covering number (with respect to metric d) $\mathcal{N}(K, d, \varepsilon)$ is defined as

$$\mathcal{N}(K, d, \varepsilon) := \min\left\{n \in \mathbb{Z}, n \geq 0 : \exists x_1, \ldots, x_n \in K, \text{ such that } \min_{i \leq n} d(x, x_i) \leq \varepsilon, \forall x \in K\right\}.$$

When d is the Euclidean distance, we omit d and simply denote by $\mathcal{N}(K, \varepsilon)$ the corresponding covering number.

When two definitions agree. If T is the unit ball of some norm $\|\cdot\|_T$ on $\mathbb{R}^n$, and d_T is the distance induced by the norm $\|\cdot\|_T$, then the above two definitions agree:

$$\mathcal{N}(K, \varepsilon T) = \mathcal{N}(K, d_T, \varepsilon).$$

Volumetric estimate. The following result is standard (or follows from the well-known argument based on packing number and volume), c.f. Lemma 9.5 in Ledoux and Talagrand (1991).

Lemma 9 *Let T be a symmetric convex body in $\mathbb{R}^n$ with positive volume. We have, for all $0 < \varepsilon < 1$, that*

$$\log \mathcal{N}(T, \varepsilon T) \lesssim n \log(2/\varepsilon).$$

Combined with the chain inequality $\mathcal{N}(K_1, K_3) \leq \mathcal{N}(K_1, K_2)\mathcal{N}(K_2, K_3)$, this implies

Lemma 10 *Let K be a subset of $\mathbb{R}^n$ and T be a symmetric convex body in $\mathbb{R}^n$ with positive volume. We have, for all $0 < \varepsilon < 1$, that*

$$\log \mathcal{N}(K, \varepsilon T) \lesssim n \log(2/\varepsilon) + \log \mathcal{N}(K, T).$$

B.2. Dudley's inequality

Let (T, d) be a (pseudo-)metric space. Let $(X_t)_{t \in T}$ be a random process on T with zero mean, i.e., $\mathbb{E} X_t = 0$ for all $t \in T$. The process X_t is said to be subgaussian (with respect to d) if

$$\mathbb{P}(|X_s - X_t| > t) \leq 2 \exp\left(-\frac{t^2}{2d^2(s, t)}\right), \quad \forall s, t \in T.$$

Lemma 11 (Subgaussianity of Bernoulli process, Vershynin (2018)) *Let $T \subset \mathbb{R}^n$ and $\xi_1, \ldots, \xi_n$ be i.i.d. Rademacher random variables. Then the process*

$$X_t := \sum_{i=1}^{n} \xi_i t_i, \quad t = (t_1, \ldots, t_n) \in T$$

is subgaussian on T with respect to the Euclidean distance.

Lemma 12 (Dudley's inequality, Vershynin (2018)) *If $(X_t)_{t \in T}$ is a subgaussian process with respect to a metric d, then*

$$\mathbb{E} \sup_{t \in T} |X_t| \lesssim \inf_{t \in T} \mathbb{E}|X_t| + \int_0^\infty \sqrt{\log(\mathcal{N}(T, d, \varepsilon))} \, \mathrm{d}\varepsilon.$$

It is well-known that Dudley's inequality can be refined to a tail bound.

Lemma 13 (Dudley's inequality, tail bound, Talagrand (2021)) *If $(X_t)_{t \in T}$ is a subgaussian process with respect to a metric d, then the following holds for any $t > 0$ with probability at least $1 - 2\exp(-t^2/2)$:*

$$\sup_{t \in T} |X_t| \lesssim \inf_{t \in T} \mathbb{E}|X_t| + \int_0^\infty \sqrt{\log(\mathcal{N}(T, d, \varepsilon))} \, \mathrm{d}\varepsilon + t \cdot \mathrm{Diam}(T, d).$$

The following upper bound for subgaussian processes supported on a finite set is well-known and is a simple corollary of Dudley's inequality.

Corollary 14 *If $(X_t)_{t \in T}$ is a subgaussian process on a finite set T with respect to a pseudo-metric d. Then*

$$\sup_{t \in T} |X_t| \lesssim \inf_{t \in T} \mathbb{E}|X_t| + \mathrm{Diam}(T, d)\sqrt{\log |T|}.$$

In particular, if $T \subset \mathbb{R}^n$ and d is the Euclidean distance, then

$$\sup_{t \in T} |X_t| \lesssim \sup_{t \in T} \|t\| \cdot \sqrt{\log(|T| + 1)}.$$

Proof For the first part, note that $\mathcal{N}(T, d, \varepsilon) \leq |T|$ for any $\varepsilon > 0$, and moreover $\mathcal{N}(T, d, \varepsilon) = 1$ for $\varepsilon > \mathrm{Diam}(T, d)$. The conclusion the follows from Lemma 12.

For the second part, we apply the conclusion of the first part to the set $T \cup \{0\}$ and note that

$$\mathrm{Diam}(T \cup \{0\}) = \sup_{t_1, t_2 \in T \cup \{0\}} \|t_1 - t_2\| \leq \sup_{t_1, t_2 \in T \cup \{0\}} (\|t_1\| + \|t_2\|) \leq 2 \sup_{t \in T} \|t\|.$$

∎

Dudley's inequality has a partial inverse, which can be viewed as a useful probabilistic estimate of covering numbers.

Lemma 15 (Dual Sudakov minoration, Ledoux and Talagrand (1991)) *Let B be the unit (Euclidean) ball in of $\mathbb{R}^n$, T be a symmetric convex body in $\mathbb{R}^n$, and g be the standard n-dimensional random normal vector. Denote by $T^\circ := \{x : \langle x, y \rangle \leq 1, \forall y \in T\}$ the polar of T. Then*

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathcal{N}(B, \varepsilon T^\circ)} \lesssim \mathbb{E} \sup_{t \in T} \langle t, g \rangle.$$

20

Dudley's integral is subadditive with respect to the metric d and tensorizes well, two basic properties that will greatly simplify our proof.

Lemma 16 (Sublinearity) *Let d_1, d_2 be pseudo-metrics on some set T. Let $a, b > 0$ be real numbers. Then*

$$\int_0^\infty \sqrt{\log \mathcal{N}(T, ad_1 + bd_2, \varepsilon)}\, \mathrm{d}\varepsilon \lesssim a \int_0^\infty \sqrt{\log \mathcal{N}(T, d_1, \varepsilon)}\, \mathrm{d}\varepsilon + b \int_0^\infty \sqrt{\log \mathcal{N}(T, d_2, \varepsilon)}\, \mathrm{d}\varepsilon.$$

Lemma 17 (Tensorization) *Let (T_1, d_1), (T_2, d_2) be pseudo-metric spaces. Then $T_1 \times T_2$ can be endowed with a natural pseudo-metric, defined by*

$$d\big((t_1, t_2), (t_1', t_2')\big) = d_1(t_1, t_1') + d_2(t_2, t_2').$$

We then have

$$\int_0^\infty \sqrt{\log \mathcal{N}(T_1 \times T_2, d, \varepsilon)}\, \mathrm{d}\varepsilon \lesssim \int_0^\infty \sqrt{\log \mathcal{N}(T_1, d_1, \varepsilon)}\, \mathrm{d}\varepsilon + \int_0^\infty \sqrt{\log \mathcal{N}(T_2, d_2, \varepsilon)}\, \mathrm{d}\varepsilon.$$

Appendix C. Proof of subspace embedding with fixed nonlinearity, Lemma 6

The proof relies crucially on the idea developed in Rudelson (1999). We extend a simplified version of their idea into the Lipschitz nonlinear setting. First, we apply a standard symmetrization argument to the ℓth moment of the desired quantity, which is actually the supremum of deviation of $\|\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}_1) - \boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}_2)\|^2$ over $\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)$. This reduces the problem to finding tail bounds of a Bernoulli process. Then, we utilize the sub-Gaussian property of that Bernoulli process and give an upper bound of its associated sub-Gaussian distances, which reduces the sub-Gaussian metric to the dual norm of some polytope determined by the sampling process. This allows us to invoke dual Sudakov minoration to bound the Dudley integral of the sub-Gaussian distance, ultimately leading to a tail bound of the Bernoulli process as we desire.

C.1. Symmetrization

To simplify notations, for the rest of the proof we denote

$$v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2) := f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle).$$

Lemma 18 (Symmetrization) *Let $\xi_1, \xi_2, \ldots, \xi_m$ be i.i.d. Rademacher random variables (independent of $\boldsymbol{S}$), i.e., each ξ_i takes values -1 and 1 with probabilities $1/2$ each independently. Then*

$$\mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| \left\|\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}_1) - \boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}_2)\right\|^2 - \left\|f(\boldsymbol{X}\boldsymbol{w}_1) - f(\boldsymbol{X}\boldsymbol{w}_2)\right\|^2 \right|^\ell$$

$$\leq 2^\ell \cdot \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| \frac{1}{m} \sum_{i=1}^m \xi_i \frac{(v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i}} \right|^\ell.$$

This result is standard. For sake of completeness, we provide a proof in Appendix C.5.

