Learning Single-Index Models with Shallow Neural Networks

Alberto Bietti
New York University
Meta AI

Joan Bruna
New York University

Clayton Sanford
Columbia University

Min Jae Song
New York University

Abstract

Single-index models are a class of functions given by an unknown univariate "link" function applied to an unknown one-dimensional projection of the input. These models are particularly relevant in high dimension, when the data might present low-dimensional structure that learning algorithms should adapt to. While several statistical aspects of this model, such as the sample complexity of recovering the relevant (one-dimensional) subspace, are well-understood, they rely on tailored algorithms that exploit the specific structure of the target function. In this work, we introduce a natural class of shallow neural networks and study its ability to learn single-index models via *gradient flow*. More precisely, we consider shallow networks in which biases of the neurons are frozen at random initialization. We show that the corresponding optimization landscape is benign, which in turn leads to generalization guarantees that match the near-optimal sample complexity of dedicated semi-parametric methods.

1 Introduction

High-dimensional learning with both computational and statistical guarantees, which is particularly relevant given the current scaling trends, remains an outstanding challenge. One important question which has received considerable attention is on understanding the advantages of using non-linear learning models, such as neural networks, over more mature (from a theoretical standpoint) counterparts, such as kernel methods [56, 57, 84, 22]. Perhaps surprisingly, the question remains largely open even for shallow neural networks.

While approximation benefits of shallow neural networks over non-adaptive kernels have been known for decades [9, 69], another important piece of the theoretical puzzle was provided by [5], whose analysis hinted at an inherent *statistical* advantage of neural networks for extracting information from high-dimensional data with a "hidden" low-dimensional structure. Providing *computational* guarantees, the remaining piece of this puzzle, is still mostly unresolved.

Several computational hardness results for learning functions that can be efficiently approximated by shallow neural networks have been established in the literature [30, 41, 28, 78, 20], ruling out positive results in the general setting. On the other hand, progress has been made on the positive side [1, 2, 73] by focusing on function classes with strong structural properties, thereby showcasing the adaptive representation learning capabilities of neural networks.

This work aligns with the latter effort, and focuses on the class of *single-index models*. Single-index models are high-dimensional functions $F : \mathbb{R}^d \to \mathbb{R}$ of the form $F_*(x) = f_*(\langle \theta^*, x \rangle)$, where *both* the univariate "link" function $f_* : \mathbb{R} \to \mathbb{R}$ and relevant (one-dimensional) subspace $\theta^* \in \mathbb{S}^{d-1}$ are unknown. These models have been extensively studied in the statistics literature [45, 48, 26, 46, 33], leading to dedicated algorithmic procedures, and can be provably approximated with shallow neural networks without incurring in a curse-of-dimensionality [5]. In contrast, the analysis

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

of neural network learning using gradient-based methods has focused on the so-called "Teacher-Student" setup [43, 39, 85, 86, 10], where the link function f_* is assumed to be known and used as the activation function for the student.

The mix of a high-dimensional parametric component (the hidden direction) with a non-parametric one in low dimension (the link function) in single-index models naturally suggests a shallow neural network architecture where the inner weights are shared and "active", while the biases are "lazy" [23]. We instantiate such an architecture by *freezing* the biases at random initialization, and analyze gradient descent on the free parameters in the continuous-time limit.

Our main results establish that as soon as the width N of the network is larger than a quantity which depends solely on smoothness properties of the (univariate) link function f_*, gradient flow recovers the unknown direction θ^* with near optimal sample complexity $O(d^s)$, where s is the so-called *information-exponent* of the link function [10] (at least when $s \geq 3$, see Theorem 6.1 for the formal result), and approximates the univariate link function f_* near-optimally (see Corollary 6.4). The information exponent roughly captures the signal strength, which here refers to the alignment between the network direction θ and the hidden direction θ^*, at typical initializations.

The success of gradient flow relies on the benign optimization landscape of the empirical loss, though the presence of degenerate saddles necessitates a careful analysis leveraging uniform convergence of the empirical landscape [60, 35]. We show that gradient flow over our proposed neural network architecture solves two distinct problems—univariate non-parametric kernel ridge regression and non-convex optimization in high dimension—*simultaneously* and *efficiently*, cementing its role as a versatile algorithm for high-dimensional learning. We illustrate our theoretical results with experiments in Section A.

2 Related Work

Single-index and multi-index models. A useful modeling assumption in high-dimensional regression is that the regression function $F(x) = \mathbb{E}[y|x]$ only depends on one or a few directions. This leads to the single-index model $F(x) = f_*(\langle \theta^*, x\rangle)$, and multi-index model $F(x) = f_*(\langle \theta_1^*, x\rangle, \ldots, \langle \theta_k^*, x\rangle)$, with k typically much smaller than the dimension. Such models have a long history in the statistics literature and different methods exist for various estimation problems, including projection pursuit [36, 47], slicing [52], gradient-based estimators [53], and moment-based estimators [26]. When the function f_* is also to be estimated, we face a semi-parametric problem involving parameter recovery of θ^* and non-parametric estimation of f_*. Our work is closely related to [33], which also characterizes the population landscape of certain objectives by leveraging Gaussian data. Multi-index models are also studied in [5] in the context of shallow neural networks, where it is shown that certain models of infinite-width shallow networks can adapt to such low-dimensional structure, though no tractable algorithms are introduced.

The works [8, 19, 66] show that certain neural networks trained close to initialization can learn certain sparse polynomials which take the form of multi-index models, but such networks do not directly aim to learn target directions. Recently, [1] studied the learnability of functions on the hypercube by shallow neural networks with stochastic gradient descent and introduces the *merged staircase property*, which provides necessary and sufficient conditions for learnability with linear sample complexity $n = O(d)$. While they learn a broader class of functions (including multi-index model) for a more efficient sample complexity regime ($O(d)$ vs $O(d^s)$), their setup is restricted to simple discrete data distributions, while our work captures the regime of semi-parametric estimation by considering Gaussian data without the sparsity requirements on F implied by their merged staircase property.

Concurrently to our work, [4] and [27] studied the learnability of certain single and multi-index models on Gaussian data with shallow networks, by performing a single gradient step on the first layer before fitting the second layer. While the single step is sufficient to provide a separation from kernel methods in these works, we show that optimizing both layers jointly until convergence (for a more simplistic architecture) can significantly improve the rates, by fully decoupling the non-parametric learning part from the high-dimensional inference of the hidden direction. Finally, recently [64] studied the ability of shallow neural networks to learn certain single and multi-index models, showing in particular that SGD-trained ReLU networks can learn single-index functions with monotonic index function (corresponding in our setting to $s = 1$) with linear (up to logarithmic factors) sam-

ple complexity. Our results therefore extend such positive guarantees to a broader class of index functions with arbitrary information exponent.

Teacher–student models. In the context of neural networks, several works have considered the teacher–student setting [34], where the target function F takes the form of a neural network with the same activation as the network used for learning [43, 39, 85, 77, 86, 10, 82]. In this case the problem does not involve non-parametric estimation as in our setup, but this line of work often involves studying optimization landscapes similar to ours for estimating hidden directions. In particular, the population landscape appearing in [10] is similar to ours, based on Hermite coefficients of link functions. The follow-up work [11] extends this to multiple student neuron directions, but still focuses on parametric rather than non-parametric statistical problems.

Kernels and random features. In order to obtain non-parametric estimation guarantees for learning the target function f_* of the single-index model, our work builds on the kernel methods literature for approximation and non-parametric regression [76, 13, 7], their links with neural networks [24, 5], and in particular on random feature approximation [70, 6, 72, 61].

Non-convex and non-smooth optimization landscapes. There is a vast literature studying tractable non-convex optimization landscapes, arising from high-dimensional statistics and statistical physics [59, 58, 37, 12, 79, 17, 39, 71, 55]. A particular aspect of our setup is that the optimization landscape does *not* have the strict saddle property, which is often leveraged to establish global convergence [50]. [60, 35] study concentration properties of the empirical landscape to the population one for non-convex problems including generalized linear models. Our results rely on similar concentration analyses, but depart from these previous work by also allowing optimization of the link function, and by supporting the non-smoothness arising from ReLU activations. On the algorithmic side, we consider gradient flows on non-convex and non-smooth landscapes, which require careful technical treatment, but have been studied by previous works [32, 29, 49]. We refer the interested reader to Appendix F for more details on this technical issue.

3 Preliminaries

We focus on regression problems under a single-index model with Gaussian input data. Specifically, we assume d-dimensional inputs $x \sim \gamma_d := \mathcal{N}(0, I_d)$, and labels

$$y = F^*(x) + \xi = f_*(\langle \theta^*, x \rangle) + \xi \ ,$$

where $\theta^* \in \mathbb{S}^{d-1}$ and $\xi \sim \mathcal{N}(0, \sigma^2)$ is an independent, additive Gaussian noise. The normalization $\theta^* \in \mathbb{S}^{d-1}$ is to ensure that both f_* and θ^* are identifiable.

Shallow networks and random features. We consider learning algorithms based on shallow neural networks of the form

$$G(x; c, \theta) = c^\top \Phi(\langle \theta, x \rangle) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} c_i \phi(\varepsilon_i \langle \theta, x \rangle - b_i) \ ,$$

with $\theta \in \mathbb{S}^{d-1}$, where $\phi(u) = \max\{0, u\}$ is the ReLU activation, $b_i \sim \mathcal{N}(0, \tau^2)$ (we assume $\tau > 1$) are random "bias" scalars that are frozen throughout training, and ε_i are random signs with Rademacher distribution (i.e., uniform over $\{\pm 1\}$), independent from b_i, and also frozen during training. The resulting vector of random features is thus $\Phi(u) = \frac{1}{\sqrt{N}}(\phi(\varepsilon_i u - b_i))_{i \in [N]}$.

The choice of ReLU activation is motivated by its popularity among practitioners. As we shall see, the fact that ϕ is non-smooth introduces some technical challenges, but its piece-wise linear structure enables dedicated arguments both in terms of approximation as well as in the study of the optimization landscape. In Appendix G we discuss how our main results are affected when replacing the ReLU by a smooth activation, especially when choosing it such that ϕ' is Lipschitz.

Empirical risk minimization. The supervised learning task is to estimate F^* (and therefore both f_* and θ^*) from samples $\{(x_i, y_i)\}_{i=1\dots n}$. We will focus on mean-squared error with Tychonov regularisation, determined by the following losses.

Definition 3.1 (Population risk)**.** *We define the ℓ^2-regularized population loss by*

$$L(c,\theta) = \mathop{\mathbb{E}}_{x,y}[(y - G(x;c,\theta))^2] + \lambda\|c\|^2 = \mathop{\mathbb{E}}_{x\sim\gamma_d}[(F^*(x) - G(x;c,\theta))^2] + \sigma^2 + \lambda\|c\|^2 \ .$$

Definition 3.2 (Empirical risk)**.** *We define the ℓ^2-regularized empirical loss by*

$$L_n(c,\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(c^\top\Phi(\langle\theta,x_i\rangle) - y_i\right)^2 + \lambda\|c\|^2 \ . \tag{1}$$

Hermite decomposition. Given that our data is normally distributed, we consider the family of (normalized) Hermite polynomials $\{h_j\}_{j\in\mathbb{N}}$, which form an orthonormal basis of $L^2(\gamma)$, the space of squared-integrable function under the Gaussian measure $\gamma := \mathcal{N}(0,1)$. We will denote by $f_* = \sum_j \alpha_j h_j$ the Hermite decomposition of the target link function, which we assume is in $L^2(\gamma)$ henceforth.

We apply the following useful properties of Hermite polynomials [67, Chapter 11.2]:

$$h_j' = \sqrt{j}h_{j-1} \quad \text{and} \quad \langle h_j(\langle\theta,\cdot\rangle), h_{j'}(\langle\theta',\cdot\rangle)\rangle_{\gamma_d} = \delta_{j,j'}\langle\theta,\theta'\rangle^j,$$

where $\langle\cdot,\cdot\rangle_{\gamma_d}$ is the inner product in $L^2(\gamma_d)$ and δ the Kronecker delta. We will assume throughout that $\|f_*\|_\gamma^2 = \sum_j \alpha_j^2$, $\|f_*'\|_\gamma^2 = \sum_j j\alpha_j^2$, and $\|f_*''\|_\gamma^2 = \sum_j j(j-1)\alpha_j^2$ are all finite (see Assumption 5.2). We will also consider the weighted Sobolev space $H^2(\gamma)$, which contains functions $f = \sum_j \alpha_j h_j \in L^2(\gamma)$ such that $\sum_j j^2|\alpha_j|^2 < \infty$.

Random features to Hermite coefficients. To precisely characterize the landscape of $L(c,\theta)$, we introduce notation to represent each random feature function in terms of Hermite polynomials h_j. We define the linear integral operator $\mathcal{T} : L^2(\gamma) \to \mathbb{R}^N$ by

$$(\mathcal{T}f)_i := \langle f, \phi_{b_i}^{\epsilon_i}\rangle_\gamma := \frac{1}{\sqrt{N}}\mathop{\mathbb{E}}_{z\sim\gamma}[f(z)\phi(\varepsilon_i z - b_i)] \ , \ i \in [N] \ . \tag{2}$$

Note that $\mathcal{T}$ has rank N almost surely.

The operator $\mathcal{T}$ has an adjoint $\mathcal{T}^* : \mathbb{R}^N \to L^2(\gamma)$ defined as $(\mathcal{T}^*c)(u) = \frac{1}{\sqrt{N}}\sum_{i=1}^N c_i\phi(\varepsilon_i u - b_i) = c^\top\Phi(u)$. Finally, for any $j \in \mathbb{N}$, let $\mathcal{T}_j \in \mathbb{R}^N$ defined as $\mathcal{T}_j = \mathcal{T}h_j$. We can then write down the Hermite expansion of the student network:

$$G(x;c,\theta) = c^\top\Phi(\langle\theta,x\rangle) = \sum_{j\geq 0}\langle c,\mathcal{T}_j\rangle h_j(\langle\theta,x\rangle).$$

Denoting $m = \langle\theta,\theta^*\rangle$, the regularized population objective can be expressed as

$$L(c,\theta) = \sum_j \alpha_j^2 + \sum_j \langle c,\mathcal{T}_j\rangle^2 - 2\sum_j \alpha_j\langle c,\mathcal{T}_j\rangle m^j + \lambda\|c\|^2, \tag{3}$$

where the term $\sum_j \alpha_j^2$ is a constant that can be ignored. Let $Q := \mathcal{T}\mathcal{T}^* \in \mathbb{R}^{N\times N}$ be a feature covariance matrix and $Q_\lambda = Q + \lambda I$. Note that Q_λ is positive definite for $\lambda > 0$. We define the regularized projection $\hat{P}_\lambda = \hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}$ onto the random feature space for $\hat{\Sigma} = \mathcal{T}^*\mathcal{T}$. Observe that $(\hat{P}_\lambda f)(u) = c^{*\top}\Phi(u)$, where $c^* \in \mathbb{R}^N$ is the solution to the following objective:

$$\min_{c\in\mathbb{R}^N}\left\|f - c^\top\Phi\right\|_\gamma^2 + \lambda\|c\|_2^2 \ . \tag{4}$$

Geometry on the sphere. Because the direction θ is constrained to lie on the sphere, our optimization algorithms rely on spherical (Riemannian) gradients, which are defined as follows:

$$\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta) = \Pi_{\theta^\perp}\nabla_\theta L(c,\theta) \ ,$$

where $\Pi_{\theta^\perp}v = v - \langle\theta,v\rangle\theta$. We say that (c,θ) is a *critical point* of L if $\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta) = 0$ and $\nabla_c L(c,\theta) = 0$.

4 Univariate Approximation using Random Features

Before addressing the high-dimensionality of the learning problem, we first focus on the non-parametric approximation aspects of the univariate link function. As usual, we start by deriving approximation rates of the infinitely-wide model, given by a RKHS, and then establish approximation rates for our random feature model.

Univariate RKHS. If we fix the direction θ, learning c alone may be seen as a random feature model [70] that approximates a kernel method with the following kernel.

$$\kappa(u,v) = \mathbb{E}_{b\sim\gamma_\tau,\varepsilon\sim\mathrm{Rad}}[\phi(\varepsilon u - b)\phi(\varepsilon v - b)] \ , \ u,v \in \mathbb{R} \ , \tag{5}$$

where $\gamma_\tau = \mathcal{N}(0,\tau^2)$ is the Gaussian measure on $\mathbb{R}$ with mean zero and variance τ^2, and ε is a random sign with a Rademacher distribution. The kernel κ corresponds to an RKHS $\mathcal{H}$, given by

$$\mathcal{H} := \left\{ f : \mathbb{R} \to \mathbb{R} \mid \ f(u) = \frac{1}{\sqrt{2}} \int [c_+(b)\phi(u-b) + c_-(b)\phi(-u-b)]\, d\gamma_\tau(b), \ c_+, \, c_- \in L^2(\gamma_\tau) \right\} \ .$$

The following lemma characterizes the corresponding RKHS norm $\|\cdot\|_{\mathcal{H}}$, and follows from Theorem B.8, by noting that $\kappa(u,v) = \langle \psi(u), \psi(v)\rangle_{L^2(\gamma_\tau)^2}$, with $\psi(u) = \frac{1}{\sqrt{2}}(\phi(u-\cdot), \phi(-u-\cdot))$.

Lemma 4.1 (RKHS norm)**.** *The RKHS norm in $\mathcal{H}$ is given by*

$$\|f\|_{\mathcal{H}}^2 = \inf \left\{ \|c_+\|_{\gamma_\tau}^2 + \|c_-\|_{\gamma_\tau}^2 ; \ f(u) = \frac{1}{\sqrt{2}} \int [c_+(b)\phi(u-b) + c_-(b)\phi(-u-b)]\, d\gamma_\tau(b) \right\} \ . \tag{6}$$

The choice of ReLU for the activation function gives us more explicit control over the RKHS norm, based on Sobolev representations, as already exploited by several works [68, 5, 75].

Lemma 4.2 (RKHS norm bound)**.** *Let $f \in H^2(\gamma) \cap C^2(\mathbb{R})$ and $\tau > 1$. If f and f' both have polynomial growth and $\int \frac{|f''(t)|^2}{\gamma_\tau(t)} dt < \infty$, then $f \in \mathcal{H}$ with*

$$\|f\|_{\mathcal{H}}^2 \le 6\tau \left(\int \frac{|f''(t)|^2}{\gamma_\tau(t)} dt + \|f\|_\gamma^2 + 6\|f'\|_\gamma^2 + 2\langle f, f''\rangle_\gamma \right) \ . \tag{7}$$

The proof is in Appendix C.1.

RKHS approximation properties. Let $A(f,\lambda)$ be the (regularized) L^2 approximation error for functions in the space $\mathcal{H}$ with respect to the target function f and measure γ. Formally,

$$A(f,\lambda) := \min_{g\in\mathcal{H}} \|f-g\|_\gamma^2 + \lambda\|g\|_{\mathcal{H}}^2 \ .$$

We will now show that the approximation error of the RKHS corresponding to an infinite number of random features can be bounded in terms of the regularization λ and the γ-norm of the second derivative of the target function. For that purpose, we consider the following 'source' condition to ensure a polynomial approximation error in λ.

Assumption 4.3 (Containment in $L^4(\gamma)$)**.** *Let $\mathcal{F} = \{f \in H^2(\gamma) \mid f'' \in L^4(\gamma)\}$. We assume $f \in \mathcal{F}$ and define $K := \inf\left\{B \ge 1 \mid \|f''\|_{L^4(\gamma)} \le B\|f''\|_{L^2(\gamma)}\right\}$.*

Assumption 4.3 provides a sufficient condition for approximating f with functions in the RKHS. The family of approximants $\{h_M \in \mathcal{H} \mid M > 0\}$ we use in Lemma 4.4 are exactly equal to f on $[-M, M]$ and are linear outside of $[-M, M]$. The L^4 assumption on f'' ensures control over the RKHS norm of h_M. Note that by Jensen's inequality, $L^4(\gamma) \subset L^2(\gamma)$, so K is always well-defined for $f'' \in L^4(\gamma)$. Sigmoidal functions, compactly supported smooth functions, and, more generally, functions with polynomial growth satisfy Assumption 4.3.

Lemma 4.4 (RKHS approximation error)**.** *Let $\lambda \in (0,1)$ and $f \in \mathcal{F}$. Then, there exists a universal constant $C > 0$ such that*

$$A(f,\lambda) \le C\Big(\tau^{1+\beta}\|f''\|_4^2 \cdot \lambda^\beta + \lambda C_f^2\Big) \ , \tag{8}$$

where $\beta = \frac{1-1/\tau^2}{3+1/\tau^2}$ and $C_f = \max\{\|f\|_\gamma, \|f'\|_\gamma, \|f''\|_\gamma\}$.

The proof appears in Appendix C.2. This lemma allows us to control the RKHS approximation error of a target function in terms of their Hermite decompositions. The main technical difficulty is that the RKHS integral operator Σ does not diagonalise in the Hermite basis; we address this with a dedicated argument exploiting the RKHS Sobolev representation of Lemma 4.2. The assumption that $f'' \in L^4(\gamma)$ (Assumption 4.3) is sufficient for our purposes but not necessary for polynomial-in-λ approximation rates. In Section H, we show that the ReLU function $\phi(t) = \max(0, t)$, which is Lipschitz but not in $H^2(\gamma)$, as the *target* satisfies $A(\phi, \lambda) \lesssim \tau^2\lambda^{2/3}$ using a direct argument. Extending the class of functions approximable by $\mathcal{H}$ with polynomial-in-λ rate is an interesting future direction.

Random feature approximation. We now consider (finite) random feature approximations to functions in the RKHS. Lemma C.2 shows that the best possible loss of a linear combination of sufficiently many finite features is bounded above by the best approximation with infinitely many features with high probability. More specifically, as long as $N \gtrsim \lambda^{-1}$, the random feature approximation error behaves like the RKHS approximation error. The proof leverages the 'degrees of freedom' of the kernel and closely tracks [6].

5 Population Landscape under Frozen Random Biases

To understand optimization and generalization properties of gradient flow on the empirical loss $L_n(c, \theta)$, we first the study optimization landscape of the population loss $L(c, \theta)$. We characterize the critical points of the population loss and show that gradient flow on a shallow neural network of sufficient width N converges only when its direction vector θ is either parallel or orthogonal to the target direction θ^*. Importantly, the sufficient number of random features N depends on the ℓ^2-regularization parameter $\lambda \in (0, 1)$, but *not* on the input dimension d. In Section 6, we further show that sufficiently large n, the number of training samples, guarantees similar properties for the *empirical* landscape and thus has favorable generalization properties for most initializations.

One of the main measures of complexity for the target link function f_* is its information exponent (see e.g., [10]), defined as follows.

Definition 5.1 (Information exponent)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be any function such that $f \in L^2(\gamma)$. The* information exponent *of f, which we denote by s, is the index of the first non-zero Hermite coefficient. That is, $s := \min\{j \in \mathbb{N} : \alpha_j \neq 0\}$.*

We make the following regularity assumptions on the target link function f_* to ensure small approximation error by random features, and benign population and empirical landscape.

Assumption 5.2 (Regularity of f_*)**.** *We consider $f_* \in L^2(\gamma)$, with $f_* = \sum_j \alpha_j h_j$. Assume 1) f_* is Lipschitz, 2) $\sum_j j^4|\alpha_j|^2 < \infty$, and 3) $f_*''(z) := \sum_j \sqrt{(j+2)(j+1)}\alpha_{j+2}h_j(z)$ is in $L^4(\gamma)$ (Assumption 4.3)*

We also suppose w.l.o.g. that f_* is normalized so that $\|f_*\|_\gamma = 1$. To analyze the critical points of $L(c, \theta)$, we introduce the *projected population loss* $\bar{L}(\theta)$, which can be seen as a semiparametric least squares (SLS) objective [48].

$$\bar{L}(\theta) := \min_c L(c, \theta) \ .$$

Theorem 5.3 (Critical points of the population loss)**.** *Assume f_* satisfies Assumption 5.2 and has information exponent $s \geq 1$. For $\tau > 1$, and $\delta \in (0, 1)$, there exists $\lambda^* \leq 1$ depending only on τ and the target link function f_* and a universal constant $C > 0$ such that if*

$$\lambda < \lambda^* \qquad \text{and} \qquad N \geq \frac{C}{\lambda} \log\left(\frac{1}{\lambda\delta}\right) \tag{9}$$

then with probability $1 - \delta$ over the random biases b_j and signs ε_j, $j = 1 \dots N$, the set of first-order critical points $\Omega := \{(c, \theta) : \nabla_\theta^{\mathbb{S}^{d-1}} L(c, \theta) = 0,\ \nabla_c L(c, \theta) = 0\}$ satisfies:

1. *(orientation relative to θ^*) if $(c, \theta) \in \Omega$, then either $\theta \in \{\pm\theta^*\}$ or $\langle \theta, \theta^* \rangle = 0$.*
2. *(existence and uniqueness of c) if $\nabla_\theta^{\mathbb{S}^{d-1}} \bar{L}(\theta) = 0$, then there exists a unique $c \in \mathbb{R}^N$ such that $(c, \theta) \in \Omega$.*

Theorem 5.3 thus establishes a benign optimization landscape in the population limit, rejoining several known non-convex objectives with similar behavior, such as tensor decomposition [40] or matrix completion [38]. Importantly, this optimization landscape has the same topology as the one that arises from using the Hermite basis, the tailored choice for data generated by a single-index model in Gaussian space [33, Theorem 5], instead of random scalar features. We view this as an interesting robustness property of shallow neural networks, at least in the regime where biases are randomly frozen.

6 Empirical Landscape and Generalization Guarantees

Section 5 shows that the population landscape has a relatively simple structure given $N = \Omega_d(1)$ random features. We now study the optimization properties of its finite-sample counterpart.

We consider the estimator $\hat{F}(x) := \hat{f}(\langle x, \hat{\theta}\rangle)$, where $(\hat{f}, \hat{\theta})$ are obtained by running gradient flow on c and θ to minimize $L_n(c, \theta)$. Such strategy appears to be reasonable in light of the properties of the population landscape, since its local minimizers are also global and correspond to $\tilde{f} = \hat{P}_\lambda f^*$ and $\tilde{\theta} = \theta^*$ (Theorem 5.3). For a sufficiently large sample size n, one expects the empirical landscape L_n to concentrate around its expectation L and inherit its benign optimization properties. However, the presence of a degenerate saddle at $(c, m) = (0, 0)$ for $m = \langle \theta, \theta^*\rangle$ flattens the landscape around the "equator" $\{\theta : \langle \theta, \theta^*\rangle = 0\}$. Thus, more samples are required to ensure that gradient flow escapes from the equatorial region despite its dangerously close random initialization $|\langle \theta_0, \theta^*\rangle| = \Theta(1/\sqrt{d})$.

Prior works have obtained sample complexity of $n = O(d^s)$, where we recall that s is the information exponent of the target function f_*, for recovering θ^* either by employing a learning algorithm that explicitly learns individual Hermite polynomials [33] or by assuming that f_* is known a priori [10].[1] The intuition behind this sample complexity is roughly as follows.

- The empirical optimization landscape (when regarded as a function only of the direction θ) near the equator ($|m| \ll 1$) is of the form $L(\theta) \asymp m^s$.
- In order to certify that the optimization algorithm does *not* converge to a suboptimal critical point (i.e., $\|\nabla L(\theta)\| \le \epsilon$) on the equator, one requires that $m \ge \epsilon^{1/(s-1)}$.
- A uniform gradient convergence bound of the form $\|\nabla L(\theta) - \nabla L_n(\theta)\| = O(\sqrt{d/n})$ and the fact that $m = \Theta(1/\sqrt{d})$ at initialization together imply that $n = O(d^s)$ samples are sufficient to escape from the "influence" of the equator.

In order to repurpose these arguments to our setting, the relative scaling of the top-layer weights c relative to the direction vector θ is crucial, as has also been observed in the literature on *lazy*-vs-*rich* regimes [23, 84] in the context of overparametrized neural networks.

We consider an idealized version of Gradient Descent over the empirical loss L_n in the infinitesimally small learning rate regime. This results in a gradient flow ODE of the form:

$$\begin{aligned} \dot{c}(t) &= -\zeta(t)\nabla_c L_n(c, \theta) \\ \dot{\theta}(t) &= -\nabla_\theta^{\mathbb{S}^{d-1}} L_n(c, \theta) \ , \end{aligned} \tag{10}$$

where ζ is the relative scale between c and θ, and $\nabla_\theta^{\mathbb{S}^{d-1}}$ is the Riemannian gradient.

Specifically, we study a setting where $\zeta(t) = \mathbf{1}(t > T_0)$ for an appropriately chosen time T_0. This choice produces a two-stage gradient-flow. During the first phase, up until time T_0, we only optimize the first-layer parameter θ from a random initialization. In the second phase, the parameters c and θ are jointly optimized. Additionally, the first phase only utilizes a small fraction $\frac{N_0}{N} \ll 1$ of the random features employed in the second phase. Our procedure implements this by randomly initializing $c(0) \in \mathbb{R}^N$ as a sparse vector with N_0 non-zero components. The overall approach is described in Procedure 1.

[1] Actually, in [10] the authors obtain a slightly improved sample complexity of $\widetilde{O}(d^{s-1})$ for $s \ge 3$ by directly analyzing SGD with fixed step-size, as well as a matching lower bound (for SGD in the small step-size regime) up to polylogarithmic factors.

Our main result, proved in Appendix E, establishes that this gradient flow efficiently finds an approximate minimizer of the *population* loss, with an error (explicitly quantified as a function of n) that reveals the fundamental role of the information exponent s of f_*. On top of the regularity conditions on the target link function of Assumption 5.2, the upper bound on λ, and the lower bound on N from Eq. (9) of Theorem 5.3, the main result imposes a (compatible) upper bound on N and an appropriate choice of initial norm (ρ) and sparsity (N_0/N) for $c(0)$. In this section, we are interested in behavior as n, d, and N grow asymptotically and hence treat the target function f_* and terms derived from it (including Hermite coefficients α_j and information exponent s), along with the bias parameter τ and regularity parameter β, as constants and omit them from asymptotic notation.

Theorem 6.1 (Gradient flow finds approximate minimizers)**.** *For $\delta \in (0, 1/4)$ and f_* satisfying Assumption 5.2, suppose the following are true: (i) $\lambda = O(1)$ and $\lambda = \Omega(\sqrt{\Delta_{crit}})$, where $\Delta_{crit} := \max\{\sqrt{\frac{d+N}{n}}, (\frac{d^2}{n})^{2s/(2s-1)}\}$, (ii) $n = \widetilde{\Omega}(\max\{\frac{(d+N)d^{s-1}}{\lambda^4}, \frac{d^{(s+3)/2}}{\lambda^2}\})$, (iii) $N = \Omega(\frac{1}{\lambda}\log\frac{1}{\lambda\delta})$ & $N = \widetilde{O}(\lambda\Delta_{crit}^{-1})$, (iv) $N_0 = \Theta(\log(\frac{1}{\delta}))$, (v) $\rho = \Theta(\sqrt{N}N_0^{-(2+s)/2}(\tau^2 + \lambda N/N_0)^{-1})$, (vi) $T_0 = \tilde{\Theta}(d^{s/2-1})$, and (vii) $T_1 = \tilde{\Theta}(\frac{\lambda^4 n}{d+N})$. Then, if we run Procedure 1 for $T = T_0 + T_1$ time steps with the above parameters, with probability at least $\frac{1}{2} - \delta$ we have*

$$1 - |\langle \theta_T, \theta^* \rangle| \quad = \quad \widetilde{O}\left(\lambda^{-4} \max\left\{\frac{d+N}{n}, \frac{d^4}{n^2}\right\}\right) . \tag{11}$$

The empirical gradient flow therefore escapes the influence of the degenerate saddle with sample complexity $n = \tilde{\Theta}(d^s)$ when $\lambda = \Theta(1)$ and $s > 2$. This is an instance of gradient flow successfully optimizing a non-convex objective *without the strict saddle property* as soon as $s > 2$, building from the simpler optimization landscapes of [33, 10]. This is in contrast, for example, with spiked tensor recovery problems [12] where the signal strength is substantially weaker, leading to complexity in the optimization landscape. This sample complexity nearly matches the tight lower bound $n \gg d^{s-1}$ of [10], obtained in the case $s > 2$ and applies to SGD rather than batch gradient descent, as is our case. For $s \in \{1, 2\}$, the sample complexity becomes d^2 and $d^{2.5}$, respectively, but we note that these may be improved to d^s when using a smooth activation (see Appendix G). For $s = 2$, this is comparable to [27], which requires $\Omega(d^2)$ samples, but still above the $n \gg d \log d$ of [10].

We emphasize that the "near-optimality" of our sample complexity $n = \tilde{\Theta}(d^s)$ only pertains to gradient-based methods in small learning rate regimes [10]. In fact, alternative methods have been shown to achieve a better sample complexity of $\widetilde{O}(d^{\lceil s/2 \rceil})$ in the setting where f_* is a certain degree-s polynomial with information exponent s [19], leveraging tensor factorization tools. We leave it as an interesting open question to further understand the nature of this gap.

We note that the dependence on $d + N$ in the recovery guarantee can likely be improved to d using a more refined norm-based landscape concentration analysis. We also remark that if we chose the number of random features $N = \Theta(\frac{1}{\lambda}\log\frac{1}{\lambda\delta})$, then the requirement $\lambda = \Omega(\sqrt{\Delta_{\text{crit}}})$ for large enough sample size n imposes $\lambda \gg n^{-1/5}$. This lower bound on λ guarantees that critical points near initialization can be escaped, but may slow down learning. Nevertheless, this is sufficient to obtain an excess risk that vanishes with n, with a rate independent of d, as we now show.

Corollary 6.2 (Excess risk of Algorithm 1)**.** *Under the assumptions of Theorem 6.1, and further assuming $n \gtrsim d^3$, an appropriate choice of λ yields an excess risk guarantee of the form*

$$\|\hat{F} - F^*\|_{\gamma_d}^2 = \widetilde{O}\left(\left(\frac{d}{n}\right)^{\frac{\beta}{\beta+4}} + \left(\frac{1}{n}\right)^{\frac{\beta}{\beta+5}}\right) , \tag{12}$$

where β is defined as in Lemma 4.4.

This result indicates that the joint training is consistent, with excess risk that vanishes with a rate with explicit dependence on the ambient dimension d and the non-parametric exponent β. However, it requires a regularisation strength $\lambda = \Theta\left(\max\left\{\left(\frac{1}{n}\right)^{\frac{1}{\beta+5}}, \left(\frac{d}{n}\right)^{\frac{1}{\beta+4}}\right\}\right)$ to ensure enough gradient concentration, which happens to be larger than the optimal regularisation of the univariate kernel ridge regression associated with learning f_*. We are thus 'over-regularising' as a consequence of the joint training, resulting in a slower rate than what would be dictated by estimating θ_* and f_* separately. A simple mechanism to break this inefficiency is by considering a fine-tuning step of the second-layer terms.

Procedure 1 Gradient Flow

Require: N_0, ρ, T_0, T_1, N, and λ.
Initialize $\theta(0) \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, $c(0) \sim \mathrm{Unif}(\{c \in \mathbb{R}^N; \|c\|_2 = \rho; \|c\|_0 = N_0\})$.
Run Gradient Flow (10) with $\zeta(t) = \mathbf{1}(t > T_0)$ up to time $T = T_0 + T_1$.
Set $\hat{\theta} = \theta(T)$, $\hat{c} = c(T)$.

Procedure 2 Fine-Tuning

Require: $\hat{\theta}$ from Procedure 1, and $\lambda_{n'}$.
Set $\hat{c} = \arg\min_c L'_{n'}(c, \hat{\theta})$ as in (13).

Fine-tuning the second layer. After running Algorithm 1, we may include a final fine-tuning phase of training for second layer weights c alone, using a separate training sample (x'_i, y'_i), $i = 1, \dots, n'$ and a possibly different regularization parameter $\lambda_{n'}$. More precisely, we set

$$\hat{c} = \arg\min_c \left\{ L'_{n'}(c, \hat{\theta}) := \frac{1}{n'} \sum_{i=1}^{n'} (c^\top \Phi(\langle \hat{\theta}, x'_i \rangle) - y'_i)^2 + \lambda_{n'} \|c\|^2 \right\}, \tag{13}$$

where $\hat{\theta}$ denotes the output of the previous gradient descent phase. Note that this is a strongly convex optimization problem, and can thus be optimized efficiently using gradient methods or by solving a linear system. While this may not be needed in practice, we use a different training sample for technical reasons, namely to break the dependence between the data and the kernel, which depends on the initial training sample through $\hat{\theta}$. We note that such sample splitting strategies are commonly used in other contexts in the statistics literature (e.g., [15, 21]). We obtain the following guarantee.

Proposition 6.3 (Excess risk of fine-tuning)**.** *Let $\delta \in (0, 1/4)$. Let $m = \langle \theta^*, \hat{\theta} \rangle$, where $\hat{\theta}$ is obtained from the previous gradient descent phase, and let $\hat{c}$ be the ridge regression estimator obtained from a fresh dataset $\mathcal{D}'$ of n' samples, N random features, and regularization parameter $\lambda_{n'} := (\sigma^2\tau^2/\|f''_*\|^2_\gamma n')^{1/(\beta+1)}$, and let $\hat{F}(x) = \hat{c}^\top \Phi(\langle \hat{\theta}, x \rangle)$. Assume*

$$n' \gtrsim \max\left\{ \sigma^2\tau^2/\|f''_*\|^2_\gamma, (\|f''_*\|^2_\gamma/\sigma^2\tau^2)^{1/\beta}, \|f_*\|^2_\infty/(\sigma^2\tau^2)^{\beta/(\beta+1)} \right\} \text{ , and}$$

$$N \gtrsim C_\tau \left(n' \|f''_*\|^2_\gamma / \sigma^2\tau^2 \right)^{\frac{1}{\beta+1}} \log\left(n'^{1/(\beta+1)} \delta^{-1} \right) .$$

Then with probability at least $1 - \delta$ over the random features, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}'}[\|\hat{F} - F^*\|^2_{\gamma_d} | \hat{\theta}] \lesssim \|f''_*\|_\gamma^{\frac{2}{\beta+1}} \left(\frac{\sigma^2\tau^2}{n'} \right)^{\frac{\beta}{\beta+1}} + \|f'_*\|^2_\gamma (1 - |m|) , \tag{14}$$

where the expectation is over the n' fresh samples, and is conditioned on the previously obtained $\hat{\theta}$.

