# Deep learning: a statistical viewpoint

Peter L. Bartlett* peter@berkeley.edu
Andrea Montanari† montanar@stanford.edu
Alexander Rakhlin‡ rakhlin@mit.edu

March 17, 2021

**Abstract**

The remarkable practical success of deep learning has revealed some major surprises from a theoretical perspective. In particular, simple gradient methods easily find near-optimal solutions to non-convex optimization problems, and despite giving a near-perfect fit to training data without any explicit effort to control model complexity, these methods exhibit excellent predictive accuracy. We conjecture that specific principles underlie these phenomena: that overparametrization allows gradient methods to find interpolating solutions, that these methods implicitly impose regularization, and that overparametrization leads to benign overfitting, that is, accurate predictions despite overfitting training data. In this article, we survey recent progress in statistical learning theory that provides examples illustrating these principles in simpler settings. We first review classical uniform convergence results and why they fall short of explaining aspects of the behavior of deep learning methods. We give examples of implicit regularization in simple settings, where gradient methods lead to minimal norm functions that perfectly fit the training data. Then we review prediction methods that exhibit benign overfitting, focusing on regression problems with quadratic loss. For these methods, we can decompose the prediction rule into a simple component that is useful for prediction and a spiky component that is useful for overfitting but, in a favorable setting, does not harm prediction accuracy. We focus specifically on the linear regime for neural networks, where the network can be approximated by a linear model. In this regime, we demonstrate the success of gradient flow, and we consider benign overfitting with two-layer networks, giving an exact asymptotic analysis that precisely demonstrates the impact of overparametrization. We conclude by highlighting the key challenges that arise in extending these insights to realistic deep learning settings.

## Contents



*Departments of Statistics and EECS, UC Berkeley
†Departments of EE and Statistics, Stanford University
‡Department of Brain & Cognitive Sciences and Statistics & Data Science Center, MIT

# 1 Introduction

The past decade has witnessed dramatic advances in machine learning that have led to major breakthroughs in computer vision, speech recognition, and robotics. These achievements are based on a powerful and diverse toolbox of techniques and algorithms that now bears the name 'deep learning'; see, for example, [GBC16]. Deep learning has evolved from the decades-old methodology of neural networks: circuits of parametrized nonlinear functions, trained by gradient-based methods. Practitioners have made major architectural and algorithmic innovations, and have exploited technological advances, such as increased computing power, distributed computing architectures, and the availability of large amounts of digitized data. The 2018 Turing Award celebrated these advances, a reflection of their enormous impact [LBH15].

Broadly interpreted, deep learning can be viewed as a family of highly nonlinear statistical models that are able to encode highly nontrivial representations of data. A prototypical example is a *feed-forward neural network with $L$ layers*, which is a parametrized family of functions $\boldsymbol{x} \mapsto f(\boldsymbol{x};\boldsymbol{\theta})$ defined on $\mathbb{R}^d$ by

$$f(\boldsymbol{x};\boldsymbol{\theta}) := \sigma_L(\boldsymbol{W}_L\sigma_{L-1}(\boldsymbol{W}_{L-1}\cdots\sigma_1(\boldsymbol{W}_1\boldsymbol{x})\cdots)), \tag{1}$$

where the parameters are $\boldsymbol{\theta} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L)$ with $\boldsymbol{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and $d_0 = d$, and $\sigma_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ are fixed nonlinearities, called *activation functions*. Given a training sample $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}^{d_L}$, the parameters $\boldsymbol{\theta}$ are typically chosen by a gradient method to minimize the *empirical risk*,

$$\widehat{L}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i; \boldsymbol{\theta}), y_i),$$

where $\ell$ is a suitable loss function. The aim is to ensure that this model generalizes well, in the sense that $f(\boldsymbol{x}; \boldsymbol{\theta})$ is an accurate prediction of $y$ on a subsequent $(\boldsymbol{x}, y)$ pair. It is important to emphasize that deep learning is a data-driven approach: these are rich but generic models, and the architecture, parametrization and nonlinearities are typically chosen without reference to a specific model for the process generating the data.

While deep learning has been hugely successful in the hands of practitioners, there are significant gaps in our understanding of what makes these methods successful. Indeed, deep learning reveals some major surprises from a theoretical perspective: deep learning methods can find near-optimal solutions to highly non-convex empirical risk minimization problems, solutions that give a near-perfect fit to noisy training data, but despite making no explicit effort to control model complexity, these methods lead to excellent prediction performance in practice.

To put these properties in perspective, it is helpful to recall the three competing goals that statistical prediction methods must balance: they require expressivity, to allow the richness of real data to be effectively modelled; they must control statistical complexity, to make the best use of limited training data; and they must be computationally efficient. The classical approach to managing this trade-off involves a rich, high-dimensional model, combined with some kind of regularization, which encourages simple models but allows more complexity if that is warranted by the data. In particular, complexity is controlled so that performance on the training data, that is, the empirical risk, is representative of performance on independent test data, specifically so that the function class is simple enough that sample averages $\widehat{L}(\boldsymbol{\theta})$ converge to expectations $L(\boldsymbol{\theta}) := \mathbb{E}\ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y)$ uniformly across the function class. And prediction methods are typically formulated as convex optimization problems—for example with a convex loss $\ell$ and parameters $\boldsymbol{\theta}$ that enter linearly—which can be solved efficiently.

The deep learning revolution built on two surprising empirical discoveries that are suggestive of radically different ways of managing these trade-offs. First, deep learning exploits rich and expressive models, with many parameters, and the problem of optimizing the fit to the training data appears to simplify dramatically when the function class is rich enough, that is, when it is sufficiently overparametrized. In this regime, simple, local optimization approaches, variants of stochastic gradient methods, are extraordinarily successful at finding near-optimal fits to training data, even though the nonlinear parametrization—see equation (1)—implies that the optimization problems that these simple methods solve are notoriously non-convex. *A posteriori,* the idea that overparametrization could lead to tractability might seem natural, but it would have seemed completely foolish from the point of view of classical learning theory: the resulting models are outside the realm of uniform convergence, and therefore should not be expected to generalize well.

The second surprising empirical discovery was that these models are indeed outside the realm of uniform convergence. They are enormously complex, with many parameters, they are trained with no explicit regularization to control their statistical complexity, and they typically exhibit a near-perfect fit to noisy training data, that is, empirical risk close to zero. Nonetheless this overfitting is benign, in that they produce excellent prediction performance in a number of settings. Benign overfitting appears to contradict accepted statistical wisdom, which insists on a trade-off between the complexity of a model and its fit to the data. Indeed, the rule of thumb that models fitting noisy data too well will not generalize is found in most classical texts on statistics and machine learning [FHT01, Was13]. This viewpoint has become so prevalent that the word 'overfitting' is often taken to mean both fitting data better than should be expected and also giving poor predictive accuracy as a consequence. In this paper, we use the literal meaning of the word 'overfitting'; deep learning practice has demonstrated that poor predictive accuracy is not an inevitable consequence.

This paper reviews some initial steps towards understanding these two surprising aspects of the success of deep learning. We have two working hypotheses:

**Tractability via overparametrization.** Classically, tractable statistical learning is achieved by restricting to linearly parametrized classes of functions and convex objectives. A fundamentally new principle appears to be at work in deep learning. Although the objective is highly non-convex, we conjecture that the hardness of the optimization problem depends on the relationship between the dimension of the parameter space (the number of optimization variables) and the sample size (which, when we aim for a near-perfect fit to training data, we can think of as the number of constraints), that is, tractability is achieved if and only if we choose a model that is sufficiently under-constrained or, equivalently, overparametrized.

**Generalization via implicit regularization.** Even if overparametrized models simplify the optimization task, classically we would have believed that good generalization properties would be restricted to either an underparametrized regime or a suitably regularized regime. Statistical wisdom suggests that a method that takes advantage of too many degrees of freedom by perfectly interpolating noisy training data will be poor at predicting new outcomes. In deep learning, training algorithms appear to induce a bias that breaks the equivalence among all the models that interpolate the observed data. Because these models interpolate noisy data, the classical statistical perspective would suggest that this bias cannot provide sufficient regularization to give good generalization, but in practice it does. We conjecture that deep learning models can be decomposed into a low-complexity component for which classical uniform convergence occurs and a high-complexity component that enables a perfect fit to training data, and if the model is suitably overparameterized, this perfect fit does not have a significant impact on prediction accuracy.

As we shall see, both of these hypotheses are supported by results in specific scenarios, but there are many intriguing open questions in extending these results to realistic deep learning settings.

It is worth noting that none of the results that we review here make a case for any optimization or generalization benefits of increasing depth in deep learning. Although it is not the focus here, another important aspect of deep learning concerns how deep neural networks can effectively and parsimoniously express natural functions that are well matched to the data that arise in practice. It seems likely that depth is crucial for these issues of expressivity.

## 1.1 Overview

Section 2 starts by reviewing some results from classical statistical learning theory that are relevant to the problem of prediction with deep neural networks. It describes an explicit probabilistic formulation of prediction problems. Consistent with the data-driven perspective of deep learning, this formulation assumes little more than that the $(\boldsymbol{x}, y)$ pairs are sampled independently from a fixed probability distribution. We explain the role played by uniform bounds on deviations between risk and empirical risk,

$$\sup_{f \in \mathcal{F}} \left| L(f) - \widehat{L}(f) \right|,$$

in the analysis of the generalization question for functions chosen from a class $\mathcal{F}$. We show how a partition of a rich function class $\mathcal{F}$ into a complexity hierarchy allows regularization methods that balance the statistical complexity and the empirical risk to enjoy the best bounds on generalization implied by the uniform convergence results. We consider consequences of these results for general pattern classification problems, for easier "large margin" classification problems and for regression problems, and we give some specific examples of risk bounds for feed-forward networks. Finally, we consider the implications of these results for benign overfitting: If an algorithm chooses an interpolating function to minimize some notion of complexity, what do the uniform convergence results imply about its performance? We see that there are very specific barriers to analysis of this kind in the overfitting regime; an analysis of benign overfitting must make stronger assumptions about the process that generates the data.

In Section 3, we review results on the implicit regularization that is imposed by the algorithmic approach ubiquitous in deep learning: gradient methods. We see examples of function classes and loss functions where

gradient methods, suitably initialized, return the empirical risk minimizers that minimize certain parameter norms. While all of these examples involve parameterizations of linear functions with convex losses, we shall see in Section 5 that this linear/convex viewpoint can be important for nonconvex optimization problems that arise in neural network settings.

Section 4 reviews analyses of benign overfitting. We consider extreme cases of overfitting, where the prediction rule gives a perfect interpolating fit to noisy data. In all the cases that we review where this gives good predictive accuracy, we can view the prediction rule as a linear combination of two components: $\widehat{f} = \widehat{f}_0 + \Delta$. The first, $\widehat{f}_0$, is a simple component that is useful for prediction, and the second, $\Delta$, is a spiky component that is useful for overfitting. Classical statistical theory explains the good predictive accuracy of the simple component. The other component is not useful for prediction, but equally it is not harmful for prediction. The first example we consider is the classical Nadaraya-Watson kernel smoothing method with somewhat strange, singular kernels, which lead to an interpolating solution that, for a suitable choice of the kernel bandwidth, enjoys minimax estimation rates. In this case, we can view $\widehat{f}_0$ as the prediction of a standard kernel smoothing method and $\Delta$ as a spiky component that is harmless for prediction but allows interpolation. The other examples we consider are for high-dimensional linear regression. Here, 'linear' means linearly parameterized, which of course allows for the richness of highly nonlinear features, for instance the infinite dimensional feature vectors that arise in reproducing kernel Hilbert spaces (RKHSs). Motivated by the results of Section 3, we study the behavior of the minimum norm interpolating linear function. We see that it can be decomposed into a prediction component and an overfitting component, with the split determined by the eigenvalues of the data covariance matrix. The prediction component corresponds to a high-variance subspace and the overfitting component to the orthogonal, low-variance subspace. For sub-Gaussian features, benign overfitting occurs if and only if the high-variance subspace is low-dimensional (that is, the prediction component is simple enough for the corresponding subspace of functions to exhibit uniform convergence) and the low-variance subspace has high effective dimension and suitably low energy. In that case, we see a *self-induced regularization*: the projection of the data on the low-variance subspace is well-conditioned, just as it would be if a certain level of statistical regularization were imposed, so that even though this subspace allows interpolation, it does not significantly deteriorate the predictive accuracy. (Notice that this self-induced regularization is a consequence of the decay of eigenvalues of the covariance matrix, and should not be confused with the implicit regularization, which is a consequence of the gradient optimization method and leads to the minimum norm interpolant.) Using direct arguments that avoid the sub-Gaussian assumption, we see similar behavior of the minimum norm interpolant in certain infinite-dimensional RKHSs, including an example of an RKHS with fixed input dimension where benign overfitting cannot occur and examples of RKHSs where it does occur for suitably increasing input dimension, again corresponding to decompositions into a simple subspace—in this case, a subspace of polynomials, with dimension low enough for uniform convergence—and a complex high-dimensional orthogonal subspace that allows benign overfitting.

In Section 5, we consider a specific regime where overparametrization allows a non-convex empirical risk minimization problem to be solved efficiently by gradient methods: a linear regime, in which a parameterized function can be accurately approximated by its linearization about an initial parameter vector. For a suitable parameterization and initialization, we see that a gradient method remains in the linear regime, enjoys linear convergence of the empirical risk, and leads to a solution whose predictions are well approximated by the linearization at the initialization. In the case of two-layer networks, suitably large overparametrization and initialization suffice. On the other hand, the mean-field limit for wide two-layer networks, a limit that corresponds to a smaller—and perhaps more realistic—initialization, exhibits an essentially different behavior, highlighting the need to extend our understanding beyond linear models.

Section 6 returns to benign overfitting, focusing on the linear regime for two specific families of two-layer networks: a random features model, with randomly initialized first-layer parameters that remain constant throughout training, and a neural tangent model, corresponding to the linearization about a random initialization. Again, we see decompositions into a simple subspace (of low-degree polynomials) that is useful for prediction and a complex orthogonal subspace that allows interpolation without significantly harming prediction accuracy.

Section 7 outlines future directions. Specifically, for the two working hypotheses of tractability via overparametrization and generalization via implicit regularization, this section summarizes the insights from the examples that we have reviewed—mechanisms for implicit regularization, the role of dimension, decompositions into prediction and overfitting components, data-adaptive choices of these decompositions, and the tractability benefits of overparameterization. It also speculates on how these might extend to realistic deep learning settings.

# 2 Generalization and uniform convergence

This section reviews uniform convergence results from statistical learning theory and their implications for prediction with rich families of functions, such as those computed by neural networks. In classical statistical analyses, it is common to posit a specific probabilistic model for the process generating the data and to estimate the parameters of that model; see, for example, [BD07]. In contrast, the approach in this section is motivated by viewing neural networks as defining rich, flexible families of functions that are useful for prediction in a broad range of settings. We make only weak assumptions about the process generating the data, for example, that it is sampled independently from an unknown distribution, and we aim for the best prediction accuracy.

## 2.1 Preliminaries

Consider a prediction problem in a probabilistic setting, where we aim to use data to find a function $f$ mapping from an input space $\mathcal{X}$ (for example, a representation of images) to an output space $\mathcal{Y}$ (for example, a finite set of labels for those images). We measure the quality of the predictions that $f : \mathcal{X} \to \mathcal{Y}$ makes on an $(\boldsymbol{x}, y)$ pair using the loss $\ell(f(\boldsymbol{x}), y)$, which represents the cost of predicting $f(\boldsymbol{x})$ when the actual outcome is $y$. For example, if $f(\boldsymbol{x})$ and $y$ are real-valued, we might consider the square loss, $\ell(f(\boldsymbol{x}), y) = (f(\boldsymbol{x}) - y)^2$. We assume that we have access to a training sample of input-output pairs $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, chosen independently from a probability distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$. These data are used to choose $\widehat{f} : \mathcal{X} \to \mathcal{Y}$, and we would like $\widehat{f}$ to give good predictions of the relationship between subsequent $(\boldsymbol{x}, y)$ pairs in the sense that the *risk of* $\widehat{f}$, denoted

$$L(\widehat{f}) := \mathbb{E}\ell(\widehat{f}(\boldsymbol{x}), y),$$

is small, where $(\boldsymbol{x}, y) \sim \mathbb{P}$ and $\mathbb{E}$ denotes expectation (and if $\widehat{f}$ is random, for instance because it is chosen based on random training data, we use $L(\widehat{f})$ to denote the conditional expectation given $\widehat{f}$). We are interested in ensuring that the excess risk of $\widehat{f}$,

$$L(\widehat{f}) - \inf_f L(f),$$

is close to zero, where the infimum is over all measurable functions. Notice that we assume only that $(\boldsymbol{x}, y)$ pairs are independent and identically distributed; in particular, we do not assume any functional relationship between $\boldsymbol{x}$ and $y$.

Suppose that we choose $\widehat{f}$ from a set of functions $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$. For instance, $\mathcal{F}$ might be the set of functions computed by a deep network with a particular architecture and with particular constraints on the parameters in the network. A natural approach to using the sample to choose $\widehat{f}$ is to minimize the *empirical risk* over the class $\mathcal{F}$. Define

$$\widehat{f}_{\text{erm}} \in \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{L}(f), \tag{2}$$

where the empirical risk,

$$\widehat{L}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i),$$

is the expectation of the loss under the empirical distribution defined by the sample. Often, we consider classes of functions $\boldsymbol{x} \mapsto f(\boldsymbol{x}; \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$, and we use $L(\boldsymbol{\theta})$ and $\widehat{L}(\boldsymbol{\theta})$ to denote $L(f(\cdot; \boldsymbol{\theta}))$ and $\widehat{L}(f(\cdot; \boldsymbol{\theta}))$, respectively.

We can split the excess risk of the empirical risk minimizer $\widehat{f}_{\text{erm}}$ into two components,

$$L(\widehat{f}_{\text{erm}}) - \inf_f L(f) = \left(L(\widehat{f}_{\text{erm}}) - \inf_{f\in\mathcal{F}} L(f)\right) + \left(\inf_{f\in\mathcal{F}} L(f) - \inf_f L(f)\right), \tag{3}$$

the second reflecting how well functions in the class $\mathcal{F}$ can approximate an optimal prediction rule and the first reflecting the statistical cost of estimating such a prediction rule from the finite sample. For a more complex function class $\mathcal{F}$, we should expect the approximation error to decrease and the estimation error to increase. We focus on the estimation error, and on controlling it using uniform laws of large numbers.

## 2.2 Uniform laws of large numbers

Without any essential loss of generality, suppose that a minimizer $f^*_{\mathcal{F}} \in \arg\min_{f\in\mathcal{F}} L(f)$ exists. Then we can split the estimation error of an empirical risk minimizer $\widehat{f}_{\text{erm}}$ defined in (2) into three components:

$$\begin{aligned}
&L(\widehat{f}_{\text{erm}}) - \inf_{f\in\mathcal{F}} L(f) \\
&= L(\widehat{f}_{\text{erm}}) - L(f^*_{\mathcal{F}}) \\
&= \left[L(\widehat{f}_{\text{erm}}) - \widehat{L}(\widehat{f}_{\text{erm}})\right] + \left[\widehat{L}(\widehat{f}_{\text{erm}}) - \widehat{L}(f^*_{\mathcal{F}})\right] + \left[\widehat{L}(f^*_{\mathcal{F}}) - L(f^*_{\mathcal{F}})\right].
\end{aligned} \tag{4}$$

The second term cannot be positive since $\widehat{f}_{\text{erm}}$ minimizes empirical risk. The third term converges to zero by the law of large numbers (and if the random variable $\ell(f^*_{\mathcal{F}}(\boldsymbol{x}), y)$ is sub-Gaussian, then with probability exponentially close to 1 this term is $O(n^{-1/2})$; see, for example, [BLM13, Chapter 2] and [Ver18] for the definition of sub-Gaussian and for a review of concentration inequalities of this kind). The first term is more interesting. Since $\widehat{f}_{\text{erm}}$ is chosen using the data, $\widehat{L}(\widehat{f}_{\text{erm}})$ is a biased estimate of $L(\widehat{f}_{\text{erm}})$, and so we cannot simply apply a law of large numbers. One approach is to use the crude upper bound

$$L(\widehat{f}_{\text{erm}}) - \widehat{L}(\widehat{f}_{\text{erm}}) \le \sup_{f\in\mathcal{F}} \left|L(f) - \widehat{L}(f)\right|, \tag{5}$$

and hence bound the estimation error in terms of this uniform bound. The following theorem shows that such uniform bounds on deviations between expectations and sample averages are intimately related to a notion of complexity of the *loss class* $\ell_{\mathcal{F}} = \{(\boldsymbol{x}, y) \mapsto \ell(f(\boldsymbol{x}), y) : f \in \mathcal{F}\}$ known as the *Rademacher complexity*. For a probability distribution $\mathbb{P}$ on a measurable space $\mathcal{Z}$, a sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \sim \mathbb{P}$, and a function class $\mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$, define the Rademacher complexity of $\mathcal{G}$ as

$$R_n(\mathcal{G}) := \mathbb{E} \sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{z}_i)\right|,$$

where $\epsilon_1, \ldots, \epsilon_n \in \{\pm 1\}$ are independent and uniformly distributed.

