# Improved Bounds for Agnostic Active Learning of Single Index Models

**Aarshvi Gajjar** * AARSHVI@NYU.EDU
**Xingyu Xu** * XINGYUXU@ANDREW.CMU.EDU
**Chinmay Hegde** CHINMAY.H@NYU.EDU
**Christopher Musco** CMUSCO@NYU.EDU

## Abstract

We study active learning 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 $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^d$. Such functions are important in scientific computing, where they are used to construct surrogate models for partial differential equations (PDEs) and to approximate high-dimensional Quantities of Interest. In these applications, collecting function samples requires solving a partial differential equation, so sample-efficient active learning methods translate to reduced computational cost. Our work provides two main results. First, when $f$ is known and Lipschitz, we show that $\tilde{O}(d)$ samples collected via *statistical leverage score sampling* are sufficient to find an optimal single index model for a given target function, even in the challenging and practically important agnostic (adversarial noise) setting. This result is optimal up to logarithmic factors and improves quadratically on a recent $\tilde{O}(d^2)$ bound of Gajjar et al. (2023). Second, we show that $\tilde{O}(d^{3/2})$ samples suffice in the more difficult non-parametric setting when $f$ is *unknown*, which is the also best result known in this general setting.

**Keywords:** active learning, leverage scores, single index models, scientific computing

## 1. Introduction

Single Index Models (SIMs) 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). Moreover, they serve as foundational elements within neural networks (Kakade et al., 2011; Mei et al., 2018; Abbe et al., 2022). A single index model, $F : \mathbb{R}^d \to \mathbb{R}$ can be written as $F(\boldsymbol{x}) = f(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$, with a univariate *link* function $f : \mathbb{R} \to \mathbb{R}$ and a weight vector $\boldsymbol{w} \in \mathbb{R}^d$. In the general form, both $\boldsymbol{w}$ and $f$ are unknown, presenting a challenging non-parametric estimation problem. We aim to address this problem using minimal samples within the active learning setting, where we can query the values of $F$ at selected points from a predefined set.

Active learning is crucial in many scientific computing applications, due to repeated evaluation of computationally expensive functions. For instance, in parametric PDEs, this demand stems from solving individual PDEs associated with a large number of parameters. For these problems, several recent studies, including Geist et al. (2021), Bhattacharya et al. (2021), and Kutyniok et al. (2022), have focused on approximating parameter-to-solution

∗. Equal Contribution

maps using neural networks. Similarly, there are studies on approximating specific quantities of interest within solution maps in works such as Tripathy and Bilionis (2018), Khoo et al. (2021), Zhang et al. (2019), O'Leary-Roseberry et al. (2022). We study single index models as a simplified representation of neural networks and present a provably sample-efficient active learning method for them.

The problem of learning single index models has been extensively studied without the active learning setting, (Bietti et al., 2022; Abbe et al., 2022; Dudeja and Hsu, 2018; Damian et al., 2022). These studies focus on analysing computationally feasible algorithms like gradient descent and establishing sample complexity bounds. Since their tasks are intertwined with studying these algorithms, they also rely on stronger assumptions that our setting, such as Gaussian data or smoothness assumptions on the link function, in addition to Lipschitzness. Our work relies on no distributional assumptions, once the large data matrix has been obtained.

In the active learning setting, alternative learning algorithms have been proposed, as seen in (Cohen et al., 2011; Fornasier et al., 2012; Tyagi and Cevher, 2012). However, it's important to note that these algorithms are designed for the noiseless (realizable) setting and are not robust to noise.

Our work is a continuation of this line of research on active learning algorithms where we focus on achieving the best sample complexity from a statistical perspective, decoupled from the computational perspective. This enables us to achieve better sample complexity with fewer assumptions, which we will provide detailed explanations in the next section.

### 1.1 Contribution

We improve upon the previous works by presenting a sampling-efficient method for learning SIMs. Our method has the advantages of 1) no assumptions of smoothness on $f$ beyond Lipschitz continuity, 2) agnostic capability, i.e. robustness against any adversarial noise. 3) optimal or state-of-the-art sample complexity across different settings.

Formally, the problem we aim at solving is as follows.

**Problem 1.** *Given a data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$, query access to the target vector $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, and a function class $\mathcal{F}$, the goal is to determine the minimum number of queries needed to approximately solve the least squares objective $\sum_{j=1}^n \left|f(\langle \boldsymbol{w}, \boldsymbol{x}_j\rangle) - y_j\right|^2$ across all possible weight vectors and functions within $\mathcal{F}$.*

Our sampling technique is *leverage score* sampling. This method selects data points non-uniformly, with probability proportional to their individual statistical leverage scores. Intuitively, this method favors rows that are more influential in forming the column space of the data matrix, $\boldsymbol{X}$.

