Provable Active Learning of Neural Networks for Parametric PDEs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural networks have proven effective in constructing surrogate models for parametric partial differential equations (PDEs) and for approximating high-dimensional quantity of interest (QoI) surfaces. A major cost is training such models is collecting training data, which requires solving the target PDE for a variety of different parameter settings. Active learning and experimental design methods have the potential to reduce this cost, but are not yet widely used for training neural networks, nor do there exist methods with strong theoretical foundations.

In this work we provide evidence, both empirical and theoretical, that existing active sampling techniques can be used successfully for fitting neural network models for high-dimensional parameteric PDEs. In particular, we show the effectiveness of "coherence motivated" sampling methods (i.e., leverage score sampling), which are widely used to fit PDE surrogate models based on polynomials. We prove that leverage score sampling yields strong theoretical guarantees for fitting single neuron models, even under adversarial label noise. Our theoretical bounds apply to any single neuron model with a Lipschitz non-linearity (ReLU, sigmoid, absolute value, low-degree polynomial, etc.).

1 Introduction

In recent years, neural networks have proven broadly useful in accelerating the numerical solution of partial differential equations (PDEs). In applications to parametric PDEs, one use of neural networks is in developing surrogate models and approximations for quantity-of-interest (QoI) surfaces (for use e.g. in parameter optimization or uncertainty quantification) [Tripathy and Bilionis, 2018, Zhang et al., 2019, Khoo et al., 2021, O'Leary-Roseberry et al., 2022]. In these applications, the goal is to approximate a high-dimensional function mapping PDE input parameters to scalar values. A significant cost in training neural network approximations to such functions is the collection of training data: each training point collected requires solving the PDE for a different set of parameters chosen e.g. on a grid or at random [Adcock et al., 2022a, Cohen and DeVore, 2015] for more details.

One possible approach to reducing the cost of collecting training data is to employ active learning or experimental design methods to more intelligently choose training examples. Such methods have been employed successful in QoI approximation and surrogate modeling approaches based on more traditional models, like polynomials and sparse or structured polynomials [Chkifa et al., 2018, Cohen and DeVore, 2015, Adcock et al., 2022b, Hampton and Doostan, 2015b]. However, with some exceptions, there has been significantly less work in applying active learning methods to training neural network models for parametric PDEs [Lye et al., 2021, Pestourie et al., 2020]. Moreover, in contrast to active learning approaches for more traditional functions families, most existing methods are heuristic, and not supported by strong theoretical guarantees.

Submitted to the DLDE Workshop in the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Our Approach

We take a step towards developing theoretically sound active learning methods for approximating parametric PDEs with neural networks by focusing on the special case of "single neuron" or "single index" models[1]. Such models take the form $g(\mathbf{x}) = f(\langle \mathbf{w}, \mathbf{x} \rangle)$, where f is a scalar non-linearity, and $\mathbf{w}$ is a set of weights [Pinkus, 1997, 2015, Yehudai and Ohad, 2020, Rao et al., 2017, Candès, 2003]. Single neuron models are studied in machine learning theory as tractable examples of single-layer neural networks [Diakonikolas et al., 2020, Goel et al., 2017]. However, even these simple models are known to be adept at modeling a variety of physical phenomena [Constantine et al., 2016] and for that reason can already be used effectively in building PDE surrogate models and QoI approximations for use in uncertainty quantification, model-driven design, and data assimilation [O'Leary-Roseberry et al., 2022, Constantine et al., 2017, Cohen et al., 2012, Le Maître and Knio, 2010, Lassila and Rozza, 2010, Binev et al., 2017]. As such, they serve as a natural starting point for our work.

We frame the problem of actively learning single neuron models in the *agnostic learning* or adversarial noise setting. For a given distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$, a random vector $(\mathbf{x}, y)$ sampled from $\mathcal{D}$, and non-linearity $f : \mathbb{R} \to \mathbb{R}$, our goal is to approximately minimizes the expected squared error $\mathbb{E}_{\mathbf{x},y\sim\mathcal{D}} \left(f(\langle \mathbf{w}, \mathbf{x} \rangle) - y \right)^2$. Formally, for an error parameter Δ, we want to return some $\tilde{\mathbf{w}}$ such that:

$$\mathbb{E}_{\mathbf{x},y\sim\mathcal{D}} \left(f(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle) - y \right)^2 \leq \min_{\mathbf{w}} \mathbb{E}_{\mathbf{x},y\sim\mathcal{D}} \left(f(\langle \mathbf{w}, \mathbf{x} \rangle) - y \right)^2 + \Delta.$$

Importantly, in the agnostic setting, we make no assumption that $\mathbf{y} = f(\langle \mathbf{w}^*, \mathbf{x} \rangle)$ for some ground-truth parameter vector $\mathbf{w}^*$, nor do we assume it equals $f(\langle \mathbf{w}^*, \mathbf{x} \rangle)$ plus mean-centered noise. This is in contrast to the "realizable" setting, studied in some prior work [Tyagi and Cevher, 2012, Cohen et al., 2012] and in classical work on optimal experimental design [Pukelsheim, 2006]. The agnostic setting is more challenging, but also more appropriate for PDE applications, where the function being approximated is usually not itself of the form $f(\langle \mathbf{w}, \mathbf{x} \rangle)$. It has become the standard in work on active learning for functions not based on neural networks [Chkifa et al., 2018, Cohen and DeVore, 2015, Adcock et al., 2022b, Hampton and Doostan, 2015b]).