C.2. Bounding the Bernoulli process: the sub-Gaussian distance

Lemma 18 leads us to study the following symmetrized random process.

$$Z(\boldsymbol{w}_1, \boldsymbol{w}_2) \coloneqq \sum_{i=1}^{m} \xi_i \frac{\left(v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2)\right)^2}{p_{j_i}}, \qquad (\boldsymbol{w}_1, \boldsymbol{w}_2) \in B(R) \times B(R),$$

It is clear that the above random process conditioned on the samples $\{j_1, j_2, \ldots, j_m\}$ (fixing $\boldsymbol{S}$) is a sub-Gaussian process (Vershynin (2018)) with respect to index $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ endowed with the metric ρ_f (called the sub-Gaussian distance) on $B(R) \times B(R)$ defined as follows.

$$\rho_f\big((\boldsymbol{w}_1, \boldsymbol{w}_2), (\boldsymbol{w}_1', \boldsymbol{w}_2')\big) \coloneqq \left(\sum_{i=1}^{m} \frac{1}{p_{j_i}^2} \big(v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2)^2 - v_{j_i}(\boldsymbol{w}_1', \boldsymbol{w}_2')^2\big)^2 \right)^{1/2}. \tag{14}$$

Based on sub-Gaussianity of $Z(\boldsymbol{w}_1, \boldsymbol{w}_2)$, we shall derive a near-optimal tail bound of it using Dudley's inequality (Lemma 12). This would entail establishing an easier-to-manipulate upper bound of the metric ρ_f. Before doing so (in Lemma 19) we need to set up a few notations.

Recall that we are conditioning on $\{j_1, j_2, \cdots, j_m\}$ in this part of the proof. We define a symmetric convex body

$$\mathcal{P} = \text{conv}\left(\pm \frac{\boldsymbol{x}_{j_1}}{\sqrt{p_{j_1}}}, \pm \frac{\boldsymbol{x}_{j_2}}{\sqrt{p_{j_2}}}, \ldots \pm \frac{\boldsymbol{x}_{j_m}}{\sqrt{p_{j_m}}} \right).$$

We further denote $\mathcal{P}^\circ$ as its polar given by

$$\mathcal{P}^\circ = \{\boldsymbol{z} \in \mathbb{R}^d : \big|\langle \boldsymbol{z}, \boldsymbol{p}\rangle\big| \le 1, \forall \boldsymbol{p} \in \mathcal{P}\},$$

and denote $\|\cdot\|_{\mathcal{P}^\circ}$ as the Minkowski norm associated with $\mathcal{P}^\circ$, defined for $\boldsymbol{w} \in \mathbb{R}^d$ as

$$\|\boldsymbol{w}\|_{\mathcal{P}^\circ} \coloneqq \inf\{t > 0 : \boldsymbol{w} \in t\mathcal{P}^\circ\} = \sup_{i \in [m]} \left|\langle \boldsymbol{x}_{j_i}/\sqrt{p_{j_i}}, \boldsymbol{w}\rangle\right|.$$

Using Lipschitz continuity of f, we can establish a useful upper bound of ρ_f in terms of $\|\cdot\|_{\mathcal{P}^\circ}$, reducing the study of the sub-Gaussian distance to the dual norm associated to a certain polytope.

Lemma 19 (Bounding the sub-Gaussian distance ρ_f) *With the above notations, one has for $f \in$ Lip_L that*

$$\rho_f \le \rho,$$

where ρ is the metric on $B(R) \times B(R)$ defined by

$$\rho\big((\boldsymbol{w}_1, \boldsymbol{w}_2), (\boldsymbol{w}_1', \boldsymbol{w}_2')\big) = 4L^2 R \left\| \sum_{i=1}^{m} \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \big(\|\boldsymbol{w}_1 - \boldsymbol{w}_1'\|_{\mathcal{P}^\circ} + \|\boldsymbol{w}_2 - \boldsymbol{w}_2'\|_{\mathcal{P}^\circ}\big). \tag{15}$$

Proof The definition of ρ_f and v_{j_i} leads to the following.

$$
\begin{aligned}
|v_{j_i}^2(\boldsymbol{w}_1, \boldsymbol{w}_2) - v_{j_i}^2(\boldsymbol{w}_1', \boldsymbol{w}_2')| &= \Big|\big(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big) - \big(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1'\rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2'\rangle)\big)\Big| \\
&\quad \cdot \Big|\big(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big) + \big(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1'\rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2'\rangle)\big)\Big| \\
&\le L^2 \left(\big|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 - \boldsymbol{w}_1'\rangle\big| + \big|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 - \boldsymbol{w}_2'\rangle\big|\right) \\
&\quad \cdot \left(|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1'\rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2'\rangle|\right),
\end{aligned}
$$

where the inequality follows from f being L-Lipschitz. Next we observe that

$$\left|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 - \boldsymbol{w}_1' \rangle\right| + \left|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 - \boldsymbol{w}_2' \rangle\right| = \sqrt{p_{j_i}} \left(\left|\langle \boldsymbol{x}_{j_i}/\sqrt{p_{j_i}}, \boldsymbol{w}_1 - \boldsymbol{w}_1' \rangle\right| + \left|\langle \boldsymbol{x}_{j_i}/\sqrt{p_{j_i}}, \boldsymbol{w}_2 - \boldsymbol{w}_2' \rangle\right| \right)$$
$$\leq \sqrt{p_{j_i}} (\|\boldsymbol{w}_1 - \boldsymbol{w}_1'\|_{\mathcal{P}^\circ} + \|\boldsymbol{w}_2 - \boldsymbol{w}_2'\|_{\mathcal{P}^\circ}),$$

where the inequality follows from the definition of $\|\cdot\|_{\mathcal{P}^\circ}$. Combining the above two inequalities, one has

$$\rho_f^2\big((\boldsymbol{w}_1, \boldsymbol{w}_2), (\boldsymbol{w}_1', \boldsymbol{w}_2')\big)$$

$$\leq \sum_{i=1}^m \frac{1}{p_{j_i}^2} \cdot L^4 p_{j_i} (\|\boldsymbol{w}_1 - \boldsymbol{w}_1'\|_{\mathcal{P}^\circ} + \|\boldsymbol{w}_2 - \boldsymbol{w}_2'\|_{\mathcal{P}^\circ})^2 \cdot \big(|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1' \rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2' \rangle| \big)^2$$

$$\leq L^4 (\|\boldsymbol{w}_1 - \boldsymbol{w}_1'\|_{\mathcal{P}^\circ} + \|\boldsymbol{w}_2 - \boldsymbol{w}_2'\|_{\mathcal{P}^\circ})^2 \sum_{i=1}^m \frac{1}{p_{j_i}} \big(|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1' \rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2' \rangle| \big)^2$$

$$\leq L^4 (\|\boldsymbol{w}_1 - \boldsymbol{w}_1'\|_{\mathcal{P}^\circ} + \|\boldsymbol{w}_2 - \boldsymbol{w}_2'\|_{\mathcal{P}^\circ})^2 \cdot 4 \sum_{i=1}^m \frac{1}{p_{j_i}} \Big(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle^2 + \langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle^2 + \langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1' \rangle^2 + \langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2' \rangle^2 \Big)$$

$$\leq 16 L^4 (\|\boldsymbol{w}_1 - \boldsymbol{w}_1'\|_{\mathcal{P}^\circ} + \|\boldsymbol{w}_2 - \boldsymbol{w}_2'\|_{\mathcal{P}^\circ})^2 \sup_{\boldsymbol{w} \in B(R)} \sum_{i=1}^m \frac{1}{p_{j_i}} \langle \boldsymbol{x}_{j_i}, \boldsymbol{w} \rangle^2, \tag{16}$$

where the third inequality follows from Cauchy-Schwarz. To simplify the last factor, we note that

$$\sup_{\boldsymbol{w} \in B(R)} \sum_{i=1}^m \frac{1}{p_{j_i}} \langle \boldsymbol{x}_{j_i}, \boldsymbol{w} \rangle^2 = \sup_{\boldsymbol{w} \in B(R)} \boldsymbol{w}^\top \left(\sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right) \boldsymbol{w}$$
$$= R^2 \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|.$$

Plugging this back into (16) and taking square roots yield the claimed result. ∎

The rest of the proof of Lemma 6 goes as follows. In Section C.3, we show how to derive a bound in expectation of the supremum of $Z(\boldsymbol{w}_1, \boldsymbol{w}_2)$ using Dudley's inequality and duality of metric entropy. Once this is done, we show how to turn the bound in expectation to a tail bound using a standard argument by concentration of measure in Section C.4.

C.3. Bound in expectation via duality

We apply Dudley's inequality to the process $Z(\boldsymbol{w}_1, \boldsymbol{w}_2)$ over the set $B(R) \times B(R)$. Since $Z(\boldsymbol{w}, \boldsymbol{w}) = 0$ for any $\boldsymbol{w}$, it follows that $\inf_{\boldsymbol{w}_1 \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| = 0$. Taking expectation with respect to the Rademacher random variables $\xi_1, \xi_2, \ldots, \xi_m$, we have

$$\mathbb{E}_\xi \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| \lesssim \int_0^\infty \sqrt{\log \mathcal{N}(B(R) \times B(R), \rho_f, \varepsilon)} \, \mathrm{d}\varepsilon$$

$$\lesssim L^2 R \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \int_0^\infty \sqrt{\log \mathcal{N}(B(R), \|\cdot\|_{\mathcal{P}^\circ}, \varepsilon)} \, \mathrm{d}\varepsilon, \tag{17}$$

The second inequality can be verified using Lemma 19 along with standard properties of covering numbers (Lemma 16 and Lemma 17).