**Theorem 2.1.** *For any $\mathcal{G} \subset [0,1]^{\mathcal{Z}}$ and any probability distribution $\mathbb{P}$ on $\mathcal{Z}$,*

$$\frac{1}{2}R_n(\mathcal{G}) - \sqrt{\frac{\log 2}{2n}} \le \mathbb{E} \sup_{g\in\mathcal{G}} \left|\mathbb{E}g - \widehat{\mathbb{E}}g\right| \le 2R_n(\mathcal{G}),$$

*where $\widehat{\mathbb{E}}g = n^{-1}\sum_{i=1}^n g(\boldsymbol{z}_i)$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are chosen i.i.d. according to $\mathbb{P}$. Furthermore, with probability at least $1 - 2\exp(-2\epsilon^2 n)$ over $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$,*

$$\mathbb{E} \sup_{g\in\mathcal{G}} \left|\mathbb{E}g - \widehat{\mathbb{E}}g\right| - \epsilon \le \sup_{g\in\mathcal{G}} \left|\mathbb{E}g - \widehat{\mathbb{E}}g\right| \le \mathbb{E} \sup_{g\in\mathcal{G}} \left|\mathbb{E}g - \widehat{\mathbb{E}}g\right| + \epsilon.$$

*Thus, $R_n(\mathcal{G}) \to 0$ if and only if $\sup_{g\in\mathcal{G}} \left|\mathbb{E}g - \widehat{\mathbb{E}}g\right| \xrightarrow{as} 0$.*

See [KP00, Kol01, BBL02, BM02] and [Kol06]. This theorem shows that for bounded losses, a uniform bound

$$\sup_{f\in\mathcal{F}}\left|L(f)-\widehat{L}(f)\right|$$

on the maximal deviations between risks and empirical risks of any $f$ in $\mathcal{F}$ is tightly concentrated around its expectation, which is close to the Rademacher complexity $R_n(\ell_{\mathcal{F}})$. Thus, we can bound the excess risk of $\widehat{f}_{\text{erm}}$ in terms of the sum of the approximation error $\inf_{f\in\mathcal{F}} L(f) - \inf_f L(f)$ and this bound on the estimation error.

## 2.3 Faster rates

Although the approach (5) of bounding the deviation between the risk and empirical risk of $\widehat{f}_{\text{erm}}$ by the maximum for any $f \in \mathcal{F}$ of this deviation appears to be very coarse, there are many situations where it cannot be improved by more than a constant factor without stronger assumptions (we will see examples later in this section). However, there are situations where it can be significantly improved. As an illustration, provided $\mathcal{F}$ contains functions $f$ for which the variance of $\ell(f(\boldsymbol{x}), y)$ is positive, it is easy to see that $R_n(\ell_{\mathcal{F}}) = \Omega(n^{-1/2})$. Thus, the best bound on the estimation error implied by Theorem 2.1 must go to zero no faster than $n^{-1/2}$, but it is possible for the risk of the empirical minimizer to converge to the optimal value $L(f^*_{\mathcal{F}})$ faster than this. For example, when $\mathcal{F}$ is suitably simple, this occurs for a nonnegative bounded loss, $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,1]$, when there is a function $f^*_{\mathcal{F}}$ in $\mathcal{F}$ that gives perfect predictions, in the sense that almost surely $\ell(f^*_{\mathcal{F}}(\boldsymbol{x}), y) = 0$. In that case, the following theorem is an example that gives a faster rate in terms of the *worst-case empirical Rademacher complexity*,

$$\bar{R}_n(\mathcal{F}) = \sup_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\in\mathcal{X}} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i)\right| \middle| \boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\right].$$

Notice that, for any probability distribution on $\mathcal{X}$, $R_n(\mathcal{F}) \le \bar{R}_n(\mathcal{F})$.

**Theorem 2.2.** *There is a constant $c > 0$ such that for a bounded function class $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, for $\ell(\widehat{y}, y) = (\widehat{y}-y)^2$, and for any distribution $\mathbb{P}$ on $\mathcal{X}\times[-1,1]$, with probability at least $1-\delta$, a sample $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ satisfies for all $f \in \mathcal{F}$,*

$$L(f) \le (1+c)\widehat{L}(f) + c\,(\log n)^4\,\bar{R}_n^2(\mathcal{F}) + \frac{c\log(1/\delta)}{n}.$$

In particular, when $L(f^*_{\mathcal{F}}) = 0$, the empirical minimizer has $\widehat{L}(\widehat{f}_{\text{erm}}) = 0$, and so with high probability, $L(\widehat{f}_{\text{erm}}) = \tilde{O}\left(\bar{R}_n^2(\mathcal{F})\right)$, which can be as small as $\tilde{O}(1/n)$ for a suitably simple class $\mathcal{F}$.

Typically, faster rates like these arise when the variance of the excess loss is bounded in terms of its expectation, for instance

$$\mathbb{E}\left[\ell(f(\boldsymbol{x}), y) - \ell(f^*_{\mathcal{F}}(\boldsymbol{x}), y)\right]^2 \le c\mathbb{E}\left[\ell(f(\boldsymbol{x}), y) - \ell(f^*_{\mathcal{F}}(\boldsymbol{x}), y)\right].$$

For a bounded nonnegative loss with $L(f^*_{\mathcal{F}}) = 0$, this so-called Bernstein property is immediate, and it has been exploited in that case to give fast rates for prediction with binary-valued [VC71, VC74] and real-valued [Hau92, Pol95, BL99] function classes. Theorem 2.2, which follows from [SST10, Theorem 1] and the AM-GM inequality[1], relies on the smoothness of the quadratic loss to give a bound for that case in terms of the worst-case empirical Rademacher complexity. There has been a significant body of related work over the last thirty years. First, for quadratic loss in this well-specified setting, that is, when $f^*(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$ belongs to the class $\mathcal{F}$, faster rates have been obtained even without $L(f^*) = 0$ [vdG90]. Second, the Bernstein property can occur without the minimizer of $L$ being in $\mathcal{F}$; indeed, it arises for convex $\mathcal{F}$ with quadratic loss [LBW96] or more generally strongly convex losses [Men02], and this has been exploited to give fast rates

[1]The exponent on the log factor in Theorem 2.2 is larger than the result in the cited reference; any exponent larger than 3 suffices. See [RV06, Equation (1.4)].

based on several other notions of complexity [BBM05, Kol06, LRS15]. Recent techniques [Men20] eschew concentration bounds and hence give weaker conditions for convergence of $L(\widehat{f}_{\text{erm}})$ to $L(f^*_{\mathcal{F}})$, without the requirement that the random variables $\ell(f(\boldsymbol{x}), y)$ have light tails. Finally, while we have defined $\mathcal{F}$ as the class of functions used by the prediction method, if it is viewed instead as the benchmark (that is, the aim is to predict almost as well as the best function in $\mathcal{F}$, but the prediction method can choose a prediction rule $\widehat{f}$ that is not necessarily in $\mathcal{F}$), then similar fast rates are possible under even weaker conditions, but the prediction method must be more complicated than empirical risk minimization; see [RST17].

## 2.4 Complexity regularization

The results we have seen give bounds on the excess risk of $\widehat{f}_{\text{erm}}$ in terms of a sum of approximation error and a bound on the estimation error that depends on the complexity of the function class $\mathcal{F}$. Rather than choosing the complexity of the function class $\mathcal{F}$ in advance, we could instead split a rich class $\mathcal{F}$ into a complexity hierarchy and choose the appropriate complexity based on the data, with the aim of managing this approximation-estimation tradeoff. We might define subsets $\mathcal{F}_r$ of a rich class $\mathcal{F}$, indexed by a complexity parameter $r$. We call each $\mathcal{F}_r$ a *complexity class*, and we say that it has complexity $r$.

There are many classical examples of this approach. For instance, support vector machines (SVMs) [CV95] use a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, and the complexity class $\mathcal{F}_r$ is the subset of functions in $\mathcal{H}$ with RKHS norm no more than $r$. As another example, Lasso [Tib96] uses the set $\mathcal{F}$ of linear functions on a high-dimensional space, with the complexity classes $\mathcal{F}_r$ defined by the $\ell_1$ norm of the parameter vector. Both SVMs and Lasso manage the approximation-estimation trade-off by balancing the complexity of the prediction rule and its fit to the training data: they minimize a combination of empirical risk and some increasing function of the complexity $r$.

The following theorem gives an illustration of the effectiveness of this kind of complexity regularization. In the first part of the theorem, the complexity penalty for a complexity class is a uniform bound on deviations between expectations and sample averages for that class. We have seen that uniform deviation bounds of this kind imply upper bounds on the excess risk of the empirical risk minimizer in the class. In the second part of the theorem, the complexity penalty appears in the upper bounds on excess risk that arise in settings where faster rates are possible. In both cases, the theorem shows that when the bounds hold, choosing the best penalized empirical risk minimizer in the complexity hierarchy leads to the best of these upper bounds.

**Theorem 2.3.** *For each $\mathcal{F}_r \subseteq \mathcal{F}$, define an empirical risk minimizer*

$$\widehat{f}^r_{\text{erm}} \in \operatorname*{argmin}_{f \in \mathcal{F}_r} \widehat{L}(f).$$

*Among these, select the one with complexity $\widehat{r}$ that gives an optimal balance between the empirical risk and a complexity penalty $p_r$:*

$$\widehat{f} = \widehat{f}^{\widehat{r}}_{\text{erm}}, \qquad \widehat{r} \in \operatorname*{argmin}_r \left( \widehat{L}(\widehat{f}^r_{\text{erm}}) + p_r \right). \tag{6}$$

1. *In the event that the complexity penalties are uniform deviation bounds:*

$$\text{for all } r, \ \sup_{f \in \mathcal{F}_r} \left| L(f) - \widehat{L}(f) \right| \leq p_r, \tag{7}$$

*then we have the oracle inequality*

$$L(\widehat{f}) - \inf_f L(f) \leq \inf_r \left( \inf_{f \in \mathcal{F}_r} L(f) - \inf_f L(f) + 2p_r \right). \tag{8}$$

2. *Suppose that the complexity classes and penalties are ordered, that is,*

$$r \leq s \text{ implies } \mathcal{F}_r \subseteq \mathcal{F}_s \text{ and } p_r \leq p_s,$$

*and fix* $f_r^* \in \arg\min_{f \in \mathcal{F}_r} L(f)$. *In the event that the complexity penalties satisfy the uniform relative deviation bounds*

$$\text{for all } r, \ \sup_{f \in \mathcal{F}_r} \left( L(f) - L(f_r^*) - 2\left( \widehat{L}(f) - \widehat{L}(f_r^*) \right) \right) \leq 2p_r/7 \tag{9}$$
$$\text{and } \sup_{f \in \mathcal{F}_r} \left( \widehat{L}(f) - \widehat{L}(f_r^*) - 2\left( L(f) - L(f_r^*) \right) \right) \leq 2p_r/7,$$

*then we have the oracle inequality*

$$L(\widehat{f}) - \inf_f L(f) \leq \inf_r \left( \inf_{f \in \mathcal{F}_r} L(f) - \inf_f L(f) + 3p_r \right). \tag{10}$$

These are called *oracle inequalities* because (8) (respectively (10)) gives the error bound that follows from the best of the uniform bounds (7) (respectively (9)), as if we have access to an oracle who knows the complexity that gives the best bound. The proof of the first part is a straightforward application of the same decomposition as (4); see, for example, [BBL02]. The proof of the second part, which allows significantly smaller penalties $p_r$ when faster rates are possible, is also elementary; see [Bar08]. In both cases, the broad approach to managing the trade-off between approximation error and estimation error is qualitatively the same: having identified a complexity hierarchy $\{\mathcal{F}_r\}$ with corresponding excess risk bounds $p_r$, these results show the effectiveness of choosing from the hierarchy a function $f$ that balances the complexity penalty $p_r$ with the fit to the training data $\widehat{L}(\widehat{f}^r_{\text{erm}})$.

Later in this section, we will see examples of upper bounds on estimation error for neural network classes $\mathcal{F}_r$ indexed by a complexity parameter $r$ that depends on properties of the network, such as the size of the parameters. Thus, a prediction method that trades off the fit to the training data with these measures of complexity would satisfy an oracle inequality.

## 2.5 Computational complexity of empirical risk minimization

To this point, we have considered the statistical performance of the empirical risk minimizer $\widehat{f}_{\text{erm}}$ without considering the computational cost of solving this optimization problem. The classical cases where it can be solved efficiently involve linearly parameterized function classes, convex losses, and convex complexity penalties, so that penalized empirical risk minimization is a convex optimization problem. For instance, SVMs exploit a linear function class (an RKHS, $\mathcal{H}$), a convex loss,

$$\ell(f(\boldsymbol{x}), y) := (1 - yf(\boldsymbol{x})) \vee 0 \quad \text{for} \quad f : \mathcal{X} \to \mathbb{R} \quad \text{and} \quad y \in \{\pm 1\},$$

and a convex complexity penalty,

$$\mathcal{F}_r = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq r\}, \ p_r = r/\sqrt{n},$$

and choosing $\widehat{f}$ according to (6) corresponds to solving a quadratic program. Similarly, Lasso involves linear functions on $\mathbb{R}^d$, quadratic loss, and a convex penalty,

$$\mathcal{F}_r = \{\boldsymbol{x} \mapsto \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle : \|\boldsymbol{\beta}\|_1 \leq r\}, \ p_r = r\sqrt{\log(d)/n}.$$

Again, minimizing complexity-penalized empirical risk corresponds to solving a quadratic program.

On the other hand, the optimization problems that arise in a classification setting, where functions map to a discrete set, have a combinatorial flavor, and are often computationally hard in the worst case. For instance, empirical risk minimization over the set of linear classifiers

$$\mathcal{F} = \left\{ \boldsymbol{x} \mapsto \text{sign}\left( \langle \boldsymbol{w}, \boldsymbol{x} \rangle \right) : \boldsymbol{w} \in \mathbb{R}^d \right\}$$

is NP-hard [JP78, GJ79]. In contrast, if there is a function in this class that classifies all of the training data correctly, finding an empirical risk minimizer is equivalent to solving a linear program, which can be

solved efficiently. Another approach to simplifying the algorithmic challenge of empirical risk minimization is to replace the discrete loss for this family of thresholded linear functions with a surrogate convex loss for the family of linear functions. This is the approach used in SVMs: replacing a nonconvex loss with a convex loss allows for computational efficiency, even when there is no thresholded linear function that classifies all of the training data correctly.

However, the corresponding optimization problems for neural networks appear to be more difficult. Even when $\widehat{L}(\widehat{f}_{\text{erm}}) = 0$, various natural empirical risk minimization problems over families of neural networks are NP-hard [Jud90, BR92, DSS95], and this is still true even for convex losses [Vu98, BBD02].

In the remainder of this section, we focus on the statistical complexity of prediction problems with neural network function classes (we shall return to computational complexity considerations in Section 5). We review estimation error bounds involving these classes, focusing particularly on the Rademacher complexity. The Rademacher complexity of a loss class $\ell_{\mathcal{F}}$ can vary dramatically with the loss $\ell$. For this reason, we consider separately discrete losses, such as those used for classification, convex upper bounds on these losses, like the SVM loss and other large margin losses used for classification, and Lipschitz losses used for regression.

## 2.6 Classification

We first consider loss classes for the problem of classification. For simplicity, consider a two-class classification problem, where $\mathcal{Y} = \{\pm 1\}$, and define the $\pm 1$ loss, $\ell_{\pm 1}(\widehat{y}, y) = -y\widehat{y}$. Then for $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$, $R_n(\ell_{\mathcal{F}}) = R_n(\mathcal{F})$, since the distribution of $\epsilon_i \ell_{\pm 1}(f(\boldsymbol{x}_i), y_i) = -\epsilon_i y_i f(\boldsymbol{x}_i)$ is the same as that of $\epsilon_i f(\boldsymbol{x}_i)$. The following theorem shows that the Rademacher complexity depends on a combinatorial dimension of $\mathcal{F}$, known as the VC-dimension [VC71].

**Theorem 2.4.** *For $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ and for any distribution on $\mathcal{X}$,*

$$R_n(\mathcal{F}) \leq \sqrt{\frac{2\log(2\Pi_{\mathcal{F}}(n))}{n}},$$

*where*

$$\Pi_{\mathcal{F}}(n) = \max\left\{\left|\left\{(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)) : f \in \mathcal{F}\right\}\right| : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}\right\}.$$

*If $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$ and $n \geq d = d_{VC}(\mathcal{F})$, then*

$$\Pi_{\mathcal{F}}(n) \leq (en/d)^d,$$

*where $d_{VC}(\mathcal{F}) := \max\left\{d : \Pi_{\mathcal{F}}(d) = 2^d\right\}$. In that case, for any distribution on $\mathcal{X}$,*

$$R_n(\mathcal{F}) = O\left(\sqrt{\frac{d\log(n/d)}{n}}\right),$$

*and conversely, for some probability distribution, $R_n(\mathcal{F}) = \Omega\left(\sqrt{d/n}\right)$.*

These bounds imply that, for the worst case probability distribution, the uniform deviations between sample averages and expectations grow like $\tilde{\Theta}(\sqrt{d_{VC}(\mathcal{F})/n})$, a result of [VC71]. The log factor in the upper bound can be removed; see [Tal94]. Classification problems are an example where the crude upper bound (5) cannot be improved without stronger assumptions: the *minimax excess risk* is essentially the same as these uniform deviations. In particular, these results show that empirical risk minimization leads, for any probability distribution, to excess risk that is $O(\sqrt{d_{VC}(\mathcal{F})/n})$, but conversely, for every method that predicts a $\widehat{f} \in \mathcal{F}$, there is a probability distribution for which the excess risk is $\Omega(\sqrt{d_{VC}(\mathcal{F})/n})$ [VC74]. When there is a prediction rule in $\mathcal{F}$ that predicts perfectly, that is $L(f^*_{\mathcal{F}}) = 0$, the upper and lower bounds can be improved to $\tilde{\Theta}(d_{VC}(\mathcal{F})/n)$ [BEHW89, EHKV89].

These results show that $d_{VC}(\mathcal{F})$ is critical for uniform convergence of sample averages to probabilities, and more generally for the statistical complexity of classification with a function class $\mathcal{F}$. The following

theorem summarizes the known bounds on the VC-dimension of neural networks with various piecewise-polynomial nonlinearities. Recall that a feed-forward neural network with $L$ layers is defined by a sequence of layer widths $d_1, \ldots, d_L$ and functions $\sigma_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ for $l = 1, \ldots, L$. It is a family of $\mathbb{R}^{d_L}$-valued functions on $\mathbb{R}^d$ parameterized by $\boldsymbol{\theta} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L)$; see (1). We often consider scalar nonlinearities $\sigma : \mathbb{R} \to \mathbb{R}$ applied componentwise, that is, $\sigma_l(v)_i := \sigma(v_i)$. For instance, $\sigma$ might be the scalar nonlinearity used in the *ReLU (rectified linear unit)*, $\sigma(\alpha) = \alpha \vee 0$. We say that this family has $p$ parameters if there is a total of $p$ entries in the matrices $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L$. We say that $\sigma$ is *piecewise polynomial* if it can be written as a sum of a constant number of polynomials,

$$\sigma(x) = \sum_{i=1}^{k} \mathbf{1}\left[x \in I_i\right] p_i(x),$$

where the intervals $I_1, \ldots, I_k$ form a partition of $\mathbb{R}$ and the $p_i$ are polynomials.

**Theorem 2.5.** *Consider feed-forward neural networks $\mathcal{F}_{L,\sigma}$ with $L$ layers, scalar output (that is, $d_L = 1$), output nonlinearity $\sigma_L(\alpha) = \operatorname{sign}(\alpha)$, and scalar nonlinearity $\sigma$ at every other layer. Define*

$$d_{L,\sigma,p} = \max\left\{d_{VC}(\mathcal{F}_{L,\sigma}) : \mathcal{F}_{L,\sigma} \text{ has } p \text{ parameters}\right\}.$$

1. *For $\sigma$ piecewise constant, $d_{L,\sigma,p} = \tilde{\Theta}\left(p\right)$.*
2. *For $\sigma$ piecewise linear, $d_{L,\sigma,p} = \tilde{\Theta}\left(pL\right)$.*
3. *For $\sigma$ piecewise polynomial, $d_{L,\sigma,p} = \tilde{O}\left(pL^2\right)$.*

Part 1 is from [BH89]. The upper bound in part 2 is from [BHLM19]. The lower bound in part 2 and the bound in part 3 are from [BMM98]. There are also upper bounds for the smooth sigmoid $\sigma(\alpha) = 1/(1 + \exp(-\alpha))$ that are quadratic in $p$; see [KM97]. See Chapter 8 of [AB99] for a review.

The theorem shows that the VC-dimension of these neural networks grows at least linearly with the number of parameters in the network, and hence to achieve small excess risk or uniform convergence of sample averages to probabilities for discrete losses, the sample size must be large compared to the number of parameters in these networks.

There is an important caveat to this analysis: it captures arbitrarily fine-grained properties of real-valued functions, because the operation of thresholding these functions is very sensitive to perturbations, as the following example shows.