For simplicity, we consider the case when $\mathcal{D}$ is a uniform distribution over n points in $\mathbb{R}^d$. This is essentially without loss of generality, since any continuous distribution can be approximated by the uniform distribution over a sufficient large finite sample of $\mathbf{x}$ values. In this case, we have:

Problem 1 (Single Neuron Regression)**.** *Given a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and query access to a vector of labels, $\mathbf{y} \in \mathbb{R}^n$, for a given function $f : \mathbb{R} \to \mathbb{R}$, find a vector $\mathbf{w} \in \mathbb{R}^d$ to minimize $\|f(\mathbf{X}\mathbf{w}) - \mathbf{y}\|_2^2$ using as few queries from $\mathbf{y}$ as possible.*

When f is an identity function, Problem 1 reduces to active least squares regression, which has received a lot of recent attention in computer science and machine learning. In the agnostic setting, state-of-the-art results can be obtained via "coherence motivated" sampling, also known as "leverage score" or "effective resistance" sampling [Avron et al., 2019, Cohen and Migliorati, 2017, Rauhut and Ward, 2012, Hampton and Doostan, 2015a, Erdélyi et al., 2020, Musco et al., 2022]. The idea behind such methods is to collect samples from $\mathbf{y}$ randomly but non-uniformly, using an importance sampling distribution based on the rows of $\mathbf{X}$. More "unique" rows are selected with higher probability. Formally, rows are selected with probability proportional to their statistical leverage scores:

Definition 1 (Statistical Leverage Score)**.** *The leverage score, $\tau_i(\mathbf{X})$ of the i^{th} row, $\mathbf{x}_i$ of a matrix, $\mathbf{X} \in \mathbb{R}^{n\times d}$ is equal to:*

$$\tau_i(\mathbf{X}) = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = \max_{\mathbf{w}\in\mathbb{R}^d} \frac{[\mathbf{X}\mathbf{w}]_i^2}{\|\mathbf{X}\mathbf{w}\|_2^2}$$

We always have that $0 \leq \tau_i \leq 1$. The leverage score of a row is large (closer to 1) if that row has large inner product with some vector in $\mathbb{R}^d$ in comparison to all other rows in the matrix $\mathbf{X}$. This means that the particular row is important in formulating the row space of $\mathbf{X}$. It can be shown that when $\mathbf{X}$ has d columns leverage score sampling yields a sample complexity of $O(d \log d/ + d/\epsilon)$ to find $\hat{\mathbf{w}}$ satisfying $\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 \leq (1+\epsilon) \min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$. This is optimal up to the $\log d$ factor [Chen and Price, 2019]. Our main contribution is to establish that, when combined with a novel regularization strategy, leverage scores sampling simultaneously yields theoretical guarantees for our more general Problem 1 for a broad class of non-linearities f. We only require that f is L-Lipschitz

[1]These functions are also called "ridge functions" or "plane waves" in some communities.

for some constant L, a property that holds for most non-linearities used in practice (ReLU, absolute value, low-degree polynomials, etc.). Specifically, in the Appendix A we prove:

Theorem 1 (Main Result)**.** *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a data matrix and $\mathbf{y} \in \mathbb{R}^n$ be a label vector. Let f be an L-Lipschitz non-linearity with $f(0) = 0$ and let $OPT = \min_{\mathbf{w}} \|f(\mathbf{X}\mathbf{w}) - \mathbf{y}\|_2^2$. Let $\mathbf{S} \in \mathbb{R}^{m \times n}$ be a sampling matrix with rows selected with probability proportional to the leverage scores of $\mathbf{X}$. Let $\hat{\mathbf{w}}$ solve the following constrained optimization problem involving the sampled labels $\mathbf{S}\mathbf{y}$:*

$$\hat{\mathbf{w}} = \underset{\mathbf{w}: \|\mathbf{S}\mathbf{X}\mathbf{w}\|_2^2 \leq \frac{1}{\epsilon \cdot L^2} \|\mathbf{S}(\mathbf{y}\|_2^2}{\arg\min} \|\mathbf{S}f(\mathbf{X}\mathbf{w}) - \mathbf{S}\mathbf{y}\|_2^2. \tag{1}$$

As long as $m = O\left(\frac{d^2 \log(d/\epsilon^2)}{\epsilon^4}\right)$, then for a fixed constant C, with probability $> 9/10$,

$$\|f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{y}\|_2^2 \leq C \cdot \left(OPT + \epsilon L^2 \|\mathbf{X}\mathbf{w}^*\|_2^2\right).$$

The sampling matrix referenced in Theorem 1 is formally defined as follows:

Definition 2 (Importance Sampling Matrix)**.** *Let $p_1, \ldots, p_n \in [0, 1]$ be a given set of probabilities (so that $\sum_i p_i = 1$). A matrix $\mathbf{S}$ is an $m \times n$ importance sampling matrix if each of its rows is chosen to equal $\frac{1}{\sqrt{m \cdot p_i}} \cdot \mathbf{e}_i$ with probability proportional to p_i.*

Theorem 1 mirrors previous results in the linear setting, and in constrast to some prior work on agnostically learning single neuron models, does not require any assumptions on $\mathbf{X}$ [Diakonikolas et al., 2022, Tyagi and Cevher, 2012]. In addition to multiplicative error C, it has an additive error term of $\epsilon L^2 \|\mathbf{X}\mathbf{w}^*\|_2^2$, which we believe is necessary. Similar additive error terms arise in related work on leverage score sampling for problems like logistic regression [Munteanu et al., 2018, Mai et al., 2021]. On the other hand, we believe the d^2 dependence in our bound is not necessary, and should be improvable linear in d. The ϵ is also likely improvable.

We note that the assumption $f(0) = 0$ in Theorem 1 is without loss of generality. If $f(0)$ is non-zero, we can simply solve a transformed problem with $\mathbf{y}' = \mathbf{y} - f(0)$ and $f'(x) = f(x) - f(0)$. Finally, we note that while (1) is inherently a non-convex problem, it can be solved easily in practice using standard methods (e.g. projected gradient or stochastic gradient decent).