Decoupling the regularization parameters of the two phases (along with number of random features N) allows us to keep a large λ in the first phase, leading to fast recovery as per Theorem 6.1, while obtaining vanishing excess risk through a decreasing $\lambda_{n'}$. This is illustrated in the result on the excess risk for Algorithm 2.

Corollary 6.4 (Excess risk of Algorithm 2)**.** *Let $\delta \in (0, 1/4)$. As in Theorem 6.1, let $\mu_s = \langle h_s, \Sigma h_s \rangle > 0$, and let f_* satisfy Assumption 5.2. Let $\lambda = \Theta(1)$, and assume the following on the sample sizes and number of random features for the first phase (n, N, N_0) and fine-tuning phase (n', N'):*

$$N = N_0 = \Theta\left(\frac{1}{\lambda} \log \frac{1}{\lambda\delta} \right), \quad n = \widetilde{\Omega}\left(\max\{d^s, d^{(s+3)/2}\} \right), \quad N' = \widetilde{\Omega}\left(n'^{\frac{1}{\beta+1}} \right).$$

and let ρ be as in Theorem 6.1. With probability at least $1/2 - 2\delta$ over the initial n samples, initialization, random features, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}'}[\|\hat{F} - F_{\theta^*}\|^2_{\gamma_d}] \le \widetilde{O}\left(\max\left\{ \frac{d}{n}, \frac{d^4}{n^2} \right\} + \left(\frac{1}{n'} \right)^{\frac{\beta}{\beta+1}} \right), \tag{15}$$

where the constants in $\widetilde{O}$ do not depend on d other than through logarithmic factors.

Comparing Corollaries 6.2 and 6.4, we observe that the fine-tuning stage recovers the optimal sample complexity, where the non-parametric rate is fully independent of the ambient dimension d, while in Corollary 6.2 there is still a dependence in the constants. We make the following additional remarks:

- The time-scale separation schedule for ζ in Theorem 6.1 is sufficient but possibly not necessary. The analysis of vanilla dynamics ($\zeta(t) \equiv 1$) is challenging, since during the initial phase of training there may be adverse interaction between c and θ, which under naive analysis lead to sub-optimal sample complexity of $n \geq O(d^{2s})$. Observe that this separate analysis of 'weak' and 'strong' recovery phases of learning appears in most contemporary related work [27, 1, 4, 10, 11].
- The time discretization to turn Procedure 1 into a proper algorithm should follow from standard time discretization arguments, although the case where $\phi = \text{ReLU}$ requires special care due to the non-smoothness of the loss (see Appendix F for further discussion). In such setting, such discretization arguments do not hold for vanilla gradient descent in the worst-case [51], although these may be recovered by appropriately smoothing the objective prior to computing the gradient, or by using instead a smooth activation function (see Appendix G).

7 Conclusion

This work studies the ability of shallow neural networks to learn single-index models with gradient descent. Our main results are positive, and demonstrate their ability to solve a semi-parametric problem with nearly optimal guarantees. Interestingly, this success story combines elements from the feature-learning regime, i.e., the ability to efficiently identify the hidden direction in high-dimensions under a non-convex objective, with ingredients from the lazy-regime. Our technical analysis leverages tools from high-dimensional probability (such as uniform gradient concentration) and RKHS approximation, and complements the growing body of theoretical work on the efficacy of gradient methods for non-convex objectives. We have followed the standard approach of first establishing benign topological properties of the population loss, and then extending them to the empirical loss. There are nonetheless several unanswered questions that our work has not addressed.

Weaker regularity and discrete-time analysis. Our approximation rate for ReLU as the *target* (see Appendix H) suggests that the polynomial-in-λ approximation rate may be extended to function classes beyond $\mathcal{F} \subset H^2(\gamma)$, such as Lipschitz functions with smooth tail behavior. Thus, it would be interesting to extend our empirical landscape concentration results to such functions satisfying weaker regularity assumptions, which currently rely on certain polynomial decay of the Hermite coefficients (see Assumption 5.2). Additionally, by using a smooth activation function (see Appendix G), our GF dynamics can be discretized and turned into GD with analogous sample and time complexity. In that context, a natural goal is to compare quantitatively the differences between GD with multiple passes over the training data and SGD by adapting tools from [10, 11].

Trainable biases and untied directions. Our proposed neural network architecture is non-standard, in the sense that its biases are frozen at initialization and all neurons share the same inner weight. For the purposes of learning single-index models, removing these restrictions would not bring any statistical benefits. However, it would be interesting to extend our analysis to the general setting where the first layer weights are not tied and biases are not frozen.

Extension to multi-index models. Multi-index models are natural extensions in which the hidden direction θ^* is replaced by a hidden low-dimensional *subspace*. Typically, multi-index models enjoy similar statistical guarantees as single-index models [33, 5], and thus a natural question is whether the same algorithmic tools developed here extend to the multi-index setting.

Gradient dynamics without warm-start. An unsatisfactory aspect of our results is the requirement that the algorithm starts by only optimizing θ for $t < T_0$. It would be interesting to understand whether the vanilla dynamics can also succeed provably.

Acknowledgments. We are thankful to Enric Boix-Adserà, Alex Damian, Cédric Gerbelot, Daniel Hsu, Jason Lee, Theodor Misiakiewicz, Matus Telgarsky, Eric Vanden-Eijnden, and Denny Wu for useful discussions. We also thank the anonymous NeurIPS reviewers and area chair for helpful feedback. JB, AB and MJ are partially supported by NSF RI-1816753, NSF CAREER CIF 1845360, NSF CHS-1901091, NSF Scale MoDL DMS 2134216, Capital One and Samsung Electronics. CS is supported by an NSF GRFP and by NSF grants CCF-1814873 and IIS-1838154.

References

[1] E. Abbe, E. Boix-Adsera, and T. Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. *arXiv preprint arXiv:2202.08658*, 2022. 1, 2, 10

[2] Z. Allen-Zhu and Y. Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020. 1

[3] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005. 59

[4] J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *arXiv preprint arXiv:2205.01445*, 2022. 2, 10

[5] F. Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017. 1, 2, 3, 5, 10

[6] F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017. 3, 6, 24, 25

[7] F. Bach. Learning theory from first principles, 2021. 3, 25, 53, 54, 62

[8] Y. Bai and J. D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*, 2019. 2

[9] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993. 1

[10] G. Ben Arous, R. Gheissari, and A. Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research (JMLR)*, 22:106–1, 2021. 2, 3, 6, 7, 8, 10, 18

[11] G. Ben Arous, R. Gheissari, and A. Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *arXiv preprint arXiv:2206.04030*, 2022. 3, 10

[12] G. Ben Arous, S. Mei, A. Montanari, and M. Nica. The landscape of the spiked tensor model. *Communications on Pure and Applied Mathematics*, 72(11):2282–2330, 2019. 3, 8

[13] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. 3

[14] D. Bertoin, J. Bolte, S. Gerchinovitz, and E. Pauwels. Numerical influence of relu'(0) on backpropagation. *Advances in Neural Information Processing Systems*, 34:468–479, 2021. 35

[15] X. Bing, F. Bunea, S. Strimas-Mackey, and M. Wegkamp. Prediction under latent factor regression: Adaptive pcr, interpolating predictors and beyond. *Journal of Machine Learning Research*, 22(177):1–50, 2021. 9

[16] J. Borwein and A. Lewis. *Convex Analysis*. Springer, 2006. 59, 60

[17] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *Advances in Neural Information Processing Systems*, 29, 2016. 3

[18] S.-H. Chang, P. C. Cosman, and L. B. Milstein. Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011. 23

[19] M. Chen, Y. Bai, J. D. Lee, T. Zhao, H. Wang, C. Xiong, and R. Socher. Towards understanding hierarchical learning: Benefits of neural representations. *Advances in Neural Information Processing Systems*, 2020. 2, 8

[20] S. Chen, A. Gollakota, A. R. Klivans, and R. Meka. Hardness of noise-free learning for two-hidden-layer neural networks. *arXiv preprint arXiv:2202.05258*, 2022. 1

[21] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. 9

[22] L. Chizat and F. Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020. 1

[23] L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming, 2018. 2, 7

[24] Y. Cho and L. Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009. 3

[25] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 60

[26] A. S. Dalalyan, A. Juditsky, and V. Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *The Journal of Machine Learning Research*, 9:1647–1678, 2008. 1, 2

[27] A. Damian, J. Lee, and M. Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, 2022. 2, 8, 10

[28] A. Daniely and G. Vardi. Hardness of learning neural networks with natural weights. *Advances in Neural Information Processing Systems*, 33:930–940, 2020. 1

[29] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020. 3, 59

[30] I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1514–1539. PMLR, 09–12 Jul 2020. 1

[31] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.6 of 2022-06-30. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. 20

[32] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Curves of descent. *SIAM Journal on Control and Optimization*, 53(1):114–138, 2015. 3, 59

[33] R. Dudeja and D. Hsu. Learning single-index models in gaussian space. In S. Bubeck, V. Perchet, and P. 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. 1, 2, 7, 8, 10, 34

[34] A. Engel and C. Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001. 3

[35] D. J. Foster, A. Sekhari, and K. Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, 2018. 2, 3

[36] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9):881–890, 1974. 2

[37] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015. 3

[38] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016. 7

[39] R. Ge, J. D. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017. 2, 3

[40] R. Ge and T. Ma. On the optimization landscape of tensor decompositions. *Advances in Neural Information Processing Systems*, 30, 2017. 7

[41] S. Goel, A. Gollakota, Z. Jin, S. Karmalkar, and A. Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pages 3587–3596. PMLR, 2020. 1

[42] S. Goel, S. Karmalkar, and A. Klivans. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. *Advances in Neural Information Processing Systems*, 32, 2019. 65

[43] S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019. 2, 3

[44] L. Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975. 60

[45] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and semiparametric models*, volume 1. Springer, 2004. 1

[46] M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, 29(3):595–623, 2001. 1

[47] P. J. Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985. 2

[48] H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of econometrics*, 58(1-2):71–120, 1993. 1, 6

[49] Z. Ji and M. Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020. 3, 59

[50] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently, 2017. 3

[51] G. Kornowski and O. Shamir. Oracle complexity in nonsmooth nonconvex optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 324–334. Curran Associates, Inc., 2021. 10

[52] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. 2

[53] K.-C. Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992. 2

[54] K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019. 59

[55] A. Maillard, G. B. Arous, and G. Biroli. Landscape complexity for the empirical risk of generalized linear models. *arXiv preprint arXiv:1912.02143*, 2019. 3

[56] E. Malach, P. Kamath, E. Abbe, and N. Srebro. Quantifying the benefit of using differentiable learning over tangent kernels. In *International Conference on Machine Learning*, pages 7379–7389. PMLR, 2021. 1

[57] E. Malach, G. Yehudai, S. Shalev-Schwartz, and O. Shamir. The connection between approximation, depth separation and learnability in neural networks. In *Conference on Learning Theory*, pages 3265–3295. PMLR, 2021. 1

[58] S. S. Mannelli, G. Biroli, C. Cammarota, F. Krzakala, P. Urbani, and L. Zdeborová. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020. 3

[59] S. S. Mannelli, F. Krzakala, P. Urbani, and L. Zdeborova. Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models. In *international conference on machine learning*, pages 4333–4342. PMLR, 2019. 3

[60] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016. 2, 3, 19, 20, 34, 62

[61] S. Mei, T. Misiakiewicz, and A. Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. 3

[62] S. Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017. 55, 56

[63] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017. 36

[64] A. Mousavi-Hosseini, S. Park, M. Girotti, I. Mitliagkas, and M. A. Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022. 2

[65] E. Nelson. The free markoff field. *Journal of Functional Analysis*, 12(2):211–227, 1973. 30

[66] E. Nichani, Y. Bai, and J. D. Lee. Identifying good directions to escape the ntk regime and efficiently learn low-degree plus sparse polynomials. *arXiv preprint arXiv:2206.03688*, 2022. 2

[67] R. O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014. 4, 30, 61

[68] G. Ongie, R. Willett, D. Soudry, and N. Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019. 5

[69] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999. 1

[70] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. 3, 5

[71] V. Ros, G. B. Arous, G. Biroli, and C. Cammarota. Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. *Physical Review X*, 9(1):011003, 2019. 3

[72] A. Rudi and L. Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017. 3

[73] I. Safran and J. D. Lee. Optimization-based separations for neural networks. *arXiv preprint arXiv:2112.02393*, 2021. 1

[74] S. Saitoh. *Integral transforms, reproducing kernels and their applications*, volume 369. CRC Press, 1997. 20

[75] P. Savarese, I. Evron, D. Soudry, and N. Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019. 5

[76] B. Schölkopf and A. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002. 3

[77] M. Soltanolkotabi. Learning relus via gradient descent. *Advances in neural information processing systems*, 2017. 3

[78] M. J. Song, I. Zadik, and J. Bruna. On the cryptographic hardness of learning single periodic neurons. *Advances in Neural Processing Systems (NeurIPS)*, 2021. 1

[79] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018. 3

[80] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012. 55, 56

[81] L. Van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497–540, 1996. 59

[82] R. Veiga, L. Stephan, B. Loureiro, F. Krzakala, and L. Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. *arXiv preprint arXiv:2202.00293*, 2022. 3

[83] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. 18, 19, 20, 51, 57

[84] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020. 1, 7

[85] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International conference on machine learning*, pages 4140–4149. PMLR, 2017. 2, 3

[86] M. Zhou, R. Ge, and C. Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*, pages 4577–4632. PMLR, 2021. 2, 3

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Conclusions.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a purely theoretical work with no practical applications.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 5 and 6.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section D.2.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Supplemental zip file.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section A.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section A.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section A.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendices

A Numerical Experiments **18**

B Additional Preliminaries and Concentration Bounds **18**

C Proofs of Section 4 **20**

C.1 Proof of Lemma 4.2 20

C.2 Proof of Lemma 4.4 22

C.3 Random feature approximation 24

D Proofs for Section 5 **26**

D.1 Exact expression of first-order critical points 26

D.2 Proof of Theorem 5.3 26

D.3 First-order critical points of the population loss and proof of Claim D.1 27

D.4 Proof of Lemma D.2 29

D.5 Other lemmas for the proof of Theorem 5.3 30

E Proofs for Section 6 **31**

E.1 Proof of Theorem 6.1 31

E.2 Proof of Lemma E.1 33

E.3 Proof of Lemma E.2 40

E.4 Proof of Lemma E.3 47

E.5 Proof of Corollary 6.2 51

E.6 Proof of Proposition 6.3 52

E.7 Proof of Corollary 6.4 57

E.8 Omitted proofs from Section E 58

F Gradient Flow on Non-smooth Landscapes **58**

G Smooth Activation Functions **60**

H RKHS Approximation Beyond $\mathcal{F}$ **63**

H.1 ReLU Hermite coefficients 64

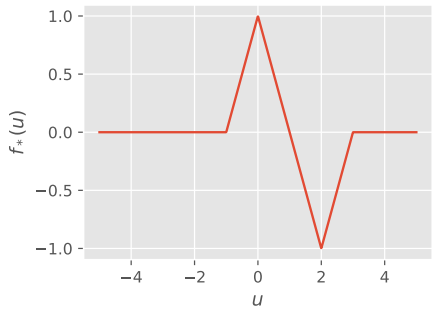

Figure 1: Piecewise linear teacher link function f_*.

A Numerical Experiments

In this section, we illustrate our results by running Algorithm 1 on a simple synthetic dataset. We consider a piecewise linear target link function f_* that is compactly supported, illustrated in Figure 1, and data generated by the model described in Section 3, with $\sigma = 0.001$. To understand behavior when varying the information exponent s, we also consider teachers $f_*^s = f_* - \sum_{j<s} < f_*, h_j > h_j$ with the low-order Hermite components are removed. We initialize the direction θ randomly on the sphere, and the parameters c_k i.i.d. with variance 1. We run gradient descent on the empirical loss for 10 000 iterations using a step-size that is 100 times larger on θ than on c, with projections of θ on the sphere after each step, and with ℓ_2 regularization on c. We start optimizing c only after 500 steps to simulate the warm-start phase. We show the loss obtained on 10K held-out test samples. For fine-tuning (denoted "ridge" in the plots), we re-optimize the output layer c exactly on the training data using ridge regression with a possibly different regularization parameter λ', We fix $N = 100$ random features, and optimize hyperparameters (λ, λ', and the step-size) on the test data. All experiments were run on CPUs. Each experiment was repeated 10 times, and the figures report either the mean and standard deviation over the 10 runs, or the best performing model out of the 10 runs.

Our experiment results are shown in Figure 2. For $s \geq 3$, only some of the 10 runs were successful in recovering the target direction, and we thus show the best performing run for such curves (indeed, our theory suggests that there may be a non-negligible probability of failure). We observe that full recovery ($|m| \to 1$) requires more samples when the dimension d increases, while the excess risk curves have approximately the same rate for large enough n, regardless of the dimension or information exponent, as predicted by our theory. The bottom plots for $d = 50$ suggest that $s = 3$ requires more samples than smaller s for perfect recovery, while the remaining curves are somewhat comparable. This similarity between $s = 1$ and $s = 2$ is reminiscent of the situation in [10], where the rates for these two cases only differ by a logarithmic factor, and suggests that it may be possible to improve the $O(d^s)$ rates in our results for $s \geq 2$.

B Additional Preliminaries and Concentration Bounds

We introduce several well-known concentration bounds that we apply throughout the appendix. Borrowing notation from [83], we first introduce notation of sub-gaussian and sub-exponential random variables, vectors, and matrices.

Definition B.1. *A real-valued random variable z is γ^2-*sub-gaussian $\|z\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}[\exp(z^2/t^2)] \leq 2\} \leq \gamma$. *Likewise, a random vector $x \in \mathbb{R}^d$ is γ^2-sub-gaussian and we denote $\|x\|_{\psi_2} \leq \gamma$ if $\|w \cdot x\|_{\psi_2} \leq \gamma$ for any fixed $w \in \mathbb{S}^{d-1}$.*

Definition B.2. *A real-valued random variable y is γ-*sub-exponential $\|y\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}[\exp(|y|/t)] \leq 2\} \leq \gamma$. *Likewise, a random vector $u \in \mathbb{R}^d$ is γ-sub-exponential and we denote $\|u\|_{\psi_1} \leq \gamma$ if $\|w \cdot u\|_{\psi_1} \leq \gamma$ for any fixed $w \in \mathbb{S}^{d-1}$.*

We note several key properties of sub-gaussian and sub-exponential random variables that we repeatedly rely on.

Fact B.3. *Let $z_1, \ldots, z_N$ and $y_1, \ldots, y_N$ be sub-gaussian and sub-exponential random variables respectively. Then the following hold for some universal constant C.*

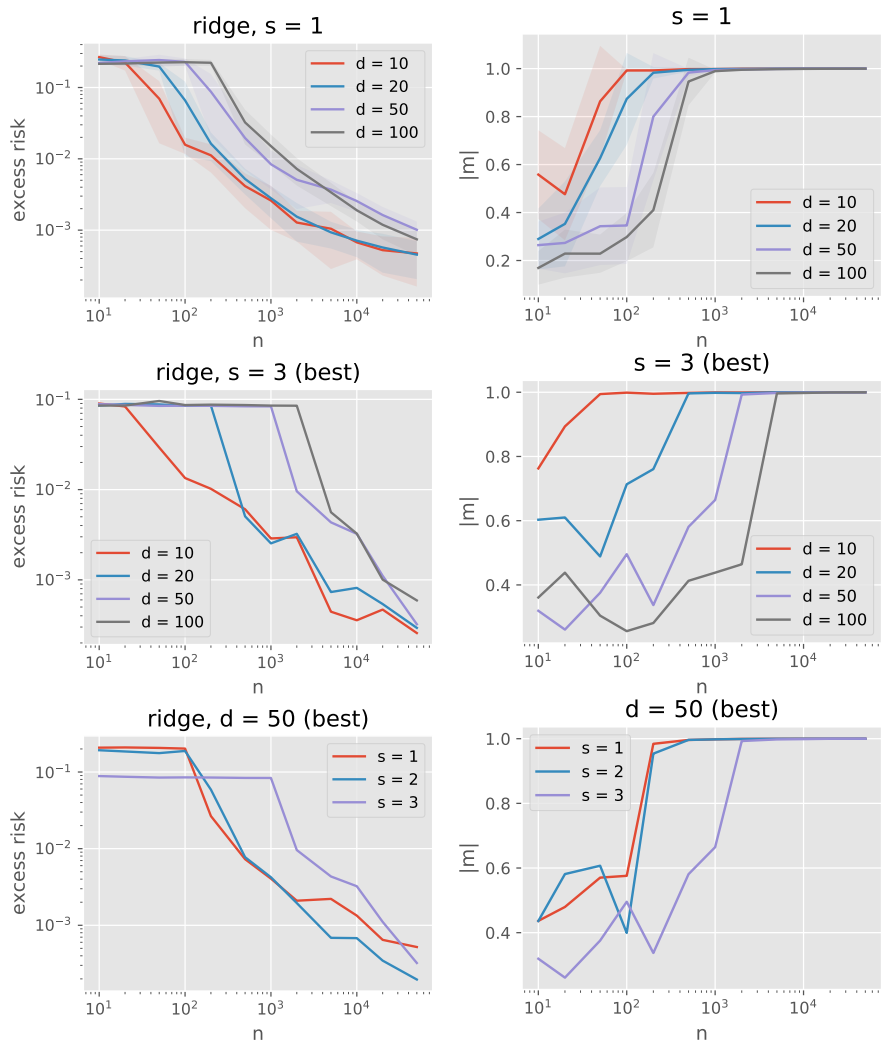

Figure 2: Excess risk $\|\hat{F} - F^*\|_{\gamma_d}^2$ with final ridge/fine-tuning step (left), and correlation $|m|$ (right) as a function of sample size n.

1. *Centering preserves sub-gaussianity and sub-exponentiality:* $\|z_1 - \mathbb{E}[z_1]\|_{\psi_2} \leq C \|z_1\|_{\psi_2}$ *and* $\|y_1 - \mathbb{E}[y_1]\|_{\psi_1} \leq C \|y_1\|_{\psi_2}$ *[83, Lemma 2.6.8 and Exercise 2.7.10].*

2. *Products of sub-gaussian random variables are sub-exponential:* $z_1 z_2$ *is sub-exponential and* $\|z_1 z_2\|_{\psi_1} \leq \|z_1\|_{\psi_2} \|z_2\|_{\psi_2}$ *[83, Lemma 2.7.7].*

3. *Sums of independent sub-gaussian random variables are sub-gaussian. If* $z_1, \ldots, z_N$ *are independent, then,*

$$\left\| \sum_{i=1}^{N} z_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^{n} \|z_i\|_{\psi_2}^2$$

   [83, Proposition 2.6.1].

4. *Sums of pairs of sub-exponential random variables are sub-exponential.* $\|y_1 + y_2\|_{\psi_1} \leq C(\|y_1\|_{\psi_1} + \|y_2\|_{\psi_1})$ *[60, Lemma 2].*

5. *Lipschitz functions preserve sub-gaussianity. For Lipschitz-continuous* $f : \mathbb{R} \to \mathbb{R}$, $\|f(z_1) - \mathbb{E}[f(z_1)]\|_{\psi_2} \leq C\mathrm{Lip}(f) \|z_1\|_{\psi_2}$.

6. *If β is a bounded random variable with $|\beta| \leq s$, then $\|\beta(z_1 - \mathbb{E}[z_1])\|_{\psi_2} \leq Cs\|z_1\|_{\psi_2}$ [60, Lemma 1].*

Theorem B.4 (Bernstein's inequality, [83, Corollary 2.8.3])**.** *For independent, mean-zero, sub-exponential random variables $x_1, \ldots, x_n$ and any $t \geq 0$,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| \geq t\right] \leq 2\exp\left(-Cn\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right)$$

for universal C and $K = \max_i \|x_i\|_{\psi_1}$.

We also include several basic facts about ϵ-covers, which are useful in several proofs.

Definition B.5. *For some compact $S \subset \mathbb{R}^d$, $\mathcal{N}_\epsilon \subseteq S$ is an ϵ-covering if for all $x \in S$, there exists $\tilde{x} \in \mathcal{N}_\epsilon$ such that $\|x - \tilde{x}\| \leq \epsilon$.*

Fact B.6 ([83, Corollary 4.2.13, Lemma 4.4.1])**.** *For all $\epsilon \in (0, 1]$, there exists an ϵ-net $\mathcal{N}_\epsilon$ for $\mathbb{S}^{d-1}$ with $|\mathcal{N}_\epsilon| \leq (\frac{3}{\epsilon})^d$. Moreover, for any $a \in \mathbb{R}^d$,*

$$\|a\| \leq \frac{1}{1-\epsilon}\max_{x \in \mathcal{N}_\epsilon}\langle a, x\rangle.$$

Lemma B.7 (Anticoncentration on the unit sphere)**.** *Let $d \in \mathbb{N}$. Let $\theta \in \mathbb{S}^{d-1}$ be any fixed unit vector and let u be a random vector drawn uniformly from $\mathbb{S}^{d-1}$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}[\langle \theta, u\rangle \leq \epsilon] \leq 4\sqrt{d}\epsilon\ .$$

Proof. We show this using an elementary argument. Let $\epsilon \in [0, 1]$. Define

$$G(\epsilon) := \int_0^\epsilon (1 - t^2)^{(d-1)/2} dt \leq \epsilon\ .$$

By Gautschi's inequality [31, Eq. 5.6.4] for the Gamma function, we have

$$G(1) = \frac{\sqrt{\pi}}{2} \cdot \frac{\Gamma((d+1)/2)}{\Gamma((d+2)/2)} \geq \frac{\sqrt{\pi}}{4} \cdot \frac{1}{\sqrt{d}}\ .$$

Thus we have the following anti-concentration bound on $\mathbb{S}^{d-1}$.

$$\mathbb{P}[|\langle \theta, u\rangle| \leq \epsilon] = G(\epsilon)/G(1) \leq \frac{4}{\sqrt{\pi}} \cdot \sqrt{d}\epsilon\ .$$

□

Finally, we recall the following result on reproducing kernel Hilbert spaces, which describes the RKHS for kernels defined from explicit features maps.

Theorem B.8 ([74, §2.1])**.** *Let $\psi : \mathcal{X} \to \mathcal{F}$ be a mapping into a Hilbert space $\mathcal{F}$, and for $x, x' \in \mathcal{X}$, define the kernel $\kappa(x, x') = \langle \psi(x), \psi(x')\rangle_{\mathcal{F}}$. The RKHS $\mathcal{H}$ of κ consists of functions of the form $f(x) = \langle g, \psi(x)\rangle_{\mathcal{F}}$, and for any $f \in \mathcal{H}$, the RKHS norm of f is defined by*

$$\|f\|_{\mathcal{H}}^2 := \inf\left\{\|g\|_{\mathcal{F}}^2\ :\ g \in \mathcal{F}\ s.t.\ f = \langle g, \psi(\cdot)\rangle_{\mathcal{F}}\right\}\ .$$

C Proofs of Section 4

C.1 Proof of Lemma 4.2

Claim C.1 (RKHS Sobolev representation)**.** *Let $f \in H^2(\gamma) \cap C^2(\mathbb{R})$ be a function such that $\lim_{t\to-\infty} f(t) = \lim_{t\to-\infty} f'(t) = 0$ and $\int_{-\infty}^{\infty}(f''(u)^2/\gamma_\tau(u))du < \infty$. Then,*

$$f(t) = \int_{-\infty}^{\infty} \frac{f''(u)}{\gamma_\tau(u)}\phi(t-u)d\gamma_\tau(u)\ . \tag{16}$$

Moreover, the RKHS norm of f is upper bounded as follows.

$$\|f\|_{\mathcal{H}}^2 \leq \int_{-\infty}^{\infty} \frac{f''(u)^2}{\gamma_\tau(u)}du\ .$$

Proof of Claim C.1. By the Fundamental Theorem of Calculus and Fubini's Theorem,

$$\begin{aligned} f(t) = \int_{-\infty}^{t} f'(w)dw = \int_{-\infty}^{\infty} f'(w)\mathbf{1}[w \le t]dw &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f''(u)\mathbf{1}[u \le w]du \right) \mathbf{1}[w \le t]dw \\ &= \int_{-\infty}^{\infty} f''(u) \left(\int_{-\infty}^{\infty} \mathbf{1}[u \le w]\mathbf{1}[w \le t]dw \right) du \\ &= \int_{-\infty}^{\infty} f''(u)\phi(t-u)du \\ &= \int_{-\infty}^{\infty} \frac{f''(u)}{\gamma_\tau(u)} \phi(t-u)d\gamma_\tau(u) \ . \end{aligned}$$

The upper bound on the RKHS norm follows from the above representation and Lemma 4.1. □

Lemma 4.2 (RKHS norm bound)**.** *Let $f \in H^2(\gamma) \cap C^2(\mathbb{R})$ and $\tau > 1$. If f and f' both have polynomial growth and $\int \frac{|f''(t)|^2}{\gamma_\tau(t)}dt < \infty$, then $f \in \mathcal{H}$ with*

$$\|f\|_{\mathcal{H}}^2 \le 6\tau \left(\int \frac{|f''(t)|^2}{\gamma_\tau(t)}dt + \|f\|_\gamma^2 + 6\|f'\|_\gamma^2 + 2\langle f, f''\rangle_\gamma \right) \ . \tag{7}$$

Proof of Lemma 4.2. For general f, the boundary conditions of Claim C.1, i.e., $\lim_{t\to-\infty} f(t) = \lim_{t\to-\infty} f'(t) = 0$, do not hold. However, we can reduce to the case considered in Claim C.1 by decomposing f into 2 parts, i.e., $f = f_1 + f_2$, where $f_1(t)$ and $f_2(-t)$ individually satisfy the assumptions of Claim C.1.

Let $\varphi(t) = \int_{-\infty}^{t} \gamma(u)du$. We decompose

$$f = f \cdot \varphi + f \cdot (1 - \varphi) := f_1 + f_2 \ .$$

For any $t \in \mathbb{R}$, $0 \le \varphi(t) \le 1$, and since we assume f has polynomial growth, $\lim_{t\to-\infty} f_1(t) = \lim_{t\to-\infty} f_1'(t) = 0$ and $\lim_{t\to\infty} f_2(t) = \lim_{t\to\infty} f_2'(t) = 0$. By Lemma 4.1, it holds that

$$\|f\|_{\mathcal{H}}^2 \le \int \left(\left(\frac{f_1''(t)}{\gamma_\tau(t)} \right)^2 + \left(\frac{f_2''(t)}{\gamma_\tau(t)} \right)^2 \right) d\gamma_\tau(t) = \int \frac{f_1''(t)^2 + f_2''(t)^2}{\gamma_\tau(t)}dt \ . \tag{17}$$

To upper bound the RHS of Eq. (17) in terms of f and its derivatives, we derive explicit expressions for f_1'' and f_2''. Since $\varphi'(t) = \gamma(t)$ and $\varphi''(t) = -t\gamma(t)$, we have

$$\begin{aligned} f_1''(t) &= f''(t) \cdot \varphi(t) + 2f'(t) \cdot \gamma(t) - f \cdot (t\gamma(t)) \\ f_2''(t) &= f''(t) \cdot (1 - \varphi(t)) - 2f'(t) \cdot \gamma(t) + f \cdot (t\gamma(t)) \ . \end{aligned}$$

Using the elementary inequality $(a_1 + a_2 + a_3)^2 \le 3(a_1^2 + a_2^2 + a_3^2)$ (from Cauchy-Schwarz) and the fact that $0 \le \varphi(t) \le 1$ and $0 \le \gamma(t) \le 1$ for all $t \in \mathbb{R}$,

$$\begin{aligned} \int \frac{f_1''(t)^2}{\gamma_\tau(t)}dt &\le \int \frac{3}{\gamma_\tau(t)} \left(f''(t)^2 + 4f'(t)^2\gamma(t)^2 + f(t)^2(t\gamma(t))^2 \right) dt \\ &\le 3 \left(\int \frac{f''(t)^2}{\gamma_\tau(t)}dt + 4\tau \int f'(t)^2 d\gamma(t) + \tau \int f(t)^2 t^2 d\gamma(t) \right) \\ &= 3 \left(\int \frac{f''(t)^2}{\gamma_\tau(t)}dt + 4\tau\|f'\|_\gamma^2 + \tau\|f \cdot t\|_\gamma^2 \right) \ . \end{aligned}$$

The same upper bound holds for f_2. Thus, from Eq. (17) and the fact that $\tau > 1$, we have

$$\|f\|_{\mathcal{H}}^2 \le 6\tau \left(\int \frac{|f''(t)|^2}{\gamma_\tau(t)}dt + 4\|f'\|_\gamma^2 + \|f \cdot t\|_\gamma^2 \right) \ .$$

It remains to upper bound $\|f \cdot t\|_\gamma$ purely in terms of f and its higher-order derivatives. Using integration by parts, we have that for any differentiable $F : \mathbb{R} \to \mathbb{R}$ with polynomial growth,

$$\int tF(t)\gamma(t)dt = -F(u)\gamma(u)\big|_{u=-\infty}^{\infty} + \int F'(t)\gamma(t)dt = \int F'(t)\gamma(t)dt \ . \tag{18}$$

Hence, by the assumption that f, f' have polynomial growth and applying the identity Eq. (18) twice, we have

$$\begin{aligned}\|f \cdot t\|_\gamma^2 = \int t \cdot (tf(t)^2)\gamma(t)dt = \int (tf(t)^2)'\gamma(t)dt &= \|f\|_\gamma^2 + 2\int t \cdot (f(t)f'(t))\,\gamma(t)dt \\ &= \|f\|_\gamma^2 + 2\int (f'(t)^2 + f(t)f''(t))\gamma(t)dt \\ &= \|f\|_\gamma^2 + 2\langle f, f''\rangle_\gamma + 2\|f'\|_\gamma^2 \ ,\end{aligned}$$

In conclusion, we have

$$\|f\|_{\mathcal{H}}^2 \le 6\tau \left(\int \frac{|f''(t)|^2}{\gamma_\tau(t)} dt + \|f\|_\gamma^2 + 6\|f'\|_\gamma^2 + 2\langle f, f''\rangle_\gamma \right) \ . \tag{19}$$

□

C.2 Proof of Lemma 4.4

Lemma 4.4 (RKHS approximation error)**.** *Let $\lambda \in (0,1)$ and $f \in \mathcal{F}$. Then, there exists a universal constant $C > 0$ such that*

$$A(f,\lambda) \le C\Big(\tau^{1+\beta}\|f''\|_4^2 \cdot \lambda^\beta + \lambda C_f^2\Big) \ , \tag{8}$$

where $\beta = \frac{1-1/\tau^2}{3+1/\tau^2}$ and $C_f = \max\{\|f\|_\gamma, \|f'\|_\gamma, \|f''\|_\gamma\}$.

Proof. We first prove the result for $f \in C^2(\mathbb{R})$ and then extend it to general $f \in H^2(\gamma)$ s.t. $f'' \in L^4(\gamma)$ using a density argument. Define the λ-regularized approximation error of $h \in \mathcal{H}$ by

$$\mathcal{E}(h) = \|f - h\|_\gamma^2 + \lambda\|h\|_{\mathcal{H}}^2 \ ,$$

and recall that $A(f,\lambda) = \min_{h\in\mathcal{H}} \mathcal{E}(h)$.

We approximate f with a single-parameter family of functions $\{h_M \mid M > 0\}$, defined by $h_M''(t) = f''(t) \cdot \mathbb{1}[|t| \le M]$. By the Fundamental Theorem of Calculus, we have

$$h_M(t) = \begin{cases} f(t) & \text{if } |t| \le M \\ f(M) + f'(M)t & \text{if } t > M \\ f(-M) + f'(-M)t & \text{if } t < -M \ . \end{cases}$$

Thus, h_M matches f exactly on $[-M, M]$ and is linear with slope $f'(M)$ (resp. $f'(-M)$) for $t \ge M$ (resp. $t \le -M$). We first show that $h_M \in \mathcal{H}$, which implies $A(f,\lambda) \le \inf_{M>0} \mathcal{E}(h_M)$, and then show that for an explicit choice of M, $\mathcal{E}(h_M)$ has the desired upper bound.

For any finite $M > 0$, h_M satisfies the assumptions of Lemma 4.2; both f and f' have polynomial growth (linear and zero growth, respectively) and

$$\int_{\mathbb{R}} h_M''(u)^2/\gamma_\tau(u) = \int_{|u|\le M} f''(u)^2/\gamma_\tau(u) \le \|f''\|_\gamma^2/(\gamma_\tau(M)\gamma(M)) < \infty \ .$$

Hence,

$$\begin{aligned}\|h_M\|_{\mathcal{H}}^2 &\le 6\tau \left(\int_{-\infty}^{\infty} \frac{{h_M}''(u)^2}{\gamma_\tau(u)} du + \|h_M\|_\gamma^2 + 6\|h_M'\|_\gamma^2 + 2\langle h_M, h_M''\rangle_\gamma \right) \\ &\le 6\tau\Big(\int_{-M}^{M} \frac{f''(u)^2}{\gamma_\tau(u)\gamma(u)} d\gamma(u) + \|h_M\|_\gamma^2 + 6\|h_M'\|_\gamma^2 + 2\|f\|_\gamma\|f''\|_\gamma \Big) \\ &\le 6\tau\Big(\frac{\|f''\|_\gamma^2}{\gamma_\tau(M)\gamma(M)} + 2(\|f\|_\gamma^2 + \|r_M\|_\gamma^2) + 12(\|f'\|_\gamma^2 + \|r_M'\|_\gamma^2) + \|f\|_\gamma^2 + \|f''\|_\gamma^2 \Big) \ ,\end{aligned} \tag{20}$$

where $r_M = f - h_M$ and we used the triangle inequality and $2ab \le a^2 + b^2$ in Eq. (20). Note that since $\tau > 1$, the first term of Eq. (20) is upper bounded by $\lesssim \tau^2 \|f''\|_\gamma^2 e^{(\tau^2+1)M^2/(2\tau^2)}$.