**Example 2.6.** *For $\alpha > 0$, define the nonlinearity $\tilde{r}(x) := (x + \alpha \sin x) \vee 0$ and the following one-parameter class of functions computed by two-layer networks with these nonlinearities:*

$$\mathcal{F}_{\tilde{r}} := \left\{x \mapsto \operatorname{sign}(\pi + \tilde{r}(wx) - \tilde{r}(wx + \pi)) : w \in \mathbb{R}\right\}.$$

*Then $d_{VC}(\mathcal{F}_{\tilde{r}}) = \infty$.*

Indeed, provided $wx \geq \alpha$, $\tilde{r}(wx) = wx + \alpha \sin(wx)$, hence $\pi + \tilde{r}(wx) - \tilde{r}(wx + \pi) = 2\alpha \sin(wx)$. This shows that the set of functions in $\mathcal{F}_{\tilde{r}}$ restricted to $\mathbb{N}$ contains

$$\left\{x \mapsto \operatorname{sign}(\sin(wx)) : w \geq \alpha\right\} = \left\{x \mapsto \operatorname{sign}(\sin(wx)) : w \geq 0\right\},$$

and the VC-dimension of the latter class of functions on $\mathbb{N}$ is infinite; see, for example, [AB99, Lemma 7.2]. Thus, with an arbitrarily small perturbation of the ReLU nonlinearity, the VC-dimension of this class changes from a small constant to infinity. See also [AB99, Theorem 7.1], which gives a similar result for a slightly perturbed version of a sigmoid nonlinearity.

As we have seen, the requirement that the sample size grows with the number of parameters is at odds with empirical experience: deep networks with far more parameters than the number of training examples routinely give good predictive accuracy. It is plausible that the algorithms used to optimize these networks are not exploiting their full expressive power. In particular, the analysis based on combinatorial dimensions

captures arbitrarily fine-grained properties of the family of real-valued functions computed by a deep network, whereas algorithms that minimize a convex loss might not be significantly affected by such fine-grained properties. Thus, we might expect that replacing the discrete loss $\ell_{\pm 1}$ with a convex surrogate, in addition to computational convenience, could lead to reduced statistical complexity. The empirical success of gradient methods with convex losses for overparameterized thresholded real-valued classifiers was observed both in neural networks [MP90], [LGT97], [CLG01] and in related classification methods [DC95], [Qui96], [Bre98]. It was noticed that classification performance can improve as the number of parameters is increased even after all training examples are classified correctly [Qui96], [Bre98].[2] These observations motivated *large margin analyses* [Bar98], [SFBL98], which reduce classification problems to regression problems.

## 2.7 Large margin classification

Although the aim of a classification problem is to minimize the expectation of a discrete loss, if we consider classifiers such as neural networks that consist of thresholded real-valued functions obtained by minimizing a surrogate loss—typically a convex function of the real-valued prediction—then it turns out that we can obtain bounds on estimation error by considering approximations of the class of real-valued functions. This is important because the statistical complexity of that function class can be considerably smaller than that of the class of thresholded functions. In effect, for a well-behaved surrogate loss, fine-grained properties of the real-valued functions are not important. If the surrogate loss $\ell$ satisfies a Lipschitz property, we can relate the Rademacher complexity of the loss class $\ell_{\mathcal{F}}$ to that of the function class $\mathcal{F}$ using the *Ledoux-Talagrand contraction inequality* [LT91, Theorem 4.12].

**Theorem 2.7.** *Suppose that, for all $y$, $\widehat{y} \mapsto \ell(\widehat{y}, y)$ is $c$-Lipschitz and satisfies $\ell(0, y) = 0$. Then $R_n(\ell_{\mathcal{F}}) \leq 2cR_n(\mathcal{F})$.*

Notice that the assumption that $\ell(0, y) = 0$ is essentially without loss of generality: adding a fixed function to $\ell_{\mathcal{F}}$ by replacing $\ell(\widehat{y}, y)$ with $\ell(\widehat{y}, y) - \ell(0, y)$ shifts the Rademacher complexity by $O\left(1/\sqrt{n}\right)$.

For classification with $y \in \{-1, 1\}$, the hinge loss $\ell(\widehat{y}, y) = (1 - y\widehat{y}) \vee 0$ used by SVMs and the logistic loss $\ell(\widehat{y}, y) = \log\left(1 + \exp(-y\widehat{y})\right)$ are examples of convex, 1-Lipschitz surrogate losses. The quadratic loss $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ and the exponential loss $\ell(\widehat{y}, y) := \exp(-y\widehat{y})$ used by AdaBoost [FS97] (see Section 3) are also convex, and they are Lipschitz when functions in $\mathcal{F}$ have bounded range.

We can write all of these surrogate losses as $\ell_\phi(\widehat{y}, y) := \phi(\widehat{y}y)$ for some function $\phi : \mathbb{R} \to [0, \infty)$. The following theorem relates the excess risk to the excess surrogate risk. It is simpler to state when $\phi$ is convex and when, rather than $\ell_{\pm 1}$, we consider a shifted, scaled version, defined as $\ell_{01}(\widehat{y}, y) := \mathbf{1}\left[\widehat{y} \neq y\right]$. We use $L_{01}(f)$ and $L_\phi(f)$ to denote $\mathbb{E}\ell_{01}(f(\boldsymbol{x}), y)$ and $\mathbb{E}\ell_\phi(f(\boldsymbol{x}), y)$ respectively.

**Theorem 2.8.** *For a convex function $\phi : \mathbb{R} \to [0, \infty)$, define $\ell_\phi(\widehat{y}, y) := \phi(\widehat{y}y)$ and $C_\theta(\alpha) := (1+\theta)\phi(\alpha)/2 + (1-\theta)\phi(-\alpha)/2$, and define $\psi_\phi : [0, 1] \to [0, \infty)$ as $\psi_\phi(\theta) := \inf\{C_\theta(\alpha) : \alpha \leq 0\} - \inf\{C_\theta(\alpha) : \alpha \in \mathbb{R}\}$. Then we have the following.*

1. *For any measurable $\widehat{f} : \mathcal{X} \to \mathbb{R}$ and any probability distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$,*

$$\psi_\phi\left(L_{01}(\widehat{f}) - \inf_f L_{01}(f)\right) \leq L_\phi(\widehat{f}) - \inf_f L_\phi(f),$$

   *where the infima are over measurable functions $f$.*

2. *For $|\mathcal{X}| \geq 2$, this inequality cannot hold if $\psi_\phi$ is replaced by any larger function:*

$$\sup_\theta \inf\Big\{L_\phi(\widehat{f}) - \inf_f L_\phi(f) - \psi_\phi(\theta) : \mathbb{P},\ \widehat{f} \text{ satisfy } L_{01}(\widehat{f}) - \inf_f L_{01}(f) = \theta\Big\} = 0.$$

[2] Both phenomena were observed more recently in neural networks; see [ZBH+17] and [NTSS17].

3. $\psi_\phi(\theta_i) \to 0$ *implies* $\theta_i \to 0$ *if and only if both* $\phi$ *is differentiable at zero and* $\phi'(0) < 0$.

For example, for the hinge loss $\phi(\alpha) = (1-\alpha) \vee 0$, the relationship between excess risk and excess $\phi$-risk is given by $\psi_\phi(\theta) = |\theta|$, for the quadratic loss $\phi(\alpha) = (1-\alpha)^2$, $\psi_\phi(\theta) = \theta^2$, and for the exponential loss $\phi(\alpha) = \exp(-\alpha)$, $\psi_\phi(\theta) = 1 - \sqrt{1-\theta^2}$. Theorem 2.8 is from [BJM06]; see also [Lin04, LV04] and [Zha04].

Using (4), (5), and Theorems 2.1 and 2.7 to bound $\mathbb{E}\ell_{\phi,\widehat{f}} - \inf_f \mathbb{E}\ell_{\phi,f}$ in terms of $R_n(\mathcal{F})$ and combining with Theorem 2.8 shows that, if $\phi$ is 1-Lipschitz then with high probability,

$$\psi_\phi\left(L_{01}(\widehat{f}) - \inf_f L_{01}(f)\right) \le 4R_n(\mathcal{F}) + O\left(\frac{1}{\sqrt{n}}\right) + \inf_{f\in\mathcal{F}} L_\phi(f) - \inf_f L_\phi(f). \tag{11}$$

Notice that in addition to the Rademacher complexity of the real-valued class $\mathcal{F}$, this bound includes an approximation error term defined in terms of the surrogate loss; the binary-valued prediction problem has been reduced to a real-valued problem.

Alternatively, we could consider more naive bounds: If a loss satisfies the pointwise inequality $\ell_{01}(\widehat{y}, y) \le \ell_\phi(\widehat{y}, y)$, then we have an upper bound on risk in terms of surrogate risk: $L_{01}(\widehat{f}) \le L_\phi(\widehat{f})$. In fact, Theorem 2.8 implies that pointwise inequalities like this are inevitable for any reasonable convex loss. Define a surrogate loss $\phi$ as *classification-calibrated* if any $f$ that minimizes the surrogate risk $L_\phi(f)$ will also minimize the classification risk $L_{01}(f)$. Then part 3 of the theorem shows that if a convex surrogate loss $\phi$ is classification-calibrated then it satisfies

$$\text{for all } \widehat{y}, y,\ \frac{\ell_\phi(\widehat{y}, y)}{\phi(0)} = \frac{\phi(\widehat{y}y)}{\phi(0)} \ge 1[\widehat{y}y \le 0] = \ell_{01}(\widehat{y}, y).$$

Thus, every classification-calibrated convex surrogate loss, suitably scaled so that $\phi(0) = 1$, is an upper bound on the discrete loss $\ell_{01}$, and hence immediately gives an upper bound on risk in terms of surrogate risk: $L_{01}(\widehat{f}) \le L_\phi(\widehat{f})$. Combining this with Theorems 2.1 and 2.7 shows that, if $\phi$ is also 1-Lipschitz then with high probability,

$$L_{01}(\widehat{f}) \le \widehat{L}_\phi(\widehat{f}) + 4R_n(\mathcal{F}) + O\left(\frac{1}{\sqrt{n}}\right). \tag{12}$$

## 2.8 Real prediction

For a real-valued function class $\mathcal{F}$, there is an analog of Theorem 2.4 with the VC-dimension of $\mathcal{F}$ replaced by the *pseudodimension* of $\mathcal{F}$, which is the VC-dimension of $\{(\boldsymbol{x}, y) \mapsto \mathbf{1}[f(\boldsymbol{x}) \ge y] : f \in \mathcal{F}\}$; see [Pol90]. Theorem 2.5 is true with the output nonlinearity $\sigma_L$ of $\mathcal{F}_{L,\sigma}$ replaced by any Lipschitz nonlinearity and with $d_{VC}$ replaced by the pseudodimension. However, using this result to obtain bounds on the excess risk of an empirical risk minimizer would again require the sample size to be large compared to the number of parameters.

Instead, we can bound $R_n(\ell_\mathcal{F})$ more directly in many cases. With a bound on $R_n(\mathcal{F})$ for a class $\mathcal{F}$ of real-valued functions computed by neural networks, we can then apply Theorem 2.7 to relate $R_n(\ell_\mathcal{F})$ to $R_n(\mathcal{F})$, provided the loss is a Lipschitz function of its first argument. This is the case, for example, for absolute loss $\ell(\widehat{y}, y) = |\widehat{y} - y|$, or for quadratic loss $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$ when $\mathcal{Y}$ and the range of functions in $\mathcal{F}$ are bounded.

The following result gives a bound on Rademacher complexity for neural networks that use a bounded, Lipschitz nonlinearity, such as the sigmoid function

$$\sigma(x) = \frac{1 - \exp(-x)}{1 + \exp(-x)}.$$

**Theorem 2.9.** *For two-layer neural networks defined on* $\mathcal{X} = [-1, 1]^d$,

$$\mathcal{F}_B = \left\{ \boldsymbol{x} \mapsto \sum_{i=1}^k b_i \sigma\left(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle\right) : \|\boldsymbol{b}\|_1 \le 1,\ \|\boldsymbol{w}_i\|_1 \le B,\ k \ge 1 \right\},$$

*where the nonlinearity $\sigma : \mathbb{R} \to [-1, 1]$ is 1-Lipschitz and has $\sigma(0) = 0$,*

$$R_n(\mathcal{F}_B) \le B\sqrt{\frac{2\log 2d}{n}}.$$

Thus, for example, applying (11) in this case with a Lipschitz convex loss $\ell_\phi$ and the corresponding $\psi_\phi$ defined by Theorem 2.8, shows that with high probability the minimizer $\widehat{f}_{\text{erm}}$ in $\mathcal{F}_B$ of $\widehat{\mathbb{E}}\ell_f$ satisfies

$$\psi_\phi\left(L_{01}(\widehat{f}_{\text{erm}}) - \inf_f L_{01}(f)\right) \le O\left(B\sqrt{\frac{\log d}{n}}\right) + \inf_{f\in\mathcal{F}_B} L_\phi(f) - \inf_f L_\phi(f).$$

If, in addition, $\ell_\phi$ is scaled so that it is an upper bound on $\ell_{01}$, applying (12) shows that with high probability every $f \in \mathcal{F}_B$ satisfies

$$L_{01}(f) \le \widehat{L}_\phi(f) + O\left(B\sqrt{\frac{\log d}{n}}\right).$$

Theorem 2.9 is from [BM02]. The proof uses the contraction inequality (Theorem 2.7) and elementary properties of Rademacher complexity.

The following theorem gives similar error bounds for networks with Lipschitz nonlinearities that, like the ReLU nonlinearity, do not necessarily have a bounded range. The definition of the function class includes deviations of the parameter matrices $\boldsymbol{W}_i$ from fixed 'centers' $\boldsymbol{M}_i$.

**Theorem 2.10.** *Consider a feed-forward network with $L$ layers, fixed vector nonlinearities $\sigma_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ and parameter $\boldsymbol{\theta} = (\boldsymbol{W}_1, \dots, \boldsymbol{W}_L)$ with $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$, for $i = 1, \dots, L$, which computes functions*

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sigma_L(\boldsymbol{W}_L\sigma_{L-1}(\boldsymbol{W}_{L-1}\cdots\sigma_1(\boldsymbol{W}_1\boldsymbol{x})\cdots)),$$

*where $d_0 = d$ and $d_L = 1$. Define $\bar{d} = d_0 \vee \cdots \vee d_L$. Fix matrices $\boldsymbol{M}_i \in \mathbb{R}^{d_i \times d_{i-1}}$, for $i = 1, \dots, L$, and define the class of functions on the unit Euclidean ball in $\mathbb{R}^d$,*

$$\mathcal{F}_r = \left\{ f(\cdot, \boldsymbol{\theta}) : \prod_{i=1}^L \|\boldsymbol{W}_i\| \left(\sum_{i=1}^L \frac{\|\boldsymbol{W}_i^\top - \boldsymbol{M}_i^\top\|_{2,1}^{2/3}}{\|\boldsymbol{W}_i\|^{2/3}}\right)^{3/2} \le r \right\},$$

*where $\|\boldsymbol{A}\|$ denotes the spectral norm of the matrix $\boldsymbol{A}$ and $\|\boldsymbol{A}\|_{2,1}$ denotes the sum of the 2-norms of its columns. If the $\sigma_i$ are all 1-Lipschitz and the surrogate loss $\ell_\phi$ is a $b$-Lipschitz upper bound on the classification loss $\ell_{01}$, then with probability at least $1 - \delta$, every $f \in \mathcal{F}_r$ has*

$$L_{01}(f) \le \widehat{L}_\phi(f) + \tilde{O}\left(\frac{rb\log\bar{d} + \sqrt{\log(1/\delta)}}{\sqrt{n}}\right).$$

Theorem 2.10 is from [BFT17]. The proof uses different techniques (covering numbers rather than the Rademacher complexity) to address the key technical difficulty, which is controlling the scale of vectors that appear throughout the network.

When the nonlinearity has a 1-*homogeneity property*, the following result gives a simple direct bound on the Rademacher complexity in terms of the Frobenius norms of the weight matrices (although it is worse than Theorem 2.10, even with $\boldsymbol{M}_i = 0$, unless the ratios $\|\boldsymbol{W}_i\|_F/\|\boldsymbol{W}_i\|$ are close to 1). We say that $\sigma : \mathbb{R} \to \mathbb{R}$ is 1-homogeneous if $\sigma(\alpha x) = \alpha\sigma(x)$ for all $x \in \mathbb{R}$ and $\alpha \ge 0$. Notice that the ReLU nonlinearity $\sigma(x) = x \vee 0$ has this property.

**Theorem 2.11.** *Let $\bar{\sigma} : \mathbb{R} \to \mathbb{R}$ be a fixed 1-homogeneous nonlinearity, and define the componentwise version $\sigma_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ via $\sigma_i(\boldsymbol{x})_j = \bar{\sigma}(\boldsymbol{x}_j)$. Consider a network with $L$ layers of these nonlinearities and parameters $\boldsymbol{\theta} = (\boldsymbol{W}_1, \dots, \boldsymbol{W}_L)$, which computes functions*

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sigma_L(\boldsymbol{W}_L\sigma_{L-1}(\boldsymbol{W}_{L-1}\cdots\sigma_1(\boldsymbol{W}_1\boldsymbol{x})\cdots)).$$

*Define the class of functions on the unit Euclidean ball in $\mathbb{R}^d$,*

$$\mathcal{F}_B = \{f(\cdot;\boldsymbol{\theta}) : \|\boldsymbol{W}_i\|_F \le B\},$$

*where $\|\boldsymbol{W}_i\|_F$ denotes the Frobenius norm of $\boldsymbol{W}_i$. Then we have*

$$R_n(\mathcal{F}_B) \lesssim \frac{\sqrt{L}B^L}{\sqrt{n}}.$$

This result is from [GRS18], which also shows that it is possible to remove the $\sqrt{L}$ factor at the cost of a worse dependence on $n$. See also [NTS15].

## 2.9 The mismatch between benign overfitting and uniform convergence

It is instructive to consider the implications of the generalization bounds we have reviewed in this section for the phenomenon of benign overfitting, which has been observed in deep learning. For concreteness, suppose that $\ell$ is the quadratic loss. Consider a neural network function $\widehat{f} \in \mathcal{F}$ chosen so that $\widehat{L}(\widehat{f}) = 0$. For an appropriate complexity hierarchy $\mathcal{F} = \bigcup_r \mathcal{F}_r$, suppose that $\widehat{f}$ is chosen to minimize the complexity $r(\widehat{f})$, defined as the smallest $r$ for which $\widehat{f} \in \mathcal{F}_r$, subject to the interpolation constraint $\widehat{L}(f) = 0$. What do the bounds based on uniform convergence imply about the excess risk $L(\widehat{f}) - \inf_{f\in\mathcal{F}} L(f)$ of this *minimum-complexity interpolant*?

Theorems 2.9, 2.10, and 2.11 imply upper bounds on risk in terms of various notions of scale of network parameters. For these bounds to be meaningful for a given probability distribution, there must be an interpolating $\widehat{f}$ for which the complexity $r(\widehat{f})$ grows suitably slowly with the sample size $n$ so that the excess risk bounds converge to zero.

An easy example is when there is an $f^* \in \mathcal{F}_r$ with $L(f^*) = 0$, where $r$ is a fixed complexity. Notice that this implies not just that the conditional expectation is in $\mathcal{F}_r$, but that there is no noise, that is, almost surely $y = f^*(\boldsymbol{x})$. In that case, if we choose $\widehat{f}$ as the interpolant $\widehat{L}(\widehat{f}) = 0$ with minimum complexity, then its complexity will certainly satisfy $r(\widehat{f}) \le r(f^*) = r$. And then as the sample size $n$ increases, $L(\widehat{f})$ will approach zero. In fact, since $\widehat{L}(\widehat{f}) = 0$, Theorem 2.2 implies a faster rate in this case: $L(\widehat{f}) = O((\log n)^4 \bar{R}_n^2(\mathcal{F}_r))$.

Theorem 2.3 shows that if we were to balance the complexity with the fit to the training data, then we can hope to enjoy excess risk as good as the best bound for any $\mathcal{F}_r$ in the complexity hierarchy. If we always choose a perfect fit to the data, there is no trade-off between complexity and empirical risk, but when there is a prediction rule $f^*$ with finite complexity and zero risk, then once the sample size is sufficiently large, the best trade-off does correspond to a perfect fit to the data. To summarize: when there is no noise, that is, when $y = f^*(x)$, and $f^* \in \mathcal{F}$, classical theory shows that a minimum-complexity interpolant $\widehat{f} \in \mathcal{F}$ will have risk $L(\widehat{f})$ converging to zero as the sample size increases.

But what if there is noise, that is, there is no deterministic relationship between $\boldsymbol{x}$ and $y$? Then it turns out that the bounds on the excess risk $L(\widehat{f}) - L(f^*_{\mathcal{F}})$ presented in this section must become vacuous: they can never decrease below a constant, no matter how large the sample size. This is because these bounds do not rely on any properties of the distribution on $\mathcal{X}$, and hence are also true in a *fixed design* setting, where the excess risk is at least the noise level.