3 Experimental Results

Leverage score sampling is already used as an active learning strategy in PDE surrogate modeling and is simple and computationally efficient to implement [Cohen and DeVore, 2015]. We applied the method to several synthetic problems, as well as a test problem on approximating a differential equation QoI surface. For all problems, leverage score sampling significantly outperforms the standard approach of choosing data uniformly at random from $\mathbf{X}$. For the synthetic data problems we let $\mathbf{X}$ either contain 10^5 random Gaussian vectors in two dimensions (Gaussian data), or the coordinates of 10^5 values in $[-1, 1]^2$ (uniform data). We also added a column of all 1's to allow

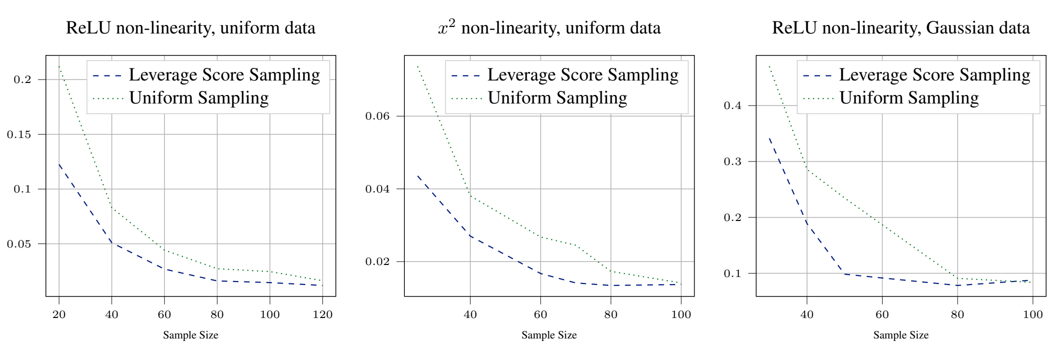
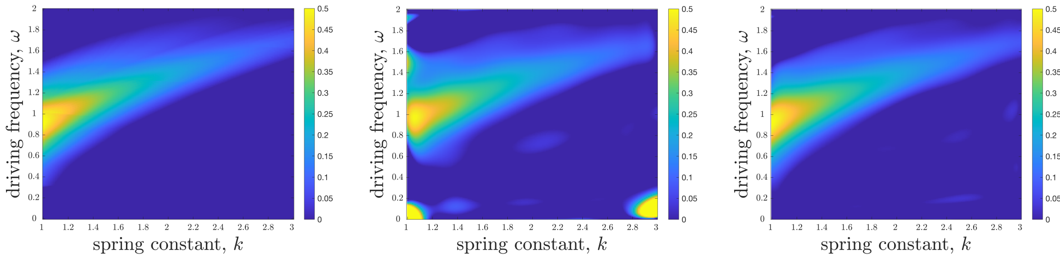


Figure 1: The three figures show the median relative error for learning the two-dimensional single neuron models $\mathrm{ReLU}(0.4x_1 + 0.4x_2 - 0.4)$, $(-0.3x_1 + 0.1x_2 + 0.1)^2$, and $\mathrm{ReLU}(0.4x_1 + 0.4x_2 - 0.6)$ corrupted with Gaussian noise $\eta_1 \sim \mathcal{N}(0, 0.05)$, $\eta_2 \sim \mathcal{N}(0, 0.05)$ and $\eta_3 \sim \mathcal{N}(0, 0.1)$. In all cases our active leverage score sampling method outperforms naive uniform sampling.



(a) True Quantity of Interest. (b) Approximation based on uniformly sampled training data. (c) Approximation based on leverage-score samples.

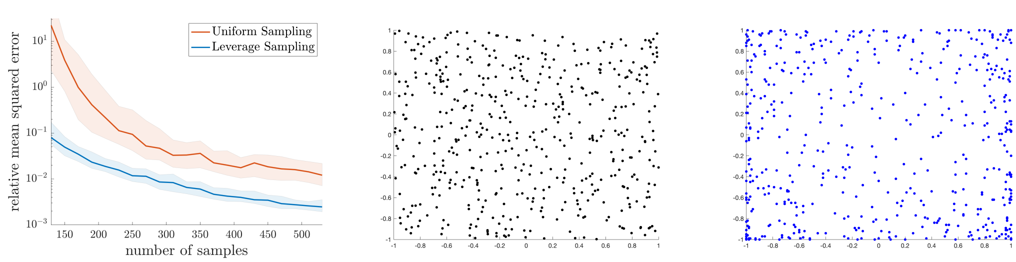
Figure 2: Plot of single neuron model fit to a QoI (maximum displacement) for a parametric ODE modeling a driven harmonic oscillator; see Equation 2. The ODE involves two free parameters: a spring constant k and a driving frequency ω. 200 training points were collected via uniform sampling (the standard approach) and our active leverage score sampling method. Evidently, leverage score sampling provides a better fit. Comparable accuracy from the uniform sampling method would require considerably more samples, and thus higher computational complexity to obtain those samples.

for a bias term. We select a ground truth $\mathbf{w}^*$ and let $\mathbf{y} = f(\mathbf{X}\mathbf{w}^*) + \mathbf{g}$, where $\mathbf{g}$ is a vector of mean-centered Gaussian noise. We ran 100 trials of leverage score and uniform sampling for various sample sizes and report median error in Figure 1. We computed $\hat{\mathbf{w}}$ by finding the optimal weights to fit our subsampled data – we found that the constraint in (1) could be dropped without hurting the performance of leverage score sampling. For the small synthetic problems we used brute force search to optimize weights to ensure a true minimum was found. Evidently, leverage scores sampling outperforms the standard approach of uniform sampling in all cases.