We now upper bound $\|r_M\|_\gamma$ and $\|r'_M\|_\gamma$. Note that both r_M and its derivative r'_M are identically zero on $[-M, M]$ and that $r''_M(t) = f''(t)$ for $|t| > M$. Thus, for $t > M$ (same holds for $t > -M$),

$$r'_M(t) = \int_M^t r''_M(u)du = \int_M^t f''(u)du \ , \text{ and } \ r_M(t) = \int_M^t r'_M(u)du \ .$$

Next, we decompose $\|r'_M\|_\gamma^2$ into two terms, the positive part $\|r'_M\|_{\gamma,+}^2$ and the negative part $\|r'_M\|_{\gamma,-}^2$. That is,

$$\|r'_M\|_\gamma^2 = \int_{-\infty}^{\infty} r'_M(u)^2 du = \int_0^\infty r'_M(u)^2 du + \int_{-\infty}^0 r'_M(u)^2 du := \|r'_M\|_{\gamma,+}^2 + \|r'_M\|_{\gamma,-}^2 \ .$$

We show an upper bound for $\|r'_M\|_{\gamma,+}^2$ with the understanding that the same upper bound applies to $\|r'_M\|_{\gamma,-}^2$. By repeated applications of Fubini's Theorem and an upper bound on the complementary error function $\mathrm{erfc}(t) = \int_t^\infty e^{-u^2/2}du \le 2e^{-t^2/2}$ [18, Theorem 1],

$$\begin{aligned}
\|r'_M\|_{\gamma,+}^2 &= \int r'_M(t)^2 \gamma(t)dt \\
&= \int \left(\int_M^\infty f''(u)\mathbb{1}[0 \le u \le t]du \int_M^\infty f''(w)\mathbb{1}[0 \le w \le t]dw \right) \gamma(t)dt \\
&\le \int_M^\infty \int_M^\infty |f''(u)||f''(w)| \left(\int \mathbb{1}[0 \le u \le t]\mathbb{1}[0 \le w \le t]\gamma(t)dt \right) dudw \\
&\le \int_M^\infty \int_M^\infty |f''(u)||f''(w)|\gamma(\max\{u,w\})dudw && \text{[18, Theorem 1]} \\
&\le \sqrt{8\pi} \left(\int_M^\infty |f''(u)|\gamma_{\sqrt{2}}(u)du \right)^2 \\
&\lesssim \left(\int_M^\infty \gamma_2(u)^{4/3}du \right)^{3/2} \cdot \left(\int (f''(u)\gamma_2(u))^4 du \right)^{1/2} && \text{[Hölder's inequality]} \\
&\lesssim \gamma_{\sqrt{3}}(M)^{3/2} \cdot \|f''\|_4^2 \ . && (21)
\end{aligned}$$

An upper bound on $\|r_M\|_{\gamma,+}^2$ follows from similar calculations.

$$\begin{aligned}
\|r_M\|_{\gamma,+}^2 &= \int_{\mathbb{R}_+} r_M^2(t)\gamma(t)dt \\
&= \int_M^\infty \int_M^\infty r'_M(u)r'_M(w) \left(\int_{\mathbb{R}_+} \mathbb{1}[0 \le u \le t] \cdot \mathbb{1}[0 \le w \le t]\gamma(t)dt \right) dudw \\
&\le 2\int_M^\infty \int_M^\infty r'_M(u)r'_M(w)\gamma(\max\{u,w\})dudw \\
&\le 2\int_M^\infty \int_M^\infty \left(\int_M^\infty |f''(\tilde{u})|\, \mathbb{1}[0 \le \tilde{u} \le u]d\tilde{u} \right) \left(\int_M^\infty |f''(\tilde{w})|\, \mathbb{1}[0 \le \tilde{w} \le w]d\tilde{w} \right) \cdot \gamma(\max\{u,w\})dudw \\
&\le 2\int_M^\infty \int_M^\infty |f''(\tilde{u})||f''(\tilde{w})| \left(\int \int \mathbb{1}[0 \le \tilde{u} \le u]\, \mathbb{1}[0 \le \tilde{w} \le w]\gamma(\max\{u,w\})dudw \right) d\tilde{u}d\tilde{w} \\
&\lesssim \int_M^\infty \int_M^\infty |f''(\tilde{u})||f''(\tilde{w})| \left(\int \mathbb{1}[0 \le \tilde{u} \le u]\gamma_{\sqrt{2}}(u)du \right) \left(\int \mathbb{1}[0 \le \tilde{w} \le w]\gamma_{\sqrt{2}}(w))dw \right) d\tilde{u}d\tilde{w} \\
&\lesssim \left(\int_M^\infty |f''(\tilde{u})|\gamma_{\sqrt{2}}(\tilde{u})d\tilde{u} \right)^2 \\
&\lesssim \left(\int_M^\infty \gamma_2(\tilde{u})^{4/3}d\tilde{u} \right)^{3/2} \cdot \left(\int (f''(\tilde{u})\gamma_2(\tilde{u}))^4 d\tilde{u} \right)^{1/2} \\
&\lesssim \gamma_{\sqrt{3}}(M)^{3/2} \cdot \|f''\|_4^2 \ ,
\end{aligned}$$

Putting everything together in Eq. (20),

$$\|h_M\|_{\mathcal{H}}^2 \lesssim \tau\Big(\frac{\|f''\|_\gamma^2}{\gamma_\tau(M)\gamma(M)} + (\|f\|_\gamma^2 + \|r_M\|_\gamma^2) + (\|f'\|_\gamma^2 + \|r_M'\|_\gamma^2) + \|f\|_\gamma^2 + \|f''\|_\gamma^2\Big)$$
$$\lesssim \tau\Big(\tau\|f''\|_\gamma^2 \cdot e^{\frac{1+\tau^2}{2\tau^2}\cdot M^2} + \|f\|_\gamma^2 + \|f'\|_\gamma^2 + \|f''\|_\gamma^2 + \|f''\|_4^2 \cdot e^{-\frac{M^2}{4}}\Big) .$$

Thus,

$$\begin{aligned}
\mathcal{E}(h_M) &= \|r_M\|_\gamma^2 + \lambda\|h_M\|_{\mathcal{H}}^2 \\
&\lesssim \|f''\|_4^2 \cdot e^{-\frac{M^2}{4}} + \lambda\tau\Big(\tau\|f''\|_\gamma^2 \cdot e^{\frac{1+\tau^2}{2\tau^2}\cdot M^2} + \|f\|_\gamma^2 + \|f'\|_\gamma^2 + \|f''\|_\gamma^2 + \|f''\|_4^2 \cdot e^{-\frac{M^2}{4}}\Big) \\
&\lesssim (1+\lambda\tau)\|f''\|_4^2 \cdot e^{-\frac{M^2}{4}} + \lambda\tau\Big(\tau\|f''\|_\gamma^2 \cdot e^{\frac{1+\tau^2}{2\tau^2}\cdot M^2} + \|f\|_\gamma^2 + \|f'\|_\gamma^2 + \|f''\|_\gamma^2\Big) \\
&\lesssim \tau\|f''\|_4^2 \cdot e^{-\frac{M^2}{4}} + \lambda\tau\Big(\tau\|f''\|_\gamma^2 \cdot e^{\frac{1+\tau^2}{2\tau^2}\cdot M^2} + \|f\|_\gamma^2 + \|f'\|_\gamma^2 + \|f''\|_\gamma^2\Big) , \qquad (22)
\end{aligned}$$

where we used the fact that $\tau > \max\{1, \lambda\tau\}$ in Eq. (22).

It remains to balance the terms in Eq. (22) by choosing an appropriate value for $M > 0$. We choose M by balancing $\tau\|f''\|_4^2 \cdot e^{-\frac{M^2}{4}}$ and $\lambda\tau^2\|f''\|_\gamma^2 \cdot e^{\frac{1+\tau^2}{2\tau^2}\cdot M^2}$.

$$\tau\|f''\|_4^2 \cdot e^{-\frac{M^2}{4}} = \lambda\tau^2\|f''\|_\gamma^2 \cdot e^{\frac{1+\tau^2}{2\tau^2}\cdot M^2} \Rightarrow M^2 = \frac{4\tau^2}{1+3\tau^2}\log\Big(\frac{\|f''\|_4^2}{\lambda\tau\|f''\|_\gamma^2}\Big) .$$

Let $\beta = 1 - \frac{1+\tau^2}{2\tau^2}\cdot\frac{4\tau^2}{1+3\tau^2} = \frac{\tau^2-1}{1+3\tau^2} = \frac{1-1/\tau^2}{3+1/\tau^2}$. Plugging the above value of M into Eq. (22),

$$\mathcal{E}(h_M) \lesssim \tau^{1+\beta}\|f''\|_4^2(\lambda\|f''\|_\gamma^2/\|f''\|_4^2)^\beta + \lambda C_f^2 \le \tau^{1+\beta}\|f''\|_4^2\lambda^\beta + \lambda C_f^2 ,$$

where $C_f = \max\{\|f\|_\gamma, \|f'\|_\gamma, \|f''\|_\gamma\}$.

Finally, let us use a density argument to extend the result to general $f \in H^2(\gamma)$ such that $f'' \in L^4(\gamma)$. We can consider a sequence $f_\rho = U_\rho f$ with $\rho \to 1$, where U_ρ is the Ornstein-Uhlenbeck semigroup given by $U_\rho f(x) = \mathbb{E}_z[f(\rho x + \sqrt{1-\rho^2}z)]$. We verify that $f_\rho \in C^2(\mathbb{R})$ for any $\rho < 1$. Moreover, from $f_\rho' = \rho U_\rho[f']$, $f_\rho'' = \rho^2 U_\rho[f'']$ and the fact that the semigroup is strongly continuous in $L^p(\gamma)$ for any $p \ge 1$, we obtain that $f_\rho \to f$, $f_\rho' \to f'$ and $f_\rho'' \to f''$ in $L^2(\gamma)$ as well as in $L^4(\gamma)$, as $\rho \to 1$. As a result,

$$\begin{aligned}
A(f,\lambda) &\le \|f - f_\rho\|_\gamma^2 + A(f_\rho,\lambda) \\
&\le \|f - f_\rho\|_\gamma^2 + C\tau^{1+\beta}\|f_\rho''\|_4^2\lambda^\beta + \lambda C_{f_\rho}^2 \to C\tau^{1+\beta}\|f''\|_4^2\lambda^\beta + \lambda C_f^2 \qquad (\text{as } \rho \to 1) .
\end{aligned}$$

□

C.3 Random feature approximation

Lemma C.2 (Random features approximation, adapted from [6, Prop 1])**.** *Let $\delta \in (0,1)$, $\tau > 1$, $\hat{\Sigma} = \mathcal{T}^*\mathcal{T}$, and let $\hat{P}_\lambda = \hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}$ be the regularized projection onto the random feature space $\hat{\mathcal{H}}$. There exists a universal constant $C > 0$ such that if $N \ge \frac{C}{\lambda}\log\frac{1}{\lambda\delta}$, then with probability at least $1-\delta$, for any $f \in L^2(\gamma)$ with $\mathbb{E}[f] = 0$, the following holds.*

$$\|(I - \hat{P}_\lambda)f\|_\gamma^2 \le 4A(f,\lambda) . \qquad (23)$$

Proof. The claim follows from Lemmas C.3 and C.4 in Section C.3. Note that the zero-mean assumption $\mathbb{E}[f] = 0$ is necessary to obtain a tight enough bound of these degrees of freedom. □

Before stating and proving the two supporting lemmas, we introduce several terms are used to study the similarity of a finite random feature model to its infinite counterpart. Let $\hat{\kappa}(u,v)$ be the random

empirical kernel associated with N random features:

$$\hat{\kappa}(u,v) = \frac{1}{N}\sum_{i=1}^{N} \phi_{b_i}^{\varepsilon_i}(u)\phi_{b_i}^{\varepsilon_i}(v) \ , \tag{24}$$

where $\phi_b^\varepsilon(u) := \phi(\varepsilon u - b)$, b_i, ε_i are i.i.d. random variables drawn from $\gamma_\tau \otimes \text{Rad}$. Its associated integral operator in $L_2(\gamma)$ is given by $\hat{\Sigma}$.

In the following, we consider Σ the integral operator corresponding to the kernel κ:

$$\Sigma f(u) = \int f(v)\kappa(u,v)d\gamma(v) \ .$$

By a technical lemma adapted from [6], the approximation error of the random feature model is controlled via the regularization parameter $\lambda > 0$.

Lemma C.3 (Random features approximation, adapted from [6])**.** *Let $\delta \in (0,1)$, and let $\mathbb{P}$ be the orthogonal projection operator on span$\{h_j, j \geq 1\}$ in $L^2(\gamma)$. Define*

$$d_{\max}(\lambda) := \sup_{b\in\mathbb{R},\varepsilon\in\{\pm 1\}} \langle \mathbb{P}\phi_b^\varepsilon, (\mathbb{P}\Sigma\mathbb{P} + \lambda I)^{-1}\mathbb{P}\phi_b^\varepsilon\rangle_\gamma. \tag{25}$$

There exists a constant $C > 0$ such that if $N \geq Cd_{\max}(\lambda)\log(d_{\max}(\lambda)/\delta)$, we have, with probability at least $1-\delta$, for any $f \in L^2(\gamma)$,

$$\|(I - \hat{P}_\lambda)\mathbb{P}f\|_\gamma^2 \leq 4A(\mathbb{P}f, \lambda) \ . \tag{26}$$

Proof. Assume for now that $f \in \mathcal{H}$ and denote $\tilde{f} := \mathbb{P}f$. Note that we have

$$\begin{aligned}
\|(I - \hat{P}_\lambda)\tilde{f}\|_\gamma^2 &= \|(I - \hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1})\tilde{f}\|_\gamma^2 \\
&= \|((\hat{\Sigma} + \lambda I) - \hat{\Sigma})(\hat{\Sigma} + \lambda I)^{-1}\tilde{f}\|_\gamma^2 \\
&= \|\lambda(\hat{\Sigma} + \lambda I)^{-1}\tilde{f}\|_\gamma^2 \\
&= \lambda^2\langle \tilde{f}, (\hat{\Sigma} + \lambda I)^{-2}\tilde{f}\rangle_\gamma \\
&\leq \lambda\langle \tilde{f}, (\hat{\Sigma} + \lambda I)^{-1}\tilde{f}\rangle_\gamma = \lambda\langle f, \mathbb{P}(\hat{\Sigma} + \lambda I)^{-1}\mathbb{P}f\rangle_\gamma \\
&= \lambda\langle f, \mathbb{P}(\mathbb{P}\hat{\Sigma}\mathbb{P} + \lambda I)^{-1}\mathbb{P}f\rangle_\gamma = \lambda\langle \tilde{f}, (\mathbb{P}\hat{\Sigma}\mathbb{P} + \lambda I)^{-1}\tilde{f}\rangle_\gamma,
\end{aligned}$$

where the last line uses the fact that $\mathbb{P}$ is a projection operator. Note that $\mathbb{P}\hat{\Sigma}\mathbb{P}$ is the integral operator of the random feature kernel with features $\mathbb{P}\phi_{b_i}^{\varepsilon_i}$. We now apply [6] to this projected kernel to control random feature approximation. From the proof of [6, Prop. 1], the following holds with probability $1-\delta$ (see [6, end of p.37]): for any $g \in L^2(\gamma)$,

$$\langle g, (\mathbb{P}\hat{\Sigma}\mathbb{P} + \lambda I)^{-1}g\rangle_\gamma \leq 4\langle g, (\mathbb{P}\Sigma\mathbb{P} + \lambda I)^{-1}g\rangle_\gamma,$$

as long as $N \geq Cd_{\max}(\lambda)\log(d_{\max}(\lambda)/\delta)$. Now note that we have

$$\lambda\langle \mathbb{P}f, (\mathbb{P}\Sigma\mathbb{P} + \lambda I)^{-1}\mathbb{P}f\rangle_\gamma = \lambda\langle \mathbb{P}f, (\Sigma + \lambda I)^{-1}\mathbb{P}f\rangle_\gamma = A(\mathbb{P}f, \lambda),$$

where the last equality follows from [7, Lemma 7.2]. Thus, we have proved the result for $f \in \mathcal{H}$. Given that (23) does not require f to be in $\mathcal{H}$, we may conclude by limiting arguments that the result holds for any f in the closure of $\mathcal{H}$, which includes $L^2(\gamma)$, since the kernel is universal, given that its associated RKHS is a weighted Sobolev Space, which is dense in $L^2(\gamma)$. □

Lemma C.4 (Degrees of freedom)**.** *We have $d_{\max}(\lambda) \leq C/\lambda$, for an absolute constant $C > 0$.*

Proof. Recall that $d_{\max}(\lambda) := \sup_{b\in\mathbb{R},\varepsilon\in\{\pm 1\}}\langle \mathbb{P}\phi_b^\varepsilon, (\mathbb{P}\Sigma\mathbb{P} + \lambda I)^{-1}\mathbb{P}\phi_b^\varepsilon\rangle_\gamma$. We consider two cases separately.

If $b \geq 0$, then we have $|\phi_b^\varepsilon(u)| \leq |u|$ for all u, thus

$$\langle \mathbb{P}\phi_b^\varepsilon, (\mathbb{P}\Sigma\mathbb{P} + \lambda I)^{-1}\mathbb{P}\phi_b^\varepsilon\rangle_\gamma \leq \frac{1}{\lambda}\|\mathbb{P}\phi_b^\varepsilon\|_\gamma^2 \leq \frac{1}{\lambda}\|\phi_b^\varepsilon\|_\gamma^2 \leq \frac{C}{\lambda},$$

with $C = 2\int_0^\infty u^2\gamma(u)du$.

If $b \le 0$, we may write

$$\phi_b^\varepsilon(u) = \phi_{-b}^{-\varepsilon}(u) + g(u),$$

with $g(u) = \epsilon u - b$ a linear function, using the relation $\max(0,u) = u + \max(0,-u)$. Then we have

$$\begin{aligned}\langle \mathbb{P}\phi_b^\varepsilon, (\mathbb{P}\Sigma\mathbb{P} + \lambda I)^{-1}\mathbb{P}\phi_b^\varepsilon\rangle_\gamma &\le \frac{1}{\lambda}\|\mathbb{P}\phi_b^\varepsilon\|_\gamma^2 \\ &\le \frac{2}{\lambda}(\|\phi_{-b}^{-\varepsilon}\|_\gamma^2 + \|\mathbb{P}g\|_\gamma^2) \le \frac{4C}{\lambda},\end{aligned}$$

with the same constant C as in the previous case, since both $\phi_{-b}^{-\varepsilon}$ and $\mathbb{P}g(u) = (g - \mathbb{E}[g])(u) = \varepsilon u$ are controlled by $u \mapsto |u|$ in absolute value. □

D Proofs for Section 5

D.1 Exact expression of first-order critical points

We first derive critical point equations for the regularized population loss.

Claim D.1 (Critical point equations)**.** *Let $(\alpha_j)_{j\in\mathbb{N}}$ be the Hermite coefficients of $f_* \in L_2(\gamma)$, let $m \in [-1,1]$, and let $g_m, \bar{g}_m \in L^2(\gamma)$ be defined by*

$$g_m(z) := \sum_{j=s}^\infty \alpha_j m^j h_j(z)\ ,\ \bar{g}_m(z) := \sum_{j=s}^\infty \alpha_j j m^{j-1} h_j(z)\ . \tag{27}$$

Then, denoting $m = \langle\theta^*, \theta\rangle$, we have

$$\bar{L}(\theta) = -\langle \hat{P}_\lambda g_m, g_m\rangle_\gamma + (1+\sigma^2) = -\|g_m\|_\gamma^2 + \langle (I - \hat{P}_\lambda)g_m, g_m\rangle_\gamma + (1+\sigma^2)\ . \tag{28}$$

Furthermore, the critical points of $L(c,\theta)$ satisfy the following equations:

$$\begin{aligned} c &= Q_\lambda^{-1}\sum_{j=s}^\infty \alpha_j m^j \mathcal{T}_j\ , \\ 0 &= -\sum_{j=s}^\infty \alpha_j^2 j m^{2j-1} + \langle (I - \hat{P}_\lambda)g_m, \bar{g}_m\rangle_\gamma\ . \end{aligned}$$

We prove Claim D.1 in Appendix D.3 by analyzing the population gradient and relating the critical points of $\bar{L}(\theta)$ to those of $L(c,\theta)$. Note that the function g_m corresponds to the minimizer of $\|f_*(\langle\theta^*,\cdot\rangle) - g(\langle\theta,\cdot\rangle)\|_{\gamma_d}$, which is essentially the optimal function we may learn from fitting the second layer c with no regularization when θ is fixed.

D.2 Proof of Theorem 5.3

The intuition behind Theorem 5.3 is as follows. We first observe that for fixed $\theta \in \mathbb{S}^{d-1}$ and $\lambda > 0$, the population loss $L(c,\theta)$ is strictly convex in c. Hence, if (c,θ) is a critical point of $L(c,\theta)$, then $L(c,\theta) = \bar{L}(\theta)$. Since $\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta) = 0$, it follows that θ must also be a critical point of the projected population loss $\bar{L}(\theta)$. Now consider the idealized (and impossible) setting in which N is large enough to exactly express any function in $L^2(\gamma)$ and there is no ℓ^2 regularization (i.e., $\lambda = 0$). Then, the projected population loss is

$$\bar{L}(\theta) = \min_{g\in L^2(\gamma)} \|f_*(\langle\theta^*,\cdot\rangle) - g(\langle\theta,\cdot\rangle)\|_{\gamma_d}^2 + \sigma^2 = -\sum_{j=s}^\infty \alpha_j^2 m^{2j} + (1+\sigma^2)\ ,$$

where $m = \langle\theta^*,\theta\rangle$. $\bar{L}(\theta)$ in the ideal case is strictly decreasing in $|m| \in (0,1]$. Using the expression for $\bar{L}(\theta)$ in Eq. (28), we observe that by setting $\lambda > 0$ sufficiently small (and N proportional to $1/\lambda$),

the projection $\hat{P}_\lambda$ onto the subspace $\hat{\mathcal{H}}$ spanned by random features approximates the identity map in the operator norm, thereby preserving the strict monotonicity of $\bar{L}(\theta)$ with respect to $|m|$. We formalize this intuition in the following proof.

We first establish the following lemma (proved in Section D.4) which controls the approximation error for the functions g_m defined in (27).

Lemma D.2 (Uniform approximation error for g_m)**.** *Under the regularity assumptions on f_* (Assumption 5.2), there exists a universal constant $C > 0$ and a constant $\tilde{K} \geq 1$ depending only on f_* and* not *on m, such that for all $|m| \leq 1$,*

$$A(g_m, \lambda) \leq C(\tau^{1+\beta}\tilde{K}^2\|g_m''\|_\gamma^2\lambda^\beta + \lambda C_{g_m}^2)\ , \tag{29}$$

where $\beta = \frac{1-1/\tau^2}{3+1/\tau^2}$ and $C_{g_m} = \max\{\|g_m\|_\gamma, \|g_m'\|_\gamma, \|g_m''\|_\gamma\}$.

Proof of Theorem 5.3. For (c, θ) to be a critical point of L, it must satisfy $\nabla_\theta^{\mathbb{S}^{d-1}} L(c, \theta) = 0$ and $\nabla_c L(\theta, c) = 0$. By Claim D.1, for any $m = \langle\theta^*, \theta\rangle \in [-1, 1]$, there is a unique $c \in \mathbb{R}^N$ such that $\nabla_c L(\theta, c) = 0$ and $\nabla_\theta^{\mathbb{S}^{d-1}} L(c, \theta) = 0$ if and only if $\theta \in \{-\theta^*, \theta^*\}$, or

$$\sum_{j=s}^{\infty} \alpha_j^2 j m^{2j-1} = \langle (I - \hat{P}_\lambda) g_m, \bar{g}_m\rangle_\gamma. \tag{30}$$

We show that the latter condition is true only if $m = 0$ by deriving a contradiction whenever $m \neq 0$. Note that by the regularity assumption on f_* (Assumption 5.2), it also holds that $\|f_*'\|_\gamma, \|f_*''\|_\gamma < \infty$. We contradict the equality in (30) with probability at least $1 - \delta$ over the randomly sampled biases $b_1, \ldots, b_N$ and signs. Let $\tilde{K} \geq 1$ be the constant from Lemma D.2 and $C_{f_*} = \max\{\|f_*\|_\gamma, \|f_*'\|_\gamma, \|f_*''\|_\gamma\}$. Define the threshold for λ by

$$\lambda^* := \left(\frac{4\sqrt{C\tau^{1+\beta}}\tilde{K}C_{f_*}^2}{\alpha_s^2 s}\right)^{-2/\beta}. \tag{31}$$

If $\lambda < \lambda^*$ and $N \geq \frac{C}{\lambda}\log\frac{1}{\lambda\delta}$, then with probability greater than $1 - \delta$ it holds

$$\begin{aligned}
\left|\langle (I - \hat{P}_\lambda) g_m, \bar{g}_m\rangle_\gamma\right| &\leq \|(I - \hat{P}_\lambda) g_m\|_\gamma \|\bar{g}_m\|_\gamma \\
&\leq 2\sqrt{A(g_m, \lambda)}\|\bar{g}_m\|_\gamma && \text{[Lemma C.2]} \\
&\leq 2\sqrt{C(\tau^{1+\beta}\tilde{K}^2\|g_m''\|_\gamma^2\lambda^\beta + \lambda C_{g_m}^2)}\|\bar{g}_m\|_\gamma && \text{[Lemma D.2]} \\
&\leq 2\lambda^{\beta/2} C_{g_m}\sqrt{2C\tau^{1+\beta}}\tilde{K}\|\bar{g}_m\|_\gamma \\
&\leq 4\lambda^{\beta/2}\,|m|^{2s-1}\sqrt{C\tau^{1+\beta}}\tilde{K}C_{f_*}^2 && \text{[Lemma D.8]} \\
&< \alpha_s^2 s\,|m|^{2s-1} \leq \sum_{j=0}^{\infty}\alpha_j^2 j\,|m|^{2j-1} = \left|\sum_{j=0}^{\infty}\alpha_j^2 j m^{2j-1}\right|, && \text{[Eq (31)]}\ ,
\end{aligned}$$

which contradicts (30). Therefore, the existence of critical points satisfying $m \notin \{-1, 0, 1\}$ is ruled out with probability at least $1 - \delta$ over the random features. □

Remark D.3 (Robust version of Theorem 5.3)**.** *The proof of Theorem 5.3 in fact implies a stronger result. It implies that if $\|\nabla^{\mathbb{S}^{d-1}}\bar{L}(\theta)\| \approx 0$, i.e., if θ is nearly a critical point for the projected population loss $\bar{L}$, then $|m| \approx 0$ or $|m| \approx 1$. This is formally stated in Lemma E.9.*

D.3 First-order critical points of the population loss and proof of Claim D.1

To characterize the critical points of $L(c, \theta)$ for fixed random features and prove Claim D.1, we derive exact expressions for L and its gradients. We observe that the population loss depends on the student direction θ only via its angle to the teacher direction θ^*.

Proposition D.4. *The ℓ^2-regularized population loss is given by*

$$L(c,\theta) = 1 + c^\top Q_\lambda c - 2\langle c, \sum_{j=s}^{\infty} \alpha_j m^j \mathcal{T}_j\rangle + \sigma^2 \ . \tag{32}$$

Proof. Recall the decomposition $c^\top \Phi(z) = \sum_{j=0}^{\infty} \langle \mathcal{T}_j, c\rangle h_j(z)$. Straightforward calculation gives

$$\begin{aligned}
L(c,\theta) &= \|(f(\langle\theta^*,\cdot\rangle) + \xi) - c^\top \Phi(\langle\theta,\cdot\rangle)\|_{\gamma_d}^2 + \lambda\|c\|^2 \\
&= \|f\|_\gamma^2 + \|c^\top\Phi\|_\gamma^2 - 2\langle f(\langle\theta^*,\cdot\rangle), c^\top\Phi(\langle\theta,\cdot\rangle)\rangle_{\gamma_d} + \sigma^2 + \lambda\|c\|^2 \\
&= 1 + \sum_{j=0}^{\infty}\langle \mathcal{T}_j, c\rangle^2 - 2\sum_{j=s}^{\infty}\langle \mathcal{T}_j, c\rangle \alpha_j \langle\theta,\theta^*\rangle^j + \sigma^2 + \lambda\,\|c\|^2 \\
&= 1 + c^\top Q_\lambda c - 2\langle c, \sum_{j=s}^{\infty}\alpha_j m^j \mathcal{T}_j\rangle + \sigma^2 \ .
\end{aligned}$$

□

Eq. (32) gives us an expression for gradients with respect to $\theta \in \mathbb{S}^{d-1}$ and $c \in \mathbb{R}^N$.

Corollary D.5 (Gradient of population loss)**.** *The (non-spherical) gradient with respect to the student direction $\theta \in \mathbb{S}^{d-1}$ and $c \in \mathbb{R}^N$ are given by*

$$\nabla_\theta L = -\left(\sum_{j=s}^{\infty}\langle \mathcal{T}_j, c\rangle j\alpha_j m^{j-1}\right)\theta^* \quad \textit{and} \quad \nabla_c L = 2\left(Q_\lambda c - \sum_{j=s}^{\infty}\alpha_j m^j \mathcal{T}_j\right) \ . \tag{33}$$

Recall that the criticality of θ depends on the *spherical* gradient being zero, not the standard one. Since the gradient $\nabla_\theta L$ is colinear with θ^*, we stipulate necessary and sufficent conditions for θ to be critical.

Corollary D.6 (Projection onto the sphere)**.** $\nabla_\theta^{\mathbb{S}^{d-1}} L = 0$ *if and only if either (i) $\theta = \theta^*$ (i.e., $m = 1$) or (ii) $\sum_{j=s}^{\infty}\langle \mathcal{T}_j, c\rangle j\alpha_j m^{j-1} = 0$.*

Identifying the critical values of θ is sufficient because the *unique* critical point for c exists as the solution to a linear system. To elaborate, $Q_\lambda \in \mathbb{R}^{N\times N}$ is non-singular for $\lambda > 0$, so if $\theta \in \mathbb{S}^{d-1}$ is fixed, then c is given as the solution to a linear system of equations: $c = Q_\lambda^{-1}(\sum_j \alpha_j m^j \mathcal{T}_j)$.

By Proposition D.4, $L(c,\theta)$ is strictly convex with respect to $c \in \mathbb{R}^N$ for any fixed $\theta \in \mathbb{S}^{d-1}$. Hence, if (c,θ) is a critical point of L, then θ must be a critical point of $\bar{L}$.

Lemma D.7 (Critical points of the *projected* population loss)**.** *Recall that $g_m(z) = \sum_{j=s}^{\infty}\alpha_j m^j h_j(z)$. Then, the projected population loss $\bar{L}$ is given by*

$$\begin{aligned}
\bar{L}(\theta) &= -\sum_{j,j'}\alpha_j\alpha_{j'} m^{j+j'}\langle h_j, \hat{P}_\lambda h_{j'}\rangle_\gamma + (1+\sigma^2) \\
&= -\langle g_m, \hat{P}_\lambda g_m\rangle_\gamma + (1+\sigma^2) \ .
\end{aligned} \tag{34}$$

Furthermore, critical points of $\bar{L}$ satisfy the following equation.

$$\sum_{j,j'}\alpha_j\alpha_{j'}(j+j')m^{j+j'-1}\langle h_j, P_\lambda h_{j'}\rangle_\gamma = 0 \ . \tag{35}$$

Proof of Lemma D.7. Because $\mathcal{T}$ is full-rank, let $\mathcal{T} = U\Lambda\mathcal{V}$ be its SVD for some $U \in \mathbb{R}^{N\times N}$, diagonal $\Lambda \in \mathbb{R}^{N\times N}$, and $\mathcal{V} : L^2(\mu) \to \mathbb{R}^N$. Then, $\mathcal{T}_j = U\Lambda\mathcal{V}h_j$, $Q_\lambda = U^\top(\Lambda^2 + \lambda I_N)U$, and $Q_\lambda^{-1} = U^\top(\Lambda^2 + \lambda I_N)^{-1}U$. Similarly, $\hat{P}_\lambda = \mathcal{V}^*\Lambda(\Lambda^2 + \lambda I_N)^{-1}\Lambda\mathcal{V}$. As a result,

$$\langle \mathcal{T}_j, Q_\lambda^{-1}\mathcal{T}_{j'}\rangle = \langle h_j, \mathcal{V}^*\Lambda(\Lambda^2 + \lambda I_N)^{-1}\Lambda\mathcal{V}h_{j'}\rangle_\gamma = \langle h_j, \hat{P}_\lambda h_{j'}\rangle_\gamma.$$

Now we plug in $c = Q_\lambda^{-1}(\sum_j \alpha_j m^j \mathcal{T}_j)$ into Eq. (32). Then, we have

$$\begin{aligned}\bar{L}(\theta) - (1+\sigma^2) &= -\left\langle \sum_j \alpha_j m^j \mathcal{T}_j, Q_\lambda^{-1} \sum_{j'} \alpha_{j'} m^{j'} \mathcal{T}_{j'} \right\rangle \\ &= -\sum_{j,j'} \alpha_j \alpha_{j'} m^j m^{j'} \left\langle \mathcal{T}_j, Q_\lambda^{-1} \sum_{j'} \mathcal{T}_{j'} \right\rangle \\ &= -\sum_{j,j'} \alpha_j \alpha_{j'} m^j m^{j'} \langle h_j, \hat{P}_\lambda h_{j'} \rangle_\gamma \ .\end{aligned}$$

Differentiating $\bar{L}(\theta)$ with respect to $m = \langle \theta, \theta^* \rangle$, we obtain the following critical point equation.

$$\sum_{j,j'} \alpha_j \alpha_{j'} (j+j') m^{j+j'-1} \langle h_j, P_\lambda h_{j'} \rangle_\gamma = 0 \ .$$

□

Claim D.1 (Critical point equations)**.** *Let $(\alpha_j)_{j\in\mathbb{N}}$ be the Hermite coefficients of $f_* \in L_2(\gamma)$, let $m \in [-1,1]$, and let $g_m, \bar{g}_m \in L^2(\gamma)$ be defined by*

$$g_m(z) := \sum_{j=s}^{\infty} \alpha_j m^j h_j(z) \ , \ \bar{g}_m(z) := \sum_{j=s}^{\infty} \alpha_j j m^{j-1} h_j(z) \ . \tag{27}$$

Then, denoting $m = \langle \theta^*, \theta \rangle$, we have

$$\bar{L}(\theta) = -\langle \hat{P}_\lambda g_m, g_m \rangle_\gamma + (1+\sigma^2) = -\|g_m\|_\gamma^2 + \langle (I - \hat{P}_\lambda) g_m, g_m \rangle_\gamma + (1+\sigma^2) \ . \tag{28}$$

Furthermore, the critical points of $L(c,\theta)$ satisfy the following equations:

$$\begin{aligned} c &= Q_\lambda^{-1} \sum_{j=s}^{\infty} \alpha_j m^j \mathcal{T}_j \ , \\ 0 &= -\sum_{j=s}^{\infty} \alpha_j^2 j m^{2j-1} + \langle (I - \hat{P}_\lambda) g_m, \bar{g}_m \rangle_\gamma \ . \end{aligned}$$

Proof of Claim D.1. As discussed above (c,θ) is a critical point of L if and only if θ is a critical point of $\bar{L}$ and $c = Q_\lambda^{-1}(\sum_j \alpha_j m^j \mathcal{T}_j)$. By applying Lemma D.7, we separate the diagonal and off-diagonal terms to rewrite Eq. (35) as

$$0 = \sum_j \alpha_j^2 (2j) m^{2j-1} - 2\langle (I - \hat{P}_\lambda) g_m, \bar{g}_m \rangle_\gamma \ .$$

Dividing both sides by 2 gives the claim. □

D.4 Proof of Lemma D.2

Lemma D.2 (Uniform approximation error for g_m)**.** *Under the regularity assumptions on f_* (Assumption 5.2), there exists a universal constant $C > 0$ and a constant $\tilde{K} \geq 1$ depending only on f_* and* not *on m, such that for all $|m| \leq 1$,*

$$A(g_m, \lambda) \leq C(\tau^{1+\beta} \tilde{K}^2 \|g_m''\|_\gamma^2 \lambda^\beta + \lambda C_{g_m}^2) \ , \tag{29}$$

where $\beta = \frac{1-1/\tau^2}{3+1/\tau^2}$ and $C_{g_m} = \max\{\|g_m\|_\gamma, \|g_m'\|_\gamma, \|g_m''\|_\gamma\}$.