**Definition 2** (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 $\tau_j(\boldsymbol{X}) \coloneqq \boldsymbol{x}_j^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_j = \sup_{\boldsymbol{w}\in\mathbb{R}^d} \frac{\langle \boldsymbol{w}, \boldsymbol{x}_j\rangle^2}{\|\boldsymbol{X}\boldsymbol{w}\|_2^2}$.*

This is a well studied sampling strategy, widely applied in active linear regression (Chen and Price (2019); Mahoney et al. (2011), etc.). Extending the application of this method to the setting of Problem 1, we show:

- A sample complexity of $\tilde{O}(d)$, when $f$ is known. This is optimal, up to logarithmic factors.
- A sample complexity of $\tilde{O}(d^{3/2})$, when $f$ remains unknown, which is the best result known in such generality.

The most closely related result in the literature to our approach is a sample complexity of $\tilde{O}(d^2)$, assuming $f$ is known, as recently shown in Gajjar et al. (2023). Our sample complexity of $\tilde{O}(d)$ is a quadratic improvement over this result. Moreover, to the best of our knowledge, the scenario where $f$ is unknown hasn't been studied at this level of generality Previous works (Damian et al., 2022; Bietti et al., 2022) indicate a sample complexity of $O(d^2)$ even with additional assumptions on $f$.

**Organization**. In Section 2, we present notations used in this paper. Following this, in Section 3, we provide a mathematical formulation of our problem, including the underlying assumptions and our targeted accuracy goal. In Section 3.3, we elaborate on the statistical leverage score sampling process, along with well-established properties of leverage score sampling. Lastly, in Section 4, we formally present our results and accompanying proof techniques. A sketch of proof and explanations can be found in the appendix.

## 2. Notation

For a natural number $n$, we let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For a vector $\boldsymbol{x}$ in $\mathbb{R}^d$, its $\ell_2$ norm is represented as $\|\boldsymbol{x}\|$. We use $\mathrm{Lip}_L$ to represent the class of $L$-Lipschitz functions on $\mathbb{R}$, more precisely, $\mathrm{Lip}_L = \{f \in \mathcal{C}(\mathbb{R}) : \left|f(x_1) - f(x_2)\right| \leq L|x_1 - x_2|, \forall x_1, x_2 \in \mathbb{R}\}$.

We extend the notation of $f(\cdot)$ to $d$ dimensional vectors: for $\boldsymbol{x} \in \mathbb{R}^d$, denote $f(\boldsymbol{x}) \in \mathbb{R}^d$ as the entrywise application of $f$ to $\boldsymbol{x}$, i.e. $f(\boldsymbol{x}) = (f(x^1), f(x^2), \ldots, f(x^d))$, where $x^k, k \in [d]$, is the $k$th element of $\boldsymbol{x}$. 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 $B(R)$. The notation $\tilde{O}$ is the big-$O$ hiding all logarithmic factors in $n$ and $d$. Moreover $a \lesssim b$ means that there exists a positive constant $C > 0$ such that $a \leq Cb$. Throughout the paper, $c$ and $C$ will denote positive universal constants that may vary upon each occurrence.

## 3. Preliminaries

We want to find a single index model that best fits a given set of data points, $(\boldsymbol{x}_j, y_j)$ for $j \in [n]$. Least squares regression is a common approach to achieve this, where the objective is to minimise the $\ell_2$ loss, over all link functions $f \in \mathcal{F}$ and weight vectors $\boldsymbol{w} \in \mathbb{R}^d$.

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

### 3.1 Assumption on the function class

Our assumption on $\mathcal{F}$ is that it is a subset of the class of $L$-Lipschitz functions, which we denote by $\mathrm{Lip}_L$. In particular, we will study two representative cases: one where the link

function $f$ is known *a priori*, i.e., $\mathcal{F} = \{f^\star\}$ for some $f^\star \in \mathrm{Lip}_L$, and another where we have no prior knowledge of $f$, i.e., $\mathcal{F} = \mathrm{Lip}_L$.

In related literature, alternative function classes, such as low degree polynomials, piecewise linear functions, or Sobolev spaces are also considered based on specific problem characteristics. We choose $\mathrm{Lip}_L$, due to its simplicity and ability to express a wide range of functions: the property of Lipschitz continuity is maintained in many practical scenarios[1].

### 3.2 Accuracy goal

Suppose that $(f^\star, \boldsymbol{w}^\star)$ minimises[2] the loss function $\mathcal{L}$ over $f \in \mathcal{F}, \boldsymbol{w} \in \mathbb{R}^d$, then we denote the optimal loss as $\mathsf{OPT}(\mathcal{F})$, defined as

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

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

**Definition 3** ($\varepsilon$-accurate solution)**.** *Fix some sufficiently large constant $C > 0$. Given some $\varepsilon > 0$, a pair $(f, \boldsymbol{w})$ with $f \in \mathcal{F}$, $\boldsymbol{w} \in \mathbb{R}^d$ is said to be an $\varepsilon$-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 appeared in the studies of leverage score sampling in other contexts (Avron et al., 2019).

### 3.3 Subsampled regression

Our sampling method is a sample-with-replacement variant of leverage score sampling. As discussed above, we assign a probability to each data point based on it's statistical leverage score, allowing us to address the regression problem using a small set of samples. This process is precisely outlined below.