For the test problem, we considered a second-order ODE modeling a damped harmonic oscillator with a sinusoidal force applied, which leads to the following set of parametric equations:

$$\frac{d^2x}{dt^2}(t) + c \cdot \frac{dx}{dt}(t) + k \cdot x(t) = f \cdot \cos(\omega t), \qquad x(0) = x_0, \qquad \frac{dy}{dt}(0) = x_1. \tag{2}$$

Here, x is the oscillators displacement, t is time, and c, k, f, ω are parameters. The choice of parameters will significantly impact the final solution. For example, if the frequency term ω is close to the resonant frequency of the oscillator, we expect the driving force to lead to large oscillations. We took as our QoI the maximum oscillator displacement after 20 seconds, approximating this value for all k and ω in the rectangle $\mathcal{U} = [1,3] \times [0,2]$. We chose to approximate the QoI (which is always positive) using a function of the form $\text{ReLU}(p(k,\omega))$, where p is a degree 12, two variate polynomial. This was accomplished by setting $\mathbf{X}$ to be a Vandermonde matrix evaluated at a grid of values on $[1,3] \times [0,2]$. We fit the QoI to this single neuron function using gradient descent implemented with a standard adaptive step-size, again dropping the constraint in (1). Results are show in Figures 2 and 3.



(a) Relative Error (b) Uniform Random Samples (c) Leverage Score Samples

Figure 3: The left plot shows sample complexity vs. relative error (median and interquartile range) for fitting the QoI visualized in Figure 2. Leverage score sampling gives roughly an order of magnitude improvement over over uniform sampling. The right plots visualize uniform vs. leverage score sampling for selecting example parameter vectors from the box $[1,3] \times [0,2]$. Our leverage score method tends to sample more heavily near the perimeter of the box to fit the single neuron model.

References

Ben Adcock, Simone Brugiapaglia, Nick Dexter, and Sebastian Morage. Deep neural networks are effective at learning high-dimensional hilbert-valued functions from limited data. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 1–36, 2022a.

Ben Adcock, Juan M. Cardenas, Nick Dexter, and Sebastian Moraga. *Towards Optimal Sampling for Learning Sparse Approximations in High Dimensions*, pages 9–77. Springer International Publishing, 2022b.

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)*, 2019.

Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Data assimilation in reduced modeling. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):1–29, 2017.

Emmanuel J. Candès. Ridgelets: estimating with ridge functions. *The Annals of Statistics*, 31(5): 1561–1599, 2003.

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

Abdellah Chkifa, Nick Dexter, Hoang Tran, and Clayton G. Webster. Polynomial approximation via compressed sensing of high-dimensional functions on lower sets. *Math. Comp.*, 87(311): 1415–1450, 2018.

Albert Cohen and Ronald DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numerica*, 24:1, 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, 2012.

Paul G. Constantine, Zachary del Rosario, and Gianluca Iaccarino. Many physical laws are ridge functions. *arXiv:1605.07974*, 2016.

Paul G. Constantine, Armin Eftekhari, Jeffrey Hokanson, and Rachel A. Ward. A near-stationary subspace for ridge approximation. *Computer Methods in Applied Mechanics and Engineering*, 326:402–421, 2017.

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)*, volume 125, pages 1452–1485, 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)*, volume 178, pages 4313–4361, 2022.

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

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)*, volume 65, pages 1004–1042, 2017.

Jerrad Hampton and Alireza Doostan. Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression. *Comput. Method. Appl. M.*, 290:73–97, 2015a.

Jerrad Hampton and Alireza Doostan. Compressive sampling of polynomial chaos expansions: Convergence analysis and sampling strategies. *Journal of Computational Physics*, 280:363–386, 2015b.

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

Toni Lassila and Gianluigi Rozza. Parametric free-form shape design with PDE models and reduced basis method. *Computer Methods in Applied Mechanics and Engineering*, 199(23):1583–1592, 2010.

Olivier P. Le Maître and Omar M. Knio. *Spectral methods for uncertainty quantification : with applications to computational fluid dynamics*. Scientific computation. Springer Netherlands, Dordrecht, New York, 2010.

Kjetil O. Lye, Siddhartha Mishra, Deep Ray, and Praveen Chandrashekar. Iterative surrogate model optimization (ISMO): An active learning algorithm for pde constrained optimization with deep neural networks. *Computer Methods in Applied Mechanics and Engineering*, 2021.

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

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

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

Raphaël Pestourie, Youssef Mroueh, Thanh V. Nguyen, Payel Das, and Steven G. Johnson. Active learning of deep surrogates for PDEs: application to metasurface design. *npj Computational Materials*, 6(1):164, 2020.

Allan Pinkus. Approximating by ridge functions. *Surface fitting and multiresolution methods*, pages 279–292, 1997.

Allan Pinkus. *Ridge functions*, volume 205. Cambridge University Press, 2015.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006. doi: 10.1137/1.9780898719109. URL `https://epubs.siam.org/doi/abs/10.1137/1.9780898719109`.

Nikhil Rao, Ravi Ganti, Laura Balzano, Rebecca Willett, and Robert Nowak. On learning high-dimensional structured single index models. In *Proceedings of the AAAI Conference on Artificial (AAAI)*, 2017.

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

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.

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).

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.

Hemant Tyagi and Volkan Cevher. Active learning of multi-index function models. In *Advances in Neural Information Processing Systems 25 (NeurIPS)*, pages 1466–1474, 2012.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012.

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

Gilad Yehudai and Shamir Ohad. Learning a single neuron with gradient methods. In *Proceedings of the 33rd Annual Conference on Computational Learning Theory (COLT)*, volume 125, pages 3756–3786, 2020.

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.