We move on to estimate the Dudley integral $\int_0^\infty \sqrt{\log \mathcal{N}(B(R), \|\cdot\|_{\mathcal{P}^\circ}, \varepsilon)}\, d\varepsilon$. Since this integral only involves the entropy in dual norm $\|\cdot\|_{\mathcal{P}^\circ}$, we control it using duality of metric entropy. In particular, one can invoke dual Sudakov minoration (Lemma 15) to prove

$$\int_0^\infty \sqrt{\log \mathcal{N}(B(R), \|\cdot\|_{\mathcal{P}^\circ}, \varepsilon)}\, d\varepsilon \lesssim \sup_{i\in[m]} \frac{\|\boldsymbol{x}_{j_i}\|}{\sqrt{p_{j_i}}} R \sqrt{\log^2 d \cdot \log m} = R\sqrt{d\log^2 d \cdot \log m}, \quad (18)$$

where the last equality uses $\sqrt{p_{j_i}} = \|\boldsymbol{x}_{j_i}\|/\sqrt{d}$ which follows from the definition $p_{j_i} = \frac{\tau_{j_i}(\boldsymbol{X})}{\sum_{j=1}^n \tau_j(\boldsymbol{X})}$ and 6. Plugging this into (17), we obtain

$$\mathbb{E}_\xi \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| \lesssim L^2 R^2 \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \sqrt{d\log^2 d \cdot \log m}. \quad (19)$$

Taking expectation w.r.t. $\boldsymbol{S}$ and noticing that, by matrix Chernoff bound (Rudelson and Vershynin, 2007),

$$\mathbb{E} \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\| \lesssim m + d\log d \lesssim m, \quad (20)$$

(where the last inequality follows from the assumption $m \gtrsim d\log^3 d$) we obtain

$$\mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| \lesssim L^2 R^2 \sqrt{md\log^2 d \cdot \log m}.$$

Under the assumption $m \gtrsim L^4 \varepsilon^{-2} d\log^3 d$, we deduce that

$$\frac{1}{m} \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| \lesssim \varepsilon R^2. \quad (21)$$

This establishes the bound in expectation as desired. Once we can show that the above also holds with high probability rather than in expectation, the proof of Lemma 6 will be completed. This will be done in the next part.

C.4. Completing the proof: tail bound via concentration of measure

This part is more or less standard applications of chaining and concentration of measure arguments (Ledoux, 2001; Talagrand, 2021). Since $Z(\boldsymbol{w}_1, \boldsymbol{w}_2)$ conditioned on $\{j_1, \ldots, j_m\}$ is sub-Gaussian with respect to metric ρ_f, we have (Talagrand, 2021)

$$\left(\mathbb{E}_\xi \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)|^\ell \right)^{1/\ell} \lesssim \mathbb{E}_\xi \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| + \sqrt{\ell} \cdot \mathrm{Diam}(B(R) \times B(R), \rho_f). \quad (22)$$

To bound the diameter $\mathrm{Diam}(B(R) \times B(R), \rho_f)$, we invoke Lemma 19 and the simple observation that for $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_1', \boldsymbol{w}_2' \in B(R)$,

$$\|\boldsymbol{w}_1 - \boldsymbol{w}_1'\|_{\mathcal{P}^\circ} + \|\boldsymbol{w}_2 - \boldsymbol{w}_2'\|_{\mathcal{P}^\circ} \leq 4R \sup_{i\in[m]} \frac{\|\boldsymbol{x}_{j_i}\|}{\sqrt{p_{j_i}}} = 4R\sqrt{d},$$

where the first inequality follows from the definition of $\|\cdot\|_{\mathcal{P}^\circ}$ and Cauchy-Schwarz, and the second inequality follows from $\sqrt{p_{j_i}} = \|\boldsymbol{x}_{j_i}\|/\sqrt{d}$ aforementioned. Plug this into Lemma 19 to obtain

$$\mathrm{Diam}(B(R) \times B(R), \rho_f) \lesssim L^2 R^2 \sqrt{d} \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2}.$$

Plug this and (19) into (22) to obtain

$$\left(\mathbb{E}_\xi \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)|^\ell \right)^{1/\ell} \lesssim L^2 R^2 \left(\sqrt{d \log^2 d \cdot \log m} + \sqrt{\ell d} \right) \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2}.$$

Applying matrix Chernoff inequality again (but in its tail bound form this time), one has

$$\left(\mathbb{E} \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{\ell/2} \right)^{2/\ell} \lesssim m + \sqrt{\ell m d \log d}.$$

Therefore

$$
\begin{aligned}
\left(\mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| Z(\boldsymbol{w}_1, \boldsymbol{w}_2) \right|^\ell \right)^{1/\ell} &= \left(\mathbb{E}_{\boldsymbol{S}} \mathbb{E}_\xi \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| Z(\boldsymbol{w}_1, \boldsymbol{w}_2) \right|^\ell \right)^{1/\ell} \\
&\lesssim \left(\mathbb{E}_{\boldsymbol{S}} \left(L^2 R^2 \left(\sqrt{d \log^2 d \cdot \log m} + \sqrt{\ell d} \right) \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \right)^\ell \right)^{1/\ell} \\
&\lesssim L^2 R^2 \left(\sqrt{d \log^2 d \cdot \log m} + \sqrt{\ell d} \right) \cdot \left(\mathbb{E} \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{\ell/2} \right)^{1/\ell} \\
&\lesssim L^2 R^2 \left(\sqrt{d \log^2 d \cdot \log m} + \sqrt{\ell d} \right) \left(m + \sqrt{\ell m d \log d} \right)^{1/2}.
\end{aligned}
$$

For the last expression, using $(a + b)^{1/2} \lesssim a^{1/2} + b^{1/2}$ for $a, b > 0$ and expanding the product, we obtain

$$
\left(\mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| Z(\boldsymbol{w}_1, \boldsymbol{w}_2) \right|^\ell \right)^{1/\ell}
\lesssim L^2 R^2 \left(\sqrt{m d \log^2 d \cdot \log m} + \ell^{\frac{1}{4}} m^{\frac{1}{4}} d^{\frac{3}{4}} \log^{\frac{5}{4}} d \cdot \log^{\frac{1}{2}} m + \ell^{\frac{1}{2}} \sqrt{m d} + \ell^{\frac{3}{4}} m^{\frac{1}{4}} d^{\frac{3}{4}} \log^{\frac{1}{4}} d \right).
$$

This gives the desired tail bound by a standard computation based on Markov's inequality, which is summarized in the following lemma.

Lemma 20 *Assume a random variable X satisfies*

$$\left(\mathbb{E}|X|^\ell\right)^{1/\ell} \le A_0 + \sum_{k=1}^{K} A_k \ell^{\alpha_k}, \quad \forall \ell \ge 1.$$

where $A_k \ge 0$, $\alpha_k \in (0,1]$ are constants. There is some constant $C > 0$ depending only on K, such that for any $\delta \in (0, 1/2)$ we have

$$\mathbb{P}\left(|X| > CA_0 + C\sum_{k=1}^{K} A_k \log^{\alpha_k}\left(\frac{1}{\delta}\right)\right) \le 1 - \delta.$$

C.5. Proof of Lemma 21

Throughout this proof, we denote

$$v(\boldsymbol{w}_1, \boldsymbol{w}_2) := f(\boldsymbol{X}\boldsymbol{w}_1) - f(\boldsymbol{X}\boldsymbol{w}_2).$$

Let $\boldsymbol{S}'$ be an independent copy of $\boldsymbol{S}$. Recalling the definition of $\boldsymbol{S}$, we denote by j_i the indices chosen by $\boldsymbol{S}$ and j_i' the indices chosen by $\boldsymbol{S}'$. We want to estimate the following

$$E := \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left|\left\|\boldsymbol{S}v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2 - \left\|v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2\right|^\ell.$$

Since $\mathbb{E}\left\|\boldsymbol{S}'v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2 - \left\|v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2 = 0$, we have

$$E = \mathbb{E}_{\boldsymbol{S}} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left|\left\|\boldsymbol{S}v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2 - \left\|v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2 - \mathbb{E}_{\boldsymbol{S}'}\left(\left\|\boldsymbol{S}'v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2 - \left\|v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2\right)\right|^\ell$$

$$= \mathbb{E}_{\boldsymbol{S}} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left|\left\|\boldsymbol{S}v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2 - \mathbb{E}_{\boldsymbol{S}'}\left\|\boldsymbol{S}'v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2\right|^\ell.$$

From the convexity of $|\cdot|^\ell$, we get

$$E \le \mathbb{E}_{\boldsymbol{S}} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \mathbb{E}_{\boldsymbol{S}'}\left|\left\|\boldsymbol{S}v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2 - \left\|\boldsymbol{S}'v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2\right|^\ell$$

$$\le \mathbb{E}_{\boldsymbol{S}, \boldsymbol{S}'} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left|\left\|\boldsymbol{S}v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2 - \left\|\boldsymbol{S}'v(\boldsymbol{w}_1, \boldsymbol{w}_2)\right\|^2\right|^\ell$$

$$= \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left|\frac{1}{m}\sum_{i=1}^{m} \frac{(v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i}} - \frac{1}{m}\sum_{i=1}^{m} \frac{(v_{j_i'}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i'}}\right|^\ell.$$

Since $\sum_{i=1}^{m} \frac{(v_{j_i'}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i'}}$ is an independent copy of $\sum_{i=1}^{m} \frac{(v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i}}$, their difference is a symmetric distribution and the same as $\sum_{i=1}^{m} \xi_i \left(\frac{(v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i}} - \frac{(v_{j_i'}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i'}}\right)$, where $\xi_1, \xi_2, \ldots, \xi_m$

are i.i.d. Rademacher random variables. Therefore

$$E \leq \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \left(\frac{(v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i}} - \frac{(v_{j'_i}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j'_i}} \right) \right|^{\ell}$$

$$\leq 2^{\ell-1} \cdot \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| \sum_{i=1}^{m} \xi_i \frac{(v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i}} \right|^{\ell} + 2^{\ell-1} \cdot \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \frac{(v_{j'_i}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j'_i}} \right|^{\ell}$$

$$= 2^{\ell} \cdot \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| \sum_{i=1}^{m} \xi_i \frac{(v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2))^2}{p_{j_i}} \right|^{\ell},$$

where the second line is the triangle inequality $|a + b|^{\ell} \leq 2^{\ell-1}(|a|^{\ell} + |b|^{\ell})$, and the last line used the observation that the two terms are identically distributed.