Proof. For simplicity, we denote the $L^p(\gamma)$ norms by $\|\cdot\|_p$. For any $\rho \in [0,1]$, we define the noise operator U_ρ by

$$U_\rho f(x) = \mathbb{E}_{z\sim\gamma}[f(\rho x + \sqrt{1-\rho^2} z)]$$

This is a reparametrisation of the *Ornstein-Uhlenbeck* semigroup,[2] and we have from [67, Prop 11.33] that $U_\rho h_j = \rho^j h_j$. In other words, the Hermite polynomials are eigenfunctions of the semigroup. As a consequence, we have from (27) that $g_m = U_m f_*$.

Let us verify that if $f_*'' \in L^4(\gamma)$ and $K = \|f_*''\|_4 / \|f_*''\|_2$, then there exists a constant $\tilde{K} \geq 1$, depending only on f_* and *not* m, such that $\|g_m''\|_4 \leq \tilde{K}\|g_m''\|_2$ for any $m \in [-1,1]$. If $f_*'' \equiv 0$, then $g_m'' \equiv 0$ for any m, and thus $\tilde{K} = 1$. Otherwise, let $\tilde{s}$ denote the information exponent of $f_*'' \neq 0$. That is, if $s \geq 2$, then $\tilde{s} = s-2$, and if $s = 1$, then $\tilde{s} = s_2 - 2$ where s_2 is the second non-zero harmonic of f_*. For $|m| \leq 1/\sqrt{3}$, we have

$$g_m'' = m^2 U_m[f_*''] = m^2 U_{\sqrt{1/3}} U_{m\sqrt{3}}[f_*''] \ .$$

Since $|m\sqrt{3}| \leq 1$, we have

$$\left\|U_{m\sqrt{3}}[f_*'']\right\|_2^2 = \sum_{j=\tilde{s}}^{\infty}(j+2)(j+1)\alpha_{j+2}^2(m\sqrt{3})^{2j} \leq (m\sqrt{3})^{2\tilde{s}} \sum_{j=\tilde{s}}^{\infty}(j+2)(j+1)\alpha_{j+2}^2 = (m\sqrt{3})^{2\tilde{s}}\|f_*''\|_2^2 \ .$$

By Nelson's Gaussian hypercontractivity [65] (reproduced in [67, Theorem 11.23]),

$$\|g_m''\|_4 = m^2 \left\|U_{\sqrt{1/3}}\Big(U_{m\sqrt{3}}[f_*'']\Big)\right\|_4 \leq m^2\|U_{m\sqrt{3}}[f_*'']\|_2 \leq |m|^{\tilde{s}+2} \cdot 3^{\tilde{s}/2}\|f_*''\|_2$$

Since $\|g_m''\|_2 \geq |m|^{\tilde{s}+2}\sqrt{(\tilde{s}+2)(\tilde{s}+1)}|\alpha_{\tilde{s}+2}|$, we obtain that

$$\sup_{|m|\leq 1/\sqrt{3}} \frac{\|g_m''\|_4}{\|g_m''\|_2} \leq \frac{3^{\tilde{s}/2}}{\sqrt{(\tilde{s}+2)(\tilde{s}+1)}|\alpha_{\tilde{s}+2}|} \cdot \|f_*''\|_2 \ .$$

Let us now consider $m \geq 1/\sqrt{3}$. Since U_ρ is an averaging operator for all $\rho \leq 1$, from Jensen's inequality (reproduced in [67, Proposition 11.15]) it holds that $\|U_\rho f\|_p \leq \|f\|_p$ for any $p \geq 1$. We thus have

$$\|g_m''\|_4 \leq m^2\|f_*''\|_4 \ .$$

and $\|g_m''\|_2 \geq 3^{-(\tilde{s}+2)/2}\sqrt{(\tilde{s}+2)(\tilde{s}+1)}|\alpha_{\tilde{s}+2}|$, therefore

$$\sup_{|m|\geq 1/\sqrt{3}} \frac{\|g_m''\|_4}{\|g_m''\|_2} \leq \frac{3^{(\tilde{s}+2)/2}\|f_*''\|_4}{\sqrt{(\tilde{s}+2)(\tilde{s}+1)}|\alpha_{\tilde{s}+2}|} \leq \frac{3^{(\tilde{s}+2)/2}K\|f_*''\|_2}{\sqrt{(\tilde{s}+2)(\tilde{s}+1)}|\alpha_{\tilde{s}+2}|} \ .$$

Hence, we set

$$\tilde{K} = \frac{3^{(\tilde{s}+2)/2}\|f_*''\| \cdot \max\{K, 1/3\}}{\sqrt{(\tilde{s}+2)(\tilde{s}+1)}|\alpha_{\tilde{s}+2}|} = \frac{3^{s/2}K\|f_*''\|}{\sqrt{(\tilde{s}+2)(\tilde{s}+1)}|\alpha_{\tilde{s}+2}|} \ .$$

By Lemma 4.4, we therefore have

$$\forall\, m \in [-1,1],\ A(g_m, \lambda) \leq C(\tau^{1+\beta}\tilde{K}^2\|g_m''\|_\gamma^2\lambda^\beta + C_{g_m}^2\lambda) \ , \tag{36}$$

where $\beta = \frac{1-1/\tau^2}{3+1/\tau^2}$, $C > 0$ is a universal constant, and $C_{g_m} = \max\{\|g_m\|_\gamma, \|g_m'\|_\gamma, \|g_m''\|_\gamma\}$.

□

D.5 Other lemmas for the proof of Theorem 5.3

Lemma D.8 (γ-norm of g_m and $\bar{g}_m$)**.** *Let $f \in H^2(\gamma)$ be such that $f', f'' \in L^2(\gamma)$ and let $s \geq 1$ be its information exponent. Furthermore, let $f_* = \sum_j \alpha_j h_j$ be the Hermite expansion of f_*, and let g_m and $\bar{g}_m$ be defined as in Theorem 5.3. Then,*

$$\|g_m\|_\gamma^2 \leq \|f_*\|_\gamma^2 m^{2s},$$
$$\|g_m''\|_\gamma^2 \leq \|f_*''\|_\gamma^2 m^{2s},$$
$$\|\bar{g}_m\|_\gamma^2 \leq \left(\|f_*''\|_\gamma^2 + \|f_*'\|_\gamma^2\right) m^{2(s-1)} \ .$$

[2] The Ornstein–Uhlenbeck semigroup $\mathcal{P}_t$ is given by $\mathcal{P}_t f(x) = \int f(e^{-t}x + \sqrt{1-e^{-2t}}z)d\gamma(z)$. We thus have $\mathcal{P}_t = U_{e^{-t}}$

Proof. By definition of g_m and Holder's inequality,

$$\|g_m\|_\gamma^2 = \sum_{j=s}^{\infty} \alpha_j^2 m^{2j} \le \|\alpha\|_2^2 \max_{j\ge s} m^{2j} = \|f_*\|_\gamma^2 m^{2s},$$

$$\|g_m''\|_\gamma^2 = \sum_{j=2}^{\infty} j(j-1)\cdot \alpha_j^2 m^{2j} \le \left(\sum_{j=2}^{\infty} j(j-1)\alpha_j^2\right) \max_{j\ge s} m^{2j} = \|f_*''\|_\gamma^2 m^{2s},$$

$$\|\bar{g}_m\|_\gamma^2 = \sum_{j=1}^{\infty} j^2\alpha_j^2 m^{2(j-1)} = \sum_{j=s}^{\infty} j(j-1)\alpha_j^2 m^{2(j-1)} + \sum_{j=s}^{\infty} j\alpha_j^2 m^{2(j-1)} \le m^{2(s-1)}\left(\|f_*''\|_\gamma^2 + \|f_*'\|_\gamma^2\right).$$

□

Corollary D.9. *Let $f \in H^2(\gamma)$ be a function satisfying assumptions of Lemma D.8, let g_m and $\bar{g}_m$ be defined as in Theorem 5.3, and let $C_{g_m} = \max\{\|g_m\|, \|g_m'\|_\gamma, \|g_m''\|_\gamma\}$. Then,*

$$C_{g_m}^2 \le \sum_{j=s}^{\infty} j^2\alpha_j^2 m^{2j} \le 2C_{g_m}^2 \quad \text{and} \quad C_{g_m}\|\bar{g}_m\|_\gamma \le \sqrt{2}\sum_{j=s}^{\infty} j^2\alpha_j^2 m^{2j-1}\ .$$

Proof. The first inequality follows from the following.

$$C_{g_m}^2 \le \sum_{j=s}^{\infty} j^2\alpha_j^2 m^{2j} = \|g_m''\|_\gamma^2 + \|g_m'\|_\gamma^2 \le 2C_{g_m}^2\ .$$

The proof of the second inequality is via straightforward algebraic manipulation.

$$\begin{aligned}
C_{g_m}\|\tilde{g}_m\|_\gamma &\le \left(\sum_{j=s}^{\infty} j^2\alpha_j^2 m^{2j}\right)^{1/2}\left(\sum_{j=s}^{\infty} j^2\alpha_j^2 m^{2(j-1)}\right)^{1/2} \\
&= \left(\sum_{j=s}^{\tilde{s}-1} j^2\alpha_j^2 m^{2j} + \sum_{j=\tilde{s}}^{\infty} j^2\alpha_j^2 m^{2j}\right)^{1/2}\left(\sum_{j=s}^{\tilde{s}-1} j^2\alpha_j^2 m^{2(j-1)} + \sum_{j=\tilde{s}}^{\infty} j^2\alpha_j^2 m^{2(j-1)}\right)^{1/2} \\
&= \left(\Big(\sum_{j=s}^{\tilde{s}-1} j^2\alpha_j^2 |m|^{2j-1}\Big)^2 + \Big(\sum_{j=s}^{\tilde{s}-1} j^2\alpha_j^2 m^{2j}\Big)\Big(\sum_{j=\tilde{s}}^{\infty} j^2\alpha_j^2 m^{2(j-1)}\Big) + \Big(\sum_{j=\tilde{s}}^{\infty} j^2\alpha_j^2 |m|^{2j-1}\Big)^2\right)^{1/2} \\
&= \left(\Big(\sum_{j=s}^{\tilde{s}-1} j^2\alpha_j^2 |m|^{2j-1}\Big)^2 + \Big(\sum_{j=s}^{\tilde{s}-1} j^2\alpha_j^2 |m|^{2j-1}\Big)\Big(\sum_{j=\tilde{s}}^{\infty} j^2\alpha_j^2 |m|^{2j-1}\Big) + \Big(\sum_{j=\tilde{s}}^{\infty} j^2\alpha_j^2 |m|^{2j-1}\Big)^2\right)^{1/2} \\
&\le \left(\Big(\sum_{j=s}^{\tilde{s}-1} j^2\alpha_j^2 |m|^{2j-1}\Big)^2 + \Big(\sum_{j=s}^{\tilde{s}-1} j^2\alpha_j^2 |m|^{2j-1}\Big)^2 + \Big(\sum_{j=\tilde{s}}^{\infty} j^2\alpha_j^2 |m|^{2j-1}\Big)^2 + \Big(\sum_{j=\tilde{s}}^{\infty} j^2\alpha_j^2 |m|^{2j-1}\Big)^2\right)^{1/2} \\
&\le \left(2\Big(\sum_{j=s}^{\tilde{s}-1} j^2\alpha_j^2 |m|^{2j-1}\Big)^2 + 2\Big(\sum_{j=\tilde{s}}^{\infty} j^2\alpha_j^2 |m|^{2j-1}\Big)^2\right)^{1/2} \\
&\le \sqrt{2}\sum_{j=s}^{\infty} j^2\alpha_j^2 |m|^{2j-1}\ .
\end{aligned}$$

where we used the inequalities $ab \le a^2+b^2$, and $\sqrt{a^2+b^2} \le a+b$ which apply to any $a, b \ge 0$. □

E Proofs for Section 6

E.1 Proof of Theorem 6.1

Theorem 6.1 (Gradient flow finds approximate minimizers)**.** *For $\delta \in (0, 1/4)$ and f_* satisfying Assumption 5.2, suppose the following are true: (i) $\lambda = O(1)$ and $\lambda = \Omega(\sqrt{\Delta_{crit}})$, where $\Delta_{crit} :=$*

$\max\{\sqrt{\frac{d+N}{n}}, (\frac{d^2}{n})^{2s/(2s-1)}\}$, *(ii)* $n = \widetilde{\Omega}(\max\{\frac{(d+N)d^{s-1}}{\lambda^4}, \frac{d^{(s+3)/2}}{\lambda^2}\})$, *(iii)* $N = \Omega(\frac{1}{\lambda}\log\frac{1}{\lambda\delta})$ & $N = \widetilde{O}(\lambda\Delta_{crit}^{-1})$, *(iv)* $N_0 = \Theta(\log(\frac{1}{\delta}))$, *(v)* $\rho = \Theta(\sqrt{N}N_0^{-(2+s)/2}(\tau^2 + \lambda N/N_0)^{-1})$, *(vi)* $T_0 = \tilde{\Theta}(d^{s/2-1})$, *and (vii)* $T_1 = \tilde{\Theta}(\frac{\lambda^4 n}{d+N})$. *Then, if we run Procedure 1 for* $T = T_0 + T_1$ *time steps with the above parameters, with probability at least* $\frac{1}{2} - \delta$ *we have*

$$1 - |\langle\theta_T, \theta^*\rangle| \quad = \quad \widetilde{O}\left(\lambda^{-4}\max\left\{\frac{d+N}{n}, \frac{d^4}{n^2}\right\}\right) . \tag{11}$$

Proof. The proof of this theorem has two separate parts: we first prove that our gradient flow procedure escapes the neighborhood of the equator, and then show that it converges to a neighborhood of the north pole. Define the set of approximate-first-order critical points of the empirical landscape in the sublevel set $L_n(c,\theta) \le \nu$.

$$\Omega_n(\epsilon_\theta, \epsilon_c, \nu) := \{(c,\theta); \|\nabla_\theta L_n(c,\theta)\| \le \epsilon_\theta,\ \|\nabla_c L_n(c,\theta)\| \le \epsilon_c; L_n(c,\theta) \le \nu\}$$

Recall from Theorem 5.3 that the structure of critical points of the population landscape $\Omega_\infty(0,0,\infty)$ has two distinct components: the equator $\mathrm{E} = \{\theta; \langle\theta,\theta^*\rangle = 0\}$ and the poles $\theta = \pm\theta^*$, leading to

$$\Omega_\infty(0,0,\infty) = \Omega_\infty(0,0,0) = \underbrace{\{(0,\theta); \theta \in \mathrm{E}\}}_{:=\Omega_\infty^{\mathrm{b}}} \ \sqcup\ \underbrace{\{(zc^*, z\theta^*); z = \{-1,+1\}\}}_{:=\Omega_\infty^{\mathrm{g}}} ,$$

with $c^* = \hat{P}_\lambda f_*$. One would expect that for n sufficiently large, these topological properties should be transferred to the empirical landscape. This intuition is indeed correct, and relies on the following uniform convergence result, proved in Appendix E.2.

Lemma E.1 (Uniform convergence of the empirical landscape)**.** *Let* $d, n, N \in \mathbb{N}$ *be such that* $d \le n$, *let* $D = \max\{d, N\}$, *let* $\delta \in (0, 1/4)$, $r \ge 1$, *and let* $\sigma^2 > 0, \tau^2 > 1$ *be the variance of label noise and random feature biases, respectively. Under Assumption 5.2, there exists a universal constant* $C_0 > 0$ *such that the following holds with probability at least* $1 - \delta$ *over the samples and random features.*

$$\sup_{\theta\in\mathbb{S}^{d-1}, \|c\|\le r} \|\nabla_\theta L_n(c,\theta) - \nabla_\theta L(c,\theta)\| \le C_{f_*} C r^2 \cdot \max\left(\sqrt{\frac{D\log(nN/\delta)}{n}}, \frac{(d\log(nN/\delta))^2}{n}\right)$$

$$\sup_{\theta\in\mathbb{S}^{d-1}, \|c\|\le r} \|\nabla_c L_n(c,\theta) - \nabla_c L(c,\theta)\| \le C^2 r \cdot \sqrt{\frac{D\log(n/\delta)}{n}} ,$$

where $C = C_0 \cdot \max\{\mathrm{Lip}(f_*), \tau\sqrt{\log(1/\delta)}, \sigma\}$, *and* $C_{f_*} = \max\{\|f_*^{(1)}\|_\gamma, \ldots, \|f_*^{(4)}\|_\gamma, 1\}$.

Equipped with this uniform gradient concentration, we can first establish the analogous classification of first-order critical points for the empirical landscape (proof in Section E.3):

Lemma E.2 (Local sharpness of the empirical landscape)**.** *Let* $d, n \in \mathbb{N}$ *be such that* $d \le n$, *let* $\delta \in (0, 1/4)$, *let* $\tilde{s} \in \mathbb{N}$ *be such that* $\tilde{s} \ge s$, *where* $s \ge 1$ *is the information exponent of* f_*, *let* $\lambda \in (0, \lambda_{\tilde{s}}^*)$ *where* $\lambda_{\tilde{s}}^* \le 1$ *depends only on* $\tilde{s}$, f_*, *and* τ, *and let* $N \in \mathbb{N}$ *be such that* $N \ge \frac{C_0}{\lambda}\log\frac{1}{\lambda\delta}$, *where* $C_0 > 0$ *is a universal constant. Furthermore, let* $D = \max\{d, N\}$ *and let* $\epsilon \in (0,1)$ *be such that* $\epsilon \le \lambda^{-2}\sqrt{d/n}$. *Then, there exists a universal constant* $C_1 > 0$ *such that for* $C = C_1 \cdot \max\{\mathrm{Lip}(f_*), \tau\sqrt{\log(1/\delta)}, \sigma\}$ *and* $\Delta = \max\left\{\sqrt{\frac{D\log(n/\delta)}{n}}, \frac{(d\log(n/\delta))^2}{n}\right\}$ *the following holds with probability at least* $1 - \delta$ *over the samples and random features.*

$$\Omega_n(\epsilon,\epsilon) = \Omega_n^{\mathrm{bad}} \sqcup \Omega_n^{\mathrm{good}} \text{ , where} \tag{37}$$

$$\Omega_n^{\mathrm{bad}} \subset \left\{(c,\theta) \,\middle|\, |m| \le \left(\frac{C_2 C^7}{\lambda^2}\cdot\Delta\right)^{\frac{1}{2\tilde{s}-1}} \ \wedge\ \|c\| \le \frac{C^3}{\lambda}\right\} ,$$

$$\Omega_n^{\mathrm{good}} \subset \left\{(c,\theta) \,\middle|\, 1 - |m| \le \frac{C_3 C^{14}}{\lambda^4}\cdot\Delta^2 \ \wedge\ \|c\| \le \frac{C^3}{\lambda}\right\} ,$$

and $C_2 = C_{f_*}/\tilde{s}\alpha_{\tilde{s}}^2$ *and* $C_3 = (2^{2\tilde{s}-1}C_2)^2$ *in the above display. Moreover,*

$$\min_{(c,\theta)\in\Omega_n^{\mathrm{bad}}} L_n(c,\theta) \geq \sigma^2 + \|f_*\|_\gamma^2 - 2\max\{\|f_*\|_\gamma^2 C_2^2 C^{14}, C^8\}\cdot \Delta_{\mathrm{crit}}/\lambda^2 \ , \tag{38}$$

where

$$\Delta_{\mathrm{crit}} = \max\left\{\sqrt{\frac{D\log(n/\delta)}{n}}, \left(\frac{d^2\log^2(n/\delta)}{n}\right)^{\frac{2\tilde{s}}{2\tilde{s}-1}}\right\} . \tag{39}$$

We consider the gradient flow procedure of Algorithm 1, that we restate here for convenience:

Procedure 3 Gradient Flow with time-scale scheduling (Restated)

Require: N_0, ρ, T_0, T_1
Initialise $\theta(0) \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, $c(0) \sim \mathrm{Unif}(\{c \in \mathbb{R}^N; \|c\|_2 = \rho; \|c\|_0 = N_0\})$.
Run Gradient Flow (10) with $\zeta(t) = \mathbf{1}(t > T_0)$ up to time $T = T_0 + T_1$.

We establish the following fact, proved in Appendix E.4:

Lemma E.3 (Gradient flow escapes the equator)**.** *Assume* $\sum_j (j+A)^k \alpha_j^2 \leq C$ *for* $A \leq s$ *and* $k \leq 3$. *With probability at least* $1/2 - 2\delta$ *over the initial condition, the draw of the data, and the draw of the random features, if* $n = \tilde{\Omega}(\max\{\lambda^{-4}(d+N)d^{s-1}, \lambda^{-2}d^{(s+3)/2}\})$ *and* $N = \Theta\left(\lambda^{-1}\log(\lambda^{-1}\delta^{-1})\right)$ *then the first phase of gradient flow with a randomly initialised* $c(0) \sim \mathrm{Unif}\{c \in \rho\mathbb{S}^{N-1}; \|c\|_0 = N_0\}$ *with* $\rho = \Theta(\sqrt{N}N_0^{-(2+s)/2}(\tau^2 + \lambda N/N_0)^{-1})$ *and* $N_0 = \Theta\left(\log\frac{1}{\delta}\right)$ *escapes the equator in time* $T_0 = \widetilde{O}\left(d^{s/2-1}\right)$.

Therefore, for any $\epsilon > 0$, Gradient Flow converges to an ϵ-approximate first-order critical point *at energy lower than* $B_{\mathrm{crit}} = \tilde{\Theta}\left(\lambda^{-2}\Delta_{crit}\right)$ in time $\widetilde{O}(\epsilon^{-2}) = \widetilde{O}(\frac{\lambda^4 n}{D})$, since gradient flow is a descent curve under our settings (see Appendix F), and therefore satisfies

$$L_n(c(T),\theta(T)) - L_n(c(0)) = -\int_0^T \|\nabla L_n(c(t),\theta(t))\|^2 dt \ . \tag{40}$$

By Lemma E.2, such critical points can only be in Ω_n^{good}, which yields the result. □

E.2 Proof of Lemma E.1

Lemma E.1 (Uniform convergence of the empirical landscape)**.** *Let* $d, n, N \in \mathbb{N}$ *be such that* $d \leq n$, *let* $D = \max\{d, N\}$, *let* $\delta \in (0, 1/4)$, $r \geq 1$, *and let* $\sigma^2 > 0, \tau^2 > 1$ *be the variance of label noise and random feature biases, respectively. Under Assumption 5.2, there exists a universal constant* $C_0 > 0$ *such that the following holds with probability at least* $1-\delta$ *over the samples and random features.*

$$\sup_{\theta\in\mathbb{S}^{d-1}, \|c\|\leq r} \|\nabla_\theta L_n(c,\theta) - \nabla_\theta L(c,\theta)\| \leq C_{f_*} C r^2 \cdot \max\left(\sqrt{\frac{D\log(nN/\delta)}{n}}, \frac{(d\log(nN/\delta))^2}{n}\right)$$

$$\sup_{\theta\in\mathbb{S}^{d-1}, \|c\|\leq r} \|\nabla_c L_n(c,\theta) - \nabla_c L(c,\theta)\| \leq C^2 r \cdot \sqrt{\frac{D\log(n/\delta)}{n}} \ ,$$

where $C = C_0 \cdot \max\{\mathrm{Lip}(f_*), \tau\sqrt{\log(1/\delta)}, \sigma\}$, *and* $C_{f_*} = \max\{\|f_*^{(1)}\|_\gamma, \ldots, \|f_*^{(4)}\|_\gamma, 1\}$.

Proof. Recall the empirical loss of equation (1) for $c \in \mathbb{R}^N$ and $\theta \in \mathbb{S}^{d-1}$:

$$L_n(c,\theta) = \frac{1}{n}\sum_{i=1}^n \ell(c,\theta;x_i,y_i) + \lambda\|c\|^2 \ ,$$

where

$$\ell(c,\theta;x,y) = \left(c^\top\Phi(\langle x,\theta\rangle) - y\right)^2 \ .$$

Uniform convergence of $\nabla_\theta L_n$. Our proof tracks closely the argument in [33, Section G.1] and [60, Theorem 1], but takes additional steps to handle the non-Lipschitzness of the sample gradient around 0. For simplicity, we write $C = \max\{\mathrm{Lip}(f_*), \tau\sqrt{\log(1/\delta)}, \sigma\}$, i.e., omit the universal constant $C_0 > 0$, with the understanding that C_0 is implicitly determined by accounting for all occurrences of $\lesssim$ in our analysis. We also repeatedly use the following elementary fact: If $C_1, C_2 \geq 1$, then $C_1 + C_2 \leq 2\max\{C_1, C_2\} \leq 2C_1C_2$. We first recall basic concentration properties of standard Gaussian random variables.

Fact E.4 (Niceness of Gaussian random variables)**.** *Let $\delta \in (0, 1/4)$, $N \in \mathbb{N}$, and let $b_1, \ldots, b_N$ be i.i.d. random variables drawn from $\mathcal{N}(0, \tau^2)$. Then, there exists a universal constant $C' > 0$ such that the following two events hold simultaneously with probability at least $1 - \delta$.*

$$\max_j |b_j| \leq C'\tau\sqrt{\log(N/\delta)} \ ,$$
$$\sum_j b_j^2 \leq N\tau^2 + C'\tau^2 \max\left\{\log(1/\delta), \sqrt{N\log(1/\delta)}\right\} \ .$$

Corollary E.5 (ℓ_2-norm of random features)**.** *Let $\delta \in (0, 1/4)$ and let $b_1, \ldots, b_N$ be i.i.d. random variables drawn from $\mathcal{N}(0, \tau^2)$. Then, there exists a universal constant $C' > 0$ such that the following holds for all $z \in \mathbb{R}$ with probability at least $1 - \delta$ over the random features,*

$$\|\Phi(z)\| \leq |z| + C'\tau(1 + \sqrt{\log(1/\delta)/N}) \leq |z| + 2C'\tau\sqrt{\log(1/\delta)} \ .$$

Proof. Let $C_1 = \max\{C, 1\}$, where $C > 0$ is the constant from Fact E.4. Then, with probability at least $1 - \delta$ over the random features,

$$\begin{aligned}
\|\Phi(z)\| &= \sqrt{\frac{1}{N}\sum_{j=1}^N \phi(\zeta_j z - b_j)^2} \\
&\leq \sqrt{\frac{1}{N}\sum_{j=1}^N (\zeta_j z - b_j)^2} \\
&= \sqrt{z^2 - \frac{2z}{N}\sum_{j=1}^N \zeta_j b_j + \frac{1}{N}\sum_{j=1}^N b_j^2} \\
&\leq \sqrt{z^2 + 2|z|C_2\tau\sqrt{\log(1/\delta)/N} + \tau^2(1 + C_1\max\{\log(1/\delta)/N, \sqrt{\log(1/\delta)/N}\})} \\
&\leq |z| + \max\{C_2^2, 2C_1\} \cdot \tau(1 + \sqrt{\log(1/\delta)/N})) \ .
\end{aligned}$$

□

Fix $c \in \mathbb{R}^N$ such that $\|c\| \leq r$. For $x \sim \mathcal{N}(0, I_d)$ consider the random vector corresponding to the samplewise gradient with respect to $\theta \in \mathbb{S}^{d-1}$.

$$\nabla_\theta \ell\left(c, \theta; x, f_*(\langle x, \theta_*\rangle)\right) = \left(c^\top \Phi'(\langle \theta, x\rangle)\right)\left(c^\top \Phi(\langle \theta, x\rangle) - f_*(\langle x, \theta_*\rangle) - \xi\right) x \ .$$

The tail of this random vector is subexponential, as stated in the following lemma.

Lemma E.6 (Sub-exponential gradients)**.** *Let $f_* : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function, let $\delta \in (0, 1/4)$, let $r \geq 1$, and let $\tau^2 > 1$ be the variance of random feature biases. Then, there exists a universal constant $C' > 0$ such that the following holds with probability at least $1 - \delta$ over the random features.*

$$\|\nabla_\theta \ell\left(c, \theta; x, f_*(\langle x, \theta^*\rangle) + \xi\right)\|_{\psi_1} \leq C'\|c\|(\mathrm{Lip}(f_*) + \sigma + \|c\|\tau(1 + \sqrt{\log(1/\delta)/N})) \leq 2C'Cr^2 \ , \tag{41}$$

where $C = \max\{\mathrm{Lip}(f_*), \tau\sqrt{\log(1/\delta)}, \sigma\}$.

Proof of Lemma E.6. Define $W = c^\top \Phi(\langle \theta, x\rangle) - f_*(\langle \theta^*, x\rangle) - \xi$. Using Fact B.3,

$$\begin{aligned}\|\nabla_\theta \ell\left(c, \theta; x, f_*(\langle x, \theta^*\rangle)\right)\|_{\psi_1} &\le \sup_{v\in\mathbb{S}^{d-1}} \left\|(c^\top \Phi'(\langle \theta, x\rangle))W\langle v, x\rangle\right\|_{\psi_1} \\ &\le \left\|(c^\top \Phi'(\langle \theta, x\rangle))W\right\|_{\psi_2} \\ &\le \|c\|\,\|W\|_{\psi_2}\ ,\end{aligned}$$

where the last inequality follows from the Cauchy-Schwartz inequality $\|c\|_1 \le \sqrt{N}\|c\|_2$ and the fact that $\Phi'(y) \in \{0, 1/\sqrt{N}\}^N$. Now denote the (correlated) Gaussian variables by $Z = \langle \theta, x\rangle$ and $Z_* = \langle \theta^*, x\rangle$. Recalling Fact B.3 and our assumption that $\mathbb{E}[f_*(Z)] = 0$,

$$\begin{aligned}\|W\|_{\psi_2} &\le \left\|c^\top \Phi(Z)\right\|_{\psi_2} + \|f_*(Z_*)\|_{\psi_2} + \|\xi\|_{\psi_2} \\ &\le \left\|c^\top \Phi(Z) - \mathbb{E}[c^\top \Phi(Z)]\right\|_{\psi_2} + \left\|\mathbb{E}[c^\top \Phi(Z)]\right\|_{\psi_2} + \|f_*(Z_*)\|_{\psi_2} + \|\xi\|_{\psi_2} \\ &\lesssim \|c\| + |\,\mathbb{E}[c^\top \Phi(Z)]| + \mathrm{Lip}(f_*) + \sigma \\ &\le \|c\|(1 + \|\,\mathbb{E}[\Phi(Z)]\|) + \mathrm{Lip}(f_*) + \sigma\ .\end{aligned}$$

By Corollary E.5, the following holds with probability at least $1-\delta$ over the random features.

$$\|\mathbb{E}[\Phi(Z)]\| \le \mathbb{E}[\|\Phi(Z)\|] \lesssim \tau(1 + \sqrt{\log(1/\delta)/N})\ .$$

□

We now prove the uniform convergence of $\nabla_\theta L_n(c, \theta)$. Let $\epsilon_\theta, \epsilon_c < 1$ be small positive integers. To make things concrete, we set $\epsilon_\theta = 1/(4n^2 N)$ and $\epsilon_c = r/n^2$. We consider ϵ-nets of $\theta \in \mathbb{S}^{d-1}$ and $c \in \mathbb{R}^N$ with $\|c\| \le r$. We denote these sets by $\mathcal{N}_\theta$ and $\mathcal{N}_c$ respectively. By Fact B.6, there exist such sets with $|\mathcal{N}_\theta| \le (3/\epsilon_\theta)^d$ and $|\mathcal{N}_c| \le (3r/\epsilon_c)^N$. We use these ϵ-nets to decompose gradient error into three terms, which we bound individually. We abuse notation and denote by $\tilde{\theta}$ the element in $\mathcal{N}_\theta$ closest to θ in the ℓ_2 norm and $\tilde{c}$ the closest element in $\mathcal{N}_c$ to c. Then, the gradient deviation term can be decomposed into

$$\begin{aligned}\sup_{\theta\in\mathbb{S}^{d-1}, \|c\|\le r} \|\nabla_\theta L_n(c, \theta) - \nabla_\theta L(c, \theta)\| \le &\sup_{\theta\in\mathbb{S}^{d-1}, \|c\|\le r} \left\|\nabla_\theta L_n(c, \theta) - \nabla_\theta L_n(\tilde{c}, \tilde{\theta})\right\| \\ &+ \sup_{\tilde{\theta}\in\mathcal{N}_\theta, \tilde{c}\in\mathcal{N}_c} \left\|\nabla_\theta L_n(\tilde{c}, \tilde{\theta}) - \nabla_\theta L(\tilde{c}, \tilde{\theta})\right\| \\ &+ \sup_{\theta\in\mathbb{S}^{d-1}, \|c\|\le r} \left\|\nabla_\theta L(c, \theta) - \nabla_\theta L(\tilde{c}, \tilde{\theta})\right\|.\end{aligned}$$

As mentioned previously, bounding the first term (with high probability) requires some care due to the non-differentiability of ReLU at the origin. The other two terms can be bounded using standard techniques. We first look at the sample gradient $\nabla_\theta \ell(c, \theta; x_i, y_i) := \nabla_\theta (y_i - c^\top \Phi(\langle x_i, \theta\rangle))^2$ to bound the "worst-case" discretization error incurred by discontinuities in the sample gradient. Then, we show that with high probability over the samples, for any $\theta \in \mathbb{S}^{d-1}$ and c, this "worst-case" discretization error occurs only for a few samples x_i. Thus, the contribution from the worst-case discretization error get averaged out in $\nabla_\theta L_n(c, \theta) - \nabla_\theta L_n(\tilde{c}, \tilde{\theta}) = \frac{1}{n}\sum_{i=1}^n (\nabla_\theta \ell(c, \theta; x_i, y_i) - \nabla_\theta \ell(\tilde{c}, \tilde{\theta}; x_i, y_i))$.

For simplicity, we set $\mathrm{ReLU}'(0) = 0$.[3] Note that $\Phi'(z) \in \{0, 1/\sqrt{N}\}^N$, so $\|\Phi'(z)\|_2 \le 1$ for any $z \in \mathbb{R}$. By Corollary E.5, the sample gradient and an upper bound for its ℓ_2 norm is given by

$$\nabla_\theta \ell_i(c, \theta; x_i, y_i) = -(c^\top \Phi'(\langle x_i, \theta\rangle))(y_i - c^\top \Phi(\langle x_i, \theta\rangle))x_i\ .$$

$$\|\nabla_\theta \ell_i(c, \theta; x_i, y_i)\| \le \|c\| \left(\mathrm{Lip}(f_*)\|x_i\| + \|c\|(\|x_i\| + C'\tau\sqrt{\log(1/\delta)})\right)\|x_i\| \lesssim Cr^2(\|x_i\|^2 + \|x_i\|)\ .$$

When discretizing $\mathbb{S}^{d-1}$, the "best-case" is when $\Phi'(\langle x_i, \theta\rangle) = \Phi'(\langle x_i, \tilde{\theta}\rangle)$ in which case the discretization error is 1-Lipschitz in $\theta - \tilde{\theta}$. On the other hand, the discretization error is not Lipschitz

[3] The particular value of $\mathrm{ReLU}'(0)$ does not matter for our proof it is $[0, 1]$. In practice, however, the particular value assigned to $\mathrm{ReLU}'(0)$ may have non-trivial implications [14].

when $\Phi'(\langle x_i, \theta\rangle) \neq \Phi'(\langle x_i, \tilde{\theta}\rangle)$, i.e., when the ReLU' sign patterns change after projecting θ onto $\mathcal{N}_\theta$. The "worst-case" discretization error corresponding to this case is upper bounded by

$$\|\nabla_\theta \ell_i(c, \theta; x_i, y_i) - \nabla_\theta \ell_i(c, \tilde{\theta}; x_i, y_i)\| \leq \|\nabla_\theta \ell_i(c, \theta; x_i, y_i)\| + \|\nabla_\theta \ell_i(c, \tilde{\theta}; x_i, y_i)\| \lesssim Cr^2(\|x_i\|^2 + \|x_i\|) \ .$$

We now show that for any fixed $\tilde{\theta} \in \mathbb{S}^{d-1}$, bottom-layer weights $b_1, \ldots, b_N$ and signs $\varepsilon_1, \ldots, \varepsilon_N$, with high probability only a small fraction of the sample gradients $\nabla_\theta \ell_i$ change ReLU' sign patterns $\Phi'(\langle x_i, \tilde{\theta}\rangle)$ within a ball of radius ϵ_θ from $\tilde{\theta}$. As a consequence, the worst-case discretization error, which comes from the non-differentiability of ReLU, does not accumulate too much for the averaged gradient $\nabla_\theta L_n(c, \theta) - \nabla_\theta L_n(c, \tilde{\theta})$. To this end, fix an arbitrary $\theta \in \mathbb{S}^{d-1}$ and let $Z_{\theta,i}$ be the indicator that projecting θ on to the ϵ-net $\mathcal{N}_\theta$ changes the sign pattern for x_i. Formally,

$$Z_{\theta,i} = \mathbb{1}\left[\exists \theta, j \in [N] \text{ s.t. } \|\tilde{\theta} - \theta\| \leq \epsilon_\theta,\ \operatorname{sign}(\varepsilon_j \langle \tilde{\theta}, x_i\rangle - b_j) \neq \operatorname{sign}(\varepsilon_j \langle \theta, x_i\rangle - b_j)\right] \ .$$

By the union bound (over $j \in [N]$) and basic properties of Gaussian random variables,

$$\begin{aligned}
\mathbb{E}[Z_{\theta,i}] &\leq N\, \mathbb{P}[\exists \theta \text{ s.t. } \|\tilde{\theta} - \theta\| \leq \epsilon_\theta,\ \operatorname{sign}(\varepsilon_1 \langle \tilde{\theta}, x_i\rangle - b_1) \neq \operatorname{sign}(\varepsilon_1 \langle \theta, x_i\rangle - b_1)] \\
&\leq N\, \mathbb{P}[|\langle \tilde{\theta}, x_i\rangle - b_1| \leq \|x_i\|\, \epsilon_\theta] \\
&= N\, \mathbb{P}[|\langle \tilde{\theta}, x_i / \|x_i\|\rangle - b_1 / \|x_i\|\,| \leq \epsilon_\theta] \\
&\leq N\, \mathop{\mathbb{E}}_t[\mathbb{P}[|\langle \tilde{\theta}, u\rangle - b_1/t| \leq \epsilon_\theta | t]] \quad \text{(denoting } t = \|x_i\| \text{ and } u = x_i / \|x_i\|) \\
&\leq N\, \mathbb{P}[|\langle \tilde{\theta}, u\rangle| \leq \epsilon_\theta] \ ,
\end{aligned}$$

where u is a random vector drawn uniformly from $\mathbb{S}^{d-1}$, and we used the fact that u and r are independent, along with the fact that the density of $\langle \tilde{\theta}, u\rangle$ is peaked around 0. The last line can be upper bounded by the surface area of an ϵ-thick strip around the equator in $\mathbb{S}^{d-1}$. This is given by the following anticoncentration bound from Section B.