A Appendix

Notation. Throughout, we use bold lower-case letters for vectors and bold upper-case letters for matrices. We let $\mathbf{e}_i$ denote the i^{th} standard basis vector (all zeros, but with a 1 in position i). The dimension of $\mathbf{e}_i$ will be clear from context. For a vector $\mathbf{y} \in \mathbb{R}^n$, $\|\mathbf{y}\|_2 = (\sum_{i=1}^n y_i^2)^{1/2}$ denotes the Euclidean norm. $\mathcal{B}^d(r)$ denotes a ball of radius r centered at 0, i.e. $\mathcal{B}^d(r) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq r\}$. For a fixed matrix $\mathbf{X}$, unobserved target vector $\mathbf{y}$, and non-linearity f, we let OPT denote $\|f(\mathbf{X}\mathbf{w}^*) - \mathbf{y}\|_2^2$ where $\mathbf{w}^* = \arg\min_{\mathbf{w}} \|f(\mathbf{X}\mathbf{w}) - \mathbf{y}\|_2$.

As mentioned, our main result is based on sampling by the leverage scores $\tau_1(\mathbf{X}), \ldots, \tau_n(\mathbf{X})$ of a matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$. For *any* full-rank $d \times d$ matrix $\mathbf{R}$, we have that $\tau_i(\mathbf{X}\mathbf{R}) = \tau_i(\mathbf{X})$. This is clear from Definition 1 and implies that τ_i only depends on the column span of $\mathbf{X}$. In our proofs, this property will allow us to easily reduce to the setting where $\mathbf{X}$ is assumed to be orthonormal. Finally, we will use the following well-known fact about using leverage score sampling to construct a "subspace embedding" for a matrix $\mathbf{X}$.

We first state an intermediate result on the solution $\hat{\mathbf{w}}$ to (1) that will be used in our main proof.

Claim 1. *With probability* $49/50$ *probability, for a fixed constant* $C > 0$*,*

$$\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}\mathbf{y}\|_2^2 \leq C \cdot \left(OPT + \epsilon L^2 \|\mathbf{X}\mathbf{w}^*\|_2^2\right).$$

Proof. Consider the case when $\|\mathbf{S}\mathbf{X}\mathbf{w}^*\|_2^2 \leq \frac{1}{\epsilon L^2}\|\mathbf{S}\mathbf{y}\|_2^2$. Then $\mathbf{w}^*$ satisfies the constraint of the above optimization problem so we have that $\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}\mathbf{y}\|_2^2 \leq \|\mathbf{S}f(\mathbf{X}\mathbf{w}^*) - \mathbf{S}\mathbf{y}\|_2^2 \leq C \cdot OPT$. The last inequality follows with probability $49/50$ via Markov's inequality since $\mathbb{E}\left[\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}\mathbf{y}\|_2^2\right] = \|f(\mathbf{X}\mathbf{w}^*) - \mathbf{y}\|_2^2 = OPT$. On the other hand, if it is not the case that $\|\mathbf{S}\mathbf{X}\mathbf{w}^*\|_2^2 \leq \frac{1}{\epsilon L^2}\|\mathbf{S}\mathbf{y}\|_2^2$, then we have that $\|\mathbf{S}\mathbf{y}\|_2^2 \leq \epsilon L^2 \cdot \|\mathbf{S}\mathbf{X}\mathbf{w}^*\|_2^2$. In this second case, we can plug in the zero vector to the above minimization problem (it clearly satisfies the constraint) and conclude again that:

$$\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}\mathbf{y}\|_2^2 \leq \|\mathbf{S}f(\mathbf{X}\mathbf{0}) - \mathbf{S}\mathbf{y}\|_2^2 = \|\mathbf{S}\mathbf{y}\|_2^2 \leq \epsilon L^2 \|\mathbf{S}\mathbf{X}\mathbf{w}^*\|_2^2 \leq 2\epsilon L^2 \|\mathbf{X}\mathbf{w}^*\|_2^2.$$

The last inequality follows from the subspace embedding inequality from Lemma 1. Not also above that we used above that $f(\mathbf{X}\mathbf{0}) = f(\mathbf{0}) = \mathbf{0}$. ☐

With Claim 1 in place, we are ready to prove our main result.

Proof of Theorem 1. First note that, without loss of generality, we can assume that $\mathbf{X}$ has orthonormal columns. In particular, if $\mathbf{X}$ is not orthonormal, we can write it as $\mathbf{X} = \mathbf{Q}\mathbf{R}$ where $\mathbf{Q} \in \mathbb{R}^{n \times \text{rank}(\mathbf{X})}$ has orthonormal columns and $\mathbf{R}$ is a square full-rank matrix. The leverage scores of $\mathbf{Q}$ are equal to those of $\mathbf{X}$. Moreover, any solution $\hat{\mathbf{w}}$ to (1) has a corresponding solution $\mathbf{R}\hat{\mathbf{w}}$ to the minimization problem if $\mathbf{X}$ were replaced by $\mathbf{Q}$. So solving the above problem is equivalent to first explicitly orthogonalizing $\mathbf{X}$ and solving the same problem.

Next, we use the fact that for any vectors $\mathbf{a}$ and $\mathbf{b}$, $\|\mathbf{a}+\mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$ to bound:

$$\begin{aligned} \|f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{y}\|_2^2 &\leq 2\,\|f(\mathbf{X}\hat{\mathbf{w}}) - f(\mathbf{X}\mathbf{w}^*)\|_2^2 + 2\,\|f(\mathbf{X}\mathbf{w}^*) - \mathbf{y}\|_2^2 \\ &\leq 2\,\|f(\mathbf{X}\hat{\mathbf{w}}) - f(\mathbf{X}\mathbf{w}^*)\|_2^2 + 2OPT. \end{aligned} \tag{3}$$

We focus on bounding the first term. To do so, we first observe that, thanks to the constraint imposed in (1), the norm of $\hat{\mathbf{w}}$ can be bounded. In particular, we claim that with probability $49/50$,

$$\|\hat{\mathbf{w}}\|_2^2 \leq \frac{100}{\epsilon L^2} \cdot \|\mathbf{y}\|_2^2. \tag{4}$$