C.6. Proof of Lemma 20

By Markov's inequality, we have, for all $t > 0$ and $\ell \geq 1$, that

$$\mathbb{P}(|X| > t) \leq t^{-\ell} \left(A_0 + \sum_{k=1}^{K} A_k \ell^{\alpha_k} \right)^{\ell}$$

$$\leq C_K^{\ell} t^{-\ell} \left(A_0^{\ell} + \sum_{k=1}^{K} A_k^{\ell} \ell^{\alpha_k \ell} \right)$$

$$= \left(\frac{C_K A_0}{t} \right)^{\ell} + \sum_{k=1}^{K} \left(\frac{C_K A_k \ell^{\alpha_k}}{t} \right)^{\ell}.$$

In particular, setting $t = 4(K+1)C_K A_0 + 4(K+1) \sum_{k=1}^{K} A_k \ell^{\alpha_k}$, we obtain

$$\mathbb{P} \left(|X| > 4(K+1)C_K A_0 + 4(K+1)C_K \sum_{k=1}^{K} A_k \ell^{\alpha_k} \right) \leq (4(K+1))^{-\ell} + \sum_{k=1}^{K} (4(K+1))^{-\ell} \leq 4^{-\ell}.$$

The desired conclusion follows from plugging in $\ell = \lceil \log(1/\delta) \rceil$.

Appendix D. Proof of main theorem with unknown nonlinearity, Theorem 3

Proof Similar to the proof of Lemma 6, we need to bound the supremum of the process

$$Z_{f_1, f_2}(\boldsymbol{w}_1, \boldsymbol{w}_2) := \sum_{i=1}^{m} \xi_i \frac{\left(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle) \right)^2}{p_{j_i}}, \qquad f_1, f_2 \in \mathrm{Lip}_L, \ (\boldsymbol{w}_1, \boldsymbol{w}_2) \in B(R) \times B(R).$$

The crucial step is our proof is to construct an appropriate discretization of Lip_L. The guideline for choosing such a net will be based on the seminal idea of chaining for Bernoulli process (Talagrand, 2021). In our setting, it suffices to use a simplified version of chaining, which we shall

present immediately after introducing the relevant notations. Define for $f_1, f_2 \in \text{Lip}_L$ the following two metrics:

$$D_\infty\bigg((f_1, f_2), (f_1', f_2')\bigg)$$

$$:= \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \sum_i \frac{1}{p_{j_i}} \left| \left(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\right)^2 - \left(f_1'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\right)^2 \right|$$

and

$$D_2\bigg((f_1, f_2), (f_1', f_2')\bigg)$$

$$:= \left(\sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \sum_i \frac{1}{p_{j_i}^2} \left| \left(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\right)^2 - \left(f_1'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\right)^2 \right|^2 \right)^{1/2}.$$

Lemma 21 (Chaining for Bernoulli process) *Let $\mathcal{N}_\Delta \subset \text{Lip}_L$ be such that $\mathcal{N}_\Delta \times \mathcal{N}_\Delta$ is a Δ-net of $\text{Lip}_L \times \text{Lip}_L$ w.r.t metric D_∞ (where Δ can be a random variable depending on $\boldsymbol{S}$; more precisely, it is measurable with respect to the σ-algebra generated by $\boldsymbol{S}$). Then*

$$\mathbb{E} \sup_{\substack{f_1, f_2 \in \text{Lip}_L \\ \boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)}} \left| \sum_{i=1}^m \frac{\xi_i}{p_{j_i}} \left(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\right)^2 \right|$$

$$\lesssim L^2 R^2 \sqrt{md\log^2 d \cdot \log m} + \mathbb{E}\Delta + \mathbb{E}\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{N}_\Delta \times \mathcal{N}_\Delta, D_2, \varepsilon)}\, d\varepsilon.$$

The proof is postponed to Section D.1.

The metrics D_∞ and D_2, which are defined on $\text{Lip}_L \times \text{Lip}_L$, are highly complicated to analyse. So, we consider the following upper bounds of them, which decompose them into simpler metrics defined on Lip_L.

Lemma 22 (Decomposing the metric D_∞ and D_2) *Let $I_{j_i} = [-R\|\boldsymbol{x}_{j_i}\|, R\|\boldsymbol{x}_{j_i}\|]$. Then*

$$D_\infty\bigg((f_1, f_2), (f_1', f_2')\bigg) \lesssim \rho_\infty(f_1, f_1') + \rho_\infty(f_2, f_2'),$$

where

$$\rho_\infty(f_1, f_2) := LR \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \left(\sum_{i=1}^m \frac{1}{p_{j_i}} \|f_1 - f_2\|_{L^\infty(I_{j_i})}^2 \right)^{1/2}.$$

On the other hand, for any $\mu > 0$, we have

$$D_2\bigg((f_1, f_2), (f_1', f_2')\bigg) \lesssim \rho_2(f_1, f_1') + \rho_2(f_2, f_2'),$$

where the universal constant hidden by $\lesssim$ is independent of μ, and the metric ρ_2 is defined by

$$\rho_2(f_1, f_2)$$

$$:= LR \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \left(\sup_{i\in[m]} \frac{\|f_1 - f_2\|_{L^\infty(I_{j_i})} \mathbf{1}_{\|\boldsymbol{x}_{j_i}\|\geq\mu}}{\sqrt{p_{j_i}}} \right) + L^2 R^2 d \left(\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\|<\mu} \right)^{1/2} \rho_\delta(f_1, f_2).$$

Here ρ_δ is the Dirac distance,

$$\rho_\delta(f_1, f_2) = \begin{cases} 0, & f_1 = f_2, \\ 1, & f_1 \neq f_2. \end{cases}$$

The proof can be found in Appendix D.2.

As an immediate application of the bounds in Lemma 22, we obtain the following simplified version of the Lemma 21.

Corollary 23 (Chaining with simplified metrics) . *Assume $\mathcal{N}_\Delta \subset \mathrm{Lip}_L$ is a Δ-net of Lip_L w.r.t metric ρ_∞ (where Δ can be a random variable depending on $\boldsymbol{S}$). Then*

$$\mathbb{E} \sup_{\substack{f_1, f_2 \in \mathrm{Lip}_L \\ \boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)}} \left| \sum_{i=1}^m \frac{\xi_i}{p_{j_i}} \left(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle) \right)^2 \right|$$
$$\lesssim L^2 R^2 \sqrt{md \log^2 d \cdot \log m} + \mathbb{E}\Delta + \mathbb{E} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{N}_\Delta, \rho_2, \varepsilon)} \, d\varepsilon.$$

We turn to construct a Δ-net for Lip_L with respect to the metric ρ_∞. In light of the definition of ρ_∞, one may try to construct an Δ-net with respect to ρ_∞ by piecewise linear functions, each of which differs with its nearest neighbors on the interval I_{j_i} by an amount proportional to $\sqrt{p_{j_i}}$, say $\eta \sqrt{p_{j_i}}$ for some $\eta > 0$ to be chosen later. As long as p_{j_i} is not too small, this is achievable. This idea culminates to the following lemma.

Lemma 24 (Construction of $\mathcal{N}_\Delta$) *Fix some $\mu, \eta \in (0, 1/2)$. Let*

$$\Delta(\mu, \eta) := \eta L^2 R^2 \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \left(md + \sum_{i=1}^m \frac{\mu^2}{p_{j_i}} \right)^{1/2}.$$

Then there exists a $\Delta(\mu, \eta)$-net of Lip_L with respect to the metric ρ_∞, such that

(i) The cardinality of $\mathcal{N}_\Delta$ is controlled:

$$\log |\mathcal{N}_\Delta| \lesssim \frac{\log(1/\mu)}{\eta}.$$

(ii) Assume $m \geq Cd \log d$. The expectation of the Dudley's integral of $\mathcal{N}_\Delta$ with respect to ρ_2 is also controlled:

$$\mathbb{E} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{N}_\Delta, \rho_2, \varepsilon)} \, d\varepsilon \lesssim L^2 R^2 \sqrt{md} \left(\log(1/\mu) + \mu \sqrt{\frac{n \log(1/\mu)}{\eta}} \right).$$

(iii) Assume $m \geq Cd \log d$. The expectation of Δ satisfies

$$\mathbb{E}\Delta(\mu, \eta) \lesssim \eta L^2 R^2 m \sqrt{d} \cdot \sqrt{1 + \frac{\mu^2 n}{d}}.$$

The proof of this lemma, which is a major technical challenge in this paper, is postponed to Appendix D.3.

Plug this into Lemma 21 to obtain

$$\mathbb{E} \sup_{\substack{f_1,f_2\in\mathrm{Lip}_L \\ \boldsymbol{w}_1,\boldsymbol{w}_2\in B(R)}} |Z_{f_1,f_2}(\boldsymbol{w}_1,\boldsymbol{w}_2)|$$

$$\lesssim L^2R^2\sqrt{md\log^2 d \cdot \log m} + \eta L^2R^2 m\sqrt{d}\cdot\sqrt{1+\frac{\mu^2 n}{d}} + L^2R^2\sqrt{md}\left(\log(1/\mu)+\mu\sqrt{\frac{n\log(1/\mu)}{\eta}}\right).$$

Setting $\mu = \frac{d^2}{nm}$ and $\eta = 1/\sqrt{m}$, we obtain

$$\mathbb{E} \sup_{\substack{f_1,f_2\in\mathrm{Lip}_L \\ \boldsymbol{w}_1,\boldsymbol{w}_2\in B(R)}} |Z_{f_1,f_2}(\boldsymbol{w}_1,\boldsymbol{w}_2)|$$

$$\lesssim L^2R^2\sqrt{md\log^2 d \cdot \log m} + L^2R^2\sqrt{md} + L^2R^2\sqrt{md}\log\left(\frac{nm}{d^2}\right).$$