**Sampling process.** For every $j \in [n]$, we assign a probability $p_j = \frac{\tau_j(\boldsymbol{X})}{\sum_{j'=1}^n \tau_{j'}(\boldsymbol{X})}$. This establishes a probability distribution, denoted as $p$ over the set $[n]$, where each index $j$ is selected with probability $p_j$. We then generate $m$ i.i.d. random indices $j_1, \ldots, j_m \sim p$, representing random variables taking values in $[n]$.

**Subsampled least-squares.** Similar to Gajjar et al. (2023), we define a sampling-and-reweighting matrix $\boldsymbol{S}$ to succinctly represent our subsampled regression problem.

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 $\frac{1}{\sqrt{m p_{j_i}}} \boldsymbol{e}_{j_i}$. The loss for the subsampled problem can be expressed as

1. For QoI estimation in PDE, $f$ is usually assumed sufficiently regular to ensure accurate estimation. In single neuron model, the most popular choices of $f$, including sigmoid function, ReLU and its variants, are all Lipschitz continuous.
2. We assume the existence of a minimiser for simplicity. Our results still hold even if no minimiser exists.

the weighted average of the individual losses. We denote $\hat{\mathcal{L}}$ as the loss of the subsampled regression problem and define it in terms of the sampling-and-reweighting matrix as follows.

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

Here $1/p_{j_i}$ can be viewed as the weight assigned to the sample $(\boldsymbol{x}_{j_i}, y_{j_i})$ in the subsampled regression problem.

One can verify that $\mathbb{E}[\boldsymbol{S}^\top \boldsymbol{S}] = \boldsymbol{I}$ and hence $\mathbb{E}\hat{\mathcal{L}} = \mathcal{L}$, justifying that $\hat{\mathcal{L}}$ is an unbiased estimator of $\mathcal{L}$. Thus it is reasonable to use $\hat{\mathcal{L}}$ as a surrogate of $\mathcal{L}$ and try to minimise $\hat{\mathcal{L}}$. After laying out the subsampled process, the main technical challenge is to quantify how large $m$ should be to ensure that minimising the subsampled loss $\hat{\mathcal{L}}$ results in an $\varepsilon$-accurate solution to (1).

## 4. Main results

As aforementioned, we will show that minimising the subsampled objective (2) will lead to an $\varepsilon$-approximate solution to (1). Moreover, our result will show that there is no need to minimise over all $\boldsymbol{w} \in \mathbb{R}^d$: we can always find an $\varepsilon$-approximate solution in a smaller region $\mathcal{R}$, originally introduced by Gajjar et al. (2023), defined as

$$\mathcal{R} \coloneqq \left\{ \boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{S}\boldsymbol{X}\boldsymbol{w}\|^2 \leq \frac{1}{\varepsilon L^2} \|\boldsymbol{S}\boldsymbol{y}\|^2 \right\}. \tag{3}$$

### 4.1 Warm up: The case of fixed $f$

Assume the link function is fixed, i.e., $\mathcal{F} = \mathcal{F}_{\mathsf{fixed}} \coloneqq \{f^\star\}$ for some $f^\star \in \mathrm{Lip}_L$.

**Theorem 4.** *Known link function Let $\hat{\boldsymbol{w}}_{\mathsf{fixed}}$ be the solution to the subsampled least square problem in this setting:*

$$\hat{\boldsymbol{w}}_{\mathsf{fixed}} \coloneqq \operatorname*{arg\,min}_{\substack{f \in \mathcal{F}_{\mathsf{fixed}} \\ \boldsymbol{w} \in \mathcal{R}}} \hat{\mathcal{L}}(f, \boldsymbol{w}) = \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathcal{R}} \hat{\mathcal{L}}(f^\star, \boldsymbol{w}). \tag{4}$$

*There exists some universal constant $C > 0$ such that the following holds. As long as $m \geq CL^4\varepsilon^{-4} d \log^3 d$, with probability at least 0.99, one has that $(f^\star, \hat{\boldsymbol{w}}_{\mathsf{fixed}})$ is an $\varepsilon$-accurate solution of (1) .*

A few remarks are in order.

**Near-optimality.** To the best of our knowledge, this is the first result establishing $\tilde{O}(d)$ sample complexity in this setting. The previous state of the art is Gajjar et al. (2023), where the sample complexity is $\tilde{O}(d^2)$. It is worth emphasizing that $\tilde{O}(d)$ is optimal up to logarithmic factors: it is clear that in general at least $d$ samples are required to find an approximate solution, since the weight $\boldsymbol{w} \in \mathbb{R}^d$ has $d$ degrees of freedom (Chen and Price, 2019).

**Proof technique.** The main ingredient is to quantify the idea that $\|\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w})-\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w}^\star)\|^2$ is close to $\|f(\boldsymbol{X}\boldsymbol{w})-f(\boldsymbol{X}\boldsymbol{w}^\star)\|^2$ in terms of a concentration inequality, which in the case of linear regression follows from matrix Chernoff bound. However, when $f$ is nonlinear, classical matrix concentration inequalities fail to provide optimal bounds, since the deviation cannot be expressed as the norm of a random matrix due to the nonlinear coupling between $\boldsymbol{S}$, $\boldsymbol{X}$ and $\boldsymbol{w}$. In order to obtain optimal results like Theorem 4, we resort to more fundamental principles in probability theory, building up a nonlinear concentration inequality using chaining and duality of metric entropy based on the idea of Rudelson (1996).