To see that this is the case, note that under our assumption that $\mathbf{X}$ is orthogonal, we have $\|\hat{\mathbf{w}}\|_2^2 = \|\mathbf{X}\hat{\mathbf{w}}\|_2^2$. We can bound $\|\mathbf{X}\hat{\mathbf{w}}\|_2^2$ as follows:

$$\begin{aligned} \|\mathbf{X}\hat{\mathbf{w}}\|_2^2 &\leq 2\,\|\mathbf{S}\mathbf{X}\hat{\mathbf{w}}\|_2^2 \quad \text{(Lemma 1)} \\ &\leq 2\frac{1}{\epsilon \cdot L^2}\,\|\mathbf{S}\mathbf{y}\|_2^2 \quad \text{(From the constraint in (1))} \\ &\leq \frac{100}{\epsilon \cdot L^2}\,\|\mathbf{y}\|_2^2 \quad \text{(Markov's inequality)} \end{aligned}$$

In the last inequality, we used that $\mathbb{E}[\|\mathbf{S}\mathbf{y}\|_2^2] = \|\mathbf{y}\|_2^2$, which holds regardless of the choice of probabilities used to construct $\mathbf{S}$. Since $\hat{\mathbf{w}}$ lies in $\mathcal{B}(R)$, where $R = \frac{100}{\epsilon L^2} \cdot \|\mathbf{y}\|_2^2$, we can apply Lemma 3 to conclude that, as long as $m \geq c\frac{d^2 \log(1/\epsilon)}{\epsilon^2}$,

$$\begin{aligned} \|f(\mathbf{X}\hat{\mathbf{w}}) - f(\mathbf{X}\mathbf{w}^*)\|_2^2 &\leq 2\,\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}f(\mathbf{X}\mathbf{w}^*)\|_2^2 + \epsilon\|\mathbf{y}\|_2^2 + \epsilon^2 L^2\,\|\mathbf{X}\mathbf{w}^*\|_2^2 \\ &\leq 4\,\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}\mathbf{y}\|_2^2 + 4\,\|\mathbf{S}f(\mathbf{X}\mathbf{w}^*) - \mathbf{S}\mathbf{y}\|_2^2 + \epsilon\|\mathbf{y}\|_2^2 + \epsilon^2 L^2\,\|\mathbf{X}\mathbf{w}^*\|_2^2 \\ &\leq 4\,\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}\mathbf{y}\|_2^2 + C \cdot OPT + \epsilon\|\mathbf{y}\|_2^2 + \epsilon^2 L^2\,\|\mathbf{X}\mathbf{w}^*\|_2^2. \end{aligned}$$

As in the proof of Claim 1, the last inequality follows with probability $49/50$ via Markov's inequality since $\mathbb{E}\left[\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}\mathbf{y}\|_2^2\right] = \|f(\mathbf{X}\mathbf{w}^*) - \mathbf{y}\|_2^2 = OPT$.

Next we apply Claim 1 to bound $\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}\mathbf{y}\|_2^2 \leq O\left(OPT + \epsilon L^2\|\mathbf{X}\mathbf{w}^*\|_2^2\right)$. So overall, we conclude that for a constant C,

$$\|f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{y}\|_2^2 \leq C \cdot \left(OPT + \epsilon L^2\|\mathbf{X}\mathbf{w}^*\|_2^2 + \epsilon\|\mathbf{y}\|_2^2\right). \tag{5}$$

By triangle inequality, we have that $\|\mathbf{y}\|_2 =\leq 2OPT + 2\|f(\mathbf{X}\mathbf{w}^*)\|_2^2 \leq 2OPT + 2L^2\|\mathbf{X}\mathbf{w}^*\|_2^2$. Using this fact and plugging (5) into (3) yields the theorem. □

A.1 Concentration Bounds

In our main proof, we use several concentration results that follow from leverage score sampling. The first is a standard "subspace embedding" for a matrix $\mathbf{X}$.

Lemma 1 (Subspace Embedding (see e.g. Theorem 17 in Woodruff [2014])**.** *Given* $\mathbf{X} \in \mathbb{R}^{n \times d}$ *with leverage scores* $\tau_1, \ldots, \tau_n$*, let* $p_i = \tau_i / \sum_i \tau_i$*. Let* $\mathbf{S} \in \mathbb{R}^{m \times n}$ *be a sampling matrix constructed as in Definition 2 using the probabilities* $p_1, \ldots, p_n$*. For any* $0 < \gamma < 1$*, as long as* $m \geq c \cdot d\log(d/\delta)/\gamma^2$ *for some fixed constant c, then with probability* $1 - \delta$ *we have that simultaneously for all* $\mathbf{w} \in \mathbb{R}^d$*,*

$$(1-\gamma)\|\mathbf{X}\mathbf{w}\|_2^2 \leq \|\mathbf{S}\mathbf{X}\mathbf{w}\|_2^2 \leq (1+\gamma)\|\mathbf{X}\mathbf{w}\|_2^2.$$

Lemma 1 establishes that, with high probability, leverage score sampling preserves the norm of any vector $\mathbf{X}\mathbf{w}$ in the column span of $\mathbf{X}$. This guarantee can be proven using an argument that reduces to a matrix Chernoff bound [Spielman and Srivastava, 2011] and is a critical component in previous active learning guarantees for leverage score sampling when fitting linear functions [Sarlos, 2006].

Our next two lemmas establish similar results to Lemma 1, but for preserving the norm of non-linear ridge functions involving $\mathbf{X}$.