From this, it can be seen that whenever $m \gtrsim L^4\varepsilon^{-2}d(\log^2 n + \log^3 d)$, we have

$$\mathbb{E} \sup_{\substack{f_1,f_2\in\mathrm{Lip}_L \\ \boldsymbol{w}_1,\boldsymbol{w}_2\in B(R)}} |Z_{f_1,f_2}(\boldsymbol{w}_1,\boldsymbol{w}_2)| \le \varepsilon R^2.$$

Similar to the proof of Lemma 6, this is close to what we desire, except that we need a tail bound. The latter can be deduced from the above using the same concentration of measure argument as in Lemma 6, which we omit here to avoid repetition. ∎

D.1. Proof of Lemma 21

We begin with an important property of D_∞ that the following error bound is true deterministically:

$$\left|Z_{f_1,f_2}(\boldsymbol{w}_1,\boldsymbol{w}_2) - Z_{f_1',f_2'}(\boldsymbol{w}_1,\boldsymbol{w}_2)\right| \le D_\infty\big((f_1,f_2),(f_1',f_2')\big).$$

This follows easily from the fact that $|\xi_i| \le 1$. Therefore, if $\mathcal{N}_\Delta \times \mathcal{N}_\Delta$ is a Δ-net of $\mathrm{Lip}_L \times \mathrm{Lip}_L$ with respect to the metric D_∞, one has

$$\sup_{f_1,f_2\in\mathrm{Lip}_L} |Z_{f_1,f_2}(\boldsymbol{w}_1,\boldsymbol{w}_2)| \le \sup_{f_1,f_2\in\mathcal{N}_\Delta} |Z_{f_1,f_2}(\boldsymbol{w}_1,\boldsymbol{w}_2)| + \Delta. \tag{23}$$

Taking supremum with respect to $\boldsymbol{w}_1,\boldsymbol{w}_2$ and then taking expectation, we obtain

$$\mathbb{E} \sup_{\substack{f_1,f_2\in\mathrm{Lip}_L \\ \boldsymbol{w}_1,\boldsymbol{w}_2\in B(R)}} |Z_{f_1,f_2}(\boldsymbol{w}_1,\boldsymbol{w}_2)| \le \mathbb{E} \sup_{\substack{f_1,f_2\in\mathcal{N}_\Delta \\ \boldsymbol{w}_1,\boldsymbol{w}_2\in B(R)}} |Z_{f_1,f_2}(\boldsymbol{w}_1,\boldsymbol{w}_2)| + \mathbb{E}\Delta. \tag{24}$$

To bound the expectation of the supremum appearing in the right hand side, we condition on $j_1,\ldots,j_m$ again and observe that $Z_{f_1,f_2}(\boldsymbol{w}_1,\boldsymbol{w}_2)$ is sub-Gaussian over index $(f_1,f_2,\boldsymbol{w}_1,\boldsymbol{w}_2) \in$

$\mathcal{N}_\Delta \times \mathcal{N}_\Delta \times B(R) \times B(R)$, endowed with metric

$$
D\big((f_1, f_2, \boldsymbol{w}_1, \boldsymbol{w}_2), (f_1', f_2', \boldsymbol{w}_1', \boldsymbol{w}_2')\big)
$$
$$
\coloneqq \left(\sum_i \frac{1}{p_{j_i}^2} \left| \big(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle)\big)^2 - \big(f_1'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1' \rangle) - f_2'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2' \rangle)\big)^2 \right|^2 \right)^{1/2} .
$$

Applying triangle inequality gives

$$
D\big((f_1, f_2, \boldsymbol{w}_1, \boldsymbol{w}_2), (f_1', f_2', \boldsymbol{w}_1', \boldsymbol{w}_2')\big)
$$
$$
\leq D\big((f_1, f_2, \boldsymbol{w}_1', \boldsymbol{w}_2'), (f_1', f_2', \boldsymbol{w}_1', \boldsymbol{w}_2')\big) + D\big((f_1, f_2, \boldsymbol{w}_1, \boldsymbol{w}_2), (f_1, f_2, \boldsymbol{w}_1', \boldsymbol{w}_2')\big).
$$

For the first term, by definition, it's easy to see that

$$
D\big((f_1, f_2, \boldsymbol{w}_1', \boldsymbol{w}_2'), (f_1', f_2', \boldsymbol{w}_1', \boldsymbol{w}_2')\big)
$$
$$
\leq \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} D\big((f_1, f_2, \boldsymbol{w}_1, \boldsymbol{w}_2), (f_1', f_2', \boldsymbol{w}_1, \boldsymbol{w}_2)\big)
$$
$$
= D_2\big((f_1, f_2), (f_1', f_2')\big),
$$

and, for the second term, it is easy to verify that

$$
D\big((f_1, f_2, \boldsymbol{w}_1, \boldsymbol{w}_2), (f_1, f_2, \boldsymbol{w}_1', \boldsymbol{w}_2')\big) = \rho_{f_1 - f_2}\big((\boldsymbol{w}_1, \boldsymbol{w}_2), (\boldsymbol{w}_1', \boldsymbol{w}_2')\big) \leq 2\rho\big((\boldsymbol{w}_1, \boldsymbol{w}_2), (\boldsymbol{w}_1', \boldsymbol{w}_2')\big),
$$

where ρ_f was defined in (15), the last inequality follows from Lemma 19 (noticing that $f_1 - f_2 \in \mathrm{Lip}_{2L}^0$).

These inequalities together imply

$$
D\big((f_1, f_2, \boldsymbol{w}_1, \boldsymbol{w}_2), (f_1', f_2', \boldsymbol{w}_1', \boldsymbol{w}_2')\big) \leq D_2\big((f_1, f_2), (f_1', f_2')\big) + 2\rho\big((\boldsymbol{w}_1, \boldsymbol{w}_2), (\boldsymbol{w}_1', \boldsymbol{w}_2')\big).
$$

Bearing the above inequality in mind, we now apply Dudley's inequality (Lemma 12) to the sub-Gaussian process $Z_{f_1, f_2}(\boldsymbol{w}_1, \boldsymbol{w}_2)$ over $f_1, f_2 \in \mathcal{N}_\Delta$, $\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)$ endowed with metric D (which is sub-Gaussian with respect to the randomness of ξ) and invoke Lemma 17 to obtain

$$
\mathbb{E}_\xi \sup_{\substack{f_1, f_2 \in \mathrm{Lip}_L \\ \boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)}} |Z_f(\boldsymbol{w}_1, \boldsymbol{w}_2)| \lesssim \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{N}_\Delta \times \mathcal{N}_\Delta, D_2, \varepsilon)} \, \mathrm{d}\varepsilon + \int_0^\infty \sqrt{\log \mathcal{N}(B(R) \times B(R), \rho, \varepsilon)} \, \mathrm{d}\varepsilon
$$

$$
\lesssim \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{N}_\Delta \times \mathcal{N}_\Delta, D_2, \varepsilon)} \, \mathrm{d}\varepsilon + L^2 R^2 \sqrt{d \log^2 d \cdot \log m} \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} ,
$$

where the last line uses (17) and (18). Put this and (24) together, and then take expectation with respect to the randomness of $j_1, \cdots, j_m$, the desired conclusion readily follows.

D.2. Proof of Lemma 22

D.2.1. CONTROLLING D_∞

First note that the term inside the summation of $D_\infty\big((f_1, f_2), (f'_1, f'_2)\big)$ can be written as

$$\left|\big(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2 - \big(f'_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f'_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2\right|$$

$$= \underbrace{\left|f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f'_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle) + f'_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\right|}_{=:T_1}$$

$$\cdot \underbrace{\left|f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle) + f'_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f'_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\right|}_{=:T_2}.$$

The first factor T_1 can be controlled in the following way. Recall that $\|\boldsymbol{w}_k\| \le R$ for $k = 1, 2$, we have $|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_k\rangle| \le R\|\boldsymbol{x}_{j_i}\|$, thus $\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_k\rangle \in I_{j_i}$. Therefore

$$\left|f_k(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_k\rangle) - f'_k(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_k\rangle)\right| \le \|f_k - f'_k\|_{L^\infty(I_{j_i})}, \quad k = 1, 2.$$

therefore

$$T_1 \le \|f_1 - f'_1\|_{L^\infty(I_{j_i})} + \|f_2 - f'_2\|_{L^\infty(I_{j_i})}.$$

The second factor T_2 can be controlled using the Lipschitz assumption, which implies for example $|f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle)| \le |f_1(0)| + L|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle| = L|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle|$ as $f_1 \in \mathrm{Lip}_L$ and similar inequalities for the other terms. Thus

$$T_2 \le 2L|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle| + 2L|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle|.$$

Combining these bounds, we obtain

$$\left|\big(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2 - \big(f'_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f'_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2\right|$$

$$\le \left(\|f_1 - f'_1\|_{L^\infty(I_{j_i})} + \|f_2 - f'_2\|_{L^\infty(I_{j_i})}\right) \cdot 2L\left(|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle|\right), \qquad (25)$$

hence

$$D_\infty\big((f_1, f_2), (f'_1, f'_2)\big)$$

$$\le 4L \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \sum_{i=1}^m \frac{1}{p_{j_i}} \left(|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle|\right) \cdot \left(\|f_1 - f'_1\|_{L^\infty(I_{j_i})} + \|f_2 - f'_2\|_{L^\infty(I_{j_i})}\right)$$

$$\le 8L \sup_{\boldsymbol{w} \in B(R)} \sum_{i=1}^m \frac{1}{p_{j_i}} |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}\rangle| \left(\|f_1 - f'_1\|_{L^\infty(I_{j_i})} + \|f_2 - f'_2\|_{L^\infty(I_{j_i})}\right)$$

$$= 8LR \sup_{\|\boldsymbol{w}\| \le 1} \sum_{i=1}^m \left|\left\langle \frac{\boldsymbol{x}_{j_i}}{\sqrt{p_{j_i}}}, \boldsymbol{w}\right\rangle\right| \cdot \frac{1}{\sqrt{p_{j_i}}} \|f_1 - f'_1\|_{L^\infty(I_{j_i})}$$

$$+ 8LR \sup_{\|\boldsymbol{w}\| \le 1} \sum_{i=1}^m \left|\left\langle \frac{\boldsymbol{x}_{j_i}}{\sqrt{p_{j_i}}}, \boldsymbol{w}\right\rangle\right| \cdot \frac{1}{\sqrt{p_{j_i}}} \|f_2 - f'_2\|_{L^\infty(I_{j_i})}.$$

The conclusion of the lemma then follows from Cauchy-Schwarz with similar procedures as in the proof of Lemma 19.