### 4.2 The case of unknown $f$

Now we consider the more general setting that $f$ is an unknown $L$-Lipschitz function, i.e., $\mathcal{F} = \mathrm{Lip}_L$.

**Theorem 5.** *Unknown link function Let $\hat{f}$ and $\hat{\boldsymbol{w}}$ be the solution to the subsampled least square problem:*

$$(\hat{f}, \hat{\boldsymbol{w}}) = \underset{\substack{f\in\mathrm{Lip}_L \\ \boldsymbol{w}\in\mathcal{R}}}{\arg\min}\, \hat{\mathcal{L}}(f, \boldsymbol{w}). \tag{5}$$

*There exists some universal constant $C > 0$ such that the following holds. As long as $m \geq CL^6\varepsilon^{-6}d^{3/2}\log(n/d)$, one has that $(\hat{f}, \hat{\boldsymbol{w}})$ is an $\varepsilon$-accurate solution of (1) with probability at least 0.99.*

**Proof technique.** Even with the nonlinear concentration inequality for a fixed $f$, handling the case of unknown $f$ remains challenging. This is because the size of $\mathrm{Lip}_L$ is infinite and necessitates an appropriate complexity measure to control this. It turns out that such a complexity measure should be based on specific properties of $\boldsymbol{S}$, differing from common measures like the VC dimension. Guided by the generic chaining theory for Bernoulli processes (Talagrand, 2021), we construct a new metric on $\mathrm{Lip}_L$, which provide fines control over the fluctuation of $\boldsymbol{S}f(\boldsymbol{X}\boldsymbol{w})$. We further upper-bound the metric entropy of $\mathrm{Lip}_L$ for this metric by carefully constructing $\varepsilon$-nets with piecewise linear functions, using properties of leverage scores.

Unlike the case of known link function, it is not clear whether the $\tilde{O}(d^{3/2})$ sample complexity in Theorem 5 is optimal, and the only lower bound we know is the same bound for the case of known link function. Closing the gap between the sample complexity in Theorem 5 and the lower bound is an interesting direction for future research.

## 5. Conclusion

We show that leverage score sampling effectively serves as an active learning strategy for learning SIMs, with optimal or state-of-the-art sample complexity across different settings. This opens two directions: 1) Finding matching lower bounds for both cases, or 2) Improving sample complexity bounds for the unknown link function case.

## 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 *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

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 SIGACT Symposium on Theory of Computing*, pages 1051–1063, 2019.

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. *Advances in Neural Information Processing Systems*, 35:9768–9783, 2022.

Xue Chen and Eric Price. Active regression via linear-sample sparsification. In *Conference on Learning Theory*, pages 663–695. PMLR, 2019.

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, December 2011.

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.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

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 Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1887–1930. PMLR, 06–09 Jul 2018.

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.

Aarshvi Gajjar, Christopher Musco, and Chinmay Hegde. Active learning for single neuron models with lipschitz non-linearities. In *International Conference on Artificial Intelligence and Statistics*, pages 4101–4113. PMLR, 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.

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

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.

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

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

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.

Mark Rudelson. Random vectors in the isotropic position, 1996.

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.

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, December 2018.

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. doi: 10.1109/ICASSP.2012.6288306.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

David P. Woodruff. Computational advertising: Techniques for targeting relevant ads. *Foundations and Trends® in Theoretical Computer Science*, 10(1-2):1–157, 2014. doi: 10.1561/0400000060.

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*, 397:108850, November 2019. doi: 10.1016/j.jcp.2019.07.048.

## Appendix A. Preliminaries

Before delving into the detailed proof, we first simplify the discussion by showing that it suffices to consider the case where $\boldsymbol{X}$ has orthonormal columns.

**Lemma 6** (Reduction to orthonormal data)**.** *If Theorem 4 and Theorem 5 hold for any orthonormal matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, then they also hold for any matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$.*

**Proof** (Sketch) The main point is that leverage score is 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

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

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 $\boldsymbol{R}\boldsymbol{w}$. ■

In virtue of Lemma 6, throughout the proof we will always assume without loss of generality that $\boldsymbol{X}$ has orthonormal columns. With this assumption, it will be convenient to note that the search region defined in (3) is close to a Euclidean ball.

**Lemma 7.** *The search region $\mathcal{R}$ defined in* (3) *satisfies $\mathcal{R} \subset B(R)$ with probability $\geq 0.999$ if $m \geq Cd\log d$, where $R = C\varepsilon^{-1/2}L^{-1}\sqrt{\mathsf{OPT} + L^2\|\boldsymbol{X}\boldsymbol{w}^\star\|^2}$.*

The proof will make use of the following classical result, which states that $\boldsymbol{S}$ is an almost isometric embedding on the column space of $\boldsymbol{X}$.