Lemma 2. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be an L-Lipschitz activation function applied entrywise to the vector* $\mathbf{X}\mathbf{w}$ *and let* $\mathbf{S} \in \mathbb{R}^{m \times n}$ *be an importance sampling matrix chosen with probabilities* $p_1, \ldots, p_n$ *where*

$p_i = \tau_i(\mathbf{X})/\operatorname{rank}(\mathbf{X})$. *As long as* $m \geq \frac{3d\log(2/\delta)}{\epsilon^2}$*, then for any fixed pair of vectors* $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$*, with probability* $\geq 1-\delta$*,*

$$\|f(\mathbf{X}\mathbf{w}_1) - f(\mathbf{X}\mathbf{w}_2)\|_2^2 - \epsilon L^2 \|\mathbf{X}\mathbf{w}_1 - \mathbf{X}\mathbf{w}_2\|_2^2 \leq \|\mathbf{S}f(\mathbf{X}\mathbf{w}_1) - \mathbf{S}f(\mathbf{X}\mathbf{w}_2)\|_2^2 \leq \|f(\mathbf{X}\mathbf{w}_1) - f(\mathbf{X}\mathbf{w}_2)\|_2^2 + \epsilon L^2 \|\mathbf{X}\mathbf{w}_1 - \mathbf{X}\mathbf{w}_2\|_2^2.$$

Proof. Let $\mathbf{x}_i$ denote the i^{th} row of $\mathbf{X}$ and let $\mathbf{u} = f(\mathbf{X}\mathbf{w}_1) - f(\mathbf{X}\mathbf{w}_2)$ and $\mathbf{v} = \mathbf{X}\mathbf{w}_1 - \mathbf{X}\mathbf{w}_2$. Since f is L-Lipschitz, for every $i \in [n]$, we have that

$$u_i = |f(\langle \mathbf{x}_i, \mathbf{w}_1\rangle) - f(\langle \mathbf{x}_i, \mathbf{w}_2\rangle)|_i \leq L \cdot |\langle \mathbf{x}_i, \mathbf{w}_1\rangle - \langle \mathbf{x}_i, \mathbf{w}_2\rangle|_i \leq L\mathbf{v}_i. \tag{6}$$

Let $j_i \in [n]$ be the index of the row from $\mathbf{X}$ selected by the i^{th} row in $\mathbf{S}$. We have that $\|\mathbf{S}\mathbf{u}\|_2^2 = \sum_{i=1}^m \frac{u_{j_i}^2}{m \cdot p_{j_i}}$, where $p_{j_i} = \tau_{j_i}(\mathbf{X})/\operatorname{rank}(\mathbf{X})$. We thus have that $\mathbb{E}\|\mathbf{S}\mathbf{u}\|_2^2 = \|\mathbf{u}\|_2^2$. Moreover, we can bound the variance in each term of the sum. In particular, we have that:

$$\operatorname{Var}\left[\frac{u_{j_i}^2}{p_{j_i}}\right] \leq \mathbb{E}\left[\left(\frac{u_{j_i}^2}{p_{j_i}}\right)^2\right] = \sum_{k=1}^n \frac{u_k^4}{p_k^2} \cdot p_k = \sum_{k=1}^n \frac{L^4 v_k^4 \operatorname{rank}(\mathbf{X})}{\tau_k(\mathbf{X})}.$$

In the last step we have used the upper bound from (6), and the fact that $p_k = \tau_k(\mathbf{X})/\operatorname{rank}(\mathbf{X})$. From the definition of leverage scores (Definition 1), and the fact that $\mathbf{v}$ lies in the span of $\mathbf{X}$, we have that $\tau_k(\mathbf{X}) \geq \frac{v_k^2}{\|\mathbf{v}\|_2^2}$. So we can further upper bound the variance as follows:

$$\operatorname{Var}\left[\frac{u_{j_i}^2}{p_{j_i}}\right] \leq L^4 \cdot \sum_{k=1}^n v_k^2 \|\mathbf{v}\|_2^2 \operatorname{rank}(\mathbf{X}) = L^4 \cdot \|\mathbf{v}\|_2^4 \cdot \operatorname{rank}(\mathbf{X}) \leq L^4 \cdot d\|\mathbf{v}\|_2^4.$$

Moreover, we have that with probability 1, $\frac{u_{j_i}^2}{p_{j_i}} \leq \max_k L^2 \cdot \frac{v_k^2 \operatorname{rank}(\mathbf{X})}{\tau_k(\mathbf{X})} \leq L^2 \cdot d\|\mathbf{v}\|_2^2$.

Finally, applying Bernstein's to the sum $\|\mathbf{S}\mathbf{u}\|_2^2 = \frac{1}{m}\sum_{i=1}^m \frac{u_{j_i}^2}{p_{j_i}}$, we have that:

$$\Pr\left[\left|\|\mathbf{S}\mathbf{u}\|_2^2 - \|\mathbf{u}\|_2^2\right| \geq t/m\right] \leq 2\exp\left(-\frac{t^2/2}{m \cdot L^4 \cdot d\|\mathbf{v}\|_2^4 + t \cdot L^2 \cdot d\|\mathbf{v}\|_2^2/3}\right).$$

Setting $m = \frac{3d\log(2/\delta)}{\epsilon^2}$ and $t = m \cdot \epsilon\|\mathbf{v}\|_2^2 \cdot L^2$ and plugging in we have:

$$\Pr\left[\left|\|\mathbf{S}\mathbf{u}\|_2^2 - \|\mathbf{u}\|_2^2\right| \geq \epsilon L^2\|\mathbf{v}\|_2^2\right] \leq 2\exp\left(-\frac{\frac{1}{2}m^2\epsilon^2\|\mathbf{v}\|_2^4 L^4}{m \cdot L^4 \cdot d\|\mathbf{v}\|_2^4 + m\epsilon L^4 \cdot d\|\mathbf{v}\|_2^4/3}\right) \leq \delta.$$