To make this precise, fix $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$ and define the *fixed design risk*

$$L_{|\boldsymbol{x}}(f) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\ell(f(\boldsymbol{x}_i), y) \middle| \boldsymbol{x} = \boldsymbol{x}_i\right].$$

Then the decomposition (4) extends to this risk: for any $\widehat{f}$ and $f^*$,

$$\begin{aligned} &L_{|\boldsymbol{x}}(\widehat{f}) - L_{|\boldsymbol{x}}(f^*) \\ &= \left[L_{|\boldsymbol{x}}(\widehat{f}) - \widehat{L}(\widehat{f})\right] + \left[\widehat{L}(\widehat{f}) - \widehat{L}(f^*)\right] + \left[\widehat{L}(f^*) - L_{|\boldsymbol{x}}(f^*)\right]. \end{aligned}$$

For a nonnegative loss, the second term is nonpositive when $\widehat{L}(\widehat{f}) = 0$, and the last term is small for any fixed $f^*$. Fix $f^*(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$, and suppose we choose $\widehat{f}$ from a class $\mathcal{F}_r$. The same proof as that of Theorem 2.1 gives a Rademacher complexity bound on the first term above, and [LT91, Theorem 4.12] implies the same contraction inequality as in Theorem 2.7 when $\widehat{y} \mapsto \ell(\widehat{y}, y)$ is $c$-Lipschitz:

$$\begin{aligned}\mathbb{E} \sup_{f \in \mathcal{F}_r} \left| L_{|\boldsymbol{x}}(f) - \widehat{L}(f) \right| &\leq 2\mathbb{E}\left[ \sup_{f \in \mathcal{F}_r} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \ell(f(\boldsymbol{x}_i), y_i) \right| \middle| \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \right] \\ &\leq 4c\bar{R}_n(\mathcal{F}_r).\end{aligned}$$

Finally, although Theorems 2.9 and Theorem 2.11 are stated as bounds on the Rademacher complexity of $\mathcal{F}_r$, they are in fact bounds on $\bar{R}_n(\mathcal{F}_r)$, the worst-case empirical Rademacher complexity of $\mathcal{F}$.

Consider the complexity hierarchy defined in Theorem 2.9 or Theorem 2.11. For the minimum-complexity interpolant $\widehat{f}$, these theorems give bounds that depend on the complexity $r(\widehat{f})$, that is, bounds of the form $L(\widehat{f}) - L(f^*) \leq B(r(\widehat{f}))$ (ignoring the fact that that the minimum complexity $r(\widehat{f})$ is random; making the bounds uniform over $r$ would give a worse bound). Then these observations imply that

$$\mathbb{E}\left[ L_{|\boldsymbol{x}}(\widehat{f}) - L_{|\boldsymbol{x}}(f^*) \right] = \mathbb{E} L_{|\boldsymbol{x}}(\widehat{f}) - L(f^*) \leq \mathbb{E} B(r(\widehat{f})).$$

But then

$$\mathbb{E} B(r(\widehat{f})) \geq \mathbb{E}\left[ L_{|\boldsymbol{x}}(\widehat{f}) - L(f^*) \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left( \widehat{f}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i) \right)^2 = L(f^*).$$

Thus, unless there is no noise, the upper bound on excess risk must be at least as big as a constant.

[BL20b] use a similar comparison between prediction problems in random design and fixed design settings to demonstrate situations where benign overfitting occurs but a general family of excess risk bounds—those that depend only on properties of $\widehat{f}$ and do not increase too quickly with sample size—must sometimes be very loose. [NK19] present a scenario where, with high probability, a classification method gives good predictive accuracy but uniform convergence bounds must fail for any function class that contains the algorithm's output. Algorithmic stability approaches—see [DW79] and [BE02]—also aim to identify sufficient conditions for closeness of risk and empirical risk, and appear to be inapplicable in the interpolation regime. These examples illustrate that to understand benign overfitting, new analysis approaches are necessary that exploit additional information. We shall review results of this kind in Section 4, for minimum-complexity interpolants in regression settings. The notion of complexity that is minimized is obviously of crucial importance here; this is the topic of the next section.

# 3 Implicit regularization

When the model $\mathcal{F}$ is complex enough to ensure zero empirical error, such as in the case of overparametrized neural networks, the set of empirical minimizers may be large. Therefore, it may very well be the case that some empirical minimizers generalize well while others do not. Optimization algorithms introduce a bias in this choice: an iterative method may converge to a solution with certain properties. Since this bias is a by-product rather than an explicitly enforced property, we follow the recent literature and call it *implicit regularization*. In subsequent sections, we shall investigate statistical consequences of such implicit regularization.

Perhaps the simplest example of implicit regularization is gradient descent on the square-loss objective with linear functions:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla \widehat{L}(\boldsymbol{\theta}_t), \qquad \widehat{L}(\boldsymbol{\theta}) = \frac{1}{n} \left\| \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y} \right\|_2^2, \quad \boldsymbol{\theta}_0 = \mathbf{0} \in \mathbb{R}^d, \tag{13}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\intercal \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} = [y_1, \ldots, y_n]^\intercal$ are the training data, and $\eta_t > 0$ is the step size. While the set of minimizers of the square-loss objective in the overparametrized ($d > n$) regime is an affine

subspace of dimension at least $d-n$, gradient descent (with any choice of step size that ensures convergence) converges to a very specific element of this subspace: the minimum-norm solution

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{\theta}\|_2 : \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle = y_i \text{ for all } i \le n \right\}. \tag{14}$$

This minimum-norm interpolant can be written in closed form as

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^{\dagger} \boldsymbol{y}, \tag{15}$$

where $\boldsymbol{X}^{\dagger}$ denotes the pseudoinverse. It can also be seen as a limit of *ridge regression*

$$\boldsymbol{\theta}_{\lambda} = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{n} \|\boldsymbol{X}\theta - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \tag{16}$$

as $\lambda \to 0^+$. The connection between minimum-norm interpolation (14) and the "ridgeless" limit of ridge regression will be fruitful in the following sections when statistical properties of these methods are analyzed and compared.

To see that the iterations in (13) converge to the minimum-norm solution, observe that the Karush-Kuhn-Tucker (KKT) conditions for the constrained optimization problem (14) are $\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$ and $\boldsymbol{\theta} + \boldsymbol{X}^{\intercal}\boldsymbol{\mu} = 0$ for Lagrange multipliers $\boldsymbol{\mu} \in \mathbb{R}^n$. Both conditions are satisfied (in finite time or in the limit) by any procedure that interpolates the data while staying in the span of the rows of $\boldsymbol{X}$, including (13). It should be clear that a similar statement holds for more general objectives $\widehat{L}(\boldsymbol{\theta}) = n^{-1} \sum_i \ell(\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle, y_i)$ under appropriate assumptions on $\ell$. Furthermore, if started from an arbitrary $\boldsymbol{\theta}_0$, gradient descent (if it converges) selects a solution that is closest to the initialization with respect to $\|\cdot\|_2$.

Boosting is another notable example of implicit regularization arising from the choice of the optimization algorithm, this time for the problem of classification. Consider the linear classification objective

$$\widehat{L}_{01}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[-y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle \ge 0\right] \tag{17}$$

where $y_1, \ldots, y_n \in \{\pm 1\}$. In the classical formulation of the boosting problem, the coordinates of vectors $\boldsymbol{x}_i$ correspond to features computed by functions in some class of base classifiers. Boosting was initially proposed as a method for minimizing empirical classification loss (17) by iteratively updating $\boldsymbol{\theta}$. In particular, AdaBoost [FS97] corresponds to *coordinate descent* on the exponential loss function

$$\boldsymbol{\theta} \mapsto \frac{1}{n} \sum_{i=1}^{n} \exp\{-y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle\} \tag{18}$$

[Bre98, Fri01]. Notably, the minimizer of this surrogate loss does not exist in the general separable case, and there are multiple directions along which the objective decreases to 0 as $\|\boldsymbol{\theta}\| \to \infty$. The AdaBoost optimization procedure and its variants were observed empirically to shift the distribution of margins (the values $y_i \langle \boldsymbol{\theta}_t, \boldsymbol{x}_i \rangle$, $i = 1, \ldots, n$) during the optimization process in the positive direction even after empirical classification error becomes zero, which in part motivated the theory of large margin classification [SFBL98]. In the separable case, convergence to the direction of the maximizing $\ell_1$ margin solution

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{\theta}\|_1 : y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle \ge 1 \text{ for all } i \le n \right\} \tag{19}$$

was shown in [ZY05] and [Tel13] assuming small enough step size, where separability means positivity of the margin

$$\max_{\|\boldsymbol{\theta}\|_1 = 1} \min_{i \in [n]} y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle . \tag{20}$$

More recently, [SHN+18] and [JT18] have shown that gradient (rather than coordinate) descent on (18) and separable data lead to a solution with direction approaching that of the maximum $\ell_2$ (rather than $\ell_1$) margin separator

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \|\boldsymbol{\theta}\|_2 : y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle \geq 1 \text{ for all } i \leq n \right\}. \tag{21}$$

We state the next theorem from [SHN+18] for the case of logistic loss, although essentially the same statement—up to a slightly modified step size upper bound—holds for any smooth loss function that has appropriate exponential-like tail behavior, including $\ell(u) = e^{-u}$ [SHN+18, JT18].

**Theorem 3.1.** *Assume the data $\boldsymbol{X}$, $\boldsymbol{y}$ are linearly separable. For logistic loss $\ell(u) = \log(1+\exp\{-u\})$, any step size $\eta \leq 8\lambda_{max}^{-1}(n^{-1}\boldsymbol{X}^\intercal\boldsymbol{X})$, and any initialization $\boldsymbol{\theta}_0$, the gradient descent iterations*

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \widehat{L}(\boldsymbol{\theta}_t), \quad \widehat{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle)$$

*satisfy $\boldsymbol{\theta}_t = \widehat{\boldsymbol{\theta}} \cdot \log t + \rho_t$ where $\widehat{\boldsymbol{\theta}}$ is the $\ell_2$ max-margin solution in (21). Furthermore, the residual grows at most as $\|\rho_t\| = O(\log\log t)$, and thus*

$$\lim_{t\to\infty} \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_2} = \frac{\widehat{\boldsymbol{\theta}}}{\|\widehat{\boldsymbol{\theta}}\|_2}.$$

These results have been extended to multi-layer fully connected neural networks and convolutional neural networks (without nonlinearities) in [GLSS18b]. On the other hand, [GLSS18a] considered the implicit bias arising from other optimization procedures, including mirror descent, steepest descent, and AdaGrad, both in the case when the global minimum is attained (as for the square loss) and when the global minimizers are at infinity (as in the classification case with exponential-like tails of the loss function). We refer to [JT19] and [NLG+19] and references therein for further studies on faster rates of convergence to the direction of the max margin solution (with more aggressive time-varying step sizes) and on milder assumptions on the loss function.

In addition to the particular optimization algorithm being employed, implicit regularization arises from the choice of model parametrization. Consider re-parametrizing the least-squares objective in (13) as

$$\min_{\boldsymbol{u}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{\theta}(\boldsymbol{u}) - \boldsymbol{y}\|_2^2, \tag{22}$$

where $\boldsymbol{\theta}(\boldsymbol{u})_i = \boldsymbol{u}_i^2$ is the coordinate-wise square. [GWB+17] show that if $\boldsymbol{\theta}_\infty(\alpha)$ is the limit point of gradient flow on (22) with initialization $\alpha\mathbf{1}$ and the limit $\widehat{\boldsymbol{\theta}} = \lim_{\alpha\to 0} \boldsymbol{\theta}_\infty(\alpha)$ exists and satisfies $\boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{y}$, then it must be that

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^d_+}{\operatorname{argmin}} \left\{ \|\boldsymbol{\theta}\|_1 : \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle = y_i \text{ for all } i \leq n \right\}. \tag{23}$$

In other words, in that case, gradient descent on the reparametrized problem with infinitesimally small step sizes and infinitesimally small initialization converges to the minimum $\ell_1$ norm solution in the original space. More generally, [GWB+17] and [LMZ18] proved an analogue of this statement for matrix-valued $\boldsymbol{\theta}$ and $\boldsymbol{x}_i$, establishing convergence to the minimum nuclear-norm solution under additional assumptions on the $\boldsymbol{x}_i$. The matrix version of the problem can be written as

$$\min_{\boldsymbol{U},\boldsymbol{V}} \sum_{i=1}^{n} \ell(\langle \boldsymbol{U}\boldsymbol{V}^\intercal, \boldsymbol{x}_i \rangle, y_i),$$

which can be viewed, in turn, as an empirical risk minimization objective for a two-layer neural network with linear activation functions.

In summary, in overparametrized problems that admit multiple minimizers of the empirical objective, the choice of the optimization method and the choice of parametrization both play crucial roles in selecting a minimizer with certain properties. As we show in the next section, these properties of the solution can ensure good generalization properties through novel mechanisms that go beyond the realm of uniform convergence.

# 4 Benign overfitting

We now turn our attention to generalization properties of specific solutions that interpolate training data. As emphasized in Section 2, mechanisms of uniform convergence alone cannot explain good statistical performance of such methods, at least in the presence of noise.

For convenience, in this section we focus our attention on regression problems with square loss $\ell(f(\boldsymbol{x}), y) = (f(\boldsymbol{x}) - y)^2$. In this case, the regression function $f^* = \mathbb{E}[y|\boldsymbol{x}]$ is a minimizer of $L(f)$, and excess loss can be written as

$$L(f) - L(f^*) = \mathbb{E}(f(\boldsymbol{x}) - f^*(\boldsymbol{x}))^2 = \|f - f^*\|^2_{L^2(\mathbb{P})}.$$

We assume that for any $\boldsymbol{x}$, conditional variance of the noise $\xi = y - f^*(\boldsymbol{x})$ is at most $\sigma_\xi^2$, and we write $\xi_i = y_i - f^*(\boldsymbol{x}_i)$.

As in the previous section, we say that a solution $\widehat{f}$ is *interpolating* if

$$\widehat{f}(\boldsymbol{x}_i) = y_i, \quad i = 1, \dots, n. \tag{24}$$

For learning rules $\widehat{f}$ expressed in closed form—such as local methods and linear and kernel regression—it is convenient to employ a bias-variance decomposition that is different from the approximation-estimation error decomposition (3) in Section 2. First, for $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^\intercal \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} = [y_1, \dots, y_n]^\intercal$, conditionally on $\boldsymbol{X}$, define

$$\widehat{\mathrm{BIAS}}^2 = \mathbb{E}_{\boldsymbol{x}} \left( f^*(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{y}} \widehat{f}(\boldsymbol{x}) \right)^2, \qquad \widehat{\mathrm{VAR}} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left( \widehat{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{y}} \widehat{f}(\boldsymbol{x}) \right)^2. \tag{25}$$

It is easy to check that

$$\mathbb{E}\|\widehat{f} - f^*\|^2_{L^2(\mathbb{P})} = \mathbb{E}_{\boldsymbol{X}} \left[ \widehat{\mathrm{BIAS}}^2 \right] + \mathbb{E}_{\boldsymbol{X}} \left[ \widehat{\mathrm{VAR}} \right]. \tag{26}$$

In this section we consider linear (in $\boldsymbol{y}$) estimators of the form $\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^n y_i \omega_i(\boldsymbol{x})$. For such estimators we have

$$\widehat{\mathrm{BIAS}}^2 = \mathbb{E}_{\boldsymbol{x}} \left( f^*(\boldsymbol{x}) - \sum_{i=1}^n f^*(\boldsymbol{x}_i) \omega_i(\boldsymbol{x}) \right)^2 \tag{27}$$

and

$$\widehat{\mathrm{VAR}} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\xi}} \left( \sum_{i=1}^n \xi_i \omega_i(\boldsymbol{x}) \right)^2 \le \sigma_\xi^2 \sum_{i=1}^n \mathbb{E}_{\boldsymbol{x}} \left( \omega_i(\boldsymbol{x}) \right)^2, \tag{28}$$

with equality if conditional noise variances are equal to $\sigma_\xi^2$ at each $\boldsymbol{x}$.

In classical statistics, the balance between bias and variance is achieved by tuning an explicit parameter. Before diving into the more unexpected interpolation results, where the behavior of bias and variance are driven by novel self-regularization phenomena, we discuss the bias-variance tradeoff in the context of one of the oldest statistical methods.

## 4.1 Local methods: Nadaraya-Watson

Consider arguably the simplest nontrivial interpolation procedure, the 1-nearest neighbour (1-NN) $\widehat{f}(\boldsymbol{x}) = y_{\mathsf{nn}(\boldsymbol{x})}$, where $\mathsf{nn}(\boldsymbol{x})$ is the index of the datapoint closest to $\boldsymbol{x}$ in Euclidean distance. While we could view $\widehat{f}$ as an empirical minimizer in some effective class $\mathcal{F}$ of possible functions (as a union for all possible $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$), this set is large and growing with $n$. Exploiting the particular form of 1-NN is, obviously, crucial. Since typical distances to the nearest neighbor in $\mathbb{R}^d$ decay as $n^{-1/d}$ for i.i.d. data, in the noiseless case ($\sigma_\xi = 0$) one can guarantee consistency and nonparametric rates of convergence of this interpolation procedure under continuity and smoothness assumptions on $f^*$ and the underlying measure. Perhaps more interesting is the case when the $\xi_i$ have non-vanishing variance. Here 1-NN is no longer consistent in general

(as can be easily seen by taking $f^* = 0$ and independent Rademacher $\xi_i$ at random $\boldsymbol{x}_i \in [0,1]$), although its asymptotic risk is at most $2L(f^*)$ [CH67]. The reason for inconsistency is insufficient averaging of the $y$-values, and this deficiency can be addressed by averaging over the $k$ nearest neighbors with $k$ growing with $n$. Classical smoothing methods generalize this idea of local averaging; however, averaging forgoes empirical fit to data in favor of estimating the regression function under smoothness assumptions. While this has been the classical view, estimation is not necessarily at odds with fitting the training data for these local methods, as we show next.

The Nadaraya-Watson (NW) smoothing estimator [Nad64, Wat64] is defined as

$$\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} y_i \omega_i(\boldsymbol{x}), \qquad \omega_i(\boldsymbol{x}) = \frac{K((\boldsymbol{x} - \boldsymbol{x}_i)/h)}{\sum_{j=1}^{n} K((\boldsymbol{x} - \boldsymbol{x}_j)/h)}, \tag{29}$$

where $K(u) : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a kernel and $h > 0$ is a bandwidth parameter. For standard kernels used in practice—such as the Gaussian, uniform, or Epanechnikov kernels—the method averages the $y$-values in a local neighborhood around $\boldsymbol{x}$, and, in general, does not interpolate. However, as noted by [DGK98], a kernel that is singular at 0 does interpolate the data. While the Hilbert kernel $K(u) = \|u\|_2^{-d}$, suggested in [DGK98], does not enjoy non-asymptotic rates of convergence, its truncated version

$$K(u) = \|u\|_2^{-a} \, \mathbf{1}\left[\|u\|_2 \leq 1\right], \; u \in \mathbb{R}^d \tag{30}$$

with a smaller power $0 < a < d/2$ was shown in [BRT19] to lead to minimax optimal rates of estimation under the corresponding smoothness assumptions. Notably, the NW estimator with the kernel in (30) is necessarily interpolating the training data for any choice of $h$.

Before stating the formal result, define the Hölder class $H(\beta, L)$, for $\beta \in (0,1]$, as the class of functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d, \qquad |f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq L \, \|\boldsymbol{x} - \boldsymbol{x}'\|_2^{\beta} .$$

The following result appears in [BRT19]; see also [BHM18]:

**Theorem 4.1.** *Let $f^* \in H(\beta, L)$ for $\beta \in (0,1]$ and $L > 0$. Suppose the marginal density $p$ of $\boldsymbol{x}$ satisfies $0 < p_{min} \leq p(\boldsymbol{x}) \leq p_{max}$ for all $\boldsymbol{x}$ in its support. Then the estimator (29) with kernel (30) satisfies*[3]

$$\mathbb{E}_{\boldsymbol{X}}\left[\widehat{\mathrm{BIAS}}^2\right] \lesssim h^{2\beta}, \qquad \mathbb{E}_{\boldsymbol{X}}\left[\widehat{\mathrm{VAR}}\right] \lesssim \sigma_\xi^2 (nh^d)^{-1} . \tag{31}$$

The result can be extended to smoothness parameters $\beta > 1$ [BRT19]. The choice of $h = n^{-1/(2\beta+d)}$ balances the two terms and leads to minimax optimal rates for Hölder classes [Tsy08].

In retrospect, Theorem 4.1 should not be surprising, and we mention it here for pedagogical purposes. It should be clear from the definition (29) that the behavior of the kernel at 0, and in particular the presence of a singularity, determines whether the estimator fits the training data exactly. This is, however, decoupled from the level of smoothing, as given by the bandwidth parameter $h$. In particular, it is the choice of $h$ alone that determines the bias-variance tradeoff, and the value of the empirical loss cannot inform us whether the estimator is over-smoothing or under-smoothing the data.

The NW estimator with the singular kernel can be also viewed as adding small "spikes" at the datapoints on top of the general smooth estimate that arises from averaging the data in a neighborhood of radius $h$. This suggests a rather obvious scheme for changing any estimator $\widehat{f}_0$ into an interpolating one by adding small deviations around the datapoints: $\widehat{f}(\boldsymbol{x}) := \widehat{f}_0(\boldsymbol{x}) + \Delta(\boldsymbol{x})$, where $\Delta(\boldsymbol{x}_j) = y_i - \widehat{f}_0(\boldsymbol{x}_j)$ but $\|\Delta\|_{L^2(\mathbb{P})} = o(1)$. The component $\widehat{f}_0$ is useful for prediction because it is smooth, whereas the spiky component $\Delta$ is useful for interpolation but does not harm the predictions of $\widehat{f}$. Such combinations have been observed experimentally in other settings and described as "spiked-smooth" estimates [WOBM17]. The examples that we see below suggest that interpolation may be easier to achieve with high-dimensional data than with low-dimensional data, and this is consistent with the requirement that the overfitting component $\Delta$ is benign: it need not be too "irregular" in high dimensions, since typical distances between datapoints in $\mathbb{R}^d$ scale at least as $n^{-1/d}$.