**Lemma 8** (Subspace embedding, Woodruff (2014))**.** *For any $\varepsilon \in (0,1)$, as long as $m \geq C\varepsilon^{-2}d\log(d/\delta)$, the following holds for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$ with probability at least $1-\delta$.*

$$(1-\varepsilon)\|\boldsymbol{X}\boldsymbol{w}_1 - \boldsymbol{X}\boldsymbol{w}_2\|^2 \leq \|\boldsymbol{S}\boldsymbol{X}\boldsymbol{w}_1 - \boldsymbol{S}\boldsymbol{X}\boldsymbol{w}_2\|^2 \leq (1+\varepsilon)\|\boldsymbol{X}\boldsymbol{w}_1 - \boldsymbol{X}\boldsymbol{w}_2\|^2. \tag{6}$$

We are now ready to prove Lemma 7.

**Proof** (Proof of Lemma 7) Throughout the proof we condition on the event that (6) holds. For any $\boldsymbol{w} \in \mathcal{R}$, first note that $\|\boldsymbol{X}\boldsymbol{w}\|^2 = \|\boldsymbol{w}\|^2$ since $\boldsymbol{X}$ has orthonormal columns. Now, applying Lemma 8, we obtain that $\|\boldsymbol{X}\boldsymbol{w}\|^2 \leq \frac{1}{2}\|\boldsymbol{S}\boldsymbol{X}\boldsymbol{w}\|^2 \leq \frac{1}{2\varepsilon L^2}\|\boldsymbol{S}\boldsymbol{y}\|^2$, from the definition of $\mathcal{R}$. Using Markov's inequality, we obtain that $\|\boldsymbol{S}\boldsymbol{y}\|^2 \leq 1000\|\boldsymbol{y}\|^2$ with probability $\geq 0.999$. Combining the results, we obtain the following

$$\|\boldsymbol{w}\|^2 = \|\boldsymbol{X}\boldsymbol{w}\|^2 \leq \frac{1}{2\varepsilon L^2} \cdot \|\boldsymbol{S}\boldsymbol{y}\|^2 \leq C\varepsilon^{-1}L^{-2}\|\boldsymbol{y}\|^2.$$

for some $C > 0$.

Now note that $\boldsymbol{y} = f(\boldsymbol{X}\boldsymbol{w}^\star) - (f(\boldsymbol{X}\boldsymbol{w}^\star) - \boldsymbol{y})$. Using an approximate triangle inequality, we find that $\|\boldsymbol{y}\|^2 \leq 2\mathsf{OPT} + 2\|f(\boldsymbol{X}\boldsymbol{w}^\star)\|^2 \leq 2\mathsf{OPT} + 2L^2\|\boldsymbol{X}\boldsymbol{w}^\star\|^2$. Plugging this into the above inequality on $\|\boldsymbol{w}\|^2$ yields the desired result. ■

## Appendix B. Proof of Theorem 4

The proof hinges crucially upon the following lemma, which can be viewed as a non-linear generalization of the classical subspace embedding lemma (Lemma 8). Proving this lemma turns out to be a major technical challenge since we are showing this result for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)$ as opposed to a similar result for a fixed pair $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ which was shown in Gajjar et al. (2023) by a straightforward application of Bernstein's Inequality.

**Lemma 9** (Non-linear subspace embedding with fixed non-linearity)**.** *Assume the input matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has orthonormal columns. For any $f^\star \in \mathrm{Lip}_L$, for any $R > 0$, as long as $m \geq CL^4\varepsilon^{-4}d(\log^3 d + \log(1/\delta))$ for some fixed constant $C > 0$, the following holds with probability $\geq 1-\delta$.*

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

A sketch of proof of this lemma will be provided in Appendix B.1. We are now ready to prove Theorem 4.

**Proof** (Proof of Theorem 4) This will follow similarly to Theorem 1 of Gajjar et al. (2023). We want to show that $(f^\star, \hat{\boldsymbol{w}}_{\mathsf{fixed}})$ is $\varepsilon$-accurate. Since

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

from triangle inequality, we get the following

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

Next, applying Lemma 9 with $R = C\varepsilon^{-1/2}L^{-1}\sqrt{\mathsf{OPT} + L^2\|\boldsymbol{X}\boldsymbol{w}^\star\|^2}$ as defined in Lemma 7, we obtain the following upper bound with probability $\geq 1-\delta$:

$$\begin{aligned}
\mathcal{L}(f^\star, \hat{\boldsymbol{w}}_{\mathsf{fixed}}) &\leq 2\big\|\boldsymbol{S}f^\star(\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{fixed}}) - \boldsymbol{S}f^\star(\boldsymbol{X}\boldsymbol{w}^\star)\big\|^2 + 2\varepsilon^2R^2 + 2\mathsf{OPT} \\
&\leq 4\hat{\mathcal{L}}(f^\star, \hat{\boldsymbol{w}}_{\mathsf{fixed}}) + 4\hat{\mathcal{L}}(f^\star, \boldsymbol{w}^\star) + 2C^2\varepsilon L^{-2}(\mathsf{OPT} + L^2\|\boldsymbol{X}\boldsymbol{w}^\star\|^2) + 2\mathsf{OPT} \\
&\leq 8\hat{\mathcal{L}}(f^\star, \boldsymbol{w}^\star) + 2C^2\varepsilon\|\boldsymbol{X}\boldsymbol{w}^\star\|^2 + 4\mathsf{OPT}, \qquad (7)
\end{aligned}$$