This completes the bound. □

Lemma 3. *Given* $\mathbf{X}$*,* f*, and* $\mathbf{y}$*, let* $\mathbf{w}^* = \arg\min_{\mathbf{w}} \|f(\mathbf{X}\mathbf{w}) - y\|_2^2$ *and let* R *be a fixed radius. Let* $\mathbf{S} \in \mathbb{R}^{m\times n}$ *be an importance sampling matrix chosen with probabilities* $p_1, \ldots, p_n$ *where* $p_i = \tau_i(\mathbf{X})/\operatorname{rank}(\mathbf{X})$*. As long as* $m \geq c\frac{d^2\log(1/\epsilon)}{\epsilon^2}$ *for* $\epsilon < 1$ *and fixed constant* c*, then with probability* $49/50$*,*

$$\|f(\mathbf{X}\hat{\mathbf{w}}) - f(\mathbf{X}\mathbf{w}^*)\|_2^2 \leq 4 \cdot \|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}f(\mathbf{X}\mathbf{w}^*)\|_2^2 + \epsilon^2 L^2 R^2 + \epsilon^2 L^2 \|\mathbf{X}\mathbf{w}^*\|_2^2$$

for all $\hat{\mathbf{w}} \in \mathcal{B}^d(R)$*.*

Proof. Let N be an (ϵR)-net in the Euclidean norm on $\mathcal{B}(R)$. I.e. for every $\mathbf{v} \in \mathcal{B}(R)$, there should be some point $\mathbf{z} \in N$ such that $\|\mathbf{z} - \mathbf{v}\|_2 \leq \epsilon R$. It is well known that such an N exists with cardinality $|N| \leq \left(1 + \frac{2}{\epsilon})\right)^d$ (see e.g. Lemma 5.2 in Vershynin [2012]). Applying Lemma 2 with $\delta = \frac{1}{50|N|}$ and combining with a union bound, we conclude that as long as $m \geq c\frac{d^2\log(1/\epsilon)}{\epsilon^4}$ for a fixed constant c, then with probability $49/50$, for all $\mathbf{z} \in N$,

$$\|f(\mathbf{X}\mathbf{z}) - f(\mathbf{X}\mathbf{w}^*)\|_2^2 \in \left[\|\mathbf{S}f(\mathbf{X}\mathbf{z}) - \mathbf{S}f(\mathbf{X}\mathbf{w}^*)\|_2^2 \pm \epsilon^2 L^2 \|\mathbf{X}\mathbf{z} - \mathbf{X}\mathbf{w}^*\|_2^2\right]. \tag{7}$$

Now, let $\mathbf{z}^*$ be the closest point to $\hat{\mathbf{w}}$ in N. I.e., $\mathbf{z}^* = \arg\min_{z \in N} \|\mathbf{z} - \hat{\mathbf{w}}\|_2$. Applying (7) and the fact that for any two vectors $\mathbf{a}, \mathbf{b}$, $\|\mathbf{a} + \mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2 + 2\|\mathbf{b}\|_2^2$, we have:

$$
\begin{aligned}
\|f(\mathbf{X}\hat{\mathbf{w}}) - f(\mathbf{X}\mathbf{w}^*)\|_2^2 &\leq 2\left\|f(\mathbf{X}\mathbf{z}^*) - f(\mathbf{X}\mathbf{w}^*)\right\|_2^2 + 2\left\|f(\mathbf{X}\hat{\mathbf{w}}) - f(\mathbf{X}\mathbf{z}^*)\right\|_2^2 \\
&\leq 2\left\|\mathbf{S}f(\mathbf{X}\mathbf{z}^*) - \mathbf{S}f(\mathbf{X}\mathbf{w}^*)\right\|_2^2 + 2\epsilon^2 L^2 \|\mathbf{X}\mathbf{z}^* - \mathbf{X}\mathbf{w}^*\|_2^2 + 2\left\|f(\mathbf{X}\hat{\mathbf{w}}) - f(\mathbf{X}\mathbf{z}^*)\right\|_2^2 \\
&\leq 4\left\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}f(\mathbf{X}\mathbf{w}^*)\right\|_2^2 + 4\left\|\mathbf{S}f(\mathbf{X}\mathbf{z}^*) - \mathbf{S}f(\mathbf{X}\hat{\mathbf{w}})\right\|_2^2 + 2\epsilon^2 L^2 \|\mathbf{X}\mathbf{z}^* - \mathbf{X}\mathbf{w}^*\|_2^2 \\
&\quad + 2\left\|f(\mathbf{X}\hat{\mathbf{w}}) - f(\mathbf{X}\mathbf{z}^*)\right\|_2^2 \\
&\leq 4\left\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}f(\mathbf{X}\mathbf{w}^*)\right\|_2^2 + 4\left\|f(\mathbf{X}\mathbf{z}^*) - f(\mathbf{X}\hat{\mathbf{w}})\right\|_2^2 + 6\epsilon^2 L^2 \|\mathbf{X}\mathbf{z}^* - \mathbf{X}\mathbf{w}^*\|_2^2 + \\
&\quad + 2L^2 \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{X}\mathbf{z}^*\|_2^2 \\
&\leq 4\left\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}f(\mathbf{X}\mathbf{w}^*)\right\|_2^2 + 4L^2\left\|\mathbf{X}\mathbf{z}^* - \mathbf{X}\hat{\mathbf{w}}\right\|_2^2 + 6\epsilon^2 L^2 (R + \|\mathbf{X}\mathbf{w}^*\|_2)^2 + \\
&\quad + 2L^2 \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{X}\mathbf{z}^*\|_2^2 \\
&\leq 4\left\|\mathbf{S}f(\mathbf{X}\hat{\mathbf{w}}) - \mathbf{S}f(\mathbf{X}\mathbf{w}^*)\right\|_2^2 + 4\epsilon^2 L^2 R^2 + 12\epsilon^2 L^2 R^2 + 12\|\mathbf{X}\mathbf{w}^*\|_2^2 + 2\epsilon^2 L^2 R^2.
\end{aligned}
$$

Combining terms and adjusting constants on ϵ yields the bound. $\square$