[3] In the remainder of this paper, the symbol $\lesssim$ denotes inequality up to a multiplicative constant.

## 4.2 Linear regression in the interpolating regime

In the previous section, we observed that the spiky part of the NW estimator, which is responsible for interpolation, does not hurt the out-of-sample performance when measured in $L^2(\mathbb{P})$. The story for minimum-norm interpolating linear and kernel regression is significantly more subtle: there is also a decomposition into a prediction component and an overfitting component, but there is no explicit parameter that trades off bias and variance. The decomposition depends on the distribution of the data, and the overfitting component provides a *self-induced regularization*[4], similar to the regularization term in ridge regression (16), and this determines the bias-variance trade-off.

Consider the problem of linear regression in the over-parametrized regime. We assume that the regression function $f^*(\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta}^*) = \langle \boldsymbol{\theta}^*, \boldsymbol{x}\rangle$ with $\boldsymbol{\theta}^*, \boldsymbol{x} \in \mathbb{R}^d$. We also assume $\mathbb{E}\boldsymbol{x} = 0$. (While we present the results for finite $d > n$, all the statements in this section hold for separable Hilbert spaces of infinite dimension.)

It is easy to see that the excess square loss can be written as

$$L(\widehat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) = \mathbb{E}\left(f(\widehat{\boldsymbol{\theta}}) - f(\boldsymbol{\theta}^*)\right)^2 = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2,$$

where we write $\|\boldsymbol{v}\|_{\Sigma}^2 := \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}$ and $\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top$. Since $d > n$, there is not enough data to learn all the $d$ directions of $\boldsymbol{\theta}^*$ reliably, unless $\boldsymbol{\Sigma}$ has favorable spectral properties. To take advantage of such properties, classical methods—as described in Section 2—resort to explicit regularization (shrinkage) or model complexity control, which inevitably comes at the expense of not fitting the noisy data exactly. In contrast, we are interested in estimates that interpolate the data. Motivated by the properties of the gradient descent method (13), we consider the minimal norm linear function that fits the data $\boldsymbol{X}$, $\boldsymbol{y}$ exactly:

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{\theta}\|_2 : \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle = y_i \text{ for all } i \le n \right\}. \tag{32}$$

The solution has a closed form and yields the estimator

$$\widehat{f}(\boldsymbol{x}) = \langle \widehat{\boldsymbol{\theta}}, \boldsymbol{x}\rangle = \langle \boldsymbol{X}^\dagger \boldsymbol{y}, \boldsymbol{x}\rangle = (\boldsymbol{X}\boldsymbol{x})^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}, \tag{33}$$

which can also be written as $\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^n y_i \omega_i(\boldsymbol{x})$, with

$$\omega_i(\boldsymbol{x}) = (\boldsymbol{x}^\top \boldsymbol{X}^\dagger)_i = (\boldsymbol{X}\boldsymbol{x})^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{e}_i. \tag{34}$$

Thus, from (27), the bias term can be written as

$$\widehat{\text{BIAS}}^2 = \mathbb{E}_{\boldsymbol{x}} \left\langle P^\perp \boldsymbol{x}, \boldsymbol{\theta}^* \right\rangle^2 = \left\| \boldsymbol{\Sigma}^{1/2} P^\perp \boldsymbol{\theta}^* \right\|_2^2, \tag{35}$$

where $P^\perp = \mathbf{I}_d - \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}$, and from (28), the variance term is

$$\widehat{\text{VAR}} \le \sigma_\xi^2 \cdot \mathbb{E}_{\boldsymbol{x}} \left\| (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}(\boldsymbol{X}\boldsymbol{x}) \right\|_2^2 = \sigma_\xi^2 \cdot \mathsf{tr}\left( (\boldsymbol{X}\boldsymbol{X}^\top)^{-2} \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top \right). \tag{36}$$

We now state our assumptions.

**Assumption 4.2.** *Suppose $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ is 1-sub-Gaussian. Without loss of generality, assume $\boldsymbol{\Sigma} = diag(\lambda_1, \dots, \lambda_d)$ with $\lambda_1 \ge \cdots \ge \lambda_d$.*

The central question now is: Are there mechanisms that can ensure small bias and variance of the minimum-norm interpolant? Surprisingly, we shall see that the answer is yes. To this end, choose an index $k \in \{1, \dots, d\}$ and consider the subspace spanned by the top $k$ eigenvectors corresponding to $\lambda_1, \dots, \lambda_k$. Write $\boldsymbol{x}^\top = [\boldsymbol{x}_{\le k}^\top, \boldsymbol{x}_{>k}^\top]$. For an appropriate choice of $k$, it turns out the decomposition of the minimum-norm

[4]This is not to be confused with *implicit regularization*, discussed in Section 3, which describes the properties of the particular empirical risk minimizer that results from the choice of an optimization algorithm. Self-induced regularization is a statistical property that also depends on the data-generating mechanism.

interpolant as $\langle \widehat{\boldsymbol{\theta}}, \boldsymbol{x} \rangle = \langle \widehat{\boldsymbol{\theta}}_{\leq k}, \boldsymbol{x}_{\leq k} \rangle + \langle \widehat{\boldsymbol{\theta}}_{>k}, \boldsymbol{x}_{>k} \rangle$ corresponds to a decomposition into a prediction component and an interpolation component. Write the data matrix as $\boldsymbol{X} = [\boldsymbol{X}_{\leq k}, \boldsymbol{X}_{>k}]$ and

$$\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}} = \boldsymbol{X}_{\leq k}\boldsymbol{X}_{\leq k}^{\mathsf{T}} + \boldsymbol{X}_{>k}\boldsymbol{X}_{>k}^{\mathsf{T}}. \tag{37}$$

Observe that if the eigenvalues of the second part were to be contained in an interval $[\gamma/c, c\gamma]$ for some $\gamma$ and a constant $c$, we could write

$$\boldsymbol{X}_{\leq k}\boldsymbol{X}_{\leq k}^{\mathsf{T}} + \gamma \boldsymbol{M}, \tag{38}$$

where $c^{-1}\mathbf{I}_n \preceq \boldsymbol{M} \preceq c\mathbf{I}_n$. If we replace $\boldsymbol{M}$ with the approximation $\mathbf{I}_n$ and substitute this expression into (33), we see that $\gamma$ would have an effect similar to *explicit* regularization through a ridge penalty: if that approximation were precise, the first $k$ components of $\widehat{\boldsymbol{\theta}}$ would correspond to

$$\widehat{\boldsymbol{\theta}}_{\leq k} = \underset{\boldsymbol{\theta} \in \mathbb{R}^k}{\operatorname{argmin}} \ \|\boldsymbol{X}_{\leq k}\, \boldsymbol{\theta} - \boldsymbol{y}\|_2^2 + \gamma \|\boldsymbol{\theta}\|_2^2 , \tag{39}$$

since this has the closed-form solution $\boldsymbol{X}_{\leq k}^{\mathsf{T}}(\boldsymbol{X}_{\leq k}\boldsymbol{X}_{\leq k}^{\mathsf{T}} + \gamma \mathbf{I}_n)^{-1}\boldsymbol{y}$. Thus, if $\gamma$ is not too large, we might expect this approximation to have a minimal impact on the bias and variance of the prediction component.

It is, therefore, natural to ask when to expect such a near-isotropic behavior arising from the "tail" features. The following lemma provides an answer to this question [BLLT20]:

**Lemma 4.3.** *Suppose coordinates of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ are independent. Then there exists a constant $c > 0$ such that, with probability at least $1 - 2\exp\{-n/c\}$,*

$$\frac{1}{c}\sum_{i>k} \lambda_i - c\lambda_{k+1} n \leq \lambda_{min}(\boldsymbol{X}_{>k}\boldsymbol{X}_{>k}^{\mathsf{T}}) \leq \lambda_{max}(\boldsymbol{X}_{>k}\boldsymbol{X}_{>k}^{\mathsf{T}}) \leq c\left(\sum_{i>k} \lambda_i + \lambda_{k+1} n\right).$$

The condition of independence of coordinates in Lemma 4.3 is satisfied for Gaussian $\boldsymbol{x}$. It can be relaxed to the following small-ball assumption:

$$\exists c > 0 : \ \mathbb{P}(c\,\|\boldsymbol{x}\|_2^2 \geq \mathbb{E}\,\|\boldsymbol{x}\|_2^2) \geq 1 - \delta. \tag{40}$$

Under this assumption, the conclusion of Lemma 4.3 still holds with probability at least $1 - 2\exp\{-n/c\} - n\delta$ [TB20].

An appealing consequence of Lemma 4.3 is the small condition number of $\boldsymbol{X}_{>k}\boldsymbol{X}_{>k}^{\mathsf{T}}$ for any $k$ such that $\sum_{i>k} \lambda_i \gtrsim \lambda_{k+1} n$. Define the *effective rank* for a given index $k$ by

$$r_k(\boldsymbol{\Sigma}) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}.$$

We see that $r_k(\boldsymbol{\Sigma}) \geq bn$ for some constant $b$ implies that the set of eigenvalues of $\boldsymbol{X}_{>k}\boldsymbol{X}_{>k}^{\mathsf{T}}$ lies in the interval $[\gamma/c, c\gamma]$ for

$$\gamma = \sum_{i>k} \lambda_i,$$

and thus the scale of the self-induced regularization in (38) is the sum of the tail eigenvalues of the covariance operator. Interestingly, the reverse implication also holds: if for some $k$ the condition number of $\boldsymbol{X}_{>k}\boldsymbol{X}_{>k}^{\mathsf{T}}$ is at most $\kappa$ with probability at least $1 - \delta$, then effective rank $r_k(\boldsymbol{\Sigma})$ is at least $c_\kappa n$ with probability at least $1 - \delta - c\exp\{-n/c\}$ for some constants $c, c_\kappa$. Therefore, the condition $r_k(\boldsymbol{\Sigma}) \gtrsim n$ characterizes the indices $k$ such that $\boldsymbol{X}_{>k}\boldsymbol{X}_{>k}^{\mathsf{T}}$ behaves as a scaling of $\mathbf{I}_d$, and the scaling is proportional to $\sum_{i>k} \lambda_i$. We may call the smallest such index $k$ *the effective dimension*, for reasons that will be clear in a bit.

How do the estimates on tail eigenvalues help in controlling the variance of the minimum-norm interpolant? Define

$$\boldsymbol{\Sigma}_{\le k} = \mathrm{diag}(\lambda_1, \dots, \lambda_k), \quad \boldsymbol{\Sigma}_{>k} = \mathrm{diag}(\lambda_{k+1}, \dots, \lambda_d).$$

Then, omitting $\sigma_\xi^2$ for the moment, the variance upper bound in (36) can be estimated by

$$\begin{aligned}\mathsf{tr}\left((\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-2}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\mathsf{T}\right) \lesssim \mathsf{tr}\left((\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-2}\boldsymbol{X}_{\le k}\boldsymbol{\Sigma}_{\le k}\boldsymbol{X}_{\le k}^\mathsf{T}\right) \\ + \mathsf{tr}\left((\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-2}\boldsymbol{X}_{>k}\boldsymbol{\Sigma}_{>k}\boldsymbol{X}_{>k}^\mathsf{T}\right).\end{aligned} \tag{41}$$

The first term is further upper-bounded by

$$\mathsf{tr}\left((\boldsymbol{X}_{\le k}\boldsymbol{X}_{\le k}^\mathsf{T})^{-2}\boldsymbol{X}_{\le k}\boldsymbol{\Sigma}_{\le k}\boldsymbol{X}_{\le k}^\mathsf{T}\right), \tag{42}$$

and its expectation corresponds to the variance of $k$-dimensional regression, which is of the order of $k/n$. On the other hand, by Bernstein's inequality, with probability at least $1 - 2\exp^{-cn}$,

$$\mathsf{tr}(\boldsymbol{X}_{>k}\boldsymbol{\Sigma}_{>k}\boldsymbol{X}_{>k}^\mathsf{T}) \lesssim n\sum_{i>k}\lambda_i^2, \tag{43}$$

so we have that the second term in (41) is, with high probability, of order at most

$$\frac{n\sum_{i>k}\lambda_i^2}{(\sum_{i>k}\lambda_i)^2}.$$

Putting these results together, we have the following theorem [TB20]:

**Theorem 4.4.** *Fix $\delta < 1/2$. Under Assumption 4.2, suppose for some $k$ the condition number of $\boldsymbol{X}_{>k}\boldsymbol{X}_{>k}^\mathsf{T}$ is at most $\kappa$ with probability at least $1-\delta$. Then*

$$\widehat{\mathrm{VAR}} \lesssim \sigma_\xi^2\kappa^2\log\left(\frac{1}{\delta}\right)\left(\frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{(\sum_{i>k}\lambda_i)^2}\right) \tag{44}$$

*with probability at least $1-2\delta$.*

We now turn to the analysis of the bias term. Since the projection operator in (35) annihilates any vector in the span of the rows of $\boldsymbol{X}$, we can write

$$\widehat{\mathrm{BIAS}}^2 = \left\|\boldsymbol{\Sigma}^{1/2}P^\perp\boldsymbol{\theta}^*\right\|_2^2 = \left\|(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}})^{1/2}P^\perp\boldsymbol{\theta}^*\right\|_2^2, \tag{45}$$

where $\widehat{\boldsymbol{\Sigma}} = n^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ is the sample covariance operator. Since projection contracts distances, we obtain an upper bound

$$\left\|(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}})^{1/2}\boldsymbol{\theta}^*\right\|_2^2 \le \|\boldsymbol{\theta}^*\|_2^2 \times \left\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}\right\|. \tag{46}$$

The rate of approximation of the covariance operator by its sample-based counterpart has been studied in [KL17], and we conclude

$$\mathrm{BIAS}^2 \lesssim \|\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}}^2 \max\left\{\sqrt{\frac{r_0(\boldsymbol{\Sigma})}{n}}, \frac{r_0(\boldsymbol{\Sigma})}{n}\right\} \tag{47}$$

(see [BLLT20] for details).

The upper bound in (47) can be sharpened significantly by analyzing the bias in the two subspaces, as proved in [TB20]:

**Theorem 4.5.** *Under the assumptions of Theorem 4.4, for $n \gtrsim \log(1/\delta)$, with probability at least $1 - 2\delta$,*

$$\widehat{\mathrm{BIAS}}^2 \lesssim \kappa^4 \left[ \left\| \boldsymbol{\theta}^*_{\leq k} \right\|^2_{\boldsymbol{\Sigma}^{-1}_{\leq k}} \left( \frac{\sum_{i>k} \lambda_i}{n} \right)^2 + \left\| \boldsymbol{\theta}^*_{>k} \right\|^2_{\boldsymbol{\Sigma}_{>k}} \right]. \tag{48}$$

The following result shows that without further assumptions, the bounds on variance and bias given in Theorems 4.4 and 4.5 cannot be improved by more than constant factors; see [BLLT20] and [TB20].

**Theorem 4.6.** *There are absolute constants $b$ and $c$ such that for Gaussian $\boldsymbol{x} \sim \mathsf{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$, with probability at least $1 - \exp(-n/c)$,*

$$\widehat{\mathrm{VAR}} \gtrsim 1 \wedge \left( \sigma_\xi^2 \left( \frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{(\sum_{i>k} \lambda_i)^2} \right) \right),$$

*where $k$ is the effective dimension, $k = \min\{l : r_l(\boldsymbol{\Sigma}) \geq bn\}$. Furthermore, for any $\boldsymbol{\theta} \in \mathbb{R}^d$, if the regression function $f^*(\cdot) = \langle \cdot, \boldsymbol{\theta}^* \rangle$, where $\theta^*_i = \epsilon_i \theta_i$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_d) \sim \mathsf{Unif}\left(\{\pm 1\}^d\right)$, then with probability at least $1 - \exp(-n/c)$,*

$$\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\mathrm{BIAS}}^2 \gtrsim \left[ \left\| \boldsymbol{\theta}^*_{\leq k} \right\|^2_{\boldsymbol{\Sigma}^{-1}_{\leq k}} \left( \frac{\sum_{i>k} \lambda_i}{n} \right)^2 + \left\| \boldsymbol{\theta}^*_{>k} \right\|^2_{\boldsymbol{\Sigma}_{>k}} \right].$$

A discussion of Theorems 4.4, 4.5 and 4.6 is in order. First, the upper and lower bounds match up to constants, and in particular both involve the decomposition of $\widehat{f}$ into a prediction component $\widehat{f}_0(\boldsymbol{x}) := \langle \widehat{\boldsymbol{\theta}}_{\leq k}, \boldsymbol{x}_{\leq k} \rangle$ and an interpolation component $\Delta(\boldsymbol{x}) := \langle \widehat{\boldsymbol{\theta}}_{>k}, \boldsymbol{x}_{>k} \rangle$ with distinct bias and variance contributions, so this decomposition is not an artifact of our analysis. Second, the $\left\| \boldsymbol{\theta}^*_{>k} \right\|^2_{\boldsymbol{\Sigma}_{>k}}$ term in the bias and the $k/n$ term in the variance for the prediction component $\widehat{f}_0$ correspond to the terms we would get by performing ordinary least-squares (OLS) restricted to the first $k$ coordinates of $\boldsymbol{\theta}$. Provided $k$ is small compared to $n$, there is enough data to estimate the signal in this $k$-dimensional component, and the bias contribution is the approximation error due to truncation at $k$. The other aspect of the interpolating component $\Delta$ that could harm prediction accuracy is its variance term. The definition of the effective dimension $k$ implies that this is no more than a constant, and it is small if the tail eigenvalues decay slowly and $d - k \gg n$, for in that case, the ratio of the squared $\ell_1$ norm to the squared $\ell_2$ norm of these eigenvalues is large compared to $n$; overparametrization is important. Finally, the bias and variance terms are similar to those that arise in ridge regression (16), with the regularization coefficient determined by the self-induced regularization. Indeed, define

$$\lambda = \frac{b}{n} \sum_{i>k} \lambda_i \tag{49}$$

for the constant $b$ in the definition of the effective dimension $k$. That definition implies that $\lambda_k \geq \lambda \geq \lambda_{k+1}$, so we can write the bias and variance terms, within constant factors, as

$$\widehat{\mathrm{BIAS}}^2 \approx \sum_{i=1}^d \theta_i^{*2} \frac{\lambda_i}{(1 + \lambda_i/\lambda)^2}, \qquad \widehat{\mathrm{VAR}} \approx \frac{\sigma_\xi^2}{n} \sum_{i=1}^d \left( \frac{\lambda_i}{\lambda + \lambda_i} \right)^2.$$

These are reminiscent of the bias and variance terms that arise in ridge regression (16). Indeed, a ridge regression estimate in a fixed design setting with $\boldsymbol{X}^\mathsf{T} \boldsymbol{X} = \mathrm{diag}(s_1, \ldots, s_d)$ has precisely these bias and variance terms with $\lambda_i$ replaced by $s_i$; see, for example, [DFKU13, Lemma 1]. In Section 4.3.3, we shall see the same bias-variance decomposition arise in a related setting, but with the dimension growing with sample size.

## 4.3 Linear regression in Reproducing Kernel Hilbert Spaces

Kernel methods are among the core algorithms in machine learning and statistics. These methods were introduced to machine learning in the pioneering work of [ABR64] as a generalization of the Perceptron algorithm to nonlinear functions by lifting the $\boldsymbol{x}$-variable to a high- or infinite-dimensional feature space. Our interest in studying kernel methods here is two-fold: on the one hand, as discussed in detail in Sections 5 and 6, sufficiently wide neural networks with random initialization stay close to a certain kernel-based solution during optimization and are essentially equivalent to a minimum-norm interpolant; on the other hand, it has been noted that kernel methods exhibit similar surprising behavior of benign interpolation to neural networks [BMM18].

A kernel method in the regression setting amounts to choosing a feature map $\boldsymbol{x} \mapsto \phi(\boldsymbol{x})$ and computing a (regularized) linear regression solution in the feature space. While Section 4.2 already addressed the question of overparametrized linear regression, the non-linear feature map $\phi(\boldsymbol{x})$ might not satisfy Assumption 4.2. In this section, we study interpolating RKHS regression estimates using a more detailed analysis of certain random kernel matrices.

Since the linear regression solution involves inner products of $\phi(\boldsymbol{x})$ and $\phi(\boldsymbol{x}')$, the feature maps do not need to be computed explicitly. Instead, kernel methods rely on a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that, in turn, corresponds to an RKHS $\mathcal{H}$. A classical method is kernel ridge regression (KRR)

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1} (f(\boldsymbol{x}_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 , \tag{50}$$

which has been extensively analyzed through the lens of bias-variance tradeoff with an appropriately tuned parameter $\lambda > 0$ [CDV07]. As $\lambda \to 0^+$, we obtain a minimum-norm interpolant

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{H}} \Big\{ \|f\|_{\mathcal{H}} : f(\boldsymbol{x}_i) = y_i \text{ for all } i \le n \Big\}, \tag{51}$$

which has the closed-form solution

$$\widehat{f}(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{X})^\top K(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{y}, \tag{52}$$

assuming $K(\boldsymbol{X}, \boldsymbol{X})$ is invertible; see (32) and (33). Here $K(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{n \times n}$ is the kernel matrix with

$$[K(\boldsymbol{X}, \boldsymbol{X})]_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) \quad \text{and} \quad K(\boldsymbol{x}, \boldsymbol{X}) = [k(\boldsymbol{x}, \boldsymbol{x}_1), \dots, k(\boldsymbol{x}, \boldsymbol{x}_n)]^\top.$$

Alternatively, we can write the solution as

$$\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} y_i \omega_i(\boldsymbol{x}) \quad \text{with} \quad \omega_i(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{X}) K(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{e}_i,$$

which makes it clear that $\omega_i(\boldsymbol{x}_j) = \mathbf{1}[i = j]$. We first describe a setting where this approach does not lead to benign overfitting.