Lemma B.7 (Anticoncentration on the unit sphere)**.** *Let $d \in \mathbb{N}$. Let $\theta \in \mathbb{S}^{d-1}$ be any fixed unit vector and let u be a random vector drawn uniformly from $\mathbb{S}^{d-1}$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}[\langle \theta, u\rangle \leq \epsilon] \leq 4\sqrt{d}\epsilon \ .$$

By Lemma B.7, if $\epsilon_\theta = 1/(4n^2 N)$, then $\mathbb{E}[Z_{\theta,i}] \leq 1/n$ and thus $\sum_{i=1}^n \mathbb{E}[Z_{\theta,i}] \leq 1$. By the independence of the samples x_i and a Chernoff bound, it holds that the probability that at least $q \geq 6$ inputs x_i satisfy the above event is upper bounded by 2^{-q} [63, Theorem 4.4]. As a result, with probability at least $1 - (3/\epsilon_\theta)^d \cdot 2^{-q}$, no more than q inputs can change sign pattern for each fixed $\tilde{\theta}$. It suffices to set $q = 4d\log(1/(\epsilon_\theta \delta)) \lesssim d\log(nN/\delta)$.

We assume that this event holds and establish the first inequality with the help of the ϵ-nets. We assume without loss of generality that for fixed $\tilde{\theta}$, the inputs that may change ReLU signs for any θ in an ϵ_θ-ball of $\tilde{\theta}$ are contained in $x_1, \ldots, x_q$. Recall that for any given θ and c, we denote by $\tilde{\theta}$ and $\tilde{c}$ the closest elements in the ϵ-nets.

$$\begin{aligned}
&\left\|\nabla_\theta L_n(c, \theta) - \nabla_\theta L_n(\tilde{c}, \tilde{\theta})\right\| \\
&\quad \leq \frac{1}{n}\sum_{i=1}^n \left\|\nabla_\theta \ell(c, \theta; x_i, y_i) - \nabla_\theta \ell(\tilde{c}, \tilde{\theta}; x_i, y_i)\right\| \\
&\quad \leq \frac{1}{n}\sum_{i=1}^n \left(\|\nabla_\theta \ell(c, \theta; x_i, y_i) - \nabla_\theta \ell(\tilde{c}, \theta; x_i, y_i)\| + \left\|\nabla_\theta \ell(\tilde{c}, \theta; x_i, y_i) - \nabla_\theta \ell(\tilde{c}, \tilde{\theta}; x_i, y_i)\right\|\right)
\end{aligned}$$

We first bound the terms involving differences between c and $\tilde{c}$. Note that by Fact E.4, $\max_i |\xi_i| \lesssim \sigma\sqrt{\log(n/\delta)}$ and $\max_i \|x_i\| \lesssim \sqrt{d\log(n/\delta)}$ with probability at least $1 - \delta/6$. Thus, we consider

events for which these conditions hold.

$$
\begin{aligned}
&\|\nabla_\theta \ell(c,\theta;x,y) - \nabla_\theta \ell(\tilde{c},\theta;x,y)\| \\
&\quad = \left\|(c^\top \Phi'(\langle\theta,x\rangle)(c^\top\Phi(\langle\theta,x\rangle)-y)x - (\tilde{c}^\top \Phi'(\langle\theta,x\rangle)(\tilde{c}^\top\Phi(\langle\theta,x\rangle)-y)x\right\| \\
&\quad \le \left\|(c^\top \Phi'(\langle\theta,x\rangle)((c-\tilde{c})^\top\Phi(\langle\theta,x\rangle)x\right\| + \left\|(c-\tilde{c})^\top \Phi'(\langle\theta,x\rangle)(\tilde{c}^\top\Phi(\langle\theta,x\rangle) - f_*(\langle x,\theta_*\rangle - \xi))x\right\| \\
&\quad \le \|c\|\epsilon_c\|\Phi(\langle\theta,x\rangle)\|\|x\| + \epsilon_c\left(\|c\|\|\Phi(\langle\theta,x\rangle)\| + \mathrm{Lip}(f_*)\|x\| + |\xi|\right)\|x\| \\
&\quad \lesssim \epsilon_c\|x\|\left(\|c\|(\|x\|+C) + C\|x\| + |\xi|\right) \\
&\quad \lesssim Cr\epsilon_c\left(\|x\|^2 + \|x\|(1+|\xi|/\sigma)\right) \\
&\quad \lesssim Cr\epsilon_c\left(\|x\|^2 + \|x\|\sqrt{\log(n/\delta)}\right) \ . \qquad (42)
\end{aligned}
$$

Next, we consider the first q terms of the differences between θ and $\tilde{\theta}$. For some $i \in [\ell]$

$$
\begin{aligned}
&\left\|\nabla_\theta \ell(\tilde{c},\theta;x_i,y_i) - \nabla_\theta \ell(\tilde{c},\tilde{\theta};x_i,y_i)\right\| \\
&\quad \le \left\|(\tilde{c}^\top \Phi'(\langle\theta,x_i\rangle)(\tilde{c}^\top\Phi(\langle\theta,x_i\rangle)-y_i)x_i - (\tilde{c}^\top \Phi'(\langle\tilde{\theta},x_i\rangle)(\tilde{c}^\top\Phi(\langle\tilde{\theta},x_i\rangle)-y_i)x_i\right\| \\
&\quad \le 2\sup_{\tilde{c},\theta}\left\|(\tilde{c}^\top \Phi'(\langle\theta,x_i\rangle)(\tilde{c}^\top\Phi(\langle\theta,x_i\rangle)-y_i)x_i\right\| \\
&\quad \le 2\sup_{\tilde{c},\theta}\|\tilde{c}\|\left(\|\tilde{c}\|\,\|\Phi(\langle\theta,x_i\rangle)\| + \mathrm{Lip}(f_*)\|x_i\| + |\xi_i|\right)\|x_i\| \\
&\quad \lesssim \sup_{\tilde{c}}\left(\left\|\tilde{c}^2\right\|\left(\|x_i\| + \tau(1+\sqrt{\log(4/\delta)/N})\right) + \|\tilde{c}\|\left(\mathrm{Lip}(f_*)\|x_i\| + |\xi_i|\right)\right)\|x_i\| \\
&\quad \lesssim r(r+\mathrm{Lip}(f_*))\|x_i\|^2 + \left(r^2\tau(1+\sqrt{\log(4/\delta)/N}) + r|\xi_i|\right)\|x_i\| \\
&\quad \lesssim Cr^2(\|x_i\|^2 + \|x_i\|(1+|\xi_i|/\sigma)) \\
&\quad \lesssim Cr^2(\|x_i\|^2 + \|x_i\|\sqrt{\log(n/\delta)}) \ . \qquad (43)
\end{aligned}
$$

Finally, we bound the remaining $n-q$ terms.

$$
\begin{aligned}
\left\|\nabla_\theta \ell(\tilde{c},\theta;x_i,y_i) - \nabla_\theta \ell(\tilde{c},\tilde{\theta};x_i,y_i)\right\| &= \left\|\tilde{c}^\top \Phi'(\langle\theta,x_i\rangle)\left(\tilde{c}^\top(\Phi(\langle\theta,x_i\rangle) - \Phi(\langle\tilde{\theta},x_i\rangle))\right)x_i\right\| \\
&\le r^2\epsilon_\theta\|x_i\|^2 \ .
\end{aligned}
$$

As a result, with probability at least $1-\delta/3$,

$$
\begin{aligned}
\left\|\nabla_\theta L_n(c,\theta) - \nabla_\theta L_n(\tilde{c},\tilde{\theta})\right\| &\lesssim Cr^2\|x_i\|\left(\left(\frac{\epsilon_c}{r} + \frac{q}{n}\right)(\|x_i\| + \sqrt{\log(n/\delta)}) + \epsilon_\theta\|x_i\|\right) \\
&\lesssim Cr^2 \cdot \frac{(d\log(n/\delta))^2}{n} \ .
\end{aligned}
$$

where we set $q = 4d\log(1/(\epsilon_\theta\delta))$ and recall that $\epsilon_c = r/n^2$, $\epsilon_\theta = 1/(4n^2N)$.

We now bound the second term $\left\|\nabla_\theta L_n(\tilde{c},\tilde{\theta}) - \nabla_\theta L(\tilde{c},\tilde{\theta})\right\|$ using a standard concentration argument over the $(1/2)$-net on $\mathbb{S}^{d-1}$, which we denote by $\mathcal{N}_{1/2}$. By Fact B.6,

$$
\sup_{\tilde{\theta}\in\mathcal{N}_\theta,\tilde{c}\in\mathcal{N}_c}\left\|\nabla_\theta L_n(\tilde{c},\tilde{\theta}) - \nabla_\theta L(\tilde{c},\tilde{\theta})\right\| \le 2\sup_{\substack{v\in\mathcal{N}_{1/2},\\ \tilde{\theta}\in\mathcal{N}_\theta,\tilde{c}\in\mathcal{N}_c}}\langle v, \nabla_\theta L_n(\tilde{c},\tilde{\theta}) - \nabla_\theta L(\tilde{c},\tilde{\theta})\rangle. \qquad (44)
$$

By Lemma E.6, the ψ_1-norm of a sample gradient is bounded by $\|\nabla_\theta \ell(\tilde{c},\tilde{\theta};x_i,y_i)\|_{\psi_1} \lesssim Cr^2$, where we recall that $C = \max\{\mathrm{Lip}(f_*), \tau\sqrt{\log(1/\delta)}, \sigma, 1\}$. Since $\nabla_\theta L_n(\tilde{c},\tilde{\theta}) = \frac{1}{n}\sum_i \nabla_\theta \ell(\tilde{c},\tilde{\theta};x_i,y_i)$ and $\nabla_\theta L(\tilde{c},\tilde{\theta}) = \mathbb{E}[\nabla_\theta \ell(\tilde{c},\tilde{\theta};x,y)]$, and $v \in \mathbb{S}^{d-1}$, the right-hand side of equation (B.6) is the supremum of averaged sub-exponential random variables. Hence, the term can be bounded with by combining a union bound and Bernstein's inequality (Theorem B.4). Recall that

$D = \max\{d, N\}$. For sufficiently large universal constants $C_0, C_1 > 0$, the following holds.

$$\begin{aligned}
&\mathbb{P}\left[2 \sup_{\tilde{\theta}, \tilde{c}, v} \langle v, \nabla_\theta L_n(\tilde{c}, \tilde{\theta}) - \nabla_\theta L(\tilde{c}, \tilde{\theta}) \rangle \geq C_0 \cdot Cr^2 \cdot \sqrt{\frac{D \log(nN/\delta)}{n}}\right] \\
&\leq 6^d \left(\frac{3}{\epsilon_\theta}\right)^d \left(\frac{3r}{\epsilon_c}\right)^N \exp\left(-C_1 \cdot n \cdot \frac{D \log(nN/\delta)}{n}\right) \\
&\leq \exp\left(d \log\left(\frac{18}{\epsilon_\theta}\right) + N \log\left(\frac{3r}{\epsilon_c}\right) - C_1 \cdot D \log(nN/\delta)\right) \\
&\leq \frac{\delta}{3} ,
\end{aligned}$$

where we used the assumption $d \leq n$ in the last inequality.

We now bound third term. Note that this term involves only populational quantities, so discontinuity of the sample gradients is not an issue here.

$$\begin{aligned}
\left\|\nabla_\theta L(c, \theta) - \nabla_\theta L(\tilde{c}, \tilde{\theta})\right\| &\leq \left\|\nabla_\theta L(c, \theta) - \nabla_\theta L(c, \tilde{\theta})\right\| + \left\|\nabla_\theta L(c, \tilde{\theta}) - \nabla_\theta L(\tilde{c}, \tilde{\theta})\right\| \\
&= \left\|\nabla_\theta L(c, \theta) - \nabla_\theta L(c, \tilde{\theta})\right\| + \left\|\mathbb{E}[\nabla_\theta \ell(c, \tilde{\theta}; x, y) - \nabla_\theta \ell(\tilde{c}, \tilde{\theta}; x, y)]\right\| .
\end{aligned}$$

We upper bound the two terms individually. By Eq. (42), the second term is bounded as follows.

$$\left\|\mathbb{E}[\nabla_\theta \ell(c, \tilde{\theta}; x, y) - \nabla_\theta \ell(\tilde{c}, \tilde{\theta}; x, y)]\right\| \lesssim Cr\epsilon_c \cdot d\sqrt{\log(n/\delta)} \lesssim Cr^2 \sqrt{\log(n/\delta)}/n .$$

For the first term, we use our regularity assumption on the target f_* (Assumption 5.2), specifically the existence of continuous higher-order derivatives. Let $f_*(z) = \sum_{j=0}^\infty \alpha_j h_j(z)$ and $c^\top \Phi(z) = \sum_{j=0}^\infty \beta_j(c) h_j(z)$, be the Hermite expansion of the target and ReLU network, respectively. Define the univariate function $g : [-1, 1] \to \mathbb{R}$ by

$$g(m; c) = \sum_{j=0}^\infty \alpha_j \beta_j(c) m^j .$$

Straightforward algebraic manipulation using orthonormality of the Hermite basis gives the following expression of the gradient with respect to $\theta \in \mathbb{S}^{d-1}$.

$$\nabla_\theta L(\theta) = g'(\langle \theta^*, \theta \rangle; c)\theta^* .$$

The following observation gives an upper bound on the Lipschitz constant of $g'(m; c)$, which depends only on $\|c\|$ and is thus constant if r is fixed.

$$\begin{aligned}
|g''(m; c)| \leq \sum_{j=2}^\infty j(j-1)|\alpha_j||\beta_j(c)| &\leq \left(\sum_{j=1}^\infty j^2(j-1)^2 \alpha_j^2\right)^{1/2} \left(\sum_{j=0}^\infty \beta_j(c)^2\right)^{1/2} \\
&\leq \left(\sum_{j=1}^\infty j^2(j-1)^2 \alpha_j^2\right)^{1/2} \|c^\top \Phi(z)\|_\gamma \\
&\leq \left(\sum_{j=1}^\infty j^2(j-1)^2 \alpha_j^2\right)^{1/2} Cr .
\end{aligned}$$

Lemma E.7 gives a simple expression for $\sum_j j^2(j-1)^2 \alpha_j^2$. Its proof can be found in Section E.8.

Lemma E.7. *Let $f : \mathbb{R} \to \mathbb{R}$ be function such that its derivatives $f^{(1)}, \ldots, f^{(4)}$ are all in $L^2(\gamma)$. Let $f(z) = \sum_{j=0}^\infty \alpha_j h_j(z)$ be the Hermite expansion of f. Then,*

$$\sum_{j=1}^\infty j^2(j-1)^2 \alpha_j^2 = \|f^{(4)}\|_\gamma^2 + 4\|f^{(3)}\|_\gamma^2 + 2\|f^{(2)}\|_\gamma^2 .$$

Let $C_{f_*} = \max\{\|f^{(1)}\|_\gamma, \dots, \|f^{(4)}\|_\gamma, 1\}$, which is well-defined thanks to our regularity assumption on the target link function (Assumption 5.2). Then,

$$\left\|\nabla_\theta L(c,\theta) - \nabla_\theta L(c,\tilde{\theta})\right\| = |g'(\langle\theta^*,\theta\rangle) - g'(\langle\theta^*,\tilde{\theta}\rangle)| \le \sup_{m\in[-1,1]} |g''(m)|\|\theta-\tilde{\theta}\| \lesssim C_{f_*} C r \epsilon_\theta \ .$$

Putting everything together, we have that with probability at least $1-\delta$,

$$\sup_{\theta,\|c\|\le r} \|\nabla_\theta L_n(c,\theta) - \nabla_\theta L(c,\theta)\| \le C_{f_*} C r^2 \cdot \max\left(\sqrt{\frac{D\log(nN/\delta)}{n}}, \frac{(d\log(nN/\delta))^2}{n}\right) \ .$$

Uniform convergence of $\nabla_c L_n$. The proof is similar to the one for $\nabla_\theta L_n$.

Lemma E.8 ($\nabla_c \ell$ is sub-exponential)**.** *There exists a universal constant $C_0 > 0$ such that under the same assumptions as Lemma E.6, with probability at least $1-\delta$ over the random features,*

$$\|\nabla_c \ell(c,\theta;x,f_*(\langle\theta^*,x\rangle)+\xi)\|_{\psi_1} \le C^2 r \ ,$$

where we recall that $C = C_0 \cdot \max\{\mathrm{Lip}(f_*), \tau\sqrt{\log(1/\delta)}, \sigma, 1\}$.

Proof of Lemma E.8. We first let

$$\nabla_c \ell(c,\theta;x,f_*(\langle x,\theta^*\rangle)+\xi) = 2\Phi(\langle x,\theta\rangle)\left(c^\top\Phi(\langle x,\theta\rangle) - f_*(\langle x,\theta^*\rangle) - \xi\right) := 2Y_2(Y_3 - Y_1) \ ,$$

where $Y_1 = f_*(\langle x,\theta^*\rangle)+\xi$, $Y_2 = \Phi(\langle x,\theta\rangle)$ and $Y_3 = c^\top\Phi(\langle x,\theta\rangle)$. Following the proof of Lemma E.6, observe that those are sub-Gaussian random vectors, satisfying $\|Y_1\|_{\psi_2} \lesssim \mathrm{Lip}(f_*) + \sigma \lesssim C$, $\|Y_2\|_{\psi_2} \lesssim \tau\sqrt{\log(1/\delta)} \lesssim C$, and $\|Y_3\|_{\psi_2} \lesssim Cr$. Using again that the product of two sub-gaussian variables is sub-exponential and that the sum of two sub-exponential variables is sub-exponential (Fact B.3), we obtain the desired result. □

Let $\epsilon_\theta = 1/(n^2)$ and $\epsilon_c = r/n^2$ and let $\mathcal{N}_\theta$ and $\mathcal{N}_c$ be ϵ_θ and ϵ_c-nets of $\mathbb{S}^{d-1}$ and B_r^N, respectively. Also, denote by $\tilde{\theta}$ and $\tilde{c}$ the closest element in $\mathcal{N}_\theta$ and $\mathcal{N}_c$ to $\theta \in \mathbb{S}^{d-1}$ and $c \in B_r^N$. Then,

$$\begin{aligned}\sup_{\theta\in\mathbb{S}^{d-1},\|c\|\le r} \|\nabla_c L_n(c,\theta) - \nabla_c L(c,\theta)\| \le &\sup_{\theta\in\mathbb{S}^{d-1},\|c\|\le r} \left\|\nabla_c L_n(c,\theta) - \nabla_c L_n(\tilde{c},\tilde{\theta})\right\| \\ &+ \sup_{\tilde{\theta}\in\mathcal{N}_\theta,\tilde{c}\in\mathcal{N}_c} \left\|\nabla_c L_n(\tilde{c},\tilde{\theta}) - \nabla_c L(\tilde{c},\tilde{\theta})\right\| \\ &+ \sup_{\theta\in\mathbb{S}^{d-1},\|c\|\le r} \left\|\nabla_c L(c,\theta) - \nabla_c L(\tilde{c},\tilde{\theta})\right\| .\end{aligned}$$

We bound the first and third terms by bounding the discretization error for each samplewise gradient. First observe that

$$\|\nabla_c \ell(c,\theta;x,y) - \nabla_c \ell(\tilde{c},\tilde{\theta};x,y)\| \le \|\nabla_c \ell(c,\theta;x,y) - \nabla_c \ell(\tilde{c},\theta;x,y)\| + \|\nabla_c \ell(\tilde{c},\theta;x,y) - \nabla_c \ell(\tilde{c},\tilde{\theta};x,y)\| \ .$$

We bound the first term on the RHS as follows.

$$\begin{aligned}\|\nabla_c \ell(c,\theta;x,y) - \nabla_c \ell(\tilde{c},\theta;x,y)\| &\le \|\Phi(\langle\theta,x\rangle)\| \left|(\tilde{c}-c)^\top\Phi(\langle\theta,x\rangle)\right| \\ &\le \epsilon_c \|\Phi(\langle\theta,x\rangle)\|^2 \\ &\lesssim C\epsilon_c\|x\|^2 \ .\end{aligned}$$

On the other hand,

$$\begin{aligned}\|\nabla_c \ell(\tilde{c},\theta;x,y) - \nabla_c \ell(\tilde{c},\tilde{\theta};x,y)\| &\le \|\Phi(\langle\theta,x\rangle)\||\tilde{c}^\top(\Phi(\langle\theta,x\rangle) - \Phi(\langle\tilde{\theta},x\rangle))| \\ &\quad + \|\Phi(\langle\theta,x\rangle) - \Phi(\langle\tilde{\theta},x\rangle)\||\tilde{c}^\top\Phi(\langle\tilde{\theta},x\rangle))| \\ &\lesssim C\epsilon_\theta\|\tilde{c}\|\|x\| \ .\end{aligned}$$

Thus,

$$\begin{aligned}
\|\nabla_c L(c,\theta) - \nabla_c L(\tilde{c},\tilde{\theta})\| &= \|\,\mathbb{E}[\nabla_c \ell(c,\theta;x,y) - \nabla_c \ell(\tilde{c},\tilde{\theta};x,y)]\| \\
&\le \mathbb{E}[\|\nabla_c \ell(c,\theta;x,y) - \nabla_c \ell(\tilde{c},\tilde{\theta};x,y)\|] \\
&\lesssim C(\epsilon_c + \epsilon_\theta r)d\ . \\
\|\nabla_c L_n(c,\theta) - \nabla_c L_n(\tilde{c},\theta)\| &= \left\| \frac{1}{n}\sum_{i=1}^n \nabla_c \ell(c,\theta;x_i,y_i) - \nabla_c \ell(\tilde{c},\tilde{\theta};x_i,y_i) \right\| \\
&= \frac{1}{n}\sum_{i=1}^n \left\| \nabla_c \ell(c,\theta;x_i,y_i) - \nabla_c \ell(\tilde{c},\tilde{\theta};x_i,y_i) \right\| \\
&\lesssim C\cdot \epsilon_c \sup_i \|x_i\|^2 + \epsilon_\theta r \sup_i \|x_i\| \\
&\lesssim C(\epsilon_c + \epsilon_\theta r)\cdot d\log(n/\delta)\ .
\end{aligned}$$

Taking $\epsilon_c = r/n^2$ and $\epsilon_\theta = 1/n^2$, we observe that with probability $1-(2/3)\delta$,

$$\max\left\{ \left\| \nabla_c L_n(c,\theta) - \nabla_c L_n(\tilde{c},\tilde{\theta}) \right\|, \left\| \nabla_c L(c,\theta) - \nabla_c L(\tilde{c},\tilde{\theta}) \right\| \right\} \lesssim Cr\cdot \frac{\log(n/\delta)}{n}\ .$$

Note that the above upper bound is much smaller than $Cr\sqrt{d\log(n/\delta)/n}$. This indicates that most of the gradient deviation comes from the sub-exponential concentration of $\|\nabla_c L_n(\tilde{c},\tilde{\theta}) - \nabla_c L(\tilde{c},\tilde{\theta})\|$. The bound on this term is given by Bernstein's inequality (Theorem B.4) and the union bound over $\mathcal{N}_\theta \times \mathcal{N}_c$. Recall that $D = \max\{d, N\}$. Then,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\tilde{\theta},\tilde{c}} \left\| \nabla_c L_n(\tilde{c},\tilde{\theta}) - \nabla_c L(\tilde{c},\tilde{\theta}) \right\| \ge C^2 r\cdot \sqrt{\frac{D\log(n/\delta)}{n}} \right] \\
&\le 6^d \left(\frac{3}{\epsilon_\theta}\right)^d \left(\frac{3r}{\epsilon_c}\right)^N \exp\left(-C_1\cdot n\cdot \frac{D\log(n/\delta)}{n} \right) \\
&\le \exp\left(d\log\left(\frac{18}{\epsilon_\theta}\right) + N\log\left(\frac{3r}{\epsilon_c}\right) - C_1\cdot D\log(n/\delta) \right) \\
&\le \frac{\delta}{3}\ .
\end{aligned}$$

□

E.3 Proof of Lemma E.2

Lemma E.2 (Local sharpness of the empirical landscape)**.** *Let $d, n \in \mathbb{N}$ be such that $d \le n$, let $\delta \in (0, 1/4)$, let $\tilde{s} \in \mathbb{N}$ be such that $\tilde{s} \ge s$, where $s \ge 1$ is the information exponent of f_*, let $\lambda \in (0, \lambda_{\tilde{s}}^*)$ where $\lambda_{\tilde{s}}^* \le 1$ depends only on $\tilde{s}$, f_*, and τ, and let $N \in \mathbb{N}$ be such that $N \ge \frac{C_0}{\lambda}\log\frac{1}{\lambda\delta}$, where $C_0 > 0$ is a universal constant. Furthermore, let $D = \max\{d, N\}$ and let $\epsilon \in (0,1)$ be such that $\epsilon \le \lambda^{-2}\sqrt{d/n}$. Then, there exists a universal constant $C_1 > 0$ such that for $C = C_1 \cdot \max\{\mathrm{Lip}(f_*), \tau\sqrt{\log(1/\delta)}, \sigma\}$ and $\Delta = \max\left\{ \sqrt{\frac{D\log(n/\delta)}{n}}, \frac{(d\log(n/\delta))^2}{n} \right\}$ the following holds with probability at least $1-\delta$ over the samples and random features.*

$$\Omega_n(\epsilon,\epsilon) = \Omega_n^{\mathrm{bad}} \sqcup \Omega_n^{\mathrm{good}}\ , \textit{where} \tag{37}$$

$$\Omega_n^{\mathrm{bad}} \subset \left\{ (c,\theta) \;\middle|\; |m| \le \left(\frac{C_2 C^7}{\lambda^2}\cdot\Delta \right)^{\frac{1}{2\tilde{s}-1}} \ \wedge\ \|c\| \le \frac{C^3}{\lambda} \right\},$$

$$\Omega_n^{\mathrm{good}} \subset \left\{ (c,\theta) \;\middle|\; 1-|m| \le \frac{C_3 C^{14}}{\lambda^4}\cdot\Delta^2 \ \wedge\ \|c\| \le \frac{C^3}{\lambda} \right\},$$

and $C_2 = C_{f_*}/\tilde{s}\alpha_{\tilde{s}}^2$ and $C_3 = (2^{2\tilde{s}-1}C_2)^2$ in the above display. Moreover,

$$\min_{(c,\theta)\in\Omega_n^{\mathrm{bad}}} L_n(c,\theta) \ge \sigma^2 + \|f_*\|_\gamma^2 - 2\max\{\|f_*\|_\gamma^2 C_2^2 C^{14}, C^8\}\cdot\Delta_{\mathrm{crit}}/\lambda^2\ , \tag{38}$$

where

$$\Delta_{\text{crit}} = \max\left\{\sqrt{\frac{D\log(n/\delta)}{n}}, \left(\frac{d^2\log^2(n/\delta)}{n}\right)^{\frac{2\bar{s}}{2\bar{s}-1}}\right\} . \tag{39}$$

The main idea behind the proof of Lemma E.2 is that the set of near-critical points of $L(c,\theta)$, which we denote here by Ω and define as the set of (c,θ) for which $\|\nabla L(c,\theta)\| \approx 0$, can be partitioned into two disjoint sets Ω^{bad} and Ω^{good}, where Ω^{bad} is the set of points close to the equator (meaning $|\langle\theta^*,\theta\rangle| \approx 0$) and Ω^{good} is the set of points close to the poles (meaning $|\langle\theta^*,\theta\rangle| \approx 1$). Once this is established, the result follows from uniform convergence of the empirical landscape. To elaborate, with high probability over the samples and random features, any near-critical point of the empirical loss $L_n(c,\theta)$ is a near-critical point of the population loss $L(c,\theta)$ and vice versa by Lemma E.1. Thus, if (c,θ) is a near-critical point of the *empirical* loss $L_n(c,\theta)$, then it is a near-critical point of $L(c,\theta)$ as well, so topological properties of $L(c,\theta)$ dictate that either $|\langle\theta^*,\theta\rangle| \approx 0$ or ≈ 1.

The key observation $\Omega = \Omega^{\text{bad}} \cup \Omega^{\text{good}}$ follows from Lemma E.9 and E.11. Lemma E.9 states that for near-critical points of the *projected* population loss $\bar{L}(\theta)$, a robust version of Theorem D.7 holds. That is, if $\|\nabla_\theta^{\mathbb{S}^{d-1}}\bar{L}(\theta)\| \approx 0$, then $|\langle\theta^*,\theta\rangle| \approx 0$ or $|\langle\theta^*,\theta\rangle| \approx 1$. Lemma E.11 states that if (c,θ) is a near-critical point of $L(c,\theta)$, then θ is a near-critical point of the projected population loss $\bar{L}(\theta)$. That is, if $\|\nabla_c L(c,\theta)\| \approx 0$ and $\|\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta)\| \approx 0$, then $\|\nabla_\theta \bar{L}(\theta)\| \approx 0$. The formal proof is as follows.

Proof of Lemma E.2. Let $\epsilon \le \lambda^{-2}\sqrt{d/n}$ and let $(c,\theta) \in \mathbb{R}^N \times \mathbb{S}^{d-1}$ be an ϵ-approximate first-order critical point of the empirical loss L_n. We first show that with probability at least $1-\delta$ over the samples and random features,

$$\|c\| \lesssim (\text{Lip}(f_*)+\sigma)\tau\log(1/\delta)/\lambda \lesssim C^3/\lambda .$$

By definition, L_n can be expressed as

$$L_n(c,\theta) = c^\top \hat{Q}_\lambda(\theta)c - 2\langle c, \hat{Y}(\theta)\rangle + C_{\text{data}} , \tag{45}$$

where $C_{\text{data}} > 0$ is some constant independent of c and θ, and

$$\hat{Q}_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \Phi(\langle x_i,\theta\rangle)\Phi^\top(\langle x_i,\theta\rangle) + \lambda I , \ \hat{Y}(\theta) = \frac{1}{n}\sum_{i=1}^n (f_*(\langle x_i,\theta^*\rangle)+\xi_i)\Phi(\langle x_i,\theta\rangle) .$$

Using the fact that $\nabla_c L_n(c,\theta) = \hat{Q}_\lambda(\theta)c - \hat{Y}(\theta), \|\nabla_c L_n(c,\theta)\| \le \epsilon$, and $\hat{Q}_\lambda^{-1}(\theta) \preceq \lambda^{-1}I$ uniformly in θ, we obtain

$$\begin{aligned}
\|c\| &\le \|\hat{Q}_\lambda^{-1}(\theta)\|\left(\epsilon + \|\hat{Y}(\theta)\|\right) \\
&\le \lambda^{-1}\left(\epsilon + \|\hat{Y}(\theta)\|\right) \\
&\le \lambda^{-1}\left(\epsilon + \sup_{\theta\in\mathbb{S}^{d-1}}\|\hat{Y}(\theta)\|\right) .
\end{aligned}$$

Thus, it suffices to show that $\sup_\theta \|\hat{Y}(\theta)\|$ is upper bounded with high probability over the samples and random features. We achieve this using an ϵ-net argument over $\mathbb{S}^{d-1}$ and Bernstein's inequality. Let $\mathcal{N}_{1/2}$ be a $1/2$-net of $\mathbb{S}^{d-1}$ and define the random vector $Y(\theta)$

$$Y(\theta) = (f_*(\langle x,\theta^*\rangle)+\xi)\Phi(\langle x,\theta\rangle) \in \mathbb{R}^N ,$$

where $x \sim \mathcal{N}(0, I_d)$ and $\xi \sim \mathcal{N}(0,\sigma^2)$.

For any fixed $\theta \in \mathbb{S}^{d-1}$, $Y(\theta)$ is subexponentially distributed with norm $\|Y(\theta)\|_{\psi_1} \lesssim (\text{Lip}(f_*)+\sigma)\|\Phi(z)\|_{\psi_2} \lesssim (\text{Lip}(f_*)+\sigma)\tau\sqrt{\log(1/\delta)}$ (see Corollary E.5). Hence, $\|\hat{Y}(\theta)\|$ concentrates around

its expectation which is upper bounded as follows. Denote by $m = \langle \theta^*, \theta \rangle$ and let z, z' be m-correlated Gaussian random variables. Then,

$$\begin{aligned}
\mathbb{E}[\|Y(\theta)\|] &= \mathop{\mathbb{E}}_{z,z'}[|f_*(z) + \xi| \|\Phi(z')\|] \\
&\le \mathop{\mathbb{E}}_{z,z'}[|f_*(z)| \|\Phi(z')\|] + \sigma \mathop{\mathbb{E}}_{z'}[\|\Phi(z')\|] \\
&\lesssim \mathop{\mathbb{E}}_{z,z'}[\mathrm{Lip}(f_*)|z|(|z'| + \tau\sqrt{\log(1/\delta)})] + \sigma\tau\sqrt{\log(1/\delta)} \qquad \text{[Lemma E.5]} \\
&\lesssim \mathrm{Lip}(f_*)(1 + \tau\sqrt{\log(1/\delta)}) + \sigma\tau\sqrt{\log(1/\delta)} \\
&\lesssim (\mathrm{Lip}(f_*) + \sigma)\tau\sqrt{\log(1/\delta)} \ ,
\end{aligned}$$

where we used the fact that $\tau\sqrt{\log(1/\delta)} > 1$ in the last line.

By Bernstein's inequality (Theorem B.4) and the union bound over $\mathcal{N}_{1/2}$, the following holds with probability at least $1 - \delta$.

$$\begin{aligned}
\sup_{\theta \in \mathbb{S}^{d-1}} \|\hat{Y}(\theta)\| \le 2 \sup_{\tilde{\theta} \in \mathcal{N}_{1/2}} \|\hat{Y}(\tilde{\theta})\| &\lesssim (\mathrm{Lip}(f_*) + \sigma)\tau\sqrt{\log(1/\delta)}(1 + \sqrt{d\log(1/\delta)/n}) \\
&\lesssim (\mathrm{Lip}(f_*) + \sigma)\tau\log(1/\delta) \ ,
\end{aligned}$$

where we used the assumption $d \le n$ for the last inequality.

Thus, we may set $r \lesssim (\mathrm{Lip}(f_*) + \sigma)\tau\log(1/\delta)/\lambda \lesssim C^3/\lambda$ for the upper bound on $\|c\|$ in Lemma E.1. Recalling the notation $C = \max\{\mathrm{Lip}(f_*), \sigma, \tau\sqrt{\log(1/\delta)}\}$ and using the assumption $\epsilon \le \lambda^{-1}\sqrt{D/n}$, we have

$$\begin{aligned}
\|\nabla_c L(c,\theta)\| &\le \|\nabla_c L(c,\theta) - \nabla_c L_n(c,\theta)\| + \|\nabla_c L_n(c,\theta)\| \\
&\lesssim C^2 r\sqrt{\frac{D\log(n/\delta)}{n}} + \epsilon \\
&\lesssim \frac{C^5}{\lambda}\sqrt{\frac{D\log(n/\delta)}{n}} \ .
\end{aligned} \tag{46}$$

and

$$\begin{aligned}
\|\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta)\| &\le \|\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta) - \nabla_\theta^{\mathbb{S}^{d-1}} L_n(c,\theta)\| + \|\nabla_\theta^{\mathbb{S}^{d-1}} L_n(c,\theta)\| \\
&\le \|\nabla_\theta L(c,\theta) - \nabla_\theta L_n(c,\theta)\| + \|\nabla_\theta^{\mathbb{S}^{d-1}} L_n(c,\theta)\| \\
&\lesssim C_{f_*} C r^2 \cdot \max\left\{\sqrt{\frac{D\log(n/\delta)}{n}}, \frac{(d\log(n/\delta))^2}{n}\right\} + \epsilon \\
&\lesssim \frac{C_{f_*} C^7}{\lambda^2} \cdot \max\left\{\sqrt{\frac{D\log(n/\delta)}{n}}, \frac{(d\log(n/\delta))^2}{n}\right\} \ ,
\end{aligned} \tag{47}$$

where $C_{f_*} = \max\{\|f^{(1)}\|_\gamma, \ldots, \|f^{(4)}\|_\gamma\}$.