D.2.2. CONTROLLING D_2

We break the summation in the definition of D_2 into two parts:

$$D_2^2\big((f_1, f_2), (f_1', f_2')\big)$$

$$\lesssim \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \underbrace{\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| \geq \mu} \cdot \frac{1}{p_{j_i}^2} \left| \big(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2 - \big(f_1'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2 \right|^2}_{=:T_1}$$

$$+ \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \underbrace{\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| < \mu} \cdot \frac{1}{p_{j_i}^2} \left| \big(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2 - \big(f_1'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2 \right|^2}_{=:T_2}.$$

Using inequality (25), we obtain

$$T_1 \lesssim L^2 \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| \geq \mu} \frac{|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle|^2 + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle|^2}{p_{j_i}} \cdot \frac{\|f_1 - f_1'\|_{L^\infty(I_{j_i})}^2 + \|f_2 - f_2'\|_{L^\infty(I_{j_i})}^2}{p_{j_i}}$$

$$\lesssim L^2 R^2 \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\| \left(\sup_{i \in [m]} \frac{\big(\|f_1 - f_1'\|_{L^\infty(I_{j_i})}^2 + \|f_2 - f_2'\|_{L^\infty(I_{j_i})}^2\big) \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| \geq \mu}}{p_{j_i}} \right).$$

On the other hand, by Lipschitz property of f_1, we have

$$|f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle)| \leq L\|\boldsymbol{x}_{j_i}\| \cdot \|\boldsymbol{w}_1\| \leq LR\|\boldsymbol{x}_{j_i}\|,$$

since $\boldsymbol{w}_1 \in B(R)$. The same upper bound holds for $|f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)|$, $|f_1'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle)|$, and $|f_2'(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)|$. It is then easy to show

$$T_2 \lesssim \sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| < \mu} \cdot \frac{1}{p_{j_i}^2} L^4 R^4 \|\boldsymbol{x}_{j_i}\|^4 \big(\rho_\delta(f_1, f_1') + \rho_\delta(f_2, f_2')\big)$$

$$= L^4 R^4 d^2 \left(\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| < \mu} \right) \big(\rho_\delta(f_1, f_1') + \rho_\delta(f_2, f_2')\big),$$

where the last equality uses $p_{j_i} = \|\boldsymbol{x}_{j_i}\|^2/d$ again. The conclusion follows immediately from summing up the above inequalities for T_1 and T_2 and then taking square roots.

D.3. Proof of Lemma 24

We divide the proof into four steps. First, we give an explicit construction for a candidate of the net $\mathcal{N}_\Delta$ in Section D.3.1. Up to this point, we do not know whether $\mathcal{N}_\Delta$ is a net, but we can already bound the cardinality of it, thereby proving item (i) of Lemma 24. Then in Section D.3.2, we prove that $\mathcal{N}_\Delta$ is indeed a $\Delta(\mu, \eta)$ net in ρ_∞ distance. On the other hand, proving the bound of Dudley integral in item (ii) turns out the trickiest part of the proof. In Section D.3.3, we will prove this item by presenting a technique to embed $\mathcal{N}_\Delta$ to another space of Lipschitz functions (endowed with a different metric than ρ_∞), which enables to draw connections to well-known bounds on the entropy of class of Lipschitz functions. Finally, we prove item (iii) in Section D.3.4, which follows from Cauchy-Schwarz inequality.

D.3.1. CONSTRUCTION OF $\mathcal{N}_\Delta$.

For convenience, set $\eta' = \eta/4$ and

$$K = \left\lceil \frac{\log(1/\mu)}{\log(1+\eta')} \right\rceil, \qquad N = \left\lceil \frac{1}{\eta'} \right\rceil,$$

$$z_k = \begin{cases} R(1+\eta')^{-(K+N-k)}, & k = N, N+1, \cdots, N+K, \\ \frac{k}{N}z_N, & k = 0, 1, \cdots, N-1. \end{cases}$$

The construction of z_k guarantees that z_k is monotone increasing and

$$z_{k+1} = (1+\eta')z_k, \quad N \le k \le N+K-1. \tag{26}$$

and

$$\frac{1}{2}\mu R < \frac{1}{1+\eta'}\mu R \le z_N \le \mu R. \tag{27}$$

The basic idea is to construct at each point $\pm z_k$ a grid with spacing $L\eta' z_k$ for $[-Lz_k, Lz_k]$, and connect these grid points by piecewise linear functions. However, for small z_k where $k \le N$ this would be impossible without significantly enlarging the size of the net. To remedy this, for such k the spacing of the grid will be $L\eta z_N$. The precise construction is as follows. Set

$$\Sigma_0 = \left\{ \sigma = (\sigma_{-N-K}, \sigma_{-N-K+1}, \ldots, \sigma_{N+K}) : \sigma_0 = 0, \ \sigma_{\pm k} \in \mathbb{Z}, \ k \in [N+K] \right\}.$$

To each $\sigma = (\sigma_{-N-K}, \sigma_{-N-K+1}, \ldots, \sigma_{N+K}) \in \Sigma_0$, we associate a piecewise linear function defined by

$$f_\sigma(\pm z_k) = L\eta'\sigma_{\pm k}\max(z_k, z_N), \quad 0 \le k \le N+K. \tag{28}$$

The value of f_σ at a point which is not equal to any one of $\pm z_k$ is given by linear interpolation. The whole collection of $\{f_\sigma : \sigma \in \Sigma_0\}$ would be too large to work with. Fortunately, since our net is for L-Lipschitz functions rather than arbitrary functions, we can restrict our attention to a much smaller subset $\Sigma \subset \Sigma_0$. Denote

$$\Sigma = \left\{ \sigma = (\sigma_{-N-K}, \cdots, \sigma_{N+K}) \in \Sigma_0 : \sigma_0 = 0, \ |\sigma_{\pm(k+1)} - \sigma_{\pm k}| \le 2, 0 \le k \le N+K-1 \right\}. \tag{29}$$

We contend that

$$\mathcal{N}_\Delta = \{f_\sigma : \sigma \in \Sigma\}. \tag{30}$$

is a $\Delta(\mu, \eta)$-net for Lip_L w.r.t. metric ρ_∞ (note that f_σ is only $O(L)$-Lipschitz, thus $\mathcal{N}_\Delta$ may not be a subset of Lip_L, but this is of no consequence to our proof). Proving this is the goal of the next step.

We conclude this step of the proof by establishing the item (i) of Lemma 22. Suppose $\sigma \in \Sigma$. For any fixed $\sigma_{\pm k}$, there are at most five possible values of $\sigma_{\pm(k+1)}$ due to the constraint $|\sigma_{\pm(k+1)} - \sigma_k| \le 2$. But σ_0 is already fixed, thus the number of elements in Σ cannot exceed $5^{2(N+K)}$. Since $\mathcal{N}_\Delta$ is an image of Σ, we have

$$\log|\mathcal{N}_\Delta| \le \log|\Sigma| \lesssim N+K \lesssim \frac{\log(1/\mu)}{\log(1+\eta')} + \frac{1}{\eta'} \lesssim \frac{\log(1/\mu)}{\eta'} = \frac{4\log(1/\mu)}{\eta},$$

where we used $\eta' = \eta/4$ which implies $\eta' < 1/8$ and hence $\log(1+\eta') \ge \eta'/4$. This proves the item (i) of Lemma 22 as desired.

D.3.2. VERIFYING $\mathcal{N}_\Delta$ IS A $\Delta(\mu, \eta)$-NET.

For each $f \in \text{Lip}_L$, we can define $\sigma(f) \in \Sigma_0$ as

$$\sigma(f)_{\pm k} = \begin{cases} \left\lceil \frac{f(\pm z_k)}{L\eta' \max(z_k, z_N)} \right\rceil, & k \in [N + K], \\ 0, & k = 0. \end{cases}$$

By definition, one has

$$\frac{f(\pm z_k)}{L\eta' \max(z_k, z_N)} \leq \sigma(f)_{\pm k} < \frac{f(\pm z_k)}{L\eta' \max(z_k, z_N)} + 1, \quad k \in [N + K]. \tag{31}$$

It suffices to verify that $\sigma(f) \in \Sigma$ and $\rho_\infty(f, f_{\sigma(f)}) \leq \Delta(\mu, \eta)$.

Verifying $\sigma(f) \in \Sigma$. We need to show $|\sigma(f)_{\pm k}| \leq N$ and $|\sigma(f)_{\pm(k+1)} - \sigma(f)_{\pm k}| \leq 2$. The former follows easily from the definition of $\sigma(f)$ and the fact that $|f(\pm z_k)| \leq Lz_k$ (which in turn follows from $f \in \text{Lip}_L$). For the former, we distinguish three cases: (i) $k = 0$, (ii) $1 \leq k \leq N - 1$, (iii) $N \leq k \leq N + K$.

For case (i), (31) implies

$$|\sigma(f)_{\pm 1}| \leq \frac{|f(\pm z_1)|}{L\eta' z_N} + 1 \leq \frac{Lz_1}{L\eta' z_N} + 1 = \frac{1}{N\eta'} + 1 \leq 2,$$

where the equality follows from $z_1 = \frac{1}{N} z_N$, and the last inequality follows from $N = \lceil 1/\eta' \rceil \geq 1/\eta'$. Since $\sigma(f)_0 = 0$, this implies $|\sigma(f)_{\pm 1} - \sigma(f)_0| \leq 2$ as claimed.