### 4.3.1 The Laplace kernel with constant dimension

We consider the Laplace (exponential) kernel on $\mathbb{R}^d$ with parameter $\sigma > 0$:

$$k_\sigma(\boldsymbol{x}, \boldsymbol{x}') = \sigma^{-d} \exp\{- \|\boldsymbol{x} - \boldsymbol{x}'\|_2 / \sigma\}.$$

The RKHS norm corresponding to this kernel can be related to a Sobolev norm, and its RKHS has been shown [Bac17, GYK+20, CX21] to be closely related to the RKHS corresponding to the Neural Tangent Kernel (NTK), which we study in Section 6.

To motivate the lower bound, consider $d = 1$. In this case, the minimum-norm solution with the Laplace kernel corresponds to a rope hanging from nails at heights $y_i$ and locations $\boldsymbol{x}_i \in \mathbb{R}$. If points are ordered

$\boldsymbol{x}_{(1)} \le \boldsymbol{x}_{(2)} \le \ldots \le \boldsymbol{x}_{(n)}$, the form of the minimum-norm solution between two adjacent points $\boldsymbol{x}_{(i)}, \boldsymbol{x}_{(i+1)}$ is only affected by the values $y_{(i)}, y_{(i+1)}$ at these locations. As $\sigma \to \infty$, the interpolant becomes piece-wise linear, while for $\sigma \to 0$, the solution is a sum of spikes at the datapoints and zero everywhere else. In both cases, the interpolant is not consistent: the error $\mathbb{E}\|\widehat{f} - f^*\|^2_{L^2(\mathbb{P})}$ does not converge to 0 as $n$ increases. Somewhat surprisingly, there is no choice of $\sigma$ that can remedy the problem, even if $\sigma$ is chosen in a data-dependent manner.

The intuition carries over to the more general case, as long as $d$ is a constant. The following theorem appears in [RZ19]:

**Theorem 4.7.** *Suppose $f^*$ is a smooth function defined on a unit ball in $\mathbb{R}^d$. Assume the probability distribution of $\boldsymbol{x}$ has density that is bounded above and away from 0. Suppose the noise random variables $\xi_i$ are independent Rademacher.*[5] *For fixed $n$ and odd $d$, with probability at least $1 - O(n^{-1/2})$, for any choice $\sigma > 0$,*

$$\|\widehat{f} - f^*\|^2_{L^2(\mathbb{P})} = \Omega_d(1).$$

Informally, the minimum-norm interpolant with the Laplace kernel does not have the flexibility to both estimate the regression function and generate interpolating spikes with small $L^2(\mathbb{P})$ norm if the dimension $d$ is small. For high-dimensional data, however, minimum-norm interpolation with the same kernel can be more benign, as we see in the next section.

### 4.3.2 Kernels on $\mathbb{R}^d$ with $d \asymp n^\alpha$

Since $d = O(1)$ may lead to inconsistency of the minimum-norm interpolator, we consider here a scaling $d \asymp n^\alpha$ for $\alpha \in (0, 1]$. Some assumption on the independence of coordinates is needed to circumvent the lower bound of the previous section, and we assume the simplest possible scenario: each coordinate of $\boldsymbol{x} \in \mathbb{R}^d$ is independent.

**Assumption 4.8.** *Assume that $\boldsymbol{x} \sim \mathbb{P} = p^{\otimes d}$ such that $z \sim p$ is mean-zero, that for some $C > 0$ and $\nu > 1$, $\mathbb{P}(|z| \ge t) \le C(1+t)^{-\nu}$ for all $t \ge 0$, and that $p$ does not contain atoms.*

We only state the results for the inner-product kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = h\left(\frac{\langle \boldsymbol{x}, \boldsymbol{x}' \rangle}{d}\right), \qquad h(t) = \sum_{i=0}^{\infty} \alpha_i t^i, \qquad \alpha_i \ge 0$$

and remark that more general rotationally invariant kernels (including NTK: see Section 6) exhibit the same behavior under the independent-coordinate assumption [LRZ20].

For brevity, define $\boldsymbol{K} = n^{-1} K(\boldsymbol{X}, \boldsymbol{X})$. Let $\boldsymbol{r} = (r_1, \cdots, r_d) \ge 0$ be a multi-index, and write $\|\boldsymbol{r}\| = \sum_{i=1}^d r_i$. With this notation, each entry of the kernel matrix can be expanded as

$$n\boldsymbol{K}_{ij} = \sum_{\iota=0}^{\infty} \alpha_\iota \left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d}\right)^\iota = \sum_{\boldsymbol{r}} c_{\boldsymbol{r}} \alpha_{\|\boldsymbol{r}\|} p_{\boldsymbol{r}}(\boldsymbol{x}_i) p_{\boldsymbol{r}}(\boldsymbol{x}_j) / d^{\|\boldsymbol{r}\|}$$

with

$$c_{\boldsymbol{r}} = \frac{(r_1 + \cdots + r_d)!}{r_1! \cdots r_d!},$$

and the monomials are $p_{\boldsymbol{r}}(\boldsymbol{x}_i) = (\boldsymbol{x}_i[1])^{r_1} \cdots (\boldsymbol{x}_i[d])^{r_d}$ . If $h$ has infinitely many positive coefficients $\alpha$, each $\boldsymbol{x}$ is lifted to an infinite-dimensional space. However, the resulting feature map $\phi(\boldsymbol{x})$ is not (in general) sub-Gaussian. Therefore, results from Section 4.2 are not immediately applicable and a more detailed analysis that takes advantage of the structure of the feature map is needed.

As before, we separate the high-dimensional feature map into two parts, one corresponding to the prediction component, and the other corresponding to the overfitting part of the minimum-norm interpolant.

[5] $\mathbb{P}(\xi_i = \pm 1) = 1/2$.

More precisely, the truncated function $h^{\le \iota}(t) = \sum_{i=0}^{\iota} \alpha_i t^i$ leads to the degree-bounded component of the empirical kernel:

$$n\boldsymbol{K}_{ij}^{[\le \iota]} := \sum_{\|\boldsymbol{r}\| \le \iota} c_{\boldsymbol{r}} \alpha_{\|\boldsymbol{r}\|} p_{\boldsymbol{r}}(\boldsymbol{x}_i) p_{\boldsymbol{r}}(\boldsymbol{x}_j) / d^{\|\boldsymbol{r}\|}, \quad n\boldsymbol{K}^{[\le \iota]} = \Phi\Phi^\top$$

with data $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ transformed into polynomial features $\Phi \in \mathbb{R}^{n \times \binom{\iota+d}{\iota}}$ defined as

$$\Phi_{i,\boldsymbol{r}} = \left(c_{\boldsymbol{r}} \alpha_{\|\boldsymbol{r}\|}\right)^{1/2} p_{\boldsymbol{r}}(\boldsymbol{x}_i) / d^{\|\boldsymbol{r}\|/2} \ .$$

The following theorem reveals the staircase structure of the eigenvalues of the kernel, with $\Theta(d^\iota)$ eigenvalues of order $\Omega(d^{-\iota})$, as long as $n$ is large enough to sketch these directions; see [LRZ20] and [GMMM20a].

**Theorem 4.9.** *Suppose $\alpha_0, \dots, \alpha_{\iota_0} > 0$ and $d^{\iota_0} \log d = o(n)$. Under Assumption 4.8, with probability at least $1 - \exp^{-\Omega(n/d^{\iota_0})}$, for any $\iota \le \iota_0$, $\boldsymbol{K}^{[\le \iota]}$ has $\binom{\iota+d}{\iota}$ nonzero eigenvalues, all of them larger than $Cd^{-\iota}$ and the range of $\boldsymbol{K}^{[\le \iota]}$ is the span of*

$$\{(p(\boldsymbol{x}_1), \dots, p(\boldsymbol{x}_n)) : p \quad \textit{multivariable polynomial of degree at most } \iota\} \, .$$

The component $\boldsymbol{K}^{[\le \iota]}$ of the kernel matrix sketches the low-frequency component of the signal in much the same way as the corresponding $\boldsymbol{X}_{\le k}\boldsymbol{X}_{\le k}^\top$ in linear regression sketches the top $k$ directions of the population distribution (see Section 4.2).

Let us explain the key ideas behind the proof of Theorem 4.9. In correspondence with the sample covariance operator $n^{-1}\boldsymbol{X}_{\le k}^\top \boldsymbol{X}_{\le k}$ in the linear case, we define the sample covariance operator $\Theta^{[\le \iota]} := n^{-1}\Phi^\top\Phi$. If the monomials $p_{\boldsymbol{r}}(\boldsymbol{x})$ were orthogonal in $L^2(\mathbb{P})$, then we would have:

$$\mathbb{E}\left[\Theta^{[\le \iota]}\right] = \operatorname{diag}(C(0), \ \cdots, \ C(\iota')d^{-\iota'}, \ \cdots, \ \underbrace{C(\iota)d^{-\iota}}_{\binom{d+\iota-1}{d-1} \text{ such entries}} \quad )$$

where $C(\iota)$ denotes constants that depend on $\iota$. Since under our general assumptions on the distribution this orthogonality does not necessarily hold, we employ the Gram-Schmidt process on the basis $\{1, t, t^2, \dots\}$ with respect to $L^2(p)$ to produce an orthogonal polynomial basis $q_0, q_1, \dots$. This yields new features

$$\Psi_{i,\boldsymbol{r}} = \left(c_{\boldsymbol{r}} \alpha_{\|\boldsymbol{r}\|}\right)^{1/2} q_{\boldsymbol{r}}(\boldsymbol{x}_i) / d^{\|\boldsymbol{r}\|/2}, \quad q_{\boldsymbol{r}}(\boldsymbol{x}) = \prod_{j \in [d]} q_{r_j}(\boldsymbol{x}[j]).$$

As shown in [LRZ20], these features are weakly dependent and the orthogonalization process does not distort the eigenvalues of the covariance matrix by more than a multiplicative constant. A small-ball method [KM15] can then be used to prove the lower bound for the eigenvalues of $\Psi\Psi^\top$ and thus establish Theorem 4.9.

We now turn to variance and bias calculations. The analogue of (36) becomes

$$\widehat{\mathrm{VAR}} \le \sigma_\xi^2 \cdot \mathbb{E}_{\boldsymbol{x}} \left\| K(\boldsymbol{X}, \boldsymbol{X})^{-1} K(\boldsymbol{X}, \boldsymbol{x}) \right\|_2^2 \tag{53}$$

and, similarly to (37), we split the kernel matrix into two parts, according to the degree $\iota$.

The following theorem establishes an upper bound on (53) [LRZ20]:

**Theorem 4.10.** *Under Assumption 4.8 and the additional assumption of sub-Gaussianity of the distribution $p$ for the coordinates of $\boldsymbol{x}$, if $\alpha_1, \dots, \alpha_\iota > 0$, there exists $\iota' \ge 2\iota + 3$ with $\alpha_{\iota'} > 0$, and $d^\iota \log d \lesssim n \lesssim d^{\iota+1}$, then with probability at least $1 - \exp^{-\Omega(n/d^\iota)}$,*

$$\widehat{\mathrm{VAR}} \lesssim \sigma_\xi^2 \cdot \left( \frac{d^\iota}{n} + \frac{n}{d^{\iota+1}} \right). \tag{54}$$

Notice that the behavior of the upper bound changes as $n$ increases from $d^\iota$ to $d^{\iota+1}$. At $d \asymp n^\iota$, variance is large since there is not enough data to reliably estimate all the $d^\iota$ directions in the feature space. As $n$ increases, variance in the first $d^\iota$ directions decreases; new directions in the data appear (those corresponding to monomials of degree $\iota+1$, with smaller population eigenvalues) but cannot be reliably estimated. This second part of (54) grows linearly with $n$, similarly to the second term in (44). The split between these two terms occurs at the effective dimension defined in Section 4.2.

Two aspects of the *multiple-descent* behavior of the upper bound (54) should be noted. First, variance is small when $d^\iota \ll n \ll d^{\iota+1}$, between the peaks; second, the valleys become deeper as $d$ becomes larger, with variance at most $d^{-1/2}$ at $n = d^{\iota+1/2}$.

We complete the discussion of this section by exhibiting one possible upper bound on the bias term [LRZ20]:

**Theorem 4.11.** *Assume the regression function can be written as*

$$f^*(\boldsymbol{x}) = \int k(\boldsymbol{x}, \boldsymbol{z}) \rho_*(\boldsymbol{z}) \mathbb{P}(d\boldsymbol{z}) \quad \textit{with} \quad \int \rho_*^4(\boldsymbol{z}) \mathbb{P}(d\boldsymbol{z}) \le c.$$

*Let Assumption 4.8 hold, and suppose* $\sup_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}) \lesssim 1$. *Then*

$$\widehat{\mathrm{BIAS}}^2 \lesssim \delta^{-1/2} \left( \mathbb{E}_{\boldsymbol{x}} \left\| K(\boldsymbol{X}, \boldsymbol{X})^{-1} K(\boldsymbol{X}, \boldsymbol{x}) \right\|_2^2 + \frac{1}{n} \right) \tag{55}$$

*with probability at least* $1 - \delta$. *The above expectation is precisely* $\widehat{\mathrm{VAR}}/\sigma_\xi^2$ *and can be bounded as in Theorem 4.10.*

### 4.3.3 Kernels on $\mathbb{R}^d$ with $d \asymp n$

We now turn our attention to the regime $d \asymp n$ and investigate the behavior of minimum norm interpolants in the RKHS in this high-dimensional setting. Random kernel matrices in the $d \asymp n$ regime have been extensively studied in the last ten years. As shown in [EK10], under assumptions specified below, the kernel matrix can be approximated in operator norm by

$$K(\boldsymbol{X}, \boldsymbol{X}) \approx c_1 \frac{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}}{d} + c_2 \mathbf{I}_n,$$

that is, a linear kernel plus a scaling of the identity. While this equivalence can be viewed as a negative result about the utility of kernels in the $d \asymp n$ regime, the term $c_2 \mathbf{I}_n$ provides implicit regularization for the minimum-norm interpolant in the RKHS [LR20].

We make the following assumptions.

**Assumption 4.12.** *We assume that coordinates of* $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ *are independent, with zero mean and unit variance, so that* $\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^\mathsf{T}$. *Further assume there are constants* $0 < \eta, M < \infty$, *such that the following hold.*

(a) *For all* $i \le d$, $\mathbb{E}[|\boldsymbol{z}_i|^{8+\eta}] \le M$.

(b) $\|\boldsymbol{\Sigma}\| \le M$, $d^{-1} \sum_{i=1}^d \lambda_i^{-1} \le M$, *where* $\lambda_1, \dots, \lambda_d$ *are the eigenvalues of* $\boldsymbol{\Sigma}$.

Note that, for $i \ne j$, the rescaled scalar products $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle / d$ are typically of order $1/\sqrt{d}$. We can therefore approximate the kernel function by its Taylor expansion around 0. To this end, define

$$\alpha := h(0) + h''(0) \frac{\mathsf{tr}(\boldsymbol{\Sigma}^2)}{2d^2}, \quad \beta := h'(0),$$
$$\gamma := \frac{1}{h'(0)} \left[ h(\mathsf{tr}(\boldsymbol{\Sigma})/d) - h(0) - h'(0)\mathsf{tr}(\boldsymbol{\Sigma}/d) \right].$$

Under Assumption 4.12, a variant of a result of [EK10] implies that for some $c_0 \in (0, 1/2)$, the following holds with high probability

$$\left\| K(\boldsymbol{X}, \boldsymbol{X}) - K^{\text{lin}}(\boldsymbol{X}, \boldsymbol{X}) \right\| \lesssim d^{-c_0} \tag{56}$$

where

$$K^{\text{lin}}(\boldsymbol{X}, \boldsymbol{X}) = \beta \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \beta\gamma \mathbf{I}_n + \alpha \mathbf{1}\mathbf{1}^\top. \tag{57}$$

To make the self-induced regularization due to the ridge apparent, we develop an upper bound on the variance of the minimum-norm interpolant in (53). Up to an additive diminishing factor, this expression can be replaced by

$$\sigma_\xi^2 \cdot \mathsf{tr}\left( (\boldsymbol{X}\boldsymbol{X}^\top + d\gamma \mathbf{I}_n)^{-2} \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top \right), \tag{58}$$

where we assumed without loss of generality that $\alpha = 0$. Comparing to (41), we observe that here implicit regularization arises due to the 'curvature' of the kernel, in addition to any favorable tail behavior in the spectrum of $\boldsymbol{X}\boldsymbol{X}^\top$. Furthermore, this regularization arises under rather weak assumptions on the random variables even if Assumption 4.2 is not satisfied. A variant of the development in [LR20] yields a more interpretable upper bound of

$$\widehat{\text{VAR}} \lesssim \sigma_\xi^2 \cdot \frac{1}{\gamma} \left( \frac{k}{n} + \lambda_{k+1} \right) \tag{59}$$

for any $k \geq 1$ [Lia20]; the proof is in the Supplementary Material. Furthermore, a high probability bound on the bias

$$\widehat{\text{BIAS}}^2 \lesssim \|f^*\|_{\mathcal{H}}^2 \cdot \inf_{0 \leq k \leq n} \left\{ \frac{1}{n} \sum_{j>k} \lambda_j(\frac{1}{d}\boldsymbol{X}\boldsymbol{X}^\top) + \gamma + \sqrt{\frac{k}{n}} \right\} \tag{60}$$

can be established with basic tools from empirical process theory under boundedness assumptions on $\sup_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x})$ [LR20].

With more recent developments on the bias and variance of linear interpolants in [HMRT20], a significantly more precise statement can be derived for the $d \asymp n$ regime. The proof of the following theorem is in the Supplementary Material.

**Theorem 4.13.** *Let $0 < M, \eta < \infty$ be fixed constants and suppose that Assumption 4.12 holds with $M^{-1} \leq d/n \leq M$. Further assume that $h$ is continuous on $\mathbb{R}$ and smooth in a neighborhood of 0 with $h(0), h'(0) > 0$, that $\|f^*\|_{L^{4+\eta}(\mathbb{P})} \leq M$ and that the $z_i$ are $M$-subgaussian. Let $y_i = f^*(\boldsymbol{x}_i) + \xi_i$, $\mathbb{E}(\xi_i^2) = \sigma_\xi^2$, and $\boldsymbol{\beta}_0 := \boldsymbol{\Sigma}^{-1}\mathbb{E}[\boldsymbol{x} f^*(\boldsymbol{x})]$. Let $\lambda_* > 0$ be the unique positive solution of*

$$n\Big(1 - \frac{\gamma}{\lambda_*}\Big) = \mathsf{tr}\Big(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_* \mathbf{I})^{-1}\Big). \tag{61}$$

*Define $\mathscr{B}(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0)$ and $\mathscr{V}(\boldsymbol{\Sigma})$ by*

$$\mathscr{V}(\boldsymbol{\Sigma}) := \frac{\mathsf{tr}\big(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \mathbf{I})^{-2}\big)}{n - \mathsf{tr}\big(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \mathbf{I})^{-2}\big)}, \tag{62}$$

$$\mathscr{B}(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0) := \frac{\lambda_*^2 \langle \boldsymbol{\beta}_0, (\boldsymbol{\Sigma} + \lambda_* \mathbf{I})^{-2} \boldsymbol{\Sigma} \boldsymbol{\beta}_0 \rangle}{1 - n^{-1}\mathsf{tr}\big(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \mathbf{I})^{-2}\big)}. \tag{63}$$

*Finally, let $\widehat{\text{BIAS}}^2$ and $\widehat{\text{VAR}}$ denote the squared bias and variance for the minimum-norm interpolant (51). Then there exist $C, c_0 > 0$ (depending also on the constants in Assumption 4.12) such that the following*

*holds with probability at least* $1 - Cn^{-1/4}$ *(here* $\mathsf{P}_{>1}$ *denotes the projector orthogonal to affine functions in* $L^2(\mathbb{P})$*):*

$$\left|\widehat{\mathrm{BIAS}}^2 - \mathscr{B}(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0) - \|\mathsf{P}_{>1} f^*\|_{L^2}^2 (1 + \mathscr{V}(\Sigma))\right| \le C n^{-c_0}\,, \tag{64}$$

$$\left|\widehat{\mathrm{VAR}} - \sigma_\xi^2 \mathscr{V}(\boldsymbol{\Sigma})\right| \le C n^{-c_0}\,. \tag{65}$$

A few remarks are in order. First, note that the left hand side of (61) is strictly increasing in $\lambda_*$, while the right hand side is strictly decreasing. By considering the limits as $\lambda_* \to 0$ and $\lambda_* \to \infty$, it is easy to see that this equation indeed admits a unique solution. Second, the bias estimate in (60) requires $f^* \in \mathcal{H}$, while the bias calculation in (64) does not make this assumption, but instead incurs an approximation error for non-linear components of $f^*$.