Let $\tilde{\epsilon}_c = C_1(C^5/\lambda)\sqrt{D\log(n/\delta)/n}$ and $\tilde{\epsilon}_\theta = C_1(C_{f_*}C^7/\lambda^2) \cdot \max\{\sqrt{D\log(n/\delta)/n}, d^2\log^2(n/\delta)/n\}$, where $C_1 > 0$ is an appropriately chosen universal constant for Eq. (46) and (47). Then, the above inequalities can be expressed as

$$\|\nabla_c L(c,\theta)\| \le \tilde{\epsilon}_c \ \text{ and } \ \|\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta)\| \le \tilde{\epsilon}_\theta \ .$$

Since we assumed $N \ge \frac{C}{\lambda}\log\frac{1}{\lambda\delta}$, the conditions of Lemma E.11 are satisfied. Hence,

$$\|\nabla_\theta^{\mathbb{S}^{d-1}} \bar{L}(\theta)\| \lesssim C_{f_*}(\tau/\lambda)\tilde{\epsilon}_c + \epsilon_\theta \lesssim \frac{C_{f_*}C^7}{\lambda^2} \cdot \max\left\{\sqrt{\frac{D\log(n/\delta)}{n}}, \frac{(d\log(n/\delta))^2}{n}\right\} \ . \tag{48}$$

As a result, by Lemma E.9, which applies since $\lambda < \lambda^*$ and $N \geq \frac{C}{\lambda}\log\frac{1}{\lambda\delta}$, either one of the following must be true.

$$|m| \lesssim \left(\frac{C_{f_*}C^7}{\lambda^2\tilde{s}\alpha_{\tilde{s}}^2}\cdot\max\left\{\sqrt{\frac{D\log(n/\delta)}{n}}, \frac{(d\log(n/\delta))^2}{n}\right\}\right)^{\frac{1}{2\tilde{s}-1}}, \text{ or}$$

$$1-|m| \lesssim \left(\frac{2^{2\tilde{s}-1}}{\tilde{s}\alpha_{\tilde{s}}^2}\right)^2\frac{C_{f_*}^2C^{14}}{\lambda^4}\cdot\max\left\{\sqrt{\frac{D\log(n/\delta)}{n}}, \frac{(d\log(n/\delta))^2}{n}\right\}^2,$$

where we recall from Lemma E.9 that s is the information exponent of f_* and $\tilde{s}\in\mathbb{N}$ is any number satisfying $\tilde{s}\geq s$. Hence, Eq. (37) is established.

Let us now prove (38). By Lemma E.13, for any $(c,\theta)\in\Omega_n^{\text{bad}}$,

$$|L_n(c,\theta)-L(c,\theta)| \lesssim \frac{C^8}{\lambda^2}\cdot\sqrt{\frac{D\log(n/\delta)}{n}}.$$

Denoting $\Delta = \max\{\sqrt{\frac{D\log(n/\delta)}{n}}, \frac{d^2\log^2(n/\delta)}{n}\}$, and using the fact that $C>1$ and $\frac{2\tilde{s}}{2\tilde{s}-1}\leq 2$,

$$\begin{aligned}
\min_{\Omega_n^{\text{bad}}} L_n(c,\theta) &\geq \min_{\Omega_n^{\text{bad}}} L(c,\theta) - \frac{C^8}{\lambda^2}\cdot\sqrt{\frac{D\log(n/\delta)}{n}} \\
&\geq \sigma^2 + \|f_*\|_\gamma^2\cdot\min_{m\in\Omega_n^{\text{bad}}}(1-m^{2\tilde{s}}) - \frac{C^8}{\lambda^2}\cdot\sqrt{\frac{D\log(n/\delta)}{n}} \\
&\geq \sigma^2 + \|f_*\|_\gamma^2\left(1-\left(\frac{C_2C^7\Delta}{\lambda^2}\right)^{\frac{2\tilde{s}}{2\tilde{s}-1}}\right) - \frac{C^8}{\lambda^2}\cdot\sqrt{\frac{D\log(n/\delta)}{n}} \\
&\geq \sigma^2 + \|f_*\|_\gamma^2 - 2\max\{\|f_*\|_\gamma^2C_2^2C^{14}, C^8\}\cdot\frac{\Delta_{\text{crit}}}{\lambda^2}, \qquad (49)
\end{aligned}$$

where $C_2 = C_{f_*}/(\tilde{s}\alpha_{\tilde{s}}^2)$ and

$$\Delta_{\text{crit}} = \max\left\{\sqrt{\frac{D\log(n/\delta)}{n}}, \left(\frac{d^2\log^2(n/\delta)}{n}\right)^{\frac{2\tilde{s}}{2\tilde{s}-1}}\right\}.$$

This concludes the proof of the lemma. □

Lemma E.9 (Near-criticality of $\bar{L}(\theta)$)**.** *Let $\delta\in(0,1/4)$, $\tau>1$, $\beta=\frac{1-1/\tau^2}{3+1/\tau^2}$, $\lambda\in(0,1)$, and let $C_{f_*}=\max\{\|f_*\|_\gamma, \|f_*'\|_\gamma, \|f_*''\|_\gamma\}$. There exist $C_1>0$ depending only on f_* and τ, and a universal constant $C_2>0$ such that for any $\tilde{s}\geq s$, where $s\geq 1$ is the information exponent of f_*, if*

$$C_1\lambda^{\beta/2} < \min\{\tilde{s}\alpha_{\tilde{s}}^2, C_{f_*}^2/\tilde{s}\} \;\; \textit{and} \;\; N \geq \frac{C_2}{\lambda}\log\frac{1}{\lambda\delta},$$

then with probability at least $1-\delta$, the following holds for the projected loss $\bar{L}$. If $\|\nabla_\theta^{\mathbb{S}^{d-1}}\bar{L}(\theta)\|\leq\epsilon$, then $m=\langle\theta^*,\theta\rangle$ satisfies either

$$|m| \leq \left(\frac{2\epsilon}{\tilde{s}\alpha_{\tilde{s}}^2}\right)^{\frac{1}{2\tilde{s}-1}} \qquad \textit{or} \qquad 1-|m| \leq \left(\frac{2^{2\tilde{s}-1}}{\tilde{s}\alpha_{\tilde{s}}^2}\right)^2\cdot\epsilon^2.$$

Remark E.10 (Choice of $\tilde{s}$)**.** *Lemma E.9 gives us freedom over the choice of $\tilde{s}$ provided $\tilde{s}\geq s$, where s is the information exponent of the target link function f_*. If we fix $\tilde{s}=s$, then the upper bound on $C_1\lambda^{\beta/2}$ depends only on f_* and τ. Indeed, the choice $\tilde{s}=s$ implies the tightest upper bound on $|m|$ since its exponent is $1/(2\tilde{s}-1)$. Yet, extra freedom over the choice of $\tilde{s}$ allows us to apply Lemma E.9 to more general settings. For example, when the target, which we denote by f_n, potentially changes with respect to n, but converges to some limit f_* in $L^2(\gamma)$ as $n\to\infty$. In this case, the information exponent s_n of f_n is not necessarily the same as that of f_* nor do the Hermite coefficients match exactly. However, we can ensure that if f_n is sufficiently close to f_*, then $|\alpha_s(f_n)-\alpha_s(f_*)|\geq|\alpha_s(f_*)|/2$ and thus apply Lemma E.9 to f_n using quantities related to f_*.*

Proof of Lemma E.9. We use the representation of the restricted population loss $\bar{L}(\theta)$ from Lemma D.7 Eq. (34), the notation $\rho_m = 2\langle \hat{P}_\lambda g_m, \bar{g}_m\rangle$, and the definition of the Riemannian gradient to obtain

$$\nabla_\theta^{\mathbb{S}^{d-1}} \bar{L}(\theta) = \nabla_\theta \bar{L}(\theta) - \langle \nabla_\theta \bar{L}(\theta), \theta\rangle \theta = -\rho_m(\theta^* - m\theta) \ . \tag{50}$$

This yields an exact representation of the magnitude of the Riemannian gradient $\|\nabla_\theta^{\mathbb{S}^{d-1}} \bar{L}(\theta)\|^2 = \rho_m^2(1-m^2)$ that depends only on $m \in [-1,1]$. Following the proof of Theorem 5.3 for upper bounding $\left|\langle (I-\hat{P}_\lambda) g_m, \bar{g}_m\rangle_\gamma\right|$, we observe that

$$\begin{aligned}
\left|\langle (I-\hat{P}_\lambda) g_m, \bar{g}_m\rangle_\gamma\right| &\le \|(I-\hat{P}_\lambda) g_m\|_\gamma \|\bar{g}_m\|_\gamma \\
&\le 2\sqrt{A(g_m,\lambda)}\|\bar{g}_m\|_\gamma && \text{[Lemma C.2]} \\
&\le 2\sqrt{C(\tau^{1+\beta}\tilde{K}^2\|g_m''\|_\gamma^2\lambda^\beta + \lambda C_{g_m}^2)}\|\bar{g}_m\|_\gamma && \text{[Lemma D.2]} \\
&\le 2\lambda^{\beta/2} C_{g_m}\sqrt{2C\tau^{1+\beta}}\tilde{K}\|\bar{g}_m\|_\gamma \\
&\le 4\lambda^{\beta/2}\sqrt{C\tau^{1+\beta}}\tilde{K}\cdot \sum_{j=s}^\infty j^2\alpha_j^2 m^{2j-1} && \text{[Corollary D.9]} \\
&\le C'\lambda^{\beta/2}\cdot \sum_{j=s}^\infty j^2\alpha_j^2 m^{2j-1} \ ,
\end{aligned} \tag{51}$$

where $C' = 4\sqrt{C\tau^{1+\beta}\tilde{K}}$. Denoting $C_{f_*} = \max\{\|f_*\|_\gamma, \|f_*'\|_\gamma, \|f_*''\|_\gamma\}$, we observe that

$$\sum_{j=\tilde{s}}^\infty j^2\alpha_j^2 m^{2j-1} \le |m|^{2\tilde{s}-1}\sum_{j=\tilde{s}}^\infty j^2\alpha_j^2 = |m|^{2\tilde{s}-1}\left(\|f_*''\|_\gamma^2 + \|f_*''\|_\gamma^2\right) \le 2C_{f_*}^2\,|m|^{2\tilde{s}-1} \ ,$$

Thus, for λ satisfying $2C'C_{f_*}^2\lambda^{\beta/2} \le \min\{\tilde{s}\alpha_{\tilde{s}}^2, C_{f_*}^2/\tilde{s}\}$, we have

$$\begin{aligned}
|\rho_m| &\ge 2\,|\langle g_m, \bar{g}_m\rangle_\gamma| - 2|\langle (I-\hat{P}_\lambda) g_m, \bar{g}_m\rangle_\gamma| \\
&\ge 2\sum_{j=s}^\infty j\alpha_j^2\,|m|^{2j-1} - C'\lambda^{\beta/2}\Big(\sum_{j=s}^{\tilde{s}} j^2\alpha_j^2\,|m|^{2j-1} + \sum_{j=\tilde{s}}^\infty j^2\alpha_j^2\,|m|^{2j-1}\Big) \\
&\ge 2\sum_{j=s}^\infty j\alpha_j^2\,|m|^{2j-1} - \sum_{j=s}^{\tilde{s}} j\alpha_j^2\,|m|^{2j-1} - C'\lambda^{\beta/2}\sum_{j=\tilde{s}}^\infty j^2\alpha_j^2\,|m|^{2j-1} \\
&\ge 2\sum_{j=\tilde{s}}^\infty j\alpha_j^2\,|m|^{2j-1} - 2C'C_{f_*}^2\lambda^{\beta/2}\,|m|^{2\tilde{s}-1} \\
&\ge \tilde{s}\alpha_{\tilde{s}}^2\,|m|^{\tilde{s}-1} \ .
\end{aligned}$$

We now assume that $\|\nabla_\theta^{\mathbb{S}^{d-1}} \bar{L}(\theta)\| = |\rho_m|\sqrt{1-m^2} \le \epsilon$ and retrieve the claimed bounds on $|m|$. If $|m| \le 1/2$, then $|\rho_m| \le \sqrt{4/3}\epsilon$. Hence, our lower bound on $|\rho_m|$ implies that

$$|m| \le \left(\frac{2\epsilon}{\tilde{s}\alpha_{\tilde{s}}^2}\right)^{\frac{1}{2\tilde{s}-1}} .$$

If $|m| > 1/2$, then $|\rho_m| \ge (1/2)^{2\tilde{s}-1}\tilde{s}\alpha_{\tilde{s}}^2$, thus

$$1 - |m| \le \frac{\epsilon^2}{|\rho_m|^2(1+|m|)} \le \left(\frac{2^{2\tilde{s}-1}}{\tilde{s}\alpha_{\tilde{s}}^2}\right)^2 \cdot \epsilon^2 \ .$$

□

Lemma E.11 (Near-criticality of L and $\bar{L}$). *There exists a universal constant $C > 0$ such that for any $\delta \in (0,1)$ and $N \in \mathbb{N}$ satisfying $N \geq C\log(1/\delta)$, the following holds with probability at least $1-\delta$ over the random biases. For any $(c,\theta) \in \mathbb{R}^N \times \mathbb{S}^{d-1}$ such that $\|\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta)\| \leq \epsilon_\theta$ and $\|\nabla_c L(c,\theta)\| \leq \epsilon_c$, it holds*

$$\left\|\nabla_\theta^{\mathbb{S}^{d-1}} \bar{L}(\theta)\right\| \leq \frac{2m^{s-1}\sqrt{1-m^2}\tau\epsilon_c}{\lambda}\sqrt{\|f_*''\|_\gamma^2 + \|f_*'\|_\gamma^2} + \epsilon_\theta \lesssim C_{f_*}(\tau/\lambda)\epsilon_c + \epsilon_\theta \ ,$$

where $C_{f_*} = \max\{\|f_*'\|_\gamma, \|f_*''\|_\gamma\}$.

Proof of Lemma E.11. Recall from Corollary D.5 that

$$\nabla_c L(c,\theta) = 2(Q_\lambda c - \sum_j \alpha_j m^j \mathcal{T}_j) \ \text{ and } \ \nabla_\theta L(c,\theta) = -\langle c, \sum_j j\alpha_j m^{j-1}\mathcal{T}_j\rangle\theta^* \ .$$

Define $c_\theta = \arg\min_c L(c,\theta)$, so that $\bar{L}(\theta) = L(c_\theta,\theta)$. We first show that if $\|\nabla_c L(c,\theta)\| \leq \epsilon_c$, then c and c_θ are nearby. Because $\nabla_c L(c_\theta,\theta) = 0$ and $Q_\lambda \succeq \lambda I_N$,

$$\|\nabla_c L(c,\theta)\| = \|\nabla_c L(c,\theta) - \nabla_c L(c_\theta,\theta)\| = 2\,\|Q_\lambda(c-c_\theta)\| \geq 2\lambda\,\|c-c_\theta\|\,.$$

Thus, $\|c-c_\theta\| \leq \frac{\epsilon_c}{2\lambda}$. We now recall the Riemannian gradient for θ,

$$\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta) = -\langle c, \sum_j j\alpha_j m^{j-1}\mathcal{T}_j\rangle\theta^* + \langle c, \sum_j j\alpha_j m^j \mathcal{T}_j\rangle\theta,$$

and use it to bound the norm of the projected gradient.

$$\begin{aligned}
\left\|\nabla_\theta^{\mathbb{S}^{d-1}} \bar{L}(\theta)\right\| &\leq \left\|\nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta)\right\| + \left\|\nabla_\theta^{\mathbb{S}^{d-1}} L(c_\theta,\theta) - \nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta)\right\| \\
&\leq \epsilon_\theta + \left\| -\langle c_\theta - c, \sum_j j\alpha_j m^{j-1}\mathcal{T}_j\rangle\theta^* + \langle c_\theta - c, \sum_j j\alpha_j m^j\mathcal{T}_j\rangle\theta\right\| \\
&= \epsilon_\theta + \sqrt{1-m^2}\left|\langle c_\theta - c, \sum_j j\alpha_j m^{j-1}\mathcal{T}_j\rangle\right| \\
&\leq \epsilon_\theta + \sqrt{1-m^2}\,\|c_\theta - c\|\,\left\|\sum_j j\alpha_j m^{j-1}\mathcal{T}_j\right\| \\
&\leq \epsilon_\theta + \frac{\epsilon_c}{2\lambda}\sqrt{1-m^2}\left\|\sum_j j\alpha_j m^{j-1}\mathcal{T}_j\right\| \ .
\end{aligned}$$

We conclude by employing Lemmas D.8 and E.12 to obtain a bound on the final term that holds with probability at least $1-\delta$.

$$\begin{aligned}
\left\|\sum_j j\alpha_j m^{j-1}\mathcal{T}_j\right\|^2 &= \left\|\mathcal{T}\sum_j j\alpha_j m^{j-1} h_j\right\|^2 \\
&= \left\|\hat{\Sigma}^{1/2}\sum_j j\alpha_j m^{j-1}h_j\right\|^2 \leq \|\hat{\Sigma}\|_{\mathrm{op}}\left\|\sum_j j\alpha_j m^{j-1}h_j\right\|_\gamma^2 \\
&\leq \mathrm{Tr}(\hat{\Sigma})\,\|\bar{g}_m\|_\gamma^2 \leq \left(\frac{1}{2}+\tau^2\right)(\|f_*''\|_\gamma^2 + \|f_*'\|_\gamma^2)m^{2(s-1)} \ .
\end{aligned}$$

□

Lemma E.12. *There exists a universal constant $C > 0$ such that for any $\tau > 1$ and $\delta \in (0,1)$, if $N \geq C\log(1/\delta)$, then $\|\hat{\Sigma}\|_{\mathrm{op}} \leq \mathrm{Tr}(\hat{\Sigma}) \leq \tau^2 + 1/2$ with probability at least $1-\delta$.*

Proof of Lemma E.12. By definition of $\hat{\Sigma}$, $\mathrm{Tr}(\hat{\Sigma}) = \frac{1}{N}\sum_{i=1}^N \left\|\phi_{b_i}^{\varepsilon_i}\right\|_\gamma^2$, with $\phi_b^\varepsilon(u) = \phi(\varepsilon u - b)$. We compute the expectation of $\|\phi_b^\varepsilon\|_\gamma^2$ for $b \sim \gamma_\tau$ and $\varepsilon \sim$ Rad, show that it is sub-exponential,

and conclude that $\mathrm{Tr}(\hat{\Sigma})$ concentrates around its expectation. The computation of the expectation depends on elementary properties of the Gaussian distribution.

$$\mathop{\mathbb{E}}_{\substack{\varepsilon\sim\mathrm{Rad}\\ b\sim\gamma_\tau}}\left[\|\phi_b^\varepsilon\|_\gamma^2\right] = \frac{1}{2}\mathop{\mathbb{E}}_{\substack{b\sim\gamma_\tau\\ z\sim\gamma}}\left[\phi(z-b)^2\right] + \frac{1}{2}\mathop{\mathbb{E}}_{\substack{b\sim\gamma_\tau\\ z\sim\gamma}}\left[\phi(-z-b)^2\right] = \mathop{\mathbb{E}}_{u\sim\mathcal{N}(0,1+\tau^2)}\left[\phi(u)^2\right] = \frac{1}{2}(1+\tau^2).$$

Note that $\left\|\phi_b^1\right\|_\gamma$ and $\left\|\phi_b^{-1}\right\|$ are $C_1\tau^2$-subgaussian random variables for some constant C_1 because b is τ^2-subgaussian, and $\|\phi_b^\varepsilon\|_\gamma$ for fixed ε is a 1-Lipschitz function of b: $|\,\|\phi_b^\varepsilon\|_\gamma - \|\phi_{b'}^\varepsilon\|_\gamma\,| \le \|\phi_b^\varepsilon - \phi_{b'}^\varepsilon\|_\gamma \le |b-b'|$. Thus, $\|\phi_b^\varepsilon\|_\gamma = \frac{1+\varepsilon}{2}\left\|\phi_b^1\right\|_\gamma + \frac{1-\varepsilon}{2}\left\|\phi_b^{-1}\right\|_\gamma$ is $C'\tau^2$-subgaussian by Fact B.3. As a result, $\|\phi_b^\varepsilon\|_\gamma^2$ is $C'\tau^2$-subexponential, and $\mathrm{Tr}(\hat{\Sigma})$ is $\frac{C\tau^2}{N}$-subexponential. Then,

$$\mathbb{P}\left[\mathrm{Tr}(\hat{\Sigma}) \ge \mathbb{E}\left[\mathrm{Tr}(\hat{\Sigma})\right] + \frac{1}{2}\tau^2\right] \le \exp\left(-\frac{\tau^2/2}{C_2\cdot\tau^2/N}\right) = \exp\left(-\frac{N}{C_2}\right).$$

We conclude by selecting a sufficiently large N. □

Moreover, the empirical loss uniformly concentrated for $(c,\theta)\in B_N(r)\times\mathbb{S}^{d-1}$, as quantified in the following lemma, following the same strategy as our previous gradient concentration:

Lemma E.13 (Uniform convergence of empirical loss)**.** *Let $d,n,N\in\mathbb{N}$ be such $d\le n$, let $D=\max\{d,N\}$, let $\delta\in(0,1/4)$, $r\ge 1$, and let $\sigma^2>0$ $\tau^2>1$. Then, there exists a universal constant $C_0>0$ such that with probability at least $1-\delta$ over samples and random features,*

$$\sup_{\theta\in\mathbb{S}^{d-1},\|c\|\le r}|L_n(c,\theta)-L(c,\theta)| \le C^2r^2\sqrt{\frac{D\log(n/\delta)}{n}}\ ,$$

where $C = C_0\cdot\max\{\mathrm{Lip}(f_*),\tau\sqrt{\log(1/\delta)},\sigma\}$.

Proof of Lemma E.13. We use the same ϵ-net proof as that of Lemma E.1 to prove that this bound holds. As before, we first bound the sub-exponential norm of $\ell(c,\theta;x,y) = (c^\top\Phi(\langle x,\theta\rangle)-y)^2$.

Lemma E.14. *Let $f_*:\mathbb{R}\to\mathbb{R}$ be a Lipschitz function, let $\delta\in(0,1/4)$, let $r\ge 1$, and let $\tau^2>1$. Then there exists a universal constant $C'>0$ such that the following holds with probability at least $1-\delta$ over the random features.*

$$\|\ell(c,\theta;x,f_*(\langle x,\theta^*\rangle)+\xi)\|_{\psi_1} \le C'C^2r^2\ ,$$

where $C=\max\{\mathrm{Lip}(f_*),\tau\sqrt{\log(1/\delta)},\sigma\}$.

Proof of Lemma E.14. By Fact B.3, it suffices to bound $\left\|c^\top\Phi(\langle x,\theta\rangle)-f_*(\langle x,\theta_*\rangle)-\xi\right\|_{\psi_2}$. Note that this quantity identically equals $\|W\|_{\psi_2}$ for the random variable W defined in the proof of Lemma E.6. Thus, with probability at least $1-\delta$,

$$\|W\|_{\psi_2} \lesssim r\tau\left(1+\sqrt{\log(1/\delta)/N}\right) + \mathrm{Lip}(f_*) + \sigma \lesssim Cr\ .$$

□

We consider two ϵ-nets $\mathcal{N}_\theta$ and $\mathcal{N}_c$ of radii ϵ_θ and ϵ_c covering $\mathbb{S}^{d-1}$ and B_r^N respectively. We again denote by $\tilde{\theta}$ and $\tilde{c}$ the closest elements in the nets to θ and c. Then,

$$\sup_{\theta,c}|L_n(c,\theta)-L(c,\theta)| \le \sup_{\theta,c}\left|L_n(c,\theta)-L_n(\tilde{c},\tilde{\theta})\right| + \sup_{\tilde{\theta},\tilde{c}}\left|L_n(\tilde{c},\tilde{\theta})-L(\tilde{c},\tilde{\theta})\right| + \sup_{\theta,c}\left|L(c,\theta)-L(\tilde{c},\tilde{\theta})\right|.$$

We bound the first and last terms by considering the discretization error of samplewise loss.

$$\begin{aligned}\left|\ell(c,\theta;x,y)-\ell(\tilde{c},\tilde{\theta};x,y)\right| &= \left|\left(\tilde{c}^\top\Phi(\langle x,\tilde{\theta}\rangle)-y\right)^2 - \left(c^\top\Phi(\langle x,\theta\rangle)-y\right)^2\right| \\ &= \left|\tilde{c}^\top\Phi(\langle x,\tilde{\theta}\rangle)-c^\top\Phi(\langle x,\theta\rangle)\right|\left|\tilde{c}^\top\Phi(\langle x,\tilde{\theta}\rangle)+c^\top\Phi(\langle x,\theta\rangle)-2y\right|.\end{aligned}$$

We bound the first factor, relying on the event of Corollary E.5 with probability at least $1-(\delta/6)$.

$$\begin{aligned}\left|\tilde{c}^\top\Phi(\langle x,\tilde{\theta}\rangle)-c^\top\Phi(\langle x,\theta\rangle)\right| &\le \left|\tilde{c}^\top\left(\Phi(\langle x,\tilde{\theta}\rangle)-\Phi(\langle x,\theta\rangle)\right)\right|+\left|(\tilde{c}-c)^\top\Phi(\langle x,\theta\rangle)\right| \\ &\le r\left|\langle x,\tilde{\theta}-\theta\rangle\right|+\epsilon_c\left\|\Phi(\langle x,\theta\rangle)\right\| \\ &\lesssim r\epsilon_\theta\left\|x\right\|+\epsilon_c\left(\left\|x\right\|+\tau\sqrt{\log(1/\delta)}\right) \\ &\lesssim (r\epsilon_\theta+\epsilon_c)\left\|x\right\|+\epsilon_c\tau\sqrt{\log(1/\delta)}\ .\end{aligned}$$

We use the same event to bound the second factor.

$$\begin{aligned}\left|\tilde{c}^\top\Phi(\langle x,\tilde{\theta}\rangle)+c^\top\Phi(\langle x,\theta\rangle)-2y\right| &\lesssim (r+\mathrm{Lip}(f_*))\left\|x\right\|+r\tau\sqrt{\log(1/\delta)}+|\xi| \\ &\lesssim rC(\|x\|+1+|\xi|/\sigma)\ ,\end{aligned}$$

where $C=\max\{\mathrm{Lip}(f_*),\tau\sqrt{\log(1/\delta)},\sigma\}$, as defined in the Lemma statement.

Hence, by taking $\epsilon_c=r\epsilon_\theta=r/n$, we have

$$\begin{aligned}\left|\ell(c,\theta;x,y)-\ell(\tilde{c},\tilde{\theta};x,y)\right| &\lesssim \frac{1}{n}\left(r^2C\|x\|^2+(r^2C^2+r^2C)\|x\|+r^2C^2+r\|x\||\xi|/\sigma+r^2C^2|\xi|/\sigma\right) \\ &\lesssim \frac{r^2C^2}{n}\left(\|x\|^2+\|x\|(1+|\xi|/\sigma)+1\right)\ .\end{aligned}$$

By applying Fact E.4 on all ξ_i and the fact that $\|x_i\|^2\lesssim d\log(n/\delta)$ for all i with overwhelming probability, we conclude that with probability at least $1-\delta/3$

$$\sup_{\theta,c}\left|L_n(c,\theta)-L_n(\tilde{c},\tilde{\theta})\right|\lesssim C^2r^2\cdot\frac{d\log(n/\delta)}{n}\ .$$

Likewise, bounds on the expectations of $|\xi|$ and $\|x\|$ similarly give

$$\sup_{\theta,c}\left|L(c,\theta)-L(\tilde{c},\tilde{\theta})\right|\lesssim C^2r^2\cdot\frac{d}{n}\ .$$

We conclude by bounding the second term using Bernstein's inequality with the sub-exponential norm bound of Lemma E.14. Recall that $D=\max\{d,N\}$. Then, for sufficiently large C_0 (and thus sufficiently large C_1),

$$\begin{aligned}\mathbb{P}\left[\sup_{\tilde{\theta},\tilde{c}}\left|L_n(\tilde{c},\tilde{\theta})-L(\tilde{c},\tilde{\theta})\right|\ge C_0C^2r^2\sqrt{\frac{D\log(n/\delta)}{n}}\right] &\le \left(\frac{3}{\epsilon_\theta}\right)^d\left(\frac{3r}{\epsilon_c}\right)^N\exp\left(-C_1n\cdot\frac{D\log(n/\delta)}{n}\right) \\ &\le \exp\left(d\log(3n)+N\log(3n)-C_1D\log(n/\delta)\right) \\ &\le \delta/3\ .\end{aligned}$$

□

E.4 Proof of Lemma E.3

Lemma E.3 (Gradient flow escapes the equator)**.** *Assume $\sum_j(j+A)^k\alpha_j^2\le C$ for $A\le s$ and $k\le 3$. With probability at least $1/2-2\delta$ over the initial condition, the draw of the data, and the draw of the random features, if $n=\tilde{\Omega}(\max\{\lambda^{-4}(d+N)d^{s-1},\lambda^{-2}d^{(s+3)/2}\})$ and $N=\Theta\left(\lambda^{-1}\log(\lambda^{-1}\delta^{-1})\right)$ then the first phase of gradient flow with a randomly initialised $c(0)\sim\mathrm{Unif}\{c\in\rho\mathbb{S}^{N-1};\|c\|_0=N_0\}$ with $\rho=\Theta(\sqrt{N}N_0^{-(2+s)/2}(\tau^2+\lambda N/N_0)^{-1})$ and $N_0=\Theta\left(\log\frac{1}{\delta}\right)$ escapes the equator in time $T_0=\widetilde{O}\left(d^{s/2-1}\right)$.*

Recall our gradient flow dynamics in the first phase:

$$\dot{\theta}(t)=-\nabla_\theta L_n(c(0),\theta(t))\ ,$$

where $c(0) \sim \mathrm{Unif}(\{c \in \mathbb{R}^N; \|c\|_2 = \rho; \|c\|_0 = N_0\})$, $\theta(0) \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, where ρ is another parameter determining the initial norm of c.

Our goal is to show that the gradient flow trajectory is likely to cross the energy barrier $B_{\text{crit}} = \tilde{\Theta}\left(\lambda^{-2}\Delta_{\text{crit}}\right)$ and therefore avoid the bad critical points (see Eq. (38)). Denote by $m(t) = \langle \theta(t), \theta_* \rangle$ the trajectory of the correlation. We will show that, from an initial correlation $m(0) \sim 1/\sqrt{d}$, the gradient flow dynamics yield $\dot{m}(t) > 0$ for long enough to guarantee that $m(t)$ grows substantially. This will ultimately be sufficient to ensure that the loss crosses the previous energy barrier, provided $c(0)$ has an appropriate norm ρ.

Thanks to the concentration results from Lemma E.13 and Lemma E.1, we can first compute the correlation trajectory $m(t)$ for the population loss, and then extend them to the empirical gradients.

Assume that $m(0)$ and $c(0)$ are such that that $\mathrm{sign}(\alpha_s c(0)^\top \mathcal{T}_s) = \mathrm{sign}(m(0))$, which occurs with probability $1/2$ over the randomness of $c(0)$ and $m(0)$. By symmetry, we will assume $m(0) > 0$ and $\alpha_s c(0)^\top \mathcal{T}_s > 0$ for the rest of the proof. Let us express the population objective without offset as

$$L(c, \theta) = \|f_*\|_\gamma^2 + c^\top Q_\lambda c - 2m^s \bar{R} \ , \text{ with} \tag{52}$$

$$\bar{R}(m) := \alpha_s \langle c, \mathcal{T}_s \rangle + m \sum_{j \geq 0} \alpha_{j+s+1} \langle c, \mathcal{T}_{j+s+1} \rangle m^j \ . \tag{53}$$

From Lemma B.7, we know that the correlation $m(0)$ at initialization cannot be too small. More precisely,

$$\mathbb{P}\left(|m(0)| \geq \delta/\sqrt{d}\right) \leq 1 - 4\delta \ .$$

Moreover, the change in correlation according to the population gradient is given by

$$-\langle \nabla_\theta^{\mathbb{S}^{d-1}} L(c, \theta(t)), \theta_* \rangle = (1 - m^2) m^{s-1} R \ , \tag{54}$$

where we have defined

$$R(m) := s\alpha_s \langle c, \mathcal{T}_s \rangle + m \sum_{j \geq 0} (j + s + 1) \alpha_{j+s+1} \langle c, \mathcal{T}_{j+s+1} \rangle m^j \ . \tag{55}$$

The following lemma, proved below, shows there exists $\gamma = \gamma(c(0)) > 0$ and $\bar{\gamma}$ such that $R(m) > R(0)/2$ for $m \in [0, \gamma)$ and $\bar{R}(m) > \bar{R}(0)/2$ for $m \in [0, \bar{\gamma})$.

Lemma E.15. *Let* $C_{f*,\tau} = 2\tau \left(\sum_{j>0} (j+s)^2 j^2 \alpha_{j+s}^2\right)^{1/2}$ *and* $\bar{C}_{f^*,\tau} = 2\tau \left(\sum_{j>0} (j)^2 \alpha_{j+s}^2\right)^{1/2}$. *Then*

1. $R(m) > \frac{1}{2} s\alpha_s c^\top \mathcal{T}_s$ *for* $m \in [0, \gamma)$, *where*
$$\gamma \geq \frac{s\alpha_s c^\top \mathcal{T}_s}{2\rho C_{f_*,\tau}} \ ,$$
2. $\bar{R}(m) > \frac{1}{2} \alpha_s c^\top \mathcal{T}_s$ *for* $m \in [0, \bar{\gamma})$, *where*
$$\bar{\gamma} \geq \frac{\alpha_s c^\top \mathcal{T}_s}{2\rho \bar{C}_{f*,\tau}} \ .$$

In other words, the gradient flow under the population loss sees a monotonically increasing correlation m (since its time derivative under the population gradient flow is positive), until $m(t)$ reaches a value $\gamma = C\frac{\alpha_s c^\top \mathcal{T}_s}{\rho}$.

Let $\gamma_* = \min(\gamma, \bar{\gamma})$ and $\rho_0 = \rho\sqrt{\frac{N_0}{N}}$. As the correlation reaches the value $m = \gamma_*$, using Lemma E.15 to lower bound $\bar{R}$, one can verify that the population loss obeys the following upper bound:

$$L_{\text{esc}} \leq \|f_*\|_\gamma^2 + \rho^2 \langle c/\|c\|, Q_\lambda c/\|c\| \rangle - \rho\gamma_*^s \alpha_s \langle c/\|c\|, \mathcal{T}_s \rangle \tag{56}$$

$$= \|f_*\|_\gamma^2 + \lambda\rho^2 + \rho_0^2 \langle \tilde{c}, \tilde{Q}\tilde{c} \rangle - \rho_0 \gamma_*^s \alpha_s \langle \tilde{c}, \tilde{\mathcal{T}}_s \rangle \ , \tag{57}$$

where, denoting by $\mathcal{S}$ the support of c, we defined

$$\tilde{Q} = \frac{1}{N_0}\left[\langle \phi(\cdot - b_j), \phi(\cdot - b_{j'})\rangle_\gamma\right]_{j,j'\in\mathcal{S}} \in \mathbb{R}^{N_0\times N_0} \tag{58}$$

$$\tilde{\mathcal{T}}_s = \frac{1}{\sqrt{N_0}}\left[\langle h_s, \phi(\cdot - b_j)\rangle_\gamma\right]_{j\in\mathcal{S}} \in \mathbb{R}^{N_0} \tag{59}$$

$$\tilde{c} = \frac{1}{\rho}\left[c_j\right]_{j\in\mathcal{S}} \in \mathbb{R}^{N_0}. \tag{60}$$

By Lemma E.12, we have $\|\tilde{Q}\|_{\text{op}} \leq 2\tau^2$ w.p. $1-\delta$ as soon as $N_0 \gtrsim \log(1/\delta)$, so that the bound above becomes:

$$L_{\text{esc}} \leq \|f_*\|_\gamma^2 + \lambda\rho^2 + 2\tau^2\rho_0^2 - \rho_0\gamma_*^s\alpha_s\langle \tilde{c}, \tilde{\mathcal{T}}_s\rangle. \tag{61}$$

Let us now verify that the empirical correlation trajectory and loss have the same behavior. Observe that

$$\begin{aligned}
\dot{m}(t) &= -\langle \nabla_\theta^{\mathbb{S}^{d-1}} L_n(c,\theta(t)), \theta_*\rangle \\
&= -\langle \nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta(t)), \theta_*\rangle + \langle \nabla_\theta^{\mathbb{S}^{d-1}} L(c,\theta(t)) - \nabla_\theta^{\mathbb{S}^{d-1}} L_n(c,\theta(t)), \theta_*\rangle \\
&= (1-m^2)m^{s-1}R(m) + \widetilde{O}\left(\lambda^{-2}\max\left\{\sqrt{\frac{D}{n}}, \frac{d^2}{n}\right\}\right) .
\end{aligned} \tag{62}$$

From the anti-concentration Lemma B.7, it follows that whenever $n = \tilde{\Omega}(\max\{\lambda^{-4}Dd^{s-1}, \lambda^{-2}d^{\frac{s+3}{2}}\})$, with probability greater than $1-\delta$

$$(1-m(0)^2)(m(0))^{s-1}R(m(0)) \gg \widetilde{O}\left(\lambda^{-2}\max\left\{\sqrt{\frac{D}{n}}, \frac{d^2}{n}\right\}\right) \tag{63}$$

and therefore from Lemma E.15 we deduce that $\dot{m}(0) > 0$, and $m(t)$ keeps increasing at least until it reaches γ_*. From Lemma E.13, the empirical loss at this correlation level is with probability greater than $1-\delta$

$$\begin{aligned}
L_{n,\text{esc}} &\leq \|f_*\|_\gamma^2 + \sigma^2 + \lambda\rho^2 + 2\tau^2\rho_0^2 - \rho_0\gamma_*^s\alpha_s\langle \tilde{c}, \tilde{\mathcal{T}}_s\rangle + \widetilde{O}\left(\lambda^{-2}\sqrt{\frac{D}{n}}\right) \\
&= \|f_*\|_\gamma^2 + \sigma^2 + \lambda\rho^2 + 2\tau^2\rho_0^2 - C\rho_0(\alpha_s\langle \tilde{c}, \tilde{\mathcal{T}}_s\rangle)^{s+1} + \widetilde{O}\left(\lambda^{-2}\sqrt{\frac{D}{n}}\right)
\end{aligned}$$

In order to ensure that this initial training phase escapes the 'bad' empirical points near the equator $|m| \approx 0$, by Eq. (38), it is sufficient to show that

$$\lambda\rho^2 + 2\tau^2\rho_0^2 - C\rho_0(\alpha_s\langle \tilde{c}, \tilde{\mathcal{T}}_s\rangle)^{1+s} \ll -\widetilde{O}\left(\lambda^{-2}\Delta_{\text{crit}}\right), \tag{64}$$

with $\Delta_{\text{crit}} := \max\left\{\sqrt{\frac{D}{n}}, \left(\frac{d^2}{n}\right)^{\frac{2s}{2s-1}}\right\}$.