For case (ii), invoke (31) again to obtain

$$\begin{aligned} |\sigma(f)_{\pm(k+1)} - \sigma(f)_{\pm k}| &\leq \left| \frac{f(\pm z_{k+1})}{L\eta' z_N} - \frac{f(\pm z_k)}{L\eta' z_N} \right| + 1 \\ &= \frac{|f(\pm z_{k+1}) - f(\pm z_k)|}{L\eta' z_N} + 1 \\ &\leq \frac{L(z_{k+1} - z_k)}{L\eta' z_N} + 1 \\ &= \frac{1}{N\eta'} + 1 \leq 2, \end{aligned}$$

where the third line follows from f being L-Lipschitz and the last line follows from $z_{k+1} = \frac{k+1}{N} z_N$, $z_k = \frac{k}{N} z_N$.

For case (iii), (31) implies (31) implies

$$\begin{aligned} |\sigma(f)_{\pm(k+1)} - \sigma(f)_{\pm k}| &\leq \left| \frac{f(\pm z_{k+1})}{L\eta' \max(z_{k+1}, z_N)} - \frac{f(\pm z_k)}{L\eta' \max(z_k, z_N)} \right| + 1 \\ &= \left| \frac{f(\pm z_{k+1})}{L\eta'(1 + \eta')z_k} - \frac{(1 + \eta')f(\pm z_k)}{L\eta'(1 + \eta')z_k} \right| + 1 \\ &\leq \frac{|f(\pm z_{k+1}) - f(\pm z_k)|}{L\eta'(1 + \eta')z_k} + \frac{|f(\pm z_k)|}{L(1 + \eta')z_k} + 1 \\ &\leq \frac{L(z_{k+1} - z_k)}{L\eta'(1 + \eta')z_k} + \frac{Lz_k}{L(1 + \eta')z_k} + 1 \\ &= \frac{2}{1 + \eta'} + 1 < 3, \end{aligned}$$

where we used $z_{k+1} = (1 + \eta')z_k$ by (26) and the fourth line used Lipschitz continuity of f and $f(0) = 0$. This implies $|\sigma(f)_{\pm(k+1)} - \sigma(f)_{\pm k}| \leq 2$ since $\sigma(f)_k$'s are integers. In conclusion, $\sigma(f) \in \Sigma$, as desired.

Verifying $\rho_\infty(f, f_{\sigma(f)}) \leq \Delta(\mu, \eta)$. By Lemma 22, we need to bound $\|f - f_{\sigma(f)}\|_{L^\infty(I_j)}$. We begin by noting that for the maximum of the absolute value of an affine-linear function in an interval can only be attained at the endpoints of the interval. Since $f_{\sigma(f)}$ is by definition affine-linear on the interval $[z_k, z_{k+1}]$, we have

$$\sup_{z \in [z_k, z_{k+1}]} |f(z) - f_{\sigma(f)}(z)| \leq \sup_{z \in [z_k, z_{k+1}]} \max\left(|f(z) - f_{\sigma(f)}(z_k)|, |f(z) - f_{\sigma(f)}(z_{k+1})|\right).$$

From (31) and the construction of $f_{\sigma(f)}$, we have

$$|f(z_k) - f_{\sigma(f)}(z_k)| \leq L\eta' \max(z_k, z_N), \quad 0 \leq k \leq N + K.$$

On the other hand, Lipschitz continuity of f implies,

$$\sup_{z \in [z_k, z_{k+1}]} \max\left(|f(z) - f(z_k)|, |f(z) - f(z_{k+1})|\right) \leq L(z_{k+1} - z_k).$$

Combining the above three inequalities yields

$$\sup_{z \in [z_k, z_{k+1}]} |f(z) - f_{\sigma(f)}(z)| \leq L\eta' \max(z_k, z_N) + L(z_{k+1} - z_k).$$

By an argument verbatim to the above, we also have

$$\sup_{z \in [-z_{k+1}, -z_k]} |f(z) - f_{\sigma(f)}(z)| \leq L\eta' \max(z_k, z_N) + L(z_{k+1} - z_k).$$

The above two inequalities hold for all $0 \leq k \leq N + K$, therefore for any $k \in [N + K]$, we have

$$\begin{aligned}
\|f - f_{\sigma(f)}\|_{L^\infty([-z_k, z_k])} &= \max_{l \in [k-1]} \sup_{z \in [-z_{l+1}, -z_l] \cup [z_l, z_{l+1}]} |f(z) - f_{\sigma(f)}(z)| \\
&\leq \max_{l \in [k-1]} \left(L\eta' \max(z_l, z_N) + L(z_{l+1} - z_l)\right) \\
&= \begin{cases} 2L\eta' z_k, & N + 1 \leq k \leq N + K, \\ L\left(\eta' + \frac{1}{N}\right) z_N, & 1 \leq k \leq N, \end{cases}
\end{aligned} \tag{32}$$

where the last equality follows from straightforward computations by the construction of z_k's.

We are now ready to bound $\rho_\infty(f, f_\sigma(f))$. Recall Lemma 22 and the notation $I_j = [-R\|\boldsymbol{x}_j\|, R\|\boldsymbol{x}_j\|]$ there. Since we have assumed without loss of generality that $\boldsymbol{X}$ has orthonormal columns, $\boldsymbol{x}_j$ being one of its row has norm no more than 1. Therefore $R\|\boldsymbol{x}_j\| \leq R = z_{N+K}$. In virtue of (26), for any $j \in [n]$ such that $\|\boldsymbol{x}_j\| \geq z_N/R$, there is some $k_j \geq N$ such that $\frac{1}{2}z_{k_j} \leq R\|\boldsymbol{x}_j\| \leq z_{k_j}$, hence

$I_j \subset [-z_{k_j}, z_{k_j}]$. Invoking (32), we obtain, for any $j \in [n]$ such that $\|\boldsymbol{x}_j\| \geq z_N/R$, that

$$
\begin{aligned}
\frac{1}{p_j}\|f - f_{\sigma(f)}\|_{L^\infty(I_j)}^2 &\leq \frac{1}{p_j}\|f - f_{\sigma(f)}\|_{L^\infty([-z_{k_j}, z_{k_j}])}^2 \\
&\leq \frac{1}{p_j}(2L\eta' z_{k_j})^2 \\
&\leq \frac{1}{p_j}(2L\eta' \cdot 2R\|\boldsymbol{x}_j\|)^2 \\
&= \frac{d}{\|\boldsymbol{x}_j\|^2}\cdot \eta^2 L^2 R^2\|\boldsymbol{x}_j\|^2 = \eta^2 L^2 R^2 d,
\end{aligned} \tag{33}
$$

where the penultimate equality follows from $\eta' = \eta/4$ and $p_j = \|\boldsymbol{x}_j\|^2/d$. On the other hand, for $j \in [n]$ such that $\|x_j\| < z_N/R$, one may invoke (32) again to obtain

$$
\begin{aligned}
\frac{1}{p_j}\|f - f_{\sigma(f)}\|_{L^\infty(I_j)}^2 &\leq \frac{1}{p_j}\|f - f_{\sigma(f)}\|_{L^\infty([-z_N, z_N])}^2 \\
&\leq \frac{1}{p_j}\left(L\left(\eta' + \frac{1}{N}\right)z_N\right)^2 \\
&\leq \frac{1}{p_j}(2L\eta'\mu R)^2 \\
&= \frac{1}{4}\eta^2 L^2 R^2 \cdot \frac{\mu^2}{p_j},
\end{aligned}
$$

where the penultimate line follows from $N = \lceil 1/\eta' \rceil \geq 1/\eta'$ and from (27), and the last line used $\eta' = \eta/4$. Summing up the above two inequalities, we have, for any $j \in [n]$, regardless of how large $\|\boldsymbol{x}_j\|$ is, that

$$
\frac{1}{p_j}\|f - f_{\sigma(f)}\|_{L^\infty(I_j)}^2 \leq \eta^2 L^2 R^2 d + \frac{1}{4}\eta^2 L^2 R^2 \cdot \frac{\mu^2}{p_j}.
$$

Now we apply Lemma 22 to obtain

$$
\begin{aligned}
\rho_\infty(f, f_{\sigma(f)}) &\leq LR\left\|\sum_{i=1}^m \frac{1}{p_{j_i}}\boldsymbol{x}_{j_i}\boldsymbol{x}_{j_i}^\top\right\|^{1/2}\left(\sum_{i=1}^m \frac{1}{p_{j_i}}\|f - f_{\sigma(f)}\|_{L^\infty(I_{j_i})}^2\right)^{1/2} \\
&\leq LR\left\|\sum_{i=1}^m \frac{1}{p_{j_i}}\boldsymbol{x}_{j_i}\boldsymbol{x}_{j_i}^\top\right\|^{1/2}\cdot \eta LR\left(\sum_{i=1}^m\left(d + \frac{\mu^2}{4p_{j_i}}\right)\right)^{1/2} \\
&\leq \eta L^2 R^2\left\|\sum_{i=1}^m \frac{1}{p_{j_i}}\boldsymbol{x}_{j_i}\boldsymbol{x}_{j_i}^\top\right\|^{1/2}\left(md + \sum_{i=1}^m \frac{\mu^2}{p_{j_i}}\right)^{1/2} \\
&= \Delta(\mu, \eta),
\end{aligned}
$$

as claimed.