We now remark that the minimum-norm interpolant with kernel $K^{\mathrm{lin}}$ is simply ridge regression with respect to the plain covariates $\boldsymbol{X}$ and ridge penalty proportional to $\gamma$:

$$(\widehat{\theta}_0, \widehat{\boldsymbol{\theta}}) := \operatorname*{argmin}_{\theta_0, \boldsymbol{\theta}} \frac{1}{d} \left\|\boldsymbol{y} - \theta_0 - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 + \gamma \|\boldsymbol{\theta}\|_2^2\,. \tag{66}$$

The intuition is that the minimum-norm interpolant for the original kernel takes the form $\widehat{f}(\boldsymbol{x}) = \widehat{\theta}_0 + \langle \widehat{\boldsymbol{\theta}}, \boldsymbol{x}\rangle + \Delta(\boldsymbol{x})$. Here $\widehat{\theta}_0 + \langle \widehat{\boldsymbol{\theta}}, \boldsymbol{x}\rangle$ is a simple component, and $\Delta(\boldsymbol{x})$ is an overfitting component: a function that is small in $L^2(\mathbb{P})$ but allows interpolation of the data.

The characterization in (61), (62), and (63) can be shown to imply upper bounds that are related to the analysis in Section 4.2.

**Corollary 4.14.** *Under the assumptions of Theorem 4.13, further assume that* $f^*(\boldsymbol{x}) = \langle \boldsymbol{\beta}_0, \boldsymbol{x}\rangle$ *is linear and that there is an integer* $k \in \mathbb{N}$*, and a constant* $c_* > 0$ *such that* $r_k(\boldsymbol{\Sigma}) + (n\gamma/c_*\lambda_{k+1}) \ge (1 + c_*)n$*. Then there exists* $c_0 \in (0, 1/2)$ *such that, with high probability, the following hold as long as the right-hand side is less than one:*

$$\widehat{\mathrm{BIAS}}^2 \le 4\Big(\gamma + \frac{1}{n}\sum_{i=k+1}^{d} \lambda_i\Big)^2 \|\boldsymbol{\beta}_{0,\le k}\|_{\boldsymbol{\Sigma}^{-1}}^2 + \|\boldsymbol{\beta}_{0,>k}\|_{\boldsymbol{\Sigma}}^2 + n^{-c_0}\,, \tag{67}$$

$$\widehat{\mathrm{VAR}} \le \frac{2k\sigma_\xi^2}{n} + \frac{4n\sigma_\xi^2}{c_*} \frac{\sum_{i=k+1}^{d} \lambda_i^2}{(n\gamma/c_* + \sum_{i=k+1}^{d} \lambda_i)^2} + n^{-c_0}\,. \tag{68}$$

*Further, under the same assumptions, the effective regularization* $\lambda_*$ *(that is, the unique solution of (61)), satisfies*

$$\gamma + \frac{c_*}{1 + c_*} \frac{1}{n} \sum_{i=k+1}^{d} \lambda_i \le \lambda_* \le 2\gamma + \frac{2}{n} \sum_{i=k+1}^{d} \lambda_i\,. \tag{69}$$

Note that apart from the $n^{-c_0}$ term, (67) recovers the result of Theorem 4.5, while (68) recovers Theorem 4.4 (setting $\gamma = 0$), both with improved constants but limited to the proportional regime. We remark that analogues of Theorems 4.4, 4.5, and 4.6 for ridge regression with $\gamma \neq 0$ can be found in [TB20].

The formulas (61), (62), and (63) might seem somewhat mysterious. However, they have an appealing interpretation in terms of a simpler model that we will refer to as a 'sequence model' (this terminology comes from classical statistical estimation theory [Joh19]). As stated precisely in the remark below, the sequence model is a linear regression model in which the design matrix is deterministic (and diagonal), and the noise and regularization levels are determined via a fixed point equation.

**Remark 4.15.** *Assume without loss of generality* $\boldsymbol{\Sigma} = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$*. In the sequence model we observe* $\boldsymbol{y}^{\mathrm{seq}} \in \mathbb{R}^d$ *distributed according to*

$$\boldsymbol{y}_i^{\mathrm{seq}} = \lambda_i^{1/2} \beta_{0,i} + \frac{\tau}{\sqrt{n}} g_i\,, \quad (g_i)_{i \le d} \sim_{iid} \mathsf{N}(0, 1)\,, \tag{70}$$

*where $\tau$ is a parameter given below. We then perform ridge regression with regularization $\lambda_*$:*

$$\widehat{\boldsymbol{\beta}}^{\text{seq}}(\lambda_*) := \operatorname*{argmin}_{\boldsymbol{\beta}} \left\|\boldsymbol{y}^{\text{seq}} - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\right\|_2^2 + \lambda_*\|\boldsymbol{\beta}\|_2^2\,, \tag{71}$$

*which can be written in closed form as*

$$\widehat{\beta}_i^{\text{seq}}(\lambda_*) = \frac{\lambda_i^{1/2} y_i^{\text{seq}}}{\lambda_* + \lambda_i}\,. \tag{72}$$

*The noise level $\tau^2$ is then fixed via the condition $\tau^2 = \sigma_\xi^2 + \mathbb{E}\|\widehat{\boldsymbol{\beta}}^{\text{seq}}(\lambda_*) - \boldsymbol{\beta}_0\|_2^2$. Then under the assumption that $f^*$ is linear, Theorem 4.13 states that*

$$\mathbb{E}\{(f^*(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x}))^2 | \boldsymbol{X}\} = \mathbb{E}\|\widehat{\boldsymbol{\beta}}^{\text{seq}}(\lambda_*) - \boldsymbol{\beta}_0\|_2^2 + O(n^{-c_0}) \tag{73}$$

*with high probability.*

To conclude this section, we summarize the insights gained from the analyses of several models in the interpolation regime. First, in all cases, the interpolating solution $\widehat{f}$ can be decomposed into a prediction (or simple) component and an overfitting (or spiky) component. The latter ensures interpolation without hurting prediction accuracy. In the next section, we show, under appropriate conditions on the parameterization and the initialization, that gradient methods can be accurately approximated by their linearization, and hence can be viewed as converging to a minimum-norm linear interpolating solution despite their non-convexity. In Section 6, we return to the question of generalization, focusing specifically on two-layer neural networks in linear regimes.

# 5 Efficient optimization

The empirical risk minimization (ERM) problem is, in general, intractable even in simple cases. Section 2.5 gives examples of such hardness results. The classical approach to address this conundrum is to construct convex surrogates of the non-convex ERM problem. The problem of learning a linear classifier provides an easy-to-state—and yet subtle—example. Considering the 0-1 loss, ERM reads

$$\text{minimize} \quad \widehat{L}_{01}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[y_i\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle \le 0\right]\,. \tag{74}$$

Note however that the original problem (74) is not always intractable. If there exists $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\widehat{L}(\boldsymbol{\theta}) = 0$, then finding $\boldsymbol{\theta}$ amounts to solving a set of $n$ linear inequalities. This can be done in polynomial time. In other words, when the model is sufficiently rich to interpolate the data, an interpolator can be constructed efficiently.

In the case of linear classifiers, tractability arises because of the specific structure of the function class (which is linear in the parameters $\boldsymbol{\theta}$), but one might wonder whether it is instead a more general phenomenon. The problem of finding an interpolator can be phrased as a constraint optimization problem. Write the empirical risk as

$$\widehat{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i).$$

Then we are seeking $\boldsymbol{\theta} \in \Theta$ such that

$$\ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i) = 0 \quad \text{for all } i \le n\,. \tag{75}$$

Random constraint satisfaction problems have been studied in depth over the last twenty years, although under different distributions from those arising from neural network theory. Nevertheless, a recurring observation is that, when the number of free parameters is sufficiently large compared to the number of constraints,

these problems (which are NP-hard in the worst case) become tractable; see, for example, [FS96, AM97] and [CO10].

These remarks motivate a fascinating working hypothesis: modern neural networks are tractable *because* they are overparametrized.

Unfortunately, a satisfactory theory of this phenomenon is still lacking, with an important exception: the linear regime. This is a training regime in which the network can be approximated by a linear model, with a random featurization map associated with the training initialization. We discuss these results in Section 5.1.

While the linear theory can explain a number of phenomena observed in practical neural networks, it also misses some important properties. We will discuss these points, and results beyond the linear regime, in Section 5.2.

## 5.1 The linear regime

Consider a neural network with parameters $\boldsymbol{\theta} \in \mathbb{R}^p$: for an input $\boldsymbol{x} \in \mathbb{R}^d$ the network outputs $f(\boldsymbol{x}; \boldsymbol{\theta}) \in \mathbb{R}$. We consider training using the square loss

$$\widehat{L}(\boldsymbol{\theta}) := \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - f(\boldsymbol{x}_i; \boldsymbol{\theta})\big)^2 = \frac{1}{2n} \big\|\boldsymbol{y} - f_n(\boldsymbol{\theta})\big\|_2^2 . \tag{76}$$

Here $\boldsymbol{y} = (y_1, \dots, y_n)$ and $f_n : \mathbb{R}^p \to \mathbb{R}^n$ maps the parameter vector $\boldsymbol{\theta}$ to the evaluation of $f$ at the $n$ data points, $f_n : \boldsymbol{\theta} \mapsto (f(\boldsymbol{x}_1; \boldsymbol{\theta}), \dots, f(\boldsymbol{x}_n; \boldsymbol{\theta}))$. We minimize this empirical risk using gradient flow, with initialization $\boldsymbol{\theta}_0$:

$$\frac{\mathrm{d}\boldsymbol{\theta}_t}{\mathrm{d}t} = \frac{1}{n} \boldsymbol{D} f_n(\boldsymbol{\theta}_t)^{\mathsf{T}} (\boldsymbol{y} - f_n(\boldsymbol{\theta}_t)) . \tag{77}$$

Here $\boldsymbol{D} f_n(\boldsymbol{\theta}) \in \mathbb{R}^{n \times p}$ is the Jacobian matrix of the map $f_n$. Our focus on the square loss and continuous time is for simplicity of exposition. Results of the type presented below have been proved for more general loss functions and for discrete-time and stochastic gradient methods.

As first argued in [JGH18], in a highly overparametrized regime it can happen that $\boldsymbol{\theta}$ changes only slightly with respect to the initialization $\boldsymbol{\theta}_0$. This suggests comparing the original gradient flow with the one obtained by linearizing the right-hand side of (77) around the initialization $\boldsymbol{\theta}_0$:

$$\frac{\mathrm{d}\overline{\boldsymbol{\theta}}_t}{\mathrm{d}t} = \frac{1}{n} \boldsymbol{D} f_n(\boldsymbol{\theta}_0)^{\mathsf{T}} \big(\boldsymbol{y} - f_n(\boldsymbol{\theta}_0) - \boldsymbol{D} f_n(\boldsymbol{\theta}_0)(\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)\big) . \tag{78}$$

More precisely, this is the gradient flow for the risk function

$$\widehat{L}_{\mathrm{lin}}(\overline{\boldsymbol{\theta}}) := \frac{1}{2n} \|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0) - \boldsymbol{D} f_n(\boldsymbol{\theta}_0)(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|_2^2 , \tag{79}$$

which is obtained by replacing $f_n(\boldsymbol{\theta})$ with its first-order Taylor expansion at $\boldsymbol{\theta}_0$. Of course, $\widehat{L}_{\mathrm{lin}}(\overline{\boldsymbol{\theta}})$ is quadratic in $\overline{\boldsymbol{\theta}}$. In particular, if the Jacobian $\boldsymbol{D} f_n(\boldsymbol{\theta}_0)$ has full row rank, the set of global minimizers $\mathrm{ERM}_0 := \{\overline{\boldsymbol{\theta}} : \ \widehat{L}_{\mathrm{lin}}(\overline{\boldsymbol{\theta}}) = 0\}$ forms an affine space of dimension $p - n$. In this case, gradient flow converges to $\overline{\boldsymbol{\theta}}_\infty \in \mathrm{ERM}_0$, which—as discussed in Section 3—minimizes the $\ell_2$ distance from the initialization:

$$\overline{\boldsymbol{\theta}}_\infty := \operatorname{argmin} \Big\{ \|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 : \quad \boldsymbol{D} f_n(\boldsymbol{\theta}_0)(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{y} - f_n(\boldsymbol{\theta}_0) \Big\} . \tag{80}$$

The linear (or 'lazy' ) regime is a training regime in which $\boldsymbol{\theta}_t$ is well approximated by $\overline{\boldsymbol{\theta}}_t$ at all times. Of course if $f_n(\boldsymbol{\theta})$ is an affine function of $\boldsymbol{\theta}$, that is, if $\boldsymbol{D} f_n(\boldsymbol{\theta})$ is constant, then we have $\boldsymbol{\theta}_t = \overline{\boldsymbol{\theta}}_t$ for all times $t$. It is therefore natural to quantify deviations from linearity by defining the Lipschitz constant

$$\mathrm{Lip}(\boldsymbol{D} f_n) := \sup_{\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2} \frac{\|\boldsymbol{D} f_n(\boldsymbol{\theta}_1) - \boldsymbol{D} f_n(\boldsymbol{\theta}_2)\|}{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2} . \tag{81}$$

(For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, we define $\|\boldsymbol{A}\| := \sup_{\boldsymbol{x}\neq 0} \|\boldsymbol{A}\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2$.) It is also useful to define a population version of the last quantity. For this, we assume as usual that samples are i.i.d. draws $(\boldsymbol{x}_i)_{i\le n} \sim_{iid} \mathbb{P}$, and with a slight abuse of notation, we view $f : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta})$ as a map from $\mathbb{R}^p$ to $L^2(\mathbb{P}) := L^2(\mathbb{R}^d; \mathbb{P})$. We let $\boldsymbol{D}f(\boldsymbol{\theta})$ denote the differential of this map at $\boldsymbol{\theta}$, which is a linear operator, $\boldsymbol{D}f(\boldsymbol{\theta}) : \mathbb{R}^p \to L^2(\mathbb{P})$. The corresponding operator norm and Lipschitz constant are given by

$$\|\boldsymbol{D}f(\boldsymbol{\theta})\| := \sup_{\boldsymbol{v}\in\mathbb{R}^p\setminus\{0\}} \frac{\|\boldsymbol{D}f(\boldsymbol{\theta})\boldsymbol{v}\|_{L^2(\mathbb{P})}}{\|\boldsymbol{v}\|_2}\,, \tag{82}$$

$$\mathrm{Lip}(\boldsymbol{D}f) := \sup_{\boldsymbol{\theta}_1\neq\boldsymbol{\theta}_2} \frac{\|\boldsymbol{D}f(\boldsymbol{\theta}_1) - \boldsymbol{D}f(\boldsymbol{\theta}_2)\|}{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2}\,. \tag{83}$$

The next theorem establishes sufficient conditions for $\boldsymbol{\theta}_t$ to remain in the linear regime in terms of the singular values and Lipschitz constant of the Jacobian. Statements of this type were proved in several papers, starting with [DZPS19]; see, for example, [AZLS19, DLL+19, ZCZG20, OS20] and [LZB20]. We follow the abstract point of view in [OS19] and [COB19].

**Theorem 5.1.** *Assume*

$$\mathrm{Lip}(\boldsymbol{D}f_n)\,\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0)\|_2 < \frac{1}{4}\sigma^2_{\min}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0))\,. \tag{84}$$

*Further define*

$$\sigma_{\max} := \sigma_{\max}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)), \sigma_{\min} := \sigma_{\min}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)).$$

*Then the following hold for all $t > 0$:*

1. *The empirical risk decreases exponentially fast to 0, with rate $\lambda_0 = \sigma^2_{\min}/(2n)$:*

$$\widehat{L}(\boldsymbol{\theta}_t) \le \widehat{L}(\boldsymbol{\theta}_0)\, e^{-\lambda_0 t}\,. \tag{85}$$

2. *The parameters stay close to the initialization and are closely tracked by those of the linearized flow. Specifically, letting $L_n := \mathrm{Lip}(\boldsymbol{D}f_n)$,*

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \le \frac{2}{\sigma_{\min}}\,\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0)\|_2\,, \tag{86}$$

$$\|\boldsymbol{\theta}_t - \overline{\boldsymbol{\theta}}_t\|_2 \le \Big\{\frac{32\sigma_{\max}}{\sigma^2_{\min}}\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0)\|_2 + \frac{16L_n}{\sigma^3_{\min}}\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0)\|_2^2\Big\} \wedge \frac{180L_n\sigma^2_{\max}}{\sigma^5_{\min}}\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0)\|_2^2\,. \tag{87}$$

3. *The models constructed by gradient flow and by the linearized flow are similar on test data. Specifically, writing $f^{\mathrm{lin}}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_0) + \boldsymbol{D}f(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, we have*

$$\|f(\boldsymbol{\theta}_t) - f^{\mathrm{lin}}(\overline{\boldsymbol{\theta}}_t)\|_{L^2(\mathbb{P})} \le \Big\{4\,\mathrm{Lip}(\boldsymbol{D}f)\frac{1}{\sigma^2_{\min}} + 180\|\boldsymbol{D}f(\boldsymbol{\theta}_0)\|\frac{L_n\sigma^2_{\max}}{\sigma^5_{\min}}\Big\}\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0)\|_2^2\,. \tag{88}$$

The bounds in (85) and (86) follow from the main result of [OS19]. The coupling bounds in (87) and (88) are proved in the Supplementary Material.

A key role in this theorem is played by the singular values of the Jacobian at initialization, $\boldsymbol{D}f_n(\boldsymbol{\theta}_0)$. These can also be encoded in the kernel matrix $\boldsymbol{K}_{m,0} := \boldsymbol{D}f_n(\boldsymbol{\theta}_0)\boldsymbol{D}f_n(\boldsymbol{\theta}_0)^{\mathsf{T}} \in \mathbb{R}^{n\times n}$. The importance of

this matrix can be easily understood by writing the evolution of the predicted values $f_n^{\rm lin}(\overline{\boldsymbol{\theta}}_t) := f_n(\boldsymbol{\theta}_0) + \boldsymbol{D}f_n(\boldsymbol{\theta}_0)(\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)$. Equation (78) implies

$$\frac{\mathrm{d}f_n^{\rm lin}(\overline{\boldsymbol{\theta}}_t)}{\mathrm{d}t} = \frac{1}{n}\boldsymbol{K}_{m,0}\big(\boldsymbol{y} - f_n^{\rm lin}(\overline{\boldsymbol{\theta}}_t)\big)\,. \tag{89}$$

Equivalently, the residuals $\boldsymbol{r}_t := \boldsymbol{y} - f_n^{\rm lin}(\overline{\boldsymbol{\theta}}_t)$ are driven to zero according to $(\mathrm{d}/\mathrm{d}t)\boldsymbol{r}_t = -\boldsymbol{K}_{m,0}\boldsymbol{r}_t/n$.

Applying Theorem 5.1 requires the evaluation of the minimum and maximum singular values of the Jacobian, as well as its Lipschitz constant. As an example, we consider the case of two-layer neural networks:

$$f(\boldsymbol{x};\boldsymbol{\theta}) := \frac{\alpha}{\sqrt{m}}\sum_{j=1}^{m} b_j\sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle), \quad \boldsymbol{\theta} = (\boldsymbol{w}_1,\dots,\boldsymbol{w}_m)\,. \tag{90}$$

To simplify our task, we assume the second layer weights $\boldsymbol{b} = (b_1,\dots,b_m) \in \{+1,-1\}^m$ to be fixed with an equal number of $+1$s and $-1$s. Without loss of generality we can assume that $b_1 = \cdots = b_{m/2} = +1$ and $b_{m/2+1} = \cdots = b_m = -1$. We train the weights $\boldsymbol{w}_1,\dots,\boldsymbol{w}_m$ via gradient flow. The number of parameters is $p = md$. The scaling factor $\alpha$ allows tuning between different regimes. We consider two initializations, denoted by $\boldsymbol{\theta}_0^{(1)}$ and $\boldsymbol{\theta}_0^{(2)}$:

$$\boldsymbol{\theta}_0^{(1)}: \qquad (\boldsymbol{w}_i)_{i\le m} \sim_{i.i.d.} \mathsf{Unif}(\mathbb{S}^{d-1}); \tag{91}$$

$$\boldsymbol{\theta}_0^{(2)}: \qquad (\boldsymbol{w}_i)_{i\le m/2} \sim_{i.i.d.} \mathsf{Unif}(\mathbb{S}^{d-1}),\ \boldsymbol{w}_{m/2+i} = \boldsymbol{w}_i,\ i \le m/2, \tag{92}$$

where $\mathbb{S}^{d-1}$ denotes the unit sphere in $d$ dimensions. The important difference between these initializations is that (by the central limit theorem) $|f(\boldsymbol{x};\boldsymbol{\theta}_0^{(1)})| = \Theta(\alpha)$, while $f(\boldsymbol{x};\boldsymbol{\theta}_0^{(2)}) = 0$.