Let us now study the term $\langle \tilde{c}, \tilde{\mathcal{T}}_s\rangle$ for the choice of sparsity N_0 we picked for c. Let $\mu_s = \langle h_s, \Sigma h_s\rangle$. Observe that $\mu_s > 0$ since the kernel is universal. We have the following anti-concentration result:

Lemma E.16 (Anticoncentration of $|\langle \tilde{c}, \tilde{\mathcal{T}}_s\rangle|$)**.** *We have*

$$\mathbb{P}\left(|\langle \tilde{c}, \tilde{\mathcal{T}}_s\rangle| \geq \frac{\mu_s\delta}{8\sqrt{N_0}}\right) \geq 1 - 2e^{\frac{N_0 c\mu_s^2}{4\tau^4}} - \delta , \tag{65}$$

where the probability is over both the initial draw of c and the draw of the random features.

We obtain that $\alpha_s\langle \tilde{c}, \tilde{\mathcal{T}}_s\rangle \geq \frac{\alpha_s\mu_s\delta}{8\sqrt{N_0}}$ holds with probability close to $1/2$. Then, the condition (64) becomes

$$\lambda\rho^2 + 2\tau^2\rho_0^2 - C'\rho_0 N_0^{-\frac{1+s}{2}} = (\lambda N/N_0 + 2\tau^2)\rho_0^2 - C'\rho_0 N_0^{-\frac{1+s}{2}} \ll -\widetilde{O}(\lambda^{-2}\Delta_{\text{crit}}), \tag{66}$$

with C' a positive constant.

Taking $\rho_0 = \frac{C' N_0^{-\frac{1+s}{2}}}{2(2\tau^2+\lambda N/N_0)}$ yields the new condition

$$\frac{C'' N_0^{-(1+s)}}{2\tau^2 + \lambda N/N_0} = \widetilde{\Omega}\left(\lambda^{-2}\Delta_{\text{crit}}\right). \tag{67}$$

In particular, since we assume $\lambda^{-2}\Delta_{\text{crit}} \ll 1$, we may take $N_0 = \Theta(1)$ and

$$N = \widetilde{O}\left(\lambda \Delta_{\text{crit}}^{-1}\right)$$

to ensure (67), and consequently (64).

Finally, let us upper bound the escape time T needed to reach $m(T) = \gamma_*$. Denote $\Delta_n := \lambda^{-2}\max\left\{\sqrt{\frac{D}{n}}, \frac{d^2}{n}\right\}$. Observe that for $t \le T$,

$$\begin{aligned} \dot{m}(t) &\ge \frac{1}{2}(1-m^2)m^{s-1}s\alpha_s c^\top \mathcal{T}_s - \widetilde{O}\left(\Delta_n\right) \\ &\ge A(1-\gamma_*^2)m(t)^{s-1} - \widetilde{O}\left(\Delta_n\right) \\ &:= G(m(t)) - \widetilde{O}\left(\Delta_n\right) , \end{aligned} \tag{68}$$

where $G(u) = \tilde{A}u^{s-1}$ is convex in $[0,1]$ with

$$\tilde{A} = \frac{1}{2}(1-\gamma_*^2)s\alpha_s c^\top \mathcal{T}_s .$$

A crude bound is therefore

$$\begin{aligned} \dot{m}(t) &\ge G(m(0)) + G'(m(0))(m(t)-m(0)) - \widetilde{O}\left(\Delta_n\right) , \\ &:= A + Bm(t) , \end{aligned} \tag{69}$$

with

$$\begin{aligned} A &= G(m(0)) - G'(m(0))m(0) - \widetilde{O}\left(\Delta_n\right) = \tilde{A}m(0)^{s-1}(2-s) - \widetilde{O}\left(\Delta_n\right) \\ B &= G'(m(0)) = \tilde{A}(s-1)m(0)^{s-2} , \end{aligned} \tag{70}$$

which leads to a Gronwall-type inequality of the form

$$m(t) \ge \frac{A}{B-1}\left(e^{Bt}-1\right) + m(0) , \tag{71}$$

and therefore

$$T \le B^{-1}\log\left(\frac{\gamma_*(B-1)}{A}\right) = \tilde{A}^{-1}\widetilde{O}\left(d^{s/2-1}\right) = \widetilde{O}\left(d^{s/2-1}\right) , \tag{72}$$

since $\tilde{A} = \Theta(|c^\top \mathcal{T}_s|) = \Theta\left(N_0^{-\frac{2+s}{2}}\right) = \Theta(1)$. This concludes the proof. □

Proof of Lemma E.15. Recall that $\mathcal{T} : L^2(\gamma) \to \mathbb{R}^N$ is the operator $\mathcal{T}f = (\langle f, \phi_{b_i}^{\varepsilon_i}\rangle_\gamma)_{i=1\dots n}$. Its operator norm is bounded by Lemma E.12 with probability $1-\delta$ over the random features: $\|\mathcal{T}\|_{\text{op}} \le \|\hat{\Sigma}\|_{\text{op}}^{1/2} \le 2\tau$. Since $R(0) = s\alpha_s\langle c, \mathcal{T}_s\rangle$ and

$$\begin{aligned} R'(m) &= \left\langle c, \sum_{j>0}(j+s)j\alpha_{j+s}\mathcal{T}_{j+s}m^{j-1}\right\rangle \\ &= \left\langle c, \sum_{j>0}(j+s)j\alpha_{j+s}m^{j-1}\mathcal{T}h_{j+s}\right\rangle \\ &= \left\langle c, \mathcal{T}\left(\sum_{j>0}(j+s)jm^{j-1}\alpha_{j+s}h_{j+s}\right)\right\rangle \end{aligned} \tag{73}$$

satisfies

$$\begin{aligned}\sup_{m\in[0,1]}|R'(m)| &\leq \|\mathcal{T}\|\rho\left\|\sum_{j>0}(j+s)jm^{j-1}\alpha_{j+s}h_{j+s}\right\|_\gamma \\ &\leq 2\tau\rho\sqrt{\sum_{j>0}(j+s)^2j^2m^{2(j-1)}\alpha_{j+s}^2} \\ &\leq \rho C_{f_*,\tau}\ . \end{aligned} \tag{74}$$

Thus, if we assume that $m \leq \gamma$ as specified in the theorem statement,

$$\begin{aligned}R(m) &\geq R(0) - |R(m)-R(0)| \geq s\alpha_s\langle c,\mathcal{T}_s\rangle - \sup_{m\in[0,1]}|R'(m)|\,|m-0| \\ &\geq s\alpha_s\langle c,\mathcal{T}_s\rangle - m\rho C_{f_*} \geq \frac{1}{2}s\alpha_s\langle c,\mathcal{T}_s\rangle.\end{aligned}$$

The derivation for $\bar{R}(m)$ is analogous. □

Proof of Lemma E.16. Observe that the dot product $\langle\tilde{c},\tilde{\mathcal{T}}_s\rangle$ only depends on a subset of N_0 random features. Let us denote by

$$Z_s = \langle h_s, \sigma(\cdot - Z)\rangle_\gamma, \text{ where } Z\sim\gamma_\tau\ .$$

Observe that $Z_s = \psi(Z)$ with $\psi(x) = \langle h_s, \sigma(\cdot - x)\rangle_\gamma$ satisfying

$$|\psi'(x)| = |\langle h_s, \sigma'(\cdot - x)\rangle_\gamma| \leq \|h_s\|_\gamma\|\sigma'(\cdot - x)\|_\gamma \leq 1\ ,$$

which shows that Z_s is $\frac{1}{2\tau^2}$-subgaussian, and thus that the random vector

$$\tilde{\mathcal{T}}_s = \frac{1}{\sqrt{N_0}}(\langle h_s, \sigma(\cdot - b_j)\rangle_\gamma; j\in\text{supp}(c)) \in \mathbb{R}^{N_0}$$

has independent $\frac{N_0}{2\tau^2}$-subgaussian entries. Therefore, by Bernstein concentration [83, Theorem 3.1.1], the Euclidean norm $\|\tilde{\mathcal{T}}_s\|$ concentrates around its expectation $\mu_s = \sqrt{\langle h_s, \Sigma h_s\rangle}$ as

$$\mathbb{P}_\Phi(|\|\tilde{\mathcal{T}}_s\| - \mu_s| \geq t) \leq 2e^{-\frac{cN_0t^2}{\tau^4}}\ .$$

Finally, using again the anticoncentration of the correlation of a uniform direction with a fixed direction (Lemma B.7), we obtain with a union bound that

$$\mathbb{P}_{c,\Phi}\left(|\langle\tilde{c},\tilde{\mathcal{T}}_s\rangle| \geq \frac{\mu_s\delta}{4\sqrt{N_0}}\right) \geq 1-\delta-2e^{-\frac{cN_0\mu_s^2}{4\tau^4}}\ , \tag{75}$$

as claimed. □

E.5 Proof of Corollary 6.2

We restate Corollary 6.2 here for convenience.

Corollary 6.2 (Excess risk of Algorithm 1)**.** *Under the assumptions of Theorem 6.1, and further assuming $n \gtrsim d^3$, an appropriate choice of λ yields an excess risk guarantee of the form*

$$\|\hat{F} - F^*\|_{\gamma_d}^2 = \widetilde{O}\left(\left(\frac{d}{n}\right)^{\frac{\beta}{\beta+4}} + \left(\frac{1}{n}\right)^{\frac{\beta}{\beta+5}}\right)\ , \tag{12}$$

where β is defined as in Lemma 4.4.

Proof. Let $\hat{F}(x) = \hat{f}(\langle x,\hat{\theta}\rangle)$, and $G_{m,\hat{\theta}}(x) = g_m(\langle\hat{\theta},x\rangle) = \sum_j \alpha_j m^j h_j(\langle\hat{\theta},x\rangle)$, where $m = \langle\hat{\theta},\theta^*\rangle$. We have

$$\|\hat{F} - F^*\|_{\gamma_d}^2 \leq 2\|\hat{F} - G_{m,\hat{\theta}}\|_{\gamma_d}^2 + 2\|G_{m,\hat{\theta}} - F^*\|_{\gamma_d}^2 \tag{76}$$

$$= 2\|\hat{f} - g_m\|_\gamma^2 + 2\|G_{m,\hat{\theta}} - F^*\|_{\gamma_d}^2\ . \tag{77}$$

Denoting $c_\theta = Q_\lambda^{-1}\mathcal{T}g_m$, and considering $N = \Theta(\frac{1}{\lambda}\log\frac{1}{\lambda})$, recall that we have

$$\nabla_c L(\hat{c},\hat{\theta}) = 2Q_\lambda(\hat{c}-c_\theta), \qquad \|\nabla_c L(\hat{c},\hat{\theta})\| \le \widetilde{O}\left(\lambda^{-1}\sqrt{\frac{d+N}{n}}\right) = \widetilde{O}\left(\sqrt{\frac{d}{\lambda^2 n}+\frac{1}{\lambda^3 n}}\right), \tag{78}$$

Thus, we obtain

$$\|\hat{f}-\hat{P}_\lambda g_m\|_\gamma^2 = \|Q(\hat{c}-c_\theta)\|^2 \le \|Q_\lambda(\hat{c}-c_\theta)\|^2 = \frac{1}{4}\|\nabla_c L(\hat{c},\theta)\|^2 \le \widetilde{O}\left(\frac{d}{\lambda^2 n}+\frac{1}{\lambda^3 n}\right). \tag{79}$$

As a consequence, using Lemmas C.2 and 4.4 we obtain

$$\|\hat{f}-g_m\|_\gamma^2 \le 2\left(\|\hat{f}-\hat{P}_\lambda g_m\|_\gamma^2 + \|(I-\hat{P}_\lambda)g_m\|_\gamma^2\right) \le \widetilde{O}\left(\frac{d}{\lambda^2 n}+\frac{1}{\lambda^3 n}+2\lambda^\beta\|f_*''\|_\gamma^2\right). \tag{80}$$

On the other hand, we also have

$$\begin{aligned}
\|G_{m,\hat{\theta}} - F_*\|_{\gamma_d}^2 &= \sum_j \alpha_j^2 m^{2j} + \sum_j \alpha_j^2 - 2\sum_j m^{2j}\alpha_j^2 && (81)\\
&= \sum_j (1-m^{2j})\alpha_j^2 && (82)\\
&\le (1-|m|)\sum_j 2j\alpha_j^2 && (83)\\
&= O\left(1-|m|\right) = \widetilde{O}\left(\lambda^{-4}\frac{d+N}{n}\right) = \widetilde{O}\left(\frac{d}{\lambda^4 n}+\frac{1}{\lambda^5 n}\right). && (84)
\end{aligned}$$

where the $\widetilde{O}$ follows from Lemma E.2. We thus obtain

$$\|\hat{F}-F^*\|_{\gamma_d}^2 \le \widetilde{O}\left(\frac{d}{\lambda^2 n}+\frac{d}{\lambda^4 n}+\frac{1}{\lambda^3 n}+\frac{1}{\lambda^5 n}+\lambda^\beta\right) = \widetilde{O}\left(\frac{d}{\lambda^4 n}+\frac{1}{\lambda^5 n}+\lambda^\beta\right).$$

Setting

$$\lambda = \max\left\{\left(\frac{1}{n}\right)^{\frac{1}{\beta+5}}, \left(\frac{d}{n}\right)^{\frac{1}{\beta+4}}\right\},$$

we then have

$$\|\hat{F}-F^*\|_{\gamma_d}^2 = \widetilde{O}\left(\max\left\{\left(\frac{d}{n}\right)^{\frac{\beta}{\beta+4}}, \left(\frac{1}{n}\right)^{\frac{\beta}{\beta+5}}\right\}\right), \tag{85}$$

which establishes the desired rate. It remains to check that the upper bound $N = \tilde{O}(\lambda\Delta_{\text{crit}}^{-1})$ required by Theorem 6.1 holds with this choice of λ. Note that we have

$$\begin{aligned}
\frac{N\Delta_{\text{crit}}}{\lambda} &= \widetilde{O}\left(\frac{1}{\lambda^2}\sqrt{\frac{d+\frac{1}{\lambda}}{n}}\right)\\
&= \widetilde{O}\left(\sqrt{\frac{d}{\lambda^4 n}+\frac{1}{\lambda^5 n}}\right) = o(1),
\end{aligned}$$

where we control the quantity inside the square root in the same way as we obtained (85). The calculation above also shows that $\frac{\Delta_{\text{crit}}}{\lambda^2} = o(1)$, so that the requirement $\lambda = \Omega(\sqrt{\Delta_{\text{crit}}})$ is also satisfied. □

E.6 Proof of Proposition 6.3

Proposition 6.3 (Excess risk of fine-tuning)**.** *Let $\delta \in (0,1/4)$. Let $m = \langle\theta^*,\hat{\theta}\rangle$, where $\hat{\theta}$ is obtained from the previous gradient descent phase, and let $\hat{c}$ be the ridge regression estimator*

obtained from a fresh dataset $\mathcal{D}'$ of n' samples, N random features, and regularization parameter $\lambda_{n'} := (\sigma^2\tau^2/\|f_*''\|_\gamma^2 n')^{1/(\beta+1)}$, and let $\hat{F}(x) = \hat{c}^\top \Phi(\langle \hat{\theta}, x\rangle)$. Assume

$$n' \gtrsim \max\left\{\sigma^2\tau^2/\|f_*''\|_\gamma^2, (\|f_*''\|_\gamma^2/\sigma^2\tau^2)^{1/\beta}, \|f_*\|_\infty^2/(\sigma^2\tau^2)^{\beta/(\beta+1)}\right\} \text{ , and}$$
$$N \gtrsim C_\tau \left(n'\|f_*''\|_\gamma^2/\sigma^2\tau^2\right)^{\frac{1}{\beta+1}} \log\left(n'^{1/(\beta+1)}\delta^{-1}\right) .$$

Then with probability at least $1-\delta$ over the random features, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}'}[\|\hat{F} - F^*\|_{\gamma_d}^2 | \hat{\theta}] \lesssim \|f_*''\|_\gamma^{\frac{2}{\beta+1}} \left(\frac{\sigma^2\tau^2}{n'}\right)^{\frac{\beta}{\beta+1}} + \|f_*'\|_\gamma^2(1-|m|) , \tag{14}$$

where the expectation is over the n' fresh samples, and is conditioned on the previously obtained $\hat{\theta}$.

Proof. Let $\hat{\kappa}_\theta(x,x') = \Phi(\langle\theta,x\rangle)^\top\Phi(\langle\theta,x'\rangle)$ be the random feature kernel on $\mathbb{R}^d$ and denote by $\hat{\mathcal{H}}_\theta$ the corresponding RKHS. Let κ_θ and $\mathcal{H}_\theta$ be their infinite-width counterparts. Note that $\kappa_\theta(x,x') = \langle\varphi(\langle\theta,x\rangle),\varphi(\langle\theta,x'\rangle)\rangle_{\mathcal{H}}$, where $\varphi(u) = \kappa(u,\cdot)$ denotes the kernel mapping of $\mathcal{H}$. Then, one can easily show, e.g. using Theorem B.8, that $\mathcal{H}_\theta = \{F = f(\langle\theta,\cdot\rangle) \ : \ f \in \mathcal{H}\}$, with $\|F\|_{\mathcal{H}_\theta} = \|f\|_{\mathcal{H}}$ when $F(x) = f(\langle\theta,x\rangle)$.

Considering fresh samples (x_i,y_i), $i=1,\dots,n'$, with $y_i = F^*(x_i)+\epsilon_i$, $\mathbb{E}[\epsilon_i|x_i]=0$, $\mathrm{Var}[\epsilon_i|x_i] \le \sigma^2$, we now assume

$$c = \left(\frac{1}{n'}\sum_i \Phi(\langle\theta,x_i\rangle)\Phi(\langle\theta,x_i\rangle)^\top + \lambda I\right)^{-1} \frac{1}{n'}\sum_i y_i\Phi(\langle\theta,x_i\rangle)$$

Define

$$Q = \mathop{\mathbb{E}}_{x\sim\gamma_d}[\Phi(\langle\theta,x\rangle)\Phi(\langle\theta,x\rangle)^\top] = \mathop{\mathbb{E}}_{z\sim\gamma}[\Phi(z)\Phi(z)^\top] \in \mathbb{R}^{N\times N} ,$$

and

$$\hat{Q} = \frac{1}{n'}\sum_i \Phi(\langle\theta,x_i\rangle)\Phi(\langle\theta,x_i\rangle)^\top .$$

Assume for now that F^* belongs to $\hat{\mathcal{H}}_\theta$ and takes the form $F^*(x) = c_*^\top\Phi(\langle\theta,x\rangle) = \tilde{f}(\langle\theta,x\rangle)$. Then, we may use a variant of [7, Prop 7.2], which holds for bounded data, to our unbounded setting, by adapting the covariance concentration step [7, Proposition 7.1].

Lemma E.17 (Concentration for Covariance Operators, sub-exponential case)**.** *For $n' \ge \frac{R^2}{2\lambda}\log\frac{R^2}{\lambda}$, where R is a universal constant, with probability greater than $1 - 7\mathrm{Tr}(QQ_\lambda^{-1})\exp\left(-\frac{n'}{8R^2+2R}\right)$ it holds*

$$-\frac{1}{2}I \preceq (Q+\lambda I)^{-1/2}(Q-\hat{Q})(Q+\lambda I)^{-1/2} \preceq \frac{1}{2}I . \tag{86}$$

Then, we have for $n' \ge \frac{R^2}{2\lambda}\log\frac{R^2}{\lambda}$ and $\lambda \le R^2$, following [7, Proposition 7.1]

$$\mathbb{E}[\|F_{c,\theta} - F^*\|_{\gamma_d}^2] \le 16\frac{\sigma^2}{n'}\mathrm{Tr}(Q(Q+\lambda I)^{-1}) + 16\inf_{F\in\hat{\mathcal{H}}_\theta}\left\{\|F-F^*\|_{\gamma_d}^2 + \lambda\|F\|_{\hat{\mathcal{H}}_\theta}^2\right\} + \frac{24}{n'^2}\|F^*\|_\infty^2, \tag{87}$$

where the expectation is over the n' fresh samples.

We have the following upper bound on the first (variance) term

$$\mathrm{Tr}(Q(Q+\lambda I)^{-1}) \le \frac{\mathrm{Tr}(Q)}{\lambda} = \frac{\mathrm{Tr}(\hat{\Sigma})}{\lambda} \le \frac{\frac{1}{2}+\tau^2}{\lambda} \le \frac{2\tau^2}{\lambda}, \tag{88}$$

where we used Lemma E.12.

The approximation error may be controlled as follows:

$$\begin{aligned}
\inf_{F\in\hat{\mathcal{H}}_\theta}\left\{\|F-F^*\|_{\gamma_d}^2+\lambda\|F\|_{\hat{\mathcal{H}}_\theta}^2\right\} &= \lambda\langle c_*, Q(Q+\lambda I)^{-1}c_*\rangle \\
&= \lambda\langle \tilde{f}, (\hat{\Sigma}+\lambda I)^{-1}\tilde{f}\rangle_\gamma \\
&\le 4\lambda\langle \tilde{f}, (\Sigma+\lambda I)^{-1}\tilde{f}\rangle_\gamma \qquad \text{(by Lemma C.3)} \\
&= 4\lambda \inf_{f\in\mathcal{H}}\left\{\|f-\tilde{f}\|_\gamma^2+\|f\|_{\mathcal{H}}^2\right\} \\
&= 4\lambda \inf_{F\in\mathcal{H}_\theta}\left\{\|F-F^*\|_{\gamma_d}^2+\|F\|_{\mathcal{H}_\theta}^2\right\},
\end{aligned}$$

where we assume $N \ge Cd_{\max}(\tau,\lambda)\log(d_{\max}(\tau,\lambda)/\delta)$ in order to apply Lemma C.3.

We thus obtain

$$\mathbb{E}[\|F_{c,\theta}-F^*\|_{\gamma_d}^2] \lesssim \frac{\sigma^2\tau^2}{n'\lambda} + \inf_{F\in\mathcal{H}_\theta}\left\{\|F-F^*\|_{\gamma_d}^2+\lambda\|F\|_{\mathcal{H}_\theta}^2\right\} + \frac{1}{n'^2}\|F^*\|_\infty^2. \tag{89}$$

By limiting arguments, we may show that this holds for any F^* in the closure of $\mathcal{H}_\theta$. Now consider the true $F^*(x) = f_*(\langle\theta^*, x\rangle)$. When $\theta^* \ne \theta$, F^* does not belong to the closure of $\mathcal{H}_\theta$, but we may consider the projection $F^*_{\mathcal{H}_\theta}$ on this closure. Then, following the arguments of [7, Section 7.6.4], we obtain

$$\mathbb{E}[\|F_{c,\theta}-F^*\|_{\gamma_d}^2] \lesssim \frac{\sigma^2\tau^2}{n'\lambda} + \inf_{F\in\mathcal{H}_\theta}\left\{\|F-F^*_{\mathcal{H}_\theta}\|_{\gamma_d}^2+\lambda\|F\|_{\mathcal{H}_\theta}^2\right\} + \|F^*_{\mathcal{H}_\theta}-F^*\|_{\gamma_d}^2 + \frac{1}{n'^2}\|F^*\|_\infty^2 . \tag{90}$$

We may take $F^*_{\mathcal{H}_\theta}(x) = g(\langle\theta, x\rangle)$ for some g, since all functions in $\mathcal{H}_\theta$ and its closure take this form. Then, we may consider g of the form $g = \sum_j b_j h_j$, since such functions are dense in the closure of $\mathcal{H}_\theta$. Optimizing the approximation error $\|F^*_{\mathcal{H}_\theta} - F^*\|_{\gamma_d}$ over such g yields $b_j = \alpha_j m^j$, so that the approximation error becomes

$$\begin{aligned}
\|F^*_{\mathcal{H}_\theta}-F^*\|_{\gamma_d}^2 &= \|F^*_{\mathcal{H}_\theta}\|_{\gamma_d}^2 + \|F^*\|_{\gamma_d}^2 - 2\langle F^*_{\mathcal{H}_\theta}, F^*\rangle_{\gamma_d} \\
&= \sum_j \alpha_j^2 m^{2j} + \sum_j \alpha_j^2 - 2\sum_j \alpha_j^2 m^{2j} \\
&= \sum_j \alpha_j^2(1-m^{2j}) \\
&\le 2(1-|m|)\sum_j j\alpha_j^2 = (1-|m|)C_{f_*},
\end{aligned}$$

with $C_{f_*} = 2\|f_*'\|_\gamma^2$, by using the bound $1-m^{2j} \le (1-|m|)(1+|m|+\cdots+|m|^{2j-1}) \le 2j(1-|m|)$. We also have

$$\inf_{F\in\mathcal{H}_\theta}\left\{\|F-F^*_{\mathcal{H}_\theta}\|_{\gamma_d}^2+\lambda\|F\|_{\mathcal{H}_\theta}^2\right\} = A(g,\lambda) \lesssim \lambda^\beta\|g''\|_\gamma^2 \le \lambda^\beta\|f_*''\|_\gamma^2 . \tag{91}$$

The final bound becomes

$$\mathbb{E}[\|F_{c,\theta}-F^*\|_{\gamma_d}^2] \lesssim \frac{\sigma^2\tau^2}{n'\lambda} + \lambda^\beta\|f_*''\|_\gamma^2 + C_{f_*}(1-|m|) + \frac{1}{n'^2}\|f_*\|_\infty^2 . \tag{92}$$

Setting $\lambda = \left(\frac{\sigma^2\tau^2}{n'\|f_*''\|_\gamma^2}\right)^{\frac{1}{\beta+1}}$ yields

$$\mathbb{E}[\|F_{c,\theta}-F^*\|_{\gamma_d}^2] \lesssim \|f_*''\|_\gamma^{\frac{2}{\beta+1}}\left(\frac{\sigma^2\tau^2}{n'}\right)^{\frac{\beta}{\beta+1}} + C_{f_*}(1-|m|) + \frac{1}{n'^2}\|f_*\|_\infty^2. \tag{93}$$

The condition $n' \gtrsim R^2/\lambda$ is satisfied when

$$n' \gtrsim \left(\frac{\|f_*''\|_\gamma^2}{\sigma^2\tau^2}\right)^{1/\beta},$$

while the condition $\lambda \leq R^2$ is satisfied when

$$n' \gtrsim \frac{\sigma^2\tau^2}{\|f_*''\|_\gamma^2}.$$

The last term is negligible when

$$n' \gtrsim \frac{\|f_*\|_\infty^2}{(\sigma^2\tau^2)^{\beta/(\beta+1)}} .$$

Finally, the requirement on N scales as

$$N \geq \frac{C_\tau}{\lambda} \ln \frac{1}{\lambda\delta} \gtrsim C_\tau \left(\frac{n'\|f_*''\|_\gamma^2}{\sigma^2\tau^2} \right)^{\frac{1}{\beta+1}} \ln \left(\frac{n'^{1/(\beta+1)}}{\delta} \right) . \tag{94}$$

□

Proof of Lemma E.17. We establish this matrix concentration result using a dimension-independent matrix Bernstein inequality for subexponential and potentially unbounded random matrices, by adapting arguments of Minsker [62, Eq. (3.9)] and Tropp [80, Theorem 6.2]. The sub-exponential tail assumption is established next, in Lemma E.19.

Lemma E.18 (Dimension-independent matrix Bernstein bound)**.** *Let $X_1, \ldots, X_n$ be random i.i.d. self-adjoint operators with sub-exponential tails, in the sense that there exist self-adjoint operators A_i and $R > 0$ such that*

$$\mathbb{E}[X_i] = 0 \qquad \textit{and} \qquad \mathbb{E}[X_i^p] \preceq \frac{p!}{2} R^{p-2} A_i^2 \quad \textit{for } p = 2, 3, 4, \ldots$$

Defining the variance parameter

$$\sigma^2 := \left\| \sum_{i=1}^n A_i^2 \right\|,$$

we have the following for all $t \geq \sqrt{R^2 + 4\sigma^2}$:

$$\mathbb{P}\left[\left\| \sum_{i=1}^n X_i \right\|_{\mathrm{op}} \geq t \right] \leq \frac{7 \sum_i \mathrm{Tr}(A_i^2)}{\sigma^2} \exp\left(\frac{-t^2/2}{\sigma^2 + Rt} \right) . \tag{95}$$

Next we show that the sub-exponential bound needed for Lemma E.18 holds under our setting.

Lemma E.19 (Sub-exponential tail for covariance concentration)**.** *For $X \sim N(0, I)$, let*

$$M = Q_\lambda^{-1/2} (\Phi_\theta(X)\Phi_\theta(X)^\top - Q) Q_\lambda^{-1/2},$$

and define $B_\lambda = Q^{1/2} Q_\lambda^{-1/2}$. Then $\mathbb{E}[M] = 0$ and $\mathbb{E}[M^p] \preceq (R)^p\, p! B_\lambda B_\lambda^\top$ for $p \geq 2$ and some universal constant R.

Let us finally establish (86). By defining $X_i = Q_\lambda^{-1/2}(\Phi_\theta(x_i)\Phi_\theta(x_i)^T - Q)Q_\lambda^{-1/2}$, Lemma E.19 guarantees subexponential tails, satisfying $\mathbb{E}[X_i^p] \preceq R^p p! B_\lambda B_\lambda^\top$ for $p \geq 2$ and a universal constant R, and where $B_\lambda = Q^{1/2} Q_\lambda^{-1/2}$. Therefore, defining $A_i^2 := 2R^2 B_\lambda B_\lambda^\top$, we have

$$\mathbb{E}[X_i^p] \preceq \frac{p!}{2} R^{p-2} A_i^2 .$$

We can now apply Lemma E.18. In that case, $\sigma^2 = 2R^2 n$, and by choosing $t = n/2$ in (95), we obtain

$$\begin{aligned} \mathbb{P}\left[\left\| (Q + \lambda I)^{-1/2} (Q - \hat{Q}) (Q + \lambda I)^{-1/2} \right\|_{\mathrm{op}} \geq \frac{1}{2} \right] &\leq \frac{7 \sum_i \mathrm{Tr}(A_i^2)}{2R^2 n} \exp\left(-\frac{n^2/4}{n(2R^2 + R/2)} \right) \\ &\leq 7 \mathrm{Tr}(Q^{1/2} Q_\lambda^{-1} Q^{1/2}) \exp\left(-\frac{n}{8R^2 + 2R} \right) , \end{aligned}$$

proving (86). □

Proof of Lemma E.18. Let $S_n := \sum_{i=1}^n X_i$ and $\psi(\theta) := e^\theta - \theta - 1$. Following the argument of Theorem 3.1 of [62], we apply Markov's inequality and the monotonicity of ψ to upper-bound the probability that the $\| \sum_i X_i \|_{\mathrm{op}}$ is large for any fixed $\theta > 0$ and $t > 0$.

$$\begin{aligned}\mathbb{P}\left[\|S_n\|_{\mathrm{op}} \geq t\right] &= \mathbb{P}\left[\psi\left(\theta \|S_n\|_{\mathrm{op}}\right) \geq \psi(\theta t)\right] = \mathbb{P}\left[\|\psi\left(\theta S_n\right)\|_{\mathrm{op}} \geq \psi(\theta t)\right] \\ &\leq \mathbb{P}\left[\mathrm{Tr}(\psi\left(\theta S_n\right)) \geq \psi(\theta t)\right] \leq \frac{\mathbb{E}\left[\mathrm{Tr}(\psi(\theta S_n))\right]}{\psi(\theta t)}\,.\end{aligned}$$

We continue to adapt the argument of [62] in order to bound the numerator, taking advantage of the fact that $\mathbb{E}\, X_i = 0$ (and hence, $\mathbb{E}\, S_i = 0$). We additionally apply Jensen's inequality and Lieb's concavity theorem (Fact 5 of [62]).

$$\begin{aligned}\mathbb{E}\left[\mathrm{Tr}(\psi(\theta S_n))\right] &= \mathbb{E}\left[\mathrm{Tr}\left(\exp(\theta S_n) - S_n - I\right)\right] = \mathbb{E}\left[\mathrm{Tr}\left(\exp(\theta S_{n-1} + \log(\exp(X_n))) - I\right)\right] \\ &\leq \mathop{\mathbb{E}}_{X_1,\ldots,X_{n-1}}\left[\mathrm{Tr}\left(\exp\left(\theta S_{n-1} + \log\left(\mathop{\mathbb{E}}_{X_n}\left[\exp(\theta X_n)\right]\right)\right) - I\right)\right] \\ &\leq \mathrm{Tr}\left(\exp\left(\sum_{i=1}^n \log\left(\mathbb{E}\left[\exp(\theta X_i)\right]\right)\right) - I\right)\end{aligned}$$

According to Lemma 6.8 of [80], by scaling X_i appropriately and taking $\theta \in (0, \frac{1}{R})$, the assumptions in the theorem statement guarantee that

$$\mathbb{E}\left[\exp(\theta X_i)\right] \preccurlyeq \exp\left(\frac{(\theta R)^2}{2(1-\theta R)} \cdot \frac{1}{R^2} A_i^2\right) = \exp\left(\frac{\theta^2}{2(1-\theta R)} \cdot A_i^2\right),$$

for all $i \in [n]$. Let $B_n := \sum_{i=1}^n A_i^2$. As a result,

$$\begin{aligned}\mathbb{E}\left[\mathrm{Tr}(\psi(\theta S_n))\right] &\leq \mathrm{Tr}\left(\exp\left(\sum_{i=1}^n \frac{\theta^2}{2(1-\theta R)} \cdot A_i^2\right) - I\right) = \mathrm{Tr}\left(\sum_{k=1}^\infty \frac{1}{k!}\left(\frac{\theta^2}{2(1-\theta R)} B_n\right)^k\right) \\ &= \mathrm{Tr}\left(\frac{\theta^2}{2(1-\theta R)} B^{1/2} \sum_{k=1}^\infty \frac{1}{k!}\left(\frac{\theta^2}{2(1-\theta R)} B_n\right)^{k-1} B^{1/2}\right) \\ &\leq \mathrm{Tr}\left(\frac{\theta^2}{2(1-\theta R)} B^{1/2} \sum_{k=1}^\infty \frac{1}{k!}\left(\frac{\theta^2}{2(1-\theta R)} \|B_n\|_{\mathrm{op}}\right)^{k-1} B^{1/2}\right) \\ &= \frac{\theta^2}{2(1-\theta R)} \mathrm{Tr}\left(B\right) \sum_{k=1}^\infty \frac{1}{k!}\left(\frac{\theta^2}{2(1-\theta R)} \sigma^2\right)^{k-1} \\ &= \frac{\theta^2}{2(1-\theta R)} \mathrm{Tr}(B) \frac{\exp\left(\frac{\theta^2}{2(1-\theta R)}\sigma^2\right) - 1}{\frac{\theta^2}{2(1-\theta R)}\sigma^2} \leq \frac{\mathrm{Tr}(B)}{\sigma^2} \exp\left(\frac{\theta^2}{2(1-\theta R)}\sigma^2\right).\end{aligned}$$

We conclude by putting the terms together to simplify the expression (continuing to borrow from [62]) while letting $\theta := \frac{t}{\sigma^2 + Rt}$ and requiring that t be sufficiently large:

$$\begin{aligned}\mathbb{P}\left[\|S_n\|_{\mathrm{op}} \geq t\right] &\leq \frac{\mathrm{Tr}(B)}{\sigma^2} \exp\left(\frac{\theta^2}{2(1-\theta R)}\sigma^2\right) \cdot \frac{1}{\psi(\theta t)} \\ &\leq \frac{\mathrm{Tr}(B)}{\sigma^2} \exp\left(\frac{\theta^2}{2(1-\theta R)}\sigma^2 - \theta t\right) \cdot \frac{\exp(\theta t)}{\psi(\theta t)} \\ &\leq \frac{\mathrm{Tr}(B)}{\sigma^2} \exp\left(\theta t\left(\frac{\theta \sigma^2}{2t(1-\theta R)} - 1\right)\right)\left(1 + \frac{6}{(\theta t)^2}\right) \\ &= \frac{\mathrm{Tr}(B)}{\sigma^2} \exp\left(\frac{t^2}{\sigma^2 + Rt}\left(\frac{\sigma^2}{2(\sigma^2+Rt)} \cdot \frac{\sigma^2 + Rt}{\sigma^2} - 1\right)\right)\left(1 + \frac{6(\sigma^2+Rt)^2}{t^4}\right) \\ &= \frac{7\mathrm{Tr}(B)}{\sigma^2} \exp\left(-\frac{t^2/2}{\sigma^2 + Rt}\right).\end{aligned}$$

□

Proof of Lemma E.19. $\mathbb{E}[M] = 0$ is clear. For $p \geq 2$, we bound the moments of M by considering the subgaussianity of inner products.

We define $\Psi := Q^{-1/2}\Phi_\theta(X)$ and note that it is isotropic and subgaussian, i.e. $v^\top(\Psi - I_N)v$ is C_1-subgaussian for universal constant C_1 and any $v \in \mathbb{S}^{N-1}$. We also define $B_\lambda := Q^{1/2}Q_\lambda^{-1/2}$, so that $M = B_\lambda^\top(\Psi\Psi^\top - I_N)B_\lambda$.

Observe first that $v^\top(\Psi\Psi^\top - I_N)v$ is C_2-subexponential, i.e. $\mathbb{E}[\exp(C_2 v^\top(\Psi\Psi^\top - I_N)v)] \leq 2$, for any fixed $v \in \mathbb{S}^{N-1}$ because Ψ is subgaussian and $I_N = \mathbb{E}[\Psi\Psi^\top]$. As a result, $\mathbb{E}[\exp(C_2(\Psi\Psi^\top - I_N))] \preceq 2I_N$.