assuming $\varepsilon \leq C^{-2}L^2$, where the second line again follows from triangle inequality applied to

$$\boldsymbol{S}f^\star(\boldsymbol{X}\boldsymbol{w}^\star) - \boldsymbol{S}f^\star(\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{fixed}}) = \big(\boldsymbol{S}f^\star(\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathsf{fixed}}) - \boldsymbol{y}\big) - \big(\boldsymbol{S}f^\star(\boldsymbol{X}\boldsymbol{w}^\star) - \boldsymbol{y}\big),$$

and the third line follows from the minimality of $\hat{\boldsymbol{w}}_{\mathsf{fixed}}$.

Note further that $\hat{\mathcal{L}}(f^\star, \boldsymbol{w}^\star) \leq C \cdot \mathcal{L}(f^\star, \boldsymbol{w}^\star)$ with probability at least 0.999 by Chebyshev's inequality, since $\mathbb{E}\hat{\mathcal{L}}(f^\star, \boldsymbol{w}^\star) = \mathcal{L}(f^\star, \boldsymbol{w}^\star)$. The desired conclusion follows from this and (7) immediately, if we replace $\varepsilon$ by $\varepsilon/2C^2$. ■

### B.1 Proof sketch of Lemma 9

The proof relies crucially on the idea developed in Rudelson (1996). We essentially extend a simplified version of their idea into a nonlinear setting. First, we apply a standard

symmetrization argument to the $\ell$-th moment of the desired quantity, where we introduce i.i.d. Rademacher random variables and denote them by $\xi_1, \xi_2, \ldots, \xi_m$, where each $\xi_i$ takes values $-1$ and 1 with probabilities $1/2$ each.

$$\mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| \|\boldsymbol{S}(f(\boldsymbol{X}\boldsymbol{w}_1) - f(\boldsymbol{X}\boldsymbol{w}_2))\|^2 - \|f(\boldsymbol{X}\boldsymbol{w}_1) - f(\boldsymbol{X}\boldsymbol{w}_2)\|^2 \right|^\ell$$
$$\leq 2^\ell \cdot \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| \frac{1}{m^\ell} \sum_{i=1}^m \xi_i \frac{\big(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle)\big)^2}{p_{j_i}} \right|^\ell,$$

We denote $v_{j_i}(\boldsymbol{w}_1, \boldsymbol{w}_2) \coloneqq f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle)$ and this leads us to study the following symmetrized random process.

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

It can be seen that the above random process conditioned on the samples $\{j_1, j_2, \ldots, j_m\}$ (fixing $\boldsymbol{S}$) is a fortiori subgaussian (Vershynin (2018)) with respect to index $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ endowed with the metric $\rho$ on $B(R) \times B(R)$ defined as follows.

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

Now, consider the symmetric convex body $\mathcal{P} = \operatorname{conv}(\pm \frac{\boldsymbol{x}_{j_1}}{\sqrt{p_{j_1}}}, \pm \frac{\boldsymbol{x}_{j_2}}{\sqrt{p_{j_2}}}, \ldots \pm \frac{\boldsymbol{x}_{j_m}}{\sqrt{p_{j_m}}})$. We denote $\mathcal{P}^\circ$ as its polar given by $\mathcal{P}^\circ = \{\boldsymbol{z} : \big|\langle \boldsymbol{z}, \boldsymbol{p} \rangle\big| \leq 1, \forall \boldsymbol{p} \in \mathcal{P}\}$ and $\|\cdot\|_{\mathcal{P}^\circ}$ as the Minkowski norm associated with $\mathcal{P}^\circ$, defined for $\boldsymbol{w} \in \mathbb{R}^d$ as

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

Using Lipschitz continuity of $f$, we can establish an upper bound of $\rho$ in terms of $\|\cdot\|_{\mathcal{P}^\circ}$

$$\rho\big((\boldsymbol{w}_1, \boldsymbol{w}_2), (\boldsymbol{w}_1', \boldsymbol{w}_2')\big) \leq 4L^2 R \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \big( \|\boldsymbol{w}_1 - \boldsymbol{w}_1'\|_{\mathcal{P}^\circ} + \|\boldsymbol{w}_2 - \boldsymbol{w}_2'\|_{\mathcal{P}^\circ} \big), \quad (8)$$

Since we have established that our desired random process is subgaussian w.r.t. a metric, we can bound the expected supremum using the well known Dudley's inequality, which is stated as follows.

**Lemma 10** (Dudley's inequality)**.** *If $(X_t)_{t \in T}$ is a subgaussian process with respect to a metric $\rho$, then*

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

*where $\mathcal{N}(T, \rho, \varepsilon)$ denotes the minimum number of closed balls of radius $\varepsilon$ w.r.t. the metric $\rho$ that can cover the set $T$.*

Next, 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 $\inf_{\boldsymbol{w}_1 \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| \leq 0$.