D.3.3. BOUNDING THE DUDLEY INTEGRAL OF $\mathcal{N}_\Delta$

Recall that we need to bound $\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{N}_\Delta, \rho_2, \varepsilon)} \, d\varepsilon$, which amounts to bounding the entropy of the metric space $(\mathcal{N}_\Delta, \rho_2)$. This will be achieved by embedding this metric space into $(\mathrm{Lip}_1, L^\infty([-1, 1]))$, which is again a space of Lipschitz functions yet with metric $L^\infty([-1, 1])$ instead of ρ_∞. The entropy of this latter space has a well-known bound. The embedding is achieved by the following two lemmas.

Lemma 25 *With the notations $f_\sigma, \sigma \in \Sigma$ as in Section D.3.1, we have, for all $\sigma, \sigma' \in \Sigma$, that*

$$\rho_2(f_\sigma, f_{\sigma'}) \lesssim \eta L^2 R^2 \sqrt{d} \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \|\sigma - \sigma'\|_\infty + L^2 R^2 d \left(\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| < \mu} \right)^{1/2} \rho_\delta(\sigma, \sigma').$$

Here $\|\cdot\|_\infty$ denotes the ℓ^∞-norm of vectors, and ρ_δ is, by abuse of notation, the Dirac metric on Σ, defined as

$$\rho_\delta(\sigma, \sigma') = \begin{cases} 1, & \sigma = \sigma', \\ 0, & \sigma \neq \sigma'. \end{cases}$$

Proof By the construction of $f_\sigma, f_{\sigma'}$ in (28), we can verify in a similar way as we proved (33) that

$$\sup_{i \in [m]} \frac{\|f_\sigma - f_{\sigma'}\|_{L^\infty(I_{j_i})} \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| \geq \mu}}{\sqrt{p_{j_i}}} \lesssim \eta L R \sqrt{d} \|\sigma - \sigma'\|_\infty.$$

The conclusion then follows from the definition of ρ_2. $\blacksquare$

Lemma 26 *For $\sigma \in \Sigma$, define a piecewise linear function g_σ on $[-1, 1]$ by*

$$g_\sigma \left(\frac{\pm k}{N + K} \right) = \frac{\sigma_{\pm k}}{2(N + K)}.$$

At a point other than $\frac{\pm k}{N+K}$, the value of g_σ is determined by linear interpolation. We have $g(0) = 0$, g is 1-Lipschitz, and

$$\|\sigma - \sigma'\|_\infty = 2(N + K) \|g_\sigma - g_{\sigma'}\|_{L^\infty([-1,1])}, \quad \sigma, \sigma' \in \Sigma.$$

Proof The equality follows easily from the fact that the supremum of the absolute value of an affine-linear function in an interval can only be attained at the endpoints of the interval. In particular, for any $k \in \mathbb{Z}$, $-(N + K) \leq k \leq N + K - 1$, we have

$$\sup_{z \in [\frac{k}{N+K}, \frac{k+1}{N+K}]} |g_\sigma(z) - g_{\sigma'}(z)| = \frac{1}{2(N + K)} \max \left(|\sigma_{k+1} - \sigma'_{k+1}|, |\sigma_k - \sigma'_k| \right).$$

Taking union over all such k completes the proof for the equality. It remains to check g is 1-Lipschitz. As g is piecewise linear, it suffices to check the slope on each of the defining interval lies in $[-1, 1]$, i.e.,

$$(N + K) \left| g_\sigma \left(\frac{\pm(k+1)}{N + K} \right) - g_\sigma \left(\frac{\pm k}{N + K} \right) \right| \leq 1,$$

38

which, by definition of g_σ, is equivalent to

$$|\sigma_{\pm(k+1)} - \sigma_{\pm k}| \le 2.$$

But this is certainly true as $\sigma \in \Sigma$. This completes the proof. ∎

Using the embeddings given by Lemma 25 and Lemma 26, we can reduce the goal of bounding the entropy of $(\mathcal{N}_\Delta, \rho_2)$ to bounding the entropies of $(\mathrm{Lip}_1, L^\infty([-1,1]))$ and of (Σ, ρ_δ).

Corollary 27 *Assume $m \ge Cd \log d$. We have*

$$\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{N}_\Delta, \rho_2, \varepsilon)}\, d\varepsilon \lesssim L^2 R^2 \sqrt{d} \log\left(\frac{1}{\mu}\right) \cdot \left\|\sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \int_0^\infty \sqrt{\log \mathcal{N}(\mathrm{Lip}_1, L^\infty([-1,1]), \varepsilon)}\, d\varepsilon$$

$$+ L^2 R^2 d \left(\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| < \mu}\right)^{1/2} \int_0^\infty \sqrt{\log \mathcal{N}(\Sigma, \rho_\delta, \varepsilon)}\, d\varepsilon.$$

Proof Combining Lemma 25 and Lemma 26, we have

$$\rho_2(f_\sigma, f_{\sigma'}) \lesssim \eta(N+K) L^2 R^2 \sqrt{d} \left\|\sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \|g_\sigma - g_{\sigma'}\|_{L^\infty([-1,1])} + L^2 R^2 d \left(\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| < \mu}\right)^{1/2} \rho_\delta(\sigma, \sigma')$$

$$\lesssim L^2 R^2 \sqrt{d} \log\left(\frac{1}{\mu}\right) \cdot \left\|\sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \|g_\sigma - g_{\sigma'}\|_{L^\infty([-1,1])} + L^2 R^2 d \left(\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\| < \mu}\right)^{1/2} \rho_\delta(\sigma, \sigma'),$$

where we used $N \lesssim 1/\eta$ and $K \lesssim \frac{1}{\eta} \log \frac{1}{\mu}$. If we regard all three metrics $\rho_2(f_\sigma, f_{\sigma'})$, $\|g_\sigma - g_{\sigma'}\|_{L^\infty([-1,1])}$, $\rho_\delta(\sigma, \sigma')$ as metrics on Σ, the desired conclusion readily follows from Lemma 16. ∎

It remains to bound the entropies of $(\mathrm{Lip}_1, L^\infty([-1,1]))$ and of (Σ, ρ_δ). The former is well-known, presented in the next lemma, while the latter can be computed easily.

Lemma 28 (Entropy bound of Lipschitz function class, Talagrand (2021)) *We have*

$$\log \mathcal{N}(\mathrm{Lip}_1, L^\infty([-1,1]), \varepsilon) \lesssim \frac{1}{\varepsilon}, \quad \varepsilon > 0.$$

We have now collected all the ingredients to prove the item (ii) of Lemma 24.
Proof [Proof of the item (ii) of Lemma 24] Since the diameter of Lip_1 with respect to $L^\infty([-1,1])$ is 1, we deduce immediately from the Lemma 28 that

$$\int_0^\infty \sqrt{\log \mathcal{N}(\mathrm{Lip}_1, L^\infty([-1,1]), \varepsilon)}\, d\varepsilon = \int_0^1 \sqrt{\log \mathcal{N}(\mathrm{Lip}_1, L^\infty([-1,1]), \varepsilon)}\, d\varepsilon \lesssim \int_0^1 \sqrt{\frac{1}{\varepsilon}}\, d\varepsilon \lesssim 1.$$

Turning to bound the entropy of (Σ, ρ_δ), we simply note that $\rho_\delta \le 1$, hence

$$\int_0^\infty \sqrt{\log \mathcal{N}(\Sigma, \rho_\delta, \varepsilon)}\, d\varepsilon = \int_0^1 \sqrt{\log \mathcal{N}(\Sigma, \rho_\delta, \varepsilon)}\, d\varepsilon \lesssim \int_0^1 \sqrt{\log |\Sigma|}\, d\varepsilon \lesssim \sqrt{\frac{\log(1/\mu)}{\eta}},$$

where the last inequality used the conclusion of item (i) in Lemma 24, proved in Section D.3.1.

Plug the above two inequalities into Corollary 27 to obtain

$$\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{N}_\Delta, \rho_2, \varepsilon)}\, d\varepsilon \lesssim L^2 R^2 \sqrt{d} \log\left(\frac{1}{\mu}\right) \cdot \left\|\sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top\right\|^{1/2}$$
$$+ L^2 R^2 d \left(\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\|<\mu}\right)^{1/2} \sqrt{\frac{\log(1/\mu)}{\eta}}.$$

The desired conclusion follows taking expectations on both sides, using (20) and Cauchy-Schwarz to bound the expectation of the first term on the right hand side, and using the following fact to bound the expectation of the second term on the right hand side:

$$\mathbb{E}\left(\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\|<\mu}\right)^{1/2} \leq \left(\mathbb{E}\sum_{i=1}^m \mathbf{1}_{\|\boldsymbol{x}_{j_i}\|<\mu}\right)^{1/2} = \left(m\sum_{j=1}^n p_j \mathbf{1}_{\|\boldsymbol{x}_j\|<\mu}\right)^{1/2}$$
$$= \left(m\sum_{j=1}^n \frac{\|\boldsymbol{x}_j\|^2}{d} \mathbf{1}_{\|\boldsymbol{x}_j\|<\mu}\right)^{1/2}$$
$$\leq \left(m\sum_{j=1}^n \frac{\mu^2}{d}\right)^{1/2}$$
$$= \mu\sqrt{\frac{nm}{d}}.$$

■

D.3.4. BOUNDING $\mathbb{E}\Delta$.

Recall the Chernoff bound (20), we have by Cauchy-Schwarz inequality that

$$\mathbb{E}\Delta \lesssim \eta L^2 R^2 \sqrt{m} \left(md + \mathbb{E}\sum_{i=1}^m \frac{\mu^2}{p_{j_i}}\right)^{1/2}.$$

Note that

$$\mathbb{E}\frac{1}{p_{j_i}} = \sum_{j\in[n]} p_j \frac{1}{p_j} = n,$$

thus

$$\mathbb{E}\Delta \lesssim \eta L^2 R^2 \sqrt{m} \left(md + m\mu^2 n\right)^{1/2} \lesssim L^2 R^2 m\sqrt{d} \cdot \sqrt{1 + \frac{\mu^2 n}{d}},$$

as claimed.