Let us now write $M^p = B_\lambda^\top \tilde{M}_p B_\lambda$. We claim that $\|\mathbb{E}\,\tilde{M}_p\|_{\mathrm{op}} \leq 4p!/C_2^p$ for any $p \geq 2$. Indeed, observe that $\tilde{M}_p$ is a product of matrices of the form $T_1 = (\Psi\Psi^\top - I)$ and $T_2 = B_\lambda B_\lambda^\top$. Since $\|T_2\|_{\mathrm{op}} \leq 1$, we have

$$\|\mathbb{E}[\tilde{M}_p]\|_{\mathrm{op}} \leq \|\mathbb{E}[(\Psi\Psi^\top - I)^p]\|_{\mathrm{op}}\,.$$

Now, because any positive semi-definite matrix A satisfies $A^p \preceq p!(e^A + e^{-A})$, we bound the moment $\mathbb{E}[v^\top(\Psi\Psi^\top - I)^p v]$ following the simple argument made in [83, Proposition 2.71].

$$\begin{aligned}\mathbb{E}[v^\top(\Psi\Psi^\top - I_N)^p v] &\leq \frac{p!}{C_2^p} \cdot v^\top(\mathbb{E}[\exp(C_2(\Psi\Psi^\top - I_N))] + \mathbb{E}[\exp(-C_2(\Psi\Psi^\top - I_N))])v \\ &\leq \frac{4p!}{C_2^p},\end{aligned}$$

which shows that $\|\mathbb{E}[\tilde{M}_p]\|_{\mathrm{op}} \leq \frac{4p!}{C_2^p}$. Therefore, we have

$$\begin{aligned}\mathbb{E}[v^\top M^p v] &= \mathbb{E}[(B_\lambda v)^\top \tilde{M}_p (B_\lambda v)] \\ &\leq \frac{4p!}{C_2^p}\|B_\lambda v\|^2\,,\end{aligned}$$

which shows that $\mathbb{E}[M^p] \preceq C_p p! B_\lambda B_\lambda^\top$ for $p \geq 2$. The conclusion is immediate for a proper choice of constant R. □

E.7 Proof of Corollary 6.4

Corollary 6.4 (Excess risk of Algorithm 2)**.** *Let $\delta \in (0, 1/4)$. As in Theorem 6.1, let $\mu_s = \langle h_s, \Sigma h_s\rangle > 0$, and let f_* satisfy Assumption 5.2. Let $\lambda = \Theta(1)$, and assume the following on the sample sizes and number of random features for the first phase (n, N, N_0) and fine-tuning phase (n', N'):*

$$N = N_0 = \Theta\left(\frac{1}{\lambda}\log\frac{1}{\lambda\delta}\right), \quad n = \widetilde{\Omega}\left(\max\{d^s, d^{(s+3)/2}\}\right), \quad N' = \widetilde{\Omega}\left(n'^{\frac{1}{\beta+1}}\right).$$

and let ρ be as in Theorem 6.1. With probability at least $1/2 - 2\delta$ over the initial n samples, initialization, random features, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}'}[\|\hat{F} - F_{\theta^*}\|_{\gamma_d}^2] \leq \widetilde{O}\left(\max\left\{\frac{d}{n}, \frac{d^4}{n^2}\right\} + \left(\frac{1}{n'}\right)^{\frac{\beta}{\beta+1}}\right), \tag{15}$$

where the constants in $\widetilde{O}$ do not depend on d other than through logarithmic factors.

Proof. The result is immediate by applying Proposition 6.3 and using the following bound from Lemma E.2:

$$1 - |m| \leq \widetilde{O}\left(\lambda^{-4}\max\left\{\frac{d+N}{n}, \frac{d^4}{n^2}\right\}\right),$$

where λ is a constant as given in the statement. Note that with a constant λ as in the statement, the choice $N_0 = N \sim \lambda^{-1}$ for the first phase is sufficient for satisfying the assumptions of Theorem 6.1. □

E.8 Omitted proofs from Section E

Lemma E.7. *Let $f : \mathbb{R} \to \mathbb{R}$ be function such that its derivatives $f^{(1)}, \dots, f^{(4)}$ are all in $L^2(\gamma)$. Let $f(z) = \sum_{j=0}^{\infty} \alpha_j h_j(z)$ be the Hermite expansion of f. Then,*

$$\sum_{j=1}^{\infty} j^2(j-1)^2\alpha_j^2 = \|f^{(4)}\|_\gamma^2 + 4\|f^{(3)}\|_\gamma^2 + 2\|f^{(2)}\|_\gamma^2 \,.$$

Proof. The statement follows from straightforward, albeit tedious, algebraic manipulation.

$$\begin{aligned}
\sum_{j=1}^{\infty} j\alpha_j^2 &= \|f^{(1)}\|_\gamma^2 \\
\sum_{j=1}^{\infty} j^2\alpha_j^2 &= \sum_{j=1}^{\infty} j(j-1)\alpha_j^2 + \sum_{j=1}^{\infty} j\alpha_j^2 = \|f^{(2)}\|_\gamma^2 + \|f^{(1)}\|_\gamma^2 \\
\sum_{j=1}^{\infty} j^3\alpha_j^2 &= \sum_{j=1}^{2} j^3\alpha_j^2 + \sum_{j=3}^{\infty} j(j-1)(j-2)\alpha_j^2 + 3\sum_{j=3}^{\infty} j^2\alpha_j^2 - 2\sum_{j=3}^{\infty} j\alpha_j^2 \\
&= \sum_{j=1}^{2} j^3\alpha_j^2 + \sum_{j=3}^{\infty} j(j-1)(j-2)\alpha_j^2 + 3\left(\sum_{j=1}^{\infty} j^2\alpha_j^2 - \sum_{j=1}^{2} j^2\alpha_j^2\right) - 2\left(\sum_{j=3}^{\infty} j\alpha_j^2 - \sum_{j=1}^{2} j\alpha_j^2\right) \\
&= \|f^{(3)}\|_\gamma^2 + 3\|f^{(2)}\|_\gamma^2 + \|f^{(1)}\|_\gamma^2 + \sum_{j=1}^{2}(j^3 - 3j^2 + 2j)\alpha_j^2 \\
&= \|f^{(3)}\|_\gamma^2 + 3\|f^{(2)}\|_\gamma^2 + \|f^{(1)}\|_\gamma^2 \\
\sum_{j=1}^{\infty} j^4\alpha_j^2 &= \sum_{j=1}^{3} j^4\alpha_j^2 + \sum_{j=4}^{\infty} j(j-1)(j-2)(j-3)\alpha_j^2 + 6\sum_{j=4}^{\infty} j^3\alpha_j^2 - 11\sum_{j=4}^{\infty} j^2\alpha_j^2 + 6\sum_{j=4}^{\infty} j\alpha_j^2 \\
&= \sum_{j=1}^{3} j^4\alpha_j^2 + \|f^{(4)}\|_\gamma^2 \\
&\quad + 6\left(\sum_{j=1}^{\infty} j^3\alpha_j^2 - \sum_{j=1}^{3} j^3\alpha_j^2\right) - 11\left(\sum_{j=1}^{\infty} j^2\alpha_j^2 - \sum_{j=1}^{3} j^2\alpha^2\right) + 6\left(\sum_{j=4}^{\infty} j\alpha_j^2 - \sum_{j=1}^{3} j\alpha_j^2\right) \\
&= \|f^{(4)}\|_\gamma^2 + 6(\|f^{(3)}\|_\gamma^2 + 3\|f^{(2)}\|_\gamma^2 + \|f^{(1)}\|_\gamma^2) - 11(\|f^{(2)}\|_\gamma^2 + \|f^{(1)}\|_\gamma^2) + 6\|f^{(1)}\|_\gamma^2 \\
&\quad + \sum_{j=1}^{3}(j^4 - 6j^3 + 11j^2 - 6j)\alpha_j^2 \\
&= \|f^{(4)}\|_\gamma^2 + 6\|f^{(3)}\|_\gamma^2 + 7\|f^{(2)}\|_\gamma^2 + \|f^{(1)}\|_\gamma^2 \,.
\end{aligned}$$

Using the above expressions for series of the form $\sum_{j=1}^{\infty} j^p\alpha_j^2$ for $p = 1, 2, 3, 4$, we conclude

$$\begin{aligned}
\sum_{j=1}^{\infty} j^2(j-1)^2\alpha_j^2 &= \sum_{j=1}^{\infty}(j^4 - 2j^3 + j^2)\alpha_j^2 \\
&= \|f^{(4)}\|_\gamma^2 + 4\|f^{(3)}\|_\gamma^2 + 2\|f^{(2)}\|_\gamma^2 \,.
\end{aligned}$$

□

F Gradient Flow on Non-smooth Landscapes

As mentioned in Section 2, for our purposes we only require 1) the *existence* of a curve $z : [a, b] \to \mathbb{R}^p$ satisfying the subgradient dynamics (what we have conveniently referred to as "gradient flow"

in earlier Sections) and 2) the *descent property*, which requires that the (empirical) loss L be *non-increasing* along any such curve. We first introduce basic terminology and concepts used in non-smooth optimization.

For non-smooth objective functions defined on Euclidean domains, a subdifferential *set* $\partial L(\theta)$ is used in place of the gradient $\nabla L(\theta)$. We restrict our attention to locally Lipschitz objectives which enjoy the property that they are differentiable a.e. [16, Theorem 9.1.2]. Formally,

Definition F.1 (Clarke Subdifferential)**.** *For any locally Lipschitz function $L : \Omega \to \mathbb{R}$ with an open domain $\Omega \subseteq \mathbb{R}^p$, the* Clarke subdifferential *of L at $\theta \in \Omega$ is defined by*

$$\partial L(\theta) = \operatorname{conv}\left\{ \lim_{i\to\infty} \nabla L(\theta_i) \;\middle|\; x_i \in \Omega,\ \nabla L(\theta_i) \text{ exists},\ \lim_{i\to\infty} \theta_i = \theta \right\} .$$

We denote by $\bar{\partial} L(\theta)$ the unique min-norm element of $\partial L(\theta)$.

A curve satisfying the subgradient dynamics of a locally Lipschitz objective function L is any absolutely continuous function $z : [a, b] \to \Omega$ which satisfies the following differential *inclusion* almost everywhere.

$$\dot{z}(t) \in -\partial L(z(t)) . \tag{96}$$

Closely related to curves satisfying the subgradient dynamics are *near-steepest descent* curves, which can be defined in any locally convex metric space (see [3, 32]). For any locally convex metric space (Ω, d) and any lower semicontinuous objective function $L : \Omega \to \mathbb{R}$ satisfying very weak continuity conditions, the existence of a near-steepest curve for L emanating from any starting point $z_0 \in \Omega$ is guaranteed [32, Theorem 3.4] and along the curve the objective is non-increasing. Furthermore, if L admits the *chain rule* (Definition F.2), then these two notions of curves coincide; near-steepest descent curves satisfy the subgradient dynamics a.e. and vice versa [32, Proposition 4.10]. Thus, the chain rule guarantees the *descent property* for any curve satisfying the subgradient dynamics [54, 29, 49].

For our purposes, it suffices to show that the empirical squared loss on any ReLU network satisfies the chain rule. Previous work by [29, 49] show that the chain rule holds for the class of functions *definable on some o-minimal structure* [81]. We simply write "L is definable" in place of "L is definable in some o-minimal structure". Notably, empirical squared loss functionals on ReLU networks, which can be viewed as real-valued functions w.r.t. the network parameters, are definable. We refer to [49, Appendix B] for further technical definitions and detailed proofs, but reproduce the formal statements here for convenience (See also [29, Theorem 5.8]).

Definition F.2 (Chain rule [32, Definition 4.9])**.** *Consider a lower semicontinuous function $L : \mathbb{R}^p \to \mathbb{R}$. We say that L* admits a chain rule *if for every absolutely continuous function $z : [a, b] \to \mathbb{R}^p$ for which $L \circ z$ is non-increasing and L is subdifferentiable along z, the following equation holds for a.e. $t \in (a, b)$*

$$(L \circ z)'(t) = \langle z^*(t), \dot{z}(t) \rangle \text{ for all } z^*(t) \in \partial L(z(t)) .$$

Lemma F.3 ([49, Lemma B.2])**.** *Any empirical squared loss functionals of any ReLU network (as a function w.r.t. the network parameters θ) is definable.*

Lemma F.4 (Chain rule adapted from [49, Lemma B.9])**.** *Given locally Lipschitz definable $L : \Omega \to \mathbb{R}$ with an open domain $\Omega \subseteq \mathbb{R}^p$, for any absolutely continuous function $z : [a, b] \to \Omega$, it holds for a.e. $t \in [a, b]$ that*

$$(L \circ z)'(t) = \langle z^*(t), \dot{z}(t) \rangle , \text{ for all } z^*(t) \in \partial L(z(t)) .$$

Moreover, for the gradient inclusion

$$\dot{z}(t) \in -\partial L(z(t)) ,$$

it holds for a.e. $t \geq 0$ that $\dot{z}(t) = -\bar{\partial} L(z(t))$ and $dL(z(t))/dt = -\|\bar{\partial} L(z(t))\|_2^2$ and therefore

$$L(z(a)) - L(z(b)) = \int_a^b \|\bar{\partial} L(z(\tau))\|_2^2 d\tau .$$

Remark F.5 (Riemannian gradients)**.** *We also need to show existence and the descent property for* Riemannian *gradient flows on the unit sphere, in which the subdifferentials in the differential inclusion* (96) *are projected onto the tangent space of $z(t)$. This is because we take spherical gradients for the (shared) first layer weights. However, the desired results follow from the same theorems since the existence of a near-steepest descent curve only requires the objective function to be lower semi-continuous and satisfy some very weak continuity conditions. We can enforce any near-steepest descent curve to be contained in $\mathcal{S}^{p-1}$ by modifying the objective to $\tilde{L}(z) = L \cdot \delta_{\mathcal{S}^{p-1}}(z)$, where $\delta_{\mathcal{S}^{p-1}}(z)$ is 1 for $z \in \mathcal{S}^{p-1}$ and ∞ otherwise. By Lemma F.4 and Claim F.6 (see below), $\tilde{L}$ satisfies the chain rule for all curves $z : [a, b] \to \Omega$ contained entirely in $\mathcal{S}^{p-1}$. More precisely, the following observation implies that the chain rule holds for $\tilde{L}$ and any curve $z : [a, b] \to \mathbb{R}^p$ contained entirely in $\mathcal{S}^{p-1}$:*

$$\langle (I - z(t)z(t)^\top)z^*(t), \dot{z}(t)\rangle = \langle z^*(t), \dot{z}(t)\rangle - \langle z(t), z^*(t)\rangle\langle z(t), \dot{z}(t)\rangle = \langle z^*(t), \dot{z}(t)\rangle \ .$$

Claim F.6. *Let $z : [a, b] \to \mathbb{R}^p$ be an absolutely continuous function satisfying $\|z(t)\|_2 = 1$ for all $t \in [a, b]$. Then, the derivative $\dot{z}(t)$ exists almost all $t \in [a, b]$ and satisfies $\langle z(t), \dot{z}(t)\rangle = 0$.*

Proof. Since $[a, b]$ is compact and z is absolutely continuous, z is differentiable a.e. on $[a, b]$ by Rademacher's Theorem [16, Theorem 9.1.2]. Now consider the derivative of the *constant* function $\|z\|_2^2$. For any $t \in [a, b]$ such that $\dot{z}(t)$ exists, we have

$$\frac{d}{dt}\|z(t)\|_2^2 = 2\langle z(t), \dot{z}(t)\rangle = 0 \ .$$

□

G Smooth Activation Functions

We discuss the impact of replacing the ReLU activation by a smooth activation ϕ. This choice affects both approximation and optimization properties of the corresponding model. To illustrate this, we focus on Gaussian smoothing which we define using the Ornstein-Ulhenbeck semigroup.

Definition G.1. *For $\rho \in [0, 1]$, the Ornstein–Uhlenbeck noise operator U_ρ is defined by*

$$U_\rho f(t) = \int f(\rho t + \sqrt{1 - \rho^2}u)d\gamma(u) \ .$$

Assumption G.2 (Smoothed ReLU)**.** *Given $\rho \in [0, 1]$ and $\phi(t) = \max(0, t)$, also known as the ReLU activation, we refer to $\phi_\rho = U_\rho\phi$ as the ρ-smoothed ReLU.*

The resulting activation is akin to the so-called Exponential Linear Unit (ELU) [25]. As will be shown next, we leverage hypercontractivity properties of the Gaussian measure defining ϕ_ρ. From [44], our smoothing operator may be replaced by a more general one provided it satisfies a Log-Sobolev inequality, but such extensions are out of the present scope.

Approximation properties. Let $\rho \in [0, 1]$ and let $\mathcal{H}_\rho$ be the RKHS associated with the kernel

$$\kappa_\rho(x, x') = \mathop{\mathbb{E}}_{b\sim\gamma_{\tilde{\tau}},\varepsilon}[\phi_\rho(\varepsilon x - b)\phi_\rho(\varepsilon x' - b)] \ .$$

where $\tilde{\tau} = \rho\tau$, and for any $f \in L^2(\gamma)$, let

$$A(f, \lambda, \rho) := \inf_{h\in\mathcal{H}_\rho} \|f - h\|_\gamma^2 + \lambda\|h\|_{\mathcal{H}_\rho}^2 \ . \tag{97}$$

Recall the function space $\mathcal{F} = \{g \in H^2(\gamma) \mid g'' \in L^4(\gamma)\}$ (see Assumption 4.3). We define an alternate λ-regularized approximation error of f with respect to the image of $\mathcal{F}$ under the operator U_ρ by

$$B(f, \lambda, \rho) := \inf_{g\in\mathcal{F}} \|f - U_\rho g\|_\gamma^2 + \lambda(\|g\|_\gamma^2 + \|g''\|_4^2) \ . \tag{98}$$

The following proposition relates the approximation error achievable by $\mathcal{H}_\rho$ to that of $\mathcal{H}$.

Proposition G.3 (Approximation error in $\mathcal{H}_\rho$)**.** *Let $\tau > 1$ and $\beta = \frac{1-1/\tau^2}{3+1/\tau^2}$. Then, there exists a universal constant $C_0 > 0$ such that for any $\rho \in [0,1]$ and any $f \in L^2(\gamma)$,*

$$A(U_\rho f, \lambda, \rho) \leq A(f, \lambda) \ , \ \textit{and} \ \ A(f, \lambda, \rho) \leq C_0 \tau^{1+\beta} B(f, \lambda^\beta, \rho) \ .$$

Proof. We first consider target functions f which satisfy the source condition $f = U_\rho f_0$, where $f_0 \in \mathcal{F}$. Consider $h^* = \arg\min_{h \in \mathcal{H}} \|f_0 - h\|_\gamma^2 + \lambda \|h\|_{\mathcal{H}}^2$. We verify from the definition that $\|f_0 - h^*\|^2 \leq A(f_0, \lambda)$ and $\|h^*\|_{\mathcal{H}}^2 \leq \lambda^{-1} A(f_0, \lambda)$. Now consider $h_\rho = U_\rho h^*$. Let T_u be the translation operator $T_u f(t) = f(t-u)$. We verify that

$$U_\rho T_u f = \int f(\rho t + \sqrt{1-\rho^2} z - u) d\gamma(z) = \int f(\rho(t - (u/\rho)) + \sqrt{1-\rho^2} z) d\gamma(z) = T_{(u/\rho)} U_\rho f \ .$$

so $U_\rho T_u = T_{(u/\rho)} U_\rho$. From the RKHS representation of h^*.

$$h^*(t) = \int \phi(t-u) c(u) \gamma_\tau(u) du = \int T_u \phi(t) c(u) \gamma_\tau(u) du$$

with $\|c\|_\gamma^2 = \|h^*\|_{\mathcal{H}}^2$, we verify that

$$\begin{aligned} h_\rho(t) = U_\rho h^*(t) &= \int U_\rho T_u \phi(t) c(u) \gamma_\tau(u) du \\ &= \int T_{(u/\rho)} U_\rho \phi(t) c(u) \gamma_\tau(u) du = \int \phi_\rho(t - (u/\rho)) c(u) d\gamma_\tau(u) \\ &= \int \phi_\rho(t-u) c(\rho u) \gamma_{\tilde{\tau}}(u) du \ , \end{aligned}$$

which shows that $h_\rho \in \mathcal{H}_\rho$ since

$$\|h_\rho\|_{\mathcal{H}_\rho}^2 \leq \int c(\rho u)^2 \gamma_{\tilde{\tau}}(u) du = \int c(u)^2 \gamma_\tau(u) du = \|h^*\|_{\mathcal{H}}^2 < \infty \ .$$

Therefore, for any $\rho < 1$ and target f satisfying the source condition $f = U_\rho f_0$, we have

$$\begin{aligned} A(f, \lambda, \rho) &\leq \|f - U_\rho h^*\|_\gamma^2 + \lambda \|U_\rho h^*\|_{\mathcal{H}_\rho}^2 \\ &\leq \|U_\rho (f_0 - h^*)\|_\gamma^2 + \lambda \|h^*\|_{\mathcal{H}}^2 \\ &\leq A(f_0, \lambda) \ , \end{aligned}$$

where we used the fact that U_ρ is a contraction in $L^2(\gamma)$ for any $\rho \leq 1$ [67, Theorem 11.23].

Let us now consider a general $f \in L^2(\gamma)$.

$$\begin{aligned} A(f, \lambda, \rho) &= \inf_{h \in \mathcal{H}_\rho} \|f - h\|_\gamma^2 + \lambda \|h\|_{\mathcal{H}_\rho}^2 \\ &\leq 2 \inf_{g \in \mathcal{F}} \left(\inf_{h \in \mathcal{H}_\rho} \|f - U_\rho g\|_\gamma^2 + \|U_\rho g - h\|_\gamma^2 + \lambda \|h\|_{\mathcal{H}_\rho}^2 \right) \\ &\leq 2 \inf_{g \in \mathcal{F}} \left(\|f - U_\rho g\|_\gamma^2 + \left(\inf_{h \in \mathcal{H}_\rho} \|U_\rho g - h\|_\gamma^2 + \lambda \|h\|_{\mathcal{H}_\rho}^2 \right) \right) \\ &= 2 \inf_{g \in \mathcal{F}} \|f - U_\rho g\|_\gamma^2 + 2A(U_\rho g, \lambda, \rho) \\ &\leq 2 \inf_{g \in \mathcal{F}} \|f - U_\rho g\|_\gamma^2 + 2A(g, \lambda) \\ &\leq 2 \inf_{g \in \mathcal{F}} \|f - U_\rho g\|_\gamma^2 + C \lambda^\beta (\|g\|_\gamma^2 + \|g''\|_4^2) \\ &\leq \max\{2, C\} \tau^{1+\beta} \cdot B(f, \lambda^\beta, \rho) \ . \end{aligned}$$

where we used $A(g, \lambda) \leq C \tau^{1+\beta} \lambda^\beta (\|g\|_\gamma^2 + \|g''\|_4^2)$, where $C > 0$ is a universal constant satisfying Lemma 4.4, $\|g'\|_\gamma^2 \leq \|g\|_\gamma^2 + \|g''\|_\gamma^2$, and $\|\cdot\|_\gamma \leq \|\cdot\|_4$, which follows from Jensen's inequality. $\square$

Proposition G.3 shows that approximation properties can be transferred from $\mathcal{F}$ to $\mathcal{H}_\rho$ for target functions satisfying a certain smoothness property, which is encoded in the source condition $B(f, \lambda, \rho)$. The choice of U_ρ as the smoothing operator is motivated by its rich structure in $L_2(\gamma)$, in particular its (hyper-)contractivity. The source condition (98) can be explicitly controlled using the Hermite decomposition of f, though the $L^4(\gamma)$-norm penalty on the (weak) second derivative of the approximant $g \in \mathcal{F}$ imposes restrictions on the decay of its Hermite coefficients. We leave such analysis for future work.

Besides the RKHS approximation error, our results also require control of approximation error from using random features (Lemma C.3). We verify that the same argument (contained in Lemma C.4) can be directly applied to $\mathcal{H}_\rho$, leading to an analogous control in terms of degrees of freedom. That being said, one may be able to obtain better control of the degrees of freedom under smoothness, leading to smaller estimation error of the KRR estimator, which in general compensates for the worse approximation error via tuning the regularisation parameter λ [7, Chapter 7].

Optimization properties. Using a smooth activation function for the student network simplifies the analysis of the empirical optimization landscape since we can readily adapt the tools developed in [60]. Moreover, since the empirical loss becomes a smooth function with Lipschitz gradients, our gradient flow analysis can be discretized and thereby yield guarantees for gradient *descent*. We now verify that for any $\rho \in (0, 1)$, ϕ'_ρ is L-Lipschitz with $L \leq \sup_t |\phi''_\rho(t)|$ which we now compute.

Claim G.4 (Lipschitz constant of ϕ''_ρ)**.** *Let $\rho \in [0, 1)$ and let ϕ_ρ be the ρ-smoothed ReLU. Then,*

$$\sup_{t \in \mathbb{R}} |\phi''_\rho(t)| \leq \frac{4\rho^2}{\sqrt{1-\rho^2}} \ .$$

Proof.

$$\begin{aligned}
\phi_\rho(t) &= \int \phi(\rho t + \sqrt{1-\rho^2} u)\gamma(u) du \\
&= \frac{1}{\sqrt{1-\rho^2}} \int \phi(v)\gamma\left(\frac{v - \rho t}{\sqrt{1-\rho^2}}\right) dv \\
&= \frac{1}{\sqrt{1-\rho^2}} \int_0^\infty v \cdot \gamma\left(\frac{v - \rho t}{\sqrt{1-\rho^2}}\right) dv \ ,
\end{aligned}$$

using change of variables with $v = \rho t + \sqrt{1-\rho^2} u$. Hence,

$$\begin{aligned}
\phi''_\rho(t) &= \frac{\rho^2}{(1-\rho^2)^{3/2}} \int_0^\infty v \cdot \gamma''\left(\frac{v - \rho t}{\sqrt{1-\rho^2}}\right) dv \\
&= \frac{\rho^2}{1-\rho^2} \int_{-\frac{\rho t}{\sqrt{1-\rho^2}}}^\infty (\rho t + \sqrt{1-\rho^2} u)\gamma''(u) du \\
&= \frac{\rho^2}{1-\rho^2} \left((\rho t + \sqrt{1-\rho^2} u)\gamma'(u)|_{-\frac{\rho t}{\sqrt{1-\rho^2}}}^\infty - \sqrt{1-\rho^2} \int_{-\frac{\rho t}{\sqrt{1-\rho^2}}}^\infty \gamma'(u) du \right) \\
&= -\frac{\rho^2}{\sqrt{1-\rho^2}} \int_{-\frac{\rho t}{\sqrt{1-\rho^2}}}^\infty \gamma'(u) du \ ,
\end{aligned}$$

Thus,

$$
\begin{aligned}
|\phi_\rho''(t)| &\le \frac{\rho^2}{\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} |\gamma'(u)| du \\
&= \frac{\rho^2}{\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} |u|\gamma(u) du \\
&= \frac{2\rho^2}{\sqrt{1-\rho^2}} \left(\int_0^1 u\gamma(u) du + \int_{u\ge 1} u\gamma(u) du \right) \\
&\le \frac{2\rho^2}{\sqrt{1-\rho^2}} \left(1 - \gamma(1) + \int_{-\infty}^{\infty} u^2 \gamma(u) du \right) \\
&\le \frac{2\rho^2}{\sqrt{1-\rho^2}} \cdot (2 - \gamma(1)) \ .
\end{aligned}
$$

□

H RKHS Approximation Beyond $\mathcal{F}$

We further discuss the approximation capability of the RKHS $\mathcal{H}$. Recall that Lemma 4.4 states that functions in $\mathcal{F} = \{f \in H^2(\gamma) \mid f'' \in L^4(\gamma)\}$ can be approximated by functions in $\mathcal{H}$ at a polynomial-in-λ rate. We show that ReLU as the *target* function admits a polynomial-in-λ approximation rate (Proposition H.1). Note that $\mathrm{ReLU} \notin \mathcal{F}$ since ReLU is not even in $H^2(\gamma)$. Thus, Proposition H.1 demonstrates that containment in $\mathcal{F}$ (Assumption 5.2) is sufficient, but not necessary for polynomial approximation rates.

Proposition H.1 (RKHS approximation error for ReLU)**.** *Let $\tau > 1$ and let $\phi(t) = \max(0, t)$. Then, for any $\lambda \in (0, \lambda^*)$, where $\lambda^* < 1$ depends only on τ,*

$$A(\phi, \lambda) \le (2 + \tau^2) \cdot \lambda^{2/3} \ .$$

Proof. We directly upper bound $A(\phi, \lambda)$ by the one-parameter family of functions $\phi_\rho = U_\rho \phi$, where we recall that U_ρ is the Ornstein-Uhlenbeck operator. Define $\lambda^* \in (0, 1)$ by

$$\lambda^* = (1 - \sqrt{2/(2\tau^2 + 1)})^{3/2} \ .$$

We consider approximants ϕ_ρ such that $\rho > \sqrt{2/(2\tau^2+1)}$, which in turn satisfies $1 - \rho < (\lambda^*)^{2/3}$. We first show that for ρ sufficiently close to 1, ϕ_ρ approximates ϕ well in $L^2(\gamma)$. Then, we show that $\phi_\rho \in \mathcal{H}$ for $\rho > \sqrt{1/(2\tau^2+1)}$, and further show that $\|\phi_\rho\|_{\mathcal{H}}$ is roughly upper bounded by $1/\sqrt{1-\rho^2}$. From Corollary H.4, we know that the Hermite expansion of ϕ yields $\phi = \sum_j \alpha_j h_j$ with $|\alpha_j| \le j^{-5/4}$ for $j \ge 2$. Since Hermite polynomials are eigenfunctions of the operator U_ρ, we immediately have

$$
\begin{aligned}
\|\phi - \phi_\rho\|_\gamma^2 = \sum_{j=1}^{\infty} |\alpha_j|^2 (1-\rho^j)^2 &\le \sum_{j=1}^{\infty} |\alpha_j|^2 (1-\rho^j) \\
&= (1-\rho) \sum_{j=1}^{\infty} |\alpha_j|^2 (1 + \cdots + \rho^{j-1}) \\
&\le (1-\rho) \sum_{j=1}^{\infty} j |\alpha_j|^2 \\
&\le (1-\rho) \sum_{j=1}^{\infty} j^{-3/2} \\
&\le (1-\rho) \Big(1 + \int_1^{\infty} j^{-3/2} \Big) \\
&\le 2(1-\rho) \ .
\end{aligned}
$$

On the other hand, by definition and change-of-variables, we have

$$\begin{aligned}
\phi_\rho(t) &= \int \phi(\rho t + \sqrt{1-\rho^2}u)\gamma(u)du \\
&= \int \rho \cdot \phi\left(t + \frac{\sqrt{1-\rho^2}}{\rho}u\right)\gamma(u)du \\
&= \frac{\rho^2}{\sqrt{1-\rho^2}} \int \phi(t+b)\gamma\left(\frac{\rho}{\sqrt{1-\rho^2}} \cdot b\right)db \\
&= \int \phi(t+b)c(b)\gamma_\tau(b)db \ ,
\end{aligned}$$

where

$$c(b) = \frac{\rho^2\tau}{\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{\rho^2}{1-\rho^2} - \frac{1}{\tau^2}\right)b^2\right) \ .$$

We thus have

$$\begin{aligned}
\|\phi_\rho\|_{\mathcal{H}}^2 &\le \|c\|_{\gamma_\tau}^2 \\
&= \frac{\rho^4\tau}{\sqrt{2\pi}(1-\rho^2)} \int \exp\left(-\left(\frac{\rho^2}{1-\rho^2} - \frac{1}{2\tau^2}\right)b^2\right)db \\
&= \frac{\rho^4\tau}{\sqrt{2\pi}(1-\rho^2)} \int \exp\left(-\frac{(1+2\tau^2)\rho^2 - 1}{2\tau^2(1-\rho^2)}b^2\right)db \\
&= \frac{\rho^4\tau}{\sqrt{2\pi}(1-\rho^2)} \cdot \frac{\tau\sqrt{2\pi(1-\rho^2)}}{\sqrt{(1+2\tau^2)\rho^2 - 1}} \\
&= \frac{\rho^4\tau^2}{\sqrt{1-\rho^2}\sqrt{(1+2\tau^2)\rho^2 - 1}} \ ,
\end{aligned}$$

which implies that $\phi_\rho \in \mathcal{H}$ for any $\rho > 1/\sqrt{1+2\tau^2}$.

We balance the upper bounds of $\|\phi - \phi_\rho\|_\gamma^2$ and $\lambda\|\phi_\rho\|_{\mathcal{H}}^2$ to control $A(\phi, \lambda)$ in terms of λ. To this end, we set $\rho = 1 - \lambda^{2/3}$, where $\lambda < \lambda^*$. Then, we have

$$\|\phi - \phi_\rho\|_\gamma^2 \le 2(1-\rho) = 2\lambda^{2/3} \ ,$$

Moreover, using the fact that $1 - \rho^2 = (1-\rho)(1+\rho) \ge 1 - \rho$ and $(1+2\tau^2)\rho^2 > 2$, which follows from $\rho > 1 - (\lambda^*)^{2/3}$, we get

$$\begin{aligned}
\lambda\|\phi_\rho\|_{\mathcal{H}}^2 &\le \lambda \cdot \frac{(1-\lambda^{2/3})^4\tau^2}{\lambda^{1/3}\sqrt{(1+2\tau^2)\rho^2 - 1}} \\
&\le \lambda^{2/3} \cdot \tau^2 \ .
\end{aligned}$$

Hence, for any $\lambda \in (0, \lambda^*)$,

$$A(\phi, \lambda) \le \|\phi - \phi_\rho\|_\gamma^2 + \lambda\|\phi_\rho\|_{\mathcal{H}}^2 \le (2 + \tau^2) \cdot \lambda^{2/3} \ .$$

□

H.1 ReLU Hermite coefficients

Fact H.2. *Let $\{H_j(z)\}_{j\in\mathbb{N}}$ be the unnormalized (probabilist's) Hermite polynomials. Then,*

$$H_j(0) = \begin{cases} 0 & \text{if } j \text{ odd} \\ (-1)^{j/2}\frac{j!}{(j/2)!2^{j/2}} & \text{if } j \text{ even} \ . \end{cases}$$

Claim H.3 (ReLU Hermite coefficients [42, Claim 1])**.** *The Hermite coefficients of* $\mathrm{ReLU}(z)$ *are given by*

$$\alpha_j = \begin{cases} 1/\sqrt{2\pi} & \text{if } j = 0 \\ 1/2 & \text{if } j = 1 \\ \frac{1}{\sqrt{2\pi j!}}(H_j(0) + jH_{j-2}(0)) & \text{otherwise} \ . \end{cases}$$

Corollary H.4 (ReLU coefficient bounds for $j \geq 2$)**.**

$$\alpha_j = \frac{1}{\sqrt{2\pi j!}} \cdot (-1)^{(j-2)/2} \frac{(j-2)!j}{(j/2)!2^{j/2}} \ \text{ and } \ |\alpha_j| \leq \frac{1}{\sqrt{2\pi^{3/2}}} \cdot \frac{1}{j^{5/4}} \ .$$

Proof. Combining Fact H.2 and Claim H.3,

$$\begin{aligned} H_j(0) + jH_{j-2}(0) &= (-1)^{j/2} \left(\frac{j!}{(j/2)!2^{j/2}} - \frac{(j-2)!2j(j/2)}{(j/2)!2^{j/2}} \right) \\ &= (-1)^{j/2} \frac{(j-2)!}{(j/2)!2^{j/2}} \left(j(j-1) - j^2 \right) \\ &= (-1)^{(j-2)/2} \frac{(j-2)!j}{(j/2)!2^{j/2}} \\ \alpha_j &= \frac{1}{\sqrt{2\pi j!}} \cdot (-1)^{(j-2)/2} \frac{(j-2)!j}{(j/2)!2^{j/2}} \\ &= \frac{(-1)^{(j-2)/2}}{\sqrt{2\pi}} \cdot \frac{\sqrt{(j-2)!}\sqrt{j}}{\sqrt{j-1}(j/2)!2^{j/2}} \\ &= \frac{(-1)^{(j-2)/2}}{\sqrt{2\pi}} \cdot \frac{\sqrt{j}\sqrt{(j-2)!}}{\sqrt{j-1}j!!} \ , \end{aligned}$$

where we used the fact that $j!! = (j/2)!2^{j/2}$ for even j in the last line. It remains to evaluate the RHS. We use the following facts on double factorials.

$$\begin{aligned} \frac{j!!}{(j-1)!!} &= \frac{2^{j/2}(j/2)!}{j!/(2^{j/2}(j/2)!)} = 2^j \binom{j}{j/2}^{-1} \leq \sqrt{\pi(j+1)/2} \leq \sqrt{\pi j} \qquad (99) \\ (j-2)! &= (j-2)!!(j-3)!! = \frac{(j-1)!!j!!}{j(j-1)} \leq \frac{(j!!)^2}{j(j-1)\sqrt{\pi j}} \ , \end{aligned}$$

where in Eq. (99), we used the fact that $\binom{2k}{k} \geq 4^k/\sqrt{\pi(k+1/2)}$. Thus,

$$\begin{aligned} |\alpha_j| &\leq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{(j-1)(\pi j)^{1/4}} \\ &\leq \frac{1}{\sqrt{2\pi^{3/2}}} \cdot \frac{1}{j^{5/4}} \ . \end{aligned}$$

□