We have fixed $\boldsymbol{S}$ for obtaining the below upper bound and we are taking the expectation with respect to the Rademacher random variables $\xi_1, \xi_2, \ldots, \xi_m$.

$$\begin{aligned}\mathbb{E}_\xi \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} |Z(\boldsymbol{w}_1, \boldsymbol{w}_2)| &\lesssim \int_0^\infty \sqrt{\log \mathcal{N}(B(R) \times B(R), \rho, \varepsilon)}\, \mathrm{d}\varepsilon \\ &\lesssim L^2 R \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\|^{1/2} \int_0^\infty \sqrt{\log \mathcal{N}(B(R), \|\cdot\|_{\mathcal{P}^\circ}, \varepsilon)}\, \mathrm{d}\varepsilon,\end{aligned}$$

The second inequality can be verified using inequality (8) along with standard properties of covering numbers (e.g., tensorization inequality).

Since the integral $\int_0^\infty \sqrt{\log \mathcal{N}(B(R), \|\cdot\|_{\mathcal{P}^\circ}, \varepsilon)}\, \mathrm{d}\varepsilon$ only involves the entropy in dual norm $\|\cdot\|_{\mathcal{P}^\circ}$, we control this using duality of metric entropy. In particular, one can invoke dual Sudakov minoration (Ledoux and Talagrand, 1991) to prove

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

By Lemma 6 we may assume without loss of generality that $\boldsymbol{X}$ is orthonormal, and in this case one may show that $\sqrt{p_{j_i}} = \|\boldsymbol{x}_{j_i}\|/\sqrt{d}$. Combining these results, we obtain

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

Therefore, we establish an upper bound without any dependence on the nonlinearity $f$. Taking expectation w.r.t. $\boldsymbol{S}$, we obtain, by matrix Chernoff bound (Rudelson and Vershynin, 2007),

$$\mathbb{E} \left\| \sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i} \boldsymbol{x}_{j_i}^\top \right\| \lesssim m + d \log d. \tag{9}$$

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

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

This is close to the claimed result in the lemma, except that we need a tail bound, which can be obtained easily from the above by a concentration of measure argument (Ledoux, 2001). In fact, we may show

$$\left( \mathbb{E} \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \left| \frac{1}{m^\ell} \sum_{i=1}^m \xi_i \frac{\big(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle)\big)^2}{p_{j_i}} \right|^\ell \right)^{1/\ell} \lesssim \varepsilon^2 R^2 + L^2 R^2 \sqrt{\frac{d}{m}} \sqrt{\ell}.$$

This gives the desired subgaussian tail bound by a standard argument based on Markov's inequality (Vershynin, 2018).

## Appendix C. Proof of Theorem 5

The proof follows exactly the same line as the proof of Theorem 4, but with Lemma 9 there substituted with the following result which highlights the case of single index models where the nonlinearity could be unknown.

**Lemma 11** (Non-linear subspace embedding with unknown non-linearity)**.** *Assume the input matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has orthonormal columns. As long as $m \geq CL^6\varepsilon^{-6}d^{3/2}\log(n/d\delta)$ for some fixed constant $C > 0$, the following holds with probability $\geq 1-\delta$.*

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

**Proof** (Sketch) Similar to the proof of Lemma 9, we need to bound the supremum of the process

$$Z_f(\boldsymbol{w}_1, \boldsymbol{w}_2) \coloneqq \sum_{i=1}^m \xi_i \frac{\big(f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2}{p_{j_i}}, \qquad f \in \mathrm{Lip}_L,\ (\boldsymbol{w}_1, \boldsymbol{w}_2) \in B(R)\times B(R).$$

The crucial step is to construct an appropriate discretization, $\mathcal{N}$ of $\mathrm{Lip}_L$, to control the process for each $f \in \mathcal{N}$ using Lemma 9, and then apply a union bound. It turns out $\mathcal{N}$ should be chosen as an $\varepsilon$-net w.r.t. the following metric:

$$D_\infty(f_1, f_2) = \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2} \sum_{i\in[m]} \frac{1}{p_{j_i}} \left|\big(f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2 - \big(f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1\rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2\rangle)\big)^2\right|.$$

This metric has the important property that the following upper bound is true deterministically.

$$\big|Z_{f_1}(\boldsymbol{w}_1, \boldsymbol{w}_2) - Z_{f_2}(\boldsymbol{w}_1, \boldsymbol{w}_2)\big| \leq D_\infty(\boldsymbol{w}_1, \boldsymbol{w}_2)$$

Therefore, if $\mathcal{N}$ is a $\Delta$-net of $\mathrm{Lip}_L$ with respect to the metric $D_\infty$, one has

$$\sup_{f\in\mathrm{Lip}_L} |Z_f(\boldsymbol{w}_1, \boldsymbol{w}_2)| \leq \sup_{f\in\mathcal{N}} |Z_f(\boldsymbol{w}_1, \boldsymbol{w}_2)| + \Delta. \tag{10}$$

To control the size of the $\Delta$-net, we utilize the following bound:

**Lemma 12.** *Let $I_{j_i} = [-R\|X_{j_i}\|, R\|X_{j_i}\|]$. Then*

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

The proof is postponed to Appendix C.1. In light of this lemma, one may try to construct an $\Delta$-net with respect to $D_\infty$ by piecewise linear functions, each of which differs with its neighbor 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, and by some inequalities constraining the number of rows with small leverage scores, we can construct an $\Delta$-net of size

$$\log|\mathcal{N}| \lesssim \frac{\log(n/d)}{\eta}, \quad \Delta \coloneqq \eta L^2 R^2 \sqrt{d} \left\|\sum_{i=1}^m \frac{1}{p_{j_i}} \boldsymbol{x}_{j_i}\boldsymbol{x}_{j_i}^\top\right\|^{1/2} \left(m + \frac{1}{n}\sum_{i=1}^m \frac{1}{p_{j_i}}\right)^{1/2}.$$

As described before, we may apply Lemma 9 to each $f \in \mathcal{N}$, and then take a union bound, which leads us to

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

Plug this into (10) to obtain

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

Taking expectation w.r.t. $\boldsymbol{S}$ again and using the matrix Chernoff bound (9), we obtain

$$\mathbb{E} \sup_{\substack{f\in\mathrm{Lip}_L\\ \boldsymbol{w}_1,\boldsymbol{w}_2\in B(R)}} |Z_f(\boldsymbol{w}_1,\boldsymbol{w}_2)| \lesssim L^2R^2\sqrt{md}\left(\sqrt{\log^2 d\cdot\log m}+\sqrt{\frac{\log(n/d)}{\eta}}+\eta\sqrt{m}\right).$$

Optimising over $\eta>0$ leads to

$$\mathbb{E} \sup_{\substack{f\in\mathrm{Lip}_L\\ \boldsymbol{w}_1,\boldsymbol{w}_2\in B(R)}} |Z_f(\boldsymbol{w}_1,\boldsymbol{w}_2)| \lesssim L^2R^2\sqrt{md}\left(\sqrt{\log^2 d\cdot\log m}+m^{1/6}\log^{1/3}(n/d)\right).$$

From this, it can be seen that whenever $m \gtrsim L^6\varepsilon^{-6}d^{3/2}\log(n/d)$, we have

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

Similar to the proof of Lemma 9, this is close to what we desire, except that we need a tail bound. The latter can be deduced from the above using concentration of measure again. ■

### C.1 Proof of Lemma 12

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

$$\begin{aligned}&\left|\big(f_1(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_1\rangle)-f_1(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_2\rangle)\big)^2-\big(f_2(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_1\rangle)-f_2(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_2\rangle)\big)^2\right|\\ &=\underbrace{\left|f_1(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_1\rangle)-f_2(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_1\rangle)-f_1(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_2\rangle)+f_2(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_2\rangle)\right|}_{=:T_1}\\ &\quad\cdot\underbrace{\left|f_1(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_1\rangle)-f_1(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_2\rangle)+f_2(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_1\rangle)-f_2(\langle\boldsymbol{x}_{j_i},\boldsymbol{w}_2\rangle)\right|}_{=:T_2}.\end{aligned}$$

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

$$\left| f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_k \rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_k \rangle) \right| \leq \|f_1 - f_2\|_{L^\infty(I_{j_i})}, \quad k = 1, 2.$$

it is then clear that

$$T_1 \leq 2\|f_1 - f_2\|_{L^\infty(I_{j_i})}.$$

The second factor $T_2$ can be controlled using the Lipschitz assumption, which implies $|f_k(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f_k(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle)| \leq L|\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle|$ for $k = 1, 2$, hence

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

Combining these estimates, we obtain

$$\begin{aligned} &\left| \left( f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f_1(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle) \right)^2 - \left( f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle) - f_2(\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle) \right)^2 \right| \\ &\leq 2\|f_1 - f_2\|_{L^\infty(I_{j_i})} \cdot L \left( |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 \rangle| + |\langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_2 \rangle| \right), \end{aligned} \tag{11}$$

and thus

$$\begin{aligned} D_\infty(f_1, f_2) &\leq 4L \sup_{\boldsymbol{w}_1, \boldsymbol{w}_2 \in B(R)} \sum_i \frac{1}{p_{j_i}} \left| \langle \boldsymbol{x}_{j_i}, \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle \right| \cdot \|f_1 - f_2\|_{L^\infty(I_{j_i})} \\ &\leq 4L \sup_{\boldsymbol{w} \in B(2R)} \sum_i \frac{1}{p_{j_i}} \left| \langle \boldsymbol{x}_{j_i}, \boldsymbol{w} \rangle \right| \cdot \|f_1 - f_2\|_{L^\infty(I_{j_i})} \\ &= 8LR \sup_{\boldsymbol{w}: \|\boldsymbol{w}\| \leq 1} \sum_i \left| \left\langle \frac{\boldsymbol{x}_{j_i}}{\sqrt{p_{j_i}}}, \boldsymbol{w} \right\rangle \right| \cdot \frac{1}{\sqrt{p_{j_i}}} \|f_1 - f_2\|_{L^\infty(I_{j_i})}. \end{aligned}$$

The conclusion of the lemma then follows from Cauchy-Schwarz.