It is easy to compute the Jacobian $\boldsymbol{D}f_n(\boldsymbol{x};\boldsymbol{\theta}) \in \mathbb{R}^{n\times md}$:

$$[\boldsymbol{D}f_n(\boldsymbol{x};\boldsymbol{\theta})]_{i,(j,a)} = \frac{\alpha}{\sqrt{m}} b_j\sigma'(\langle \boldsymbol{w}_j, \boldsymbol{x}_i\rangle)\, x_{ia}\,, \quad i\in[n], (j,a)\in[m]\times[d]\,. \tag{93}$$

**Assumption 5.2.** *Let $\sigma : \mathbb{R}\to\mathbb{R}$ be a fixed activation function which we assume differentiable with bounded first and second order derivatives. Let*

$$\sigma = \sum_{\ell\ge 0}\mu_\ell(\sigma)h_\ell$$

*denote its decomposition into orthonormal Hermite polynomials. Assume $\mu_\ell(\sigma)\neq 0$ for all $\ell\le\ell_0$ for some constant $\ell_0$.*

**Lemma 5.3.** *Under Assumption 5.2, further assume $\{(\boldsymbol{x}_i, y_i)\}_{i\le n}$ to be i.i.d. with $\boldsymbol{x}_i \sim_{i.i.d.} \mathsf{N}(0,\mathbf{I}_d)$, and $y_i$ $B^2$-sub-Gaussian. Then there exist constants $C_i$, depending uniquely on $\sigma$, such that the following hold with probability at least $1 - 2\exp\{-n/C_0\}$, provided $md \ge C_0 n\log n$ and $n \le d^{\ell_0}$ (whenever not specified, these hold for both initializations $\boldsymbol{\theta}_0 \in \{\boldsymbol{\theta}_0^{(1)}, \boldsymbol{\theta}_0^{(2)}\}$):*

$$\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0^{(1)})\|_2 \le C_1\big(B+\alpha\big)\sqrt{n} \tag{94}$$

$$\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0^{(2)})\|_2 \le C_1 B\sqrt{n}\,, \tag{95}$$

$$\sigma_{\min}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)) \ge C_2\alpha\sqrt{d}\,, \tag{96}$$

$$\sigma_{\max}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)) \le C_3\alpha\big(\sqrt{n}+\sqrt{d}\big)\,, \tag{97}$$

$$\mathrm{Lip}(\boldsymbol{D}f_n) \le C_4\alpha\sqrt{\frac{d}{m}}\big(\sqrt{n}+\sqrt{d}\big)\,. \tag{98}$$

*Further*

$$\|\boldsymbol{D}f(\boldsymbol{\theta}_0)\| \le C_1'\alpha\,, \tag{99}$$

$$\mathrm{Lip}(\boldsymbol{D}f) \le C_4'\alpha\sqrt{\frac{d}{m}}\,. \tag{100}$$

Equations (94), (95) are straightforward [OS19]. The remaining inequalities are proved in the Supplementary Material. Using these estimates in Theorem 5.1, we get the following approximation theorem for two-layer neural nets.

**Theorem 5.4.** *Consider the two layer neural network of (90) under the assumptions of Lemma 5.3. Further let $\overline{\alpha} := \alpha/(1+\alpha)$ for initialization $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^{(1)}$ and $\overline{\alpha} := \alpha$ for $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^{(2)}$. Then there exist constants $C_i$, depending uniquely on $\sigma$, such that if $md \geq C_0 n \log n$, $d \leq n \leq d^{\ell_0}$ and*

$$\overline{\alpha} \geq C_0 \sqrt{\frac{n^2}{md}}\,, \tag{101}$$

*then, with probability at least $1 - 2\exp\{-n/C_0\}$, the following hold for all $t \geq 0$.*

1. *Gradient flow converges exponentially fast to a global minimizer. Specifically, letting $\lambda_* = C_1 \alpha^2 d/n$, we have*

$$\widehat{L}(\boldsymbol{\theta}_t) \leq \widehat{L}(\boldsymbol{\theta}_0)\, e^{-\lambda_* t}\,. \tag{102}$$

2. *The model constructed by gradient flow and linearized flow are similar on test data, namely*

$$\|f(\boldsymbol{\theta}_t) - f_{\rm lin}(\overline{\boldsymbol{\theta}}_t)\|_{L^2(\mathbb{P})} \leq C_1 \left\{ \frac{\alpha}{\overline{\alpha}^2}\sqrt{\frac{n^2}{md}} + \frac{1}{\overline{\alpha}^2}\sqrt{\frac{n^5}{md^4}} \right\}. \tag{103}$$

It is instructive to consider Theorem 5.4 for two different choices of $\alpha$ (a third one will be considered in Section 5.2).

For $\alpha = \Theta(1)$, we have $\overline{\alpha} = \Theta(1)$ and therefore the two initializations $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}_0^{(2)}\}$ behave similarly. In particular, condition (101) requires $md \gg n^2$: the number of network parameters must be quadratic in the sample size. This is significantly stronger than the simple condition that the network is overparametrized, namely $md \gg n$. Under the condition $md \gg n^2$ we have exponential convergence to vanishing training error, and the difference between the neural network and its linearization is bounded as in (103). This bound vanishes for $m \gg n^5/d^4$. While we do not expect this condition to be tight, it implies that, *under the choice* $\alpha = \Theta(1)$, sufficiently wide networks behave as linearly parametrized models.

For $\alpha \to \infty$, we have $\overline{\alpha} \to 1$ for initialization $\boldsymbol{\theta}_0^{(1)}$ and therefore Theorem 5.4 yields the same bounds as in the previous paragraph for this initialization. However, for the initialization $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^{(2)}$ (which is constructed so that $f(\boldsymbol{\theta}_0^{(2)}) = 0$) we have $\overline{\alpha} = \alpha$ and condition (101) is always verified as $\alpha \to \infty$. Therefore the conclusions of Theorem 5.4 apply under nearly minimal overparametrization, namely if $md \gg n \log n$. In that case, the linear model is an arbitrarily good approximation of the neural net as $\alpha$ grows: $\|f(\boldsymbol{\theta}_t) - f_{\rm lin}(\overline{\boldsymbol{\theta}}_t)\|_{L^2(\mathbb{P})} = O(1/\alpha)$. In other words, an overparametrized neural network can be trained in the linearized regime by choosing suitable initializations and suitable scaling of the parameters.

Recall that, as $t \to \infty$, $\overline{\boldsymbol{\theta}}_t$ converges to the min-norm interpolant $\overline{\boldsymbol{\theta}}_\infty$; see (80). Therefore, as long as condition (101) holds and the right-hand side of (103) is negligible, the generalization properties of the neural network are well approximated by those of min-norm interpolation in a linear model with featurization map $\boldsymbol{x} \mapsto \boldsymbol{D} f(\boldsymbol{x}; \boldsymbol{\theta}_0)$. We will study the latter in Section 6.

In the next subsection we will see that the linear theory outlined here fails to capture different training schemes in which the network weights genuinely change.

## 5.2 Beyond the linear regime?

For a given dimension $d$ and sample size $n$, we can distinguish two ways to violate the conditions for the linear regime, as stated for instance in Theorem 5.4. First, we can reduce the network size $m$. While Theorem 5.4 does not specify the minimum $m$ under which the conclusions of the theorem cease to hold, it is clear that $md \geq n$ is necessary in order for the training error to vanish as in (102).

However, even if the model is overparametrized, the same condition is violated if $\alpha$ is sufficiently small. In particular, the limit $m \to \infty$ with $\alpha = \alpha_0/\sqrt{m}$ has attracted considerable attention and is known as the mean field limit. In order to motivate the mean field analysis, we can suggestively rewrite (90) as

$$f(\boldsymbol{x};\boldsymbol{\theta}) := \alpha_0 \int b\, \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\, \widehat{\rho}(\mathrm{d}\boldsymbol{w}, \mathrm{d}b)\,, \tag{104}$$

where $\widehat{\rho} := m^{-1}\sum_{j=1}^m \delta_{\boldsymbol{w}_j, b_j}$ is the empirical distribution of neuron weights. If the weights are drawn i.i.d. from a common distribution $(\boldsymbol{w}_j, b_j) \sim \rho$, we can asymptotically replace $\widehat{\rho}$ with $\rho$ in the above expression, by the law of large numbers.

The gradient flow (77) defines an evolution over the space of neuron weights, and hence an evolution in the space of empirical distributions $\widehat{\rho}$. It is natural to ask whether this evolution admits a simple characterization. This question was first addressed by [NS17, MMN18, RVE18, SS20] and [CB18].

**Theorem 5.5.** *Initialize the weights so that $\{(\boldsymbol{w}_j, b_j)\}_{j\le m} \sim_{i.i.d.} \rho_0$ with $\rho_0$ a probability measure on $\mathbb{R}^{d+1}$. Further, assume the activation function $u \mapsto \sigma(u)$ to be differentiable with $\sigma'$ bounded and Lipschitz continuous, and assume $|b_j| \le C$ almost surely under the initialization $\rho_0$, for some constant $C$. Then, for any fixed $T \ge 0$, the following limit holds in $L^2(\mathbb{P})$, uniformly over $t \in [0, T]$:*

$$\lim_{m\to\infty} f(\boldsymbol{\theta}_{mt}) = F(\rho_t) := \alpha_0 \int b\, \sigma(\langle \boldsymbol{w}, \,\cdot\, \rangle)\, \rho_t(\mathrm{d}\boldsymbol{w}, \mathrm{d}b)\,, \tag{105}$$

*where $\rho_t$ is a probability measure on $\mathbb{R}^{d+1}$ that solves the following partial differential equation (to be interpreted in the weak sense):*

$$\partial_t \rho_t(\boldsymbol{w}, b) = \alpha_0 \nabla(\rho_t(\boldsymbol{w}, b)\nabla\Psi(\boldsymbol{w}, b; \rho_t))\,, \tag{106}$$

$$\Psi(\boldsymbol{w}, b; \rho) := \widehat{\mathbb{E}}\big\{ b\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\big(F(\boldsymbol{x};\rho_t) - y\big)\big\}\,. \tag{107}$$

*Here the gradient $\nabla$ is with respect to $(\boldsymbol{w}, b)$ (gradient in $d+1$ dimensions) if both first- and second-layer weights are trained, and only with respect to $\boldsymbol{w}$ (gradient in $d$ dimensions) if only first-layer weights are trained.*

This statement can be obtained by checking the conditions of [CB18, Theorem 2.6]. A quantitative version can be obtained for bounded $\sigma$ using Theorem 1 of [MMM19].

A few remarks are in order. First, the limit in (105) requires time to be accelerated by a factor $m$. This is to compensate for the fact that the function value is scaled by a factor $1/m$. Second, while we stated this theorem as an asymptotic result, for large $m$, the evolution described by the PDE (106) holds at any finite $m$ for the empirical measure $\widehat{\rho}_t$. In that case, the gradient of $\rho_t$ is not well defined, and it is important to interpret this equation in the weak sense [AGS08, San15]. The advantage of working with the average measure $\rho_t$ instead of the empirical one $\widehat{\rho}_t$ is that the former is deterministic and has a positive density (this has important connections to global convergence). Third, quantitative versions of this theorem were proved in [MMN18, MMM19], and generalizations to multi-layer networks in [NP20].

Mean-field theory can be used to prove global convergence results. Before discussing these results, let us emphasize that —in this regime— the weights move in a non-trivial way during training, despite the fact that the network is infinitely wide. For the sake of simplicity, we will focus on the case already treated in the previous section in which the weights $b_j \in \{+1, -1\}$ are initialized with signs in equal proportions, and are not changed during training. Let us first consider the evolution of the predicted values $F_n(\rho_t) := (F(\boldsymbol{x}_1;\rho_t), \dots, F(\boldsymbol{x}_n;\rho_t))$. Manipulating (106), we get

$$\frac{\mathrm{d}}{\mathrm{d}t} F_n(\rho_t) = -\frac{1}{n}\boldsymbol{K}_t\big(F_n(\rho_t) - \boldsymbol{y}\big)\,, \qquad \boldsymbol{K}_t = (K_t(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j\le n} \tag{108}$$

$$K_t(\boldsymbol{x}_1, \boldsymbol{x}_2) := \int \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \sigma'(\langle \boldsymbol{w}, \boldsymbol{x}_1\rangle)\sigma'(\langle \boldsymbol{w}, \boldsymbol{x}_2\rangle)\, \rho_t(\mathrm{d}b, \mathrm{d}\boldsymbol{w})\,, \tag{109}$$

In the short-time limit we recover the linearized evolution of (89) [MMM19], but the kernel $K_t$ is now changing with training (with a factor $m$ acceleration in time).

It also follows from the same characterization of Theorem 5.5 that the weight $\boldsymbol{w}_j$ of a neuron with weight $(\boldsymbol{w}_j, b_j) = (\boldsymbol{w}, b)$ moves at a speed $\widehat{\mathbb{E}}\{b\,\boldsymbol{x}\,\sigma'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)(F(\boldsymbol{x};\rho_t) - y)\}$. This implies

$$\lim_{m\to\infty} \frac{1}{m}\big\|\boldsymbol{W}_{t+s} - \boldsymbol{W}_t\big\|_F^2 = v_2(\rho_t)\, s^2 + o(s^2)\,, \tag{110}$$

$$v_2(\rho_t) := \frac{1}{n^2}\langle \boldsymbol{y} - F_n(\rho_t), \boldsymbol{K}_t(\boldsymbol{y} - F_n(\rho_t))\rangle\,. \tag{111}$$

This expression implies that the first-layer weights change significantly more than in the linear regime studied in Section 5.1. As an example, consider the setting of Lemma 5.3, namely data $(\boldsymbol{x}_i)_{i\le n} \sim_{i.i.d.} \mathsf{N}(0, \mathbf{I}_d)$, an activation function satisfying Assumption 5.2 and dimension parameters such that $md \ge Cn\log n$, $n \le d^{\ell_0}$. We further initialize $\rho_0 = \mathsf{Unif}(\mathbb{S}^{d-1}) \otimes \mathsf{Unif}(\{+1,-1\})$ (that is, the vectors $\boldsymbol{w}_j$ are uniform on the unit sphere and the weights $b_j$ are uniform in $\{+1,-1\}$). Under this initialization $\|\boldsymbol{W}_0\|_F^2 = m$ and hence (110) at $t = 0$ can be interpreted as describing the initial relative change of the first-layer weights.

Theorem 5.1 (see (86)) and Lemma 5.3 (see (94)–(96)) imply that, with high probability,

$$\sup_{t\ge 0} \frac{1}{\sqrt{m}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F \le C\frac{1}{\overline{\alpha}}\sqrt{\frac{n}{md}}\,, \tag{112}$$

where $\overline{\alpha} = \alpha/(1+\alpha)$ for initialization $\boldsymbol{\theta}_0^{(1)}$ and $\overline{\alpha} = \alpha$ for initialization $\boldsymbol{\theta}_0^{(2)}$. In the mean field regime $\overline{\alpha} \asymp \alpha \asymp 1/\sqrt{m}$ and the right hand side above is of order $\sqrt{n/d}$, and hence it does not vanish. This is not due to a weakness of the analysis. By (110), we can choose $\varepsilon$ a small enough constant so that

$$\lim_{m\to\infty}\sup_{t\ge 0} \frac{1}{\sqrt{m}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F \ge \lim_{m\to\infty} \frac{1}{\sqrt{m}}\|\boldsymbol{W}_\varepsilon - \boldsymbol{W}_0\|_F \ge \frac{1}{2}v_2(\rho_0)^{1/2}\,\varepsilon\,. \tag{113}$$

This is bounded away from 0 as long as $v_2(\rho_0)$ is non-vanishing. In order to see this, note that $\lambda_{\min}(\boldsymbol{K}_0) \ge c_0\, d$ with high probability for $c_0$ a constant (note that $\boldsymbol{K}_0$ is a kernel inner product random matrix, and hence this claim follows from the general results of [MMM21]). Noting that $F_n(\rho_0) = 0$ (because $\int b\rho_0(\mathrm{d}b, \mathrm{d}\boldsymbol{w}) = 0$), this implies, with high probability,

$$v(\rho_0) = \frac{1}{n^2}\langle \boldsymbol{y}, \boldsymbol{K}_0\boldsymbol{y}\rangle \ge \frac{c_0 d}{n^2}\|\boldsymbol{y}\|_2^2 \ge \frac{c_0' d}{n}\,. \tag{114}$$

We expect this lower bound to be tight, as can be seen by considering the pure noise case $\boldsymbol{y} \sim \mathsf{N}(0, \tau^2\mathbf{I}_n)$, which leads to $v(\rho_0) = \tau^2\mathsf{tr}(\boldsymbol{K}_0)/n^2(1 + o_n(1)) \asymp d/n$.

To summarize, (112) (setting $\alpha \asymp 1/\sqrt{m}$) and (113) conclude that, for $d \le n \le d^{\ell_0}$,

$$c_1\sqrt{\frac{d}{n}} \le \lim_{m\to\infty}\sup_{t\ge 0} \frac{1}{\sqrt{m}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F \le c_2\sqrt{\frac{n}{d}}\,, \tag{115}$$

hence the limit on the left-hand side of (113) is indeed non-vanishing as $m \to \infty$ at $n, d$ fixed. In other words, the fact that the upper bound in (112) is non-vanishing is not an artifact of the bounding technique, but a consequence of the change of training regime. We also note a gap between the upper and lower bounds in (115) when $n \gg d$: a better understanding of this quantity is an interesting open problem. In conclusion, both a linear and a nonlinear regime can be obtained in the infinite-width limit of two-layer neural networks, for different scalings of the normalization factor $\alpha$.

As mentioned above, the mean field limit can be used to prove global convergence results, both for two-layer [MMN18, CB18] and for multilayer networks [NP20]. Rather than stating these (rather technical) results formally, it is instructive to discuss the nature of fixed points of the evolution (106): this will also indicate the key role played by the support of the distribution $\rho_t$.

**Lemma 5.6.** *Assume $t \mapsto \sigma(t)$ to be differentiable with bounded derivative. Let $\widehat{L}(\rho) = \widehat{\mathbb{E}}\{[y - F(\boldsymbol{x};\rho)]^2\}$ be the empirical risk of an infinite-width network with neuron's distribution $\rho$, and define $\psi(\boldsymbol{w};\rho) := \widehat{\mathbb{E}}\{\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)[y - F(\boldsymbol{x};\rho)]\}$.*

*(a) $\rho_*$ is a global minimizer of $\widehat{L}$ if and only if $\psi(\boldsymbol{w};\rho_*) = 0$ for all $\boldsymbol{w} \in \mathbb{R}^d$.*

*(b) $\rho_*$ is a fixed point of the evolution (106) if and only if, for all $(b, \boldsymbol{w}) \in \mathrm{supp}(\rho_*)$, we have $\psi(\boldsymbol{w};\rho_*) = 0$ and $b\nabla_{\boldsymbol{w}}\psi(\boldsymbol{w};\rho_*) = 0$.*

*The same statement holds if the empirical averages above are replaced by population averages (that is, the empirical risk $\widehat{L}(\rho)$ is replaced by its population version $L_n(\rho) = \mathbb{E}\{[y - F(\boldsymbol{x};\rho)]^2\}$).*

This statement clarifies that fixed points of the gradient flow are only a 'small' superset of global minimizers, as $m \to \infty$. Consider for instance the case of an analytic activation function $t \mapsto \sigma(t)$. Let $\rho_*$ be a stationary point and assume that its support contains a sequence of distinct points $\{(b_i, \boldsymbol{w}_i)\}_{i\ge 1}$ such that $\{\boldsymbol{w}_i\}_{i\ge 1}$ has an accumulation point. Then, by condition $(b)$, $\psi(\boldsymbol{w};\rho_*) = 0$ identically and therefore $\rho_*$ is a global minimum. In other words, the only local minima correspond to $\rho_*$ supported on a set of isolated points. Global convergence proofs aim at ruling out this case.

### 5.3 Other approaches

The mean-field limit is only one of several analytical approaches that have been developed to understand training beyond the linear regime. A full survey of these directions goes beyond the scope of this review. Here we limit ourselves to highlighting a few of them that have a direct connection to the analysis in the previous section.

A natural idea is to view the linearized evolution as the first order in a Taylor expansion, and to construct higher order approximations. This can be achieved by writing an ordinary differential equation for the evolution of the kernel $\boldsymbol{K}_t$ (see (109) for the infinite-width limit). This takes the form [HY20]

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K}_t = -\frac{1}{n}\boldsymbol{K}_t^{(3)} \cdot (F_n(\rho_t) - \boldsymbol{y})\,, \tag{116}$$

where $\boldsymbol{K}_t^{(3)} \in (\mathbb{R}^n)^{\otimes 3}$ is a certain higher order kernel (an order-3 tensor), which is contracted along one direction with $(F_n(\rho_t) - \boldsymbol{y}) \in \mathbb{R}^n$. The linearized approximation amounts to replacing $\boldsymbol{K}_t^{(3)}$ with 0. A better approximation could be to replace $\boldsymbol{K}_t^{(3)}$ with its value at initialization $\boldsymbol{K}_0^{(3)}$. This construction can be repeated, leading to a hierarchy of increasingly complex (and accurate) approximations.

Other approaches towards constructing a Taylor expansion around the linearized evolutions were proposed, among others, by [DGA20] and [HN20].

Note that the linearized approximation relies on the assumption that the Jacobian $\boldsymbol{D}f_n(\boldsymbol{\theta}_0)$ is non-vanishing and well conditioned. [BL20a] propose specific neural network parametrizations in which the Jacobian at initialization vanishes, and the first non-trivial term in the Taylor expansion is quadratic. Under such initializations the gradient flow dynamics is 'purely nonlinear'.

## 6 Generalization in the linear regime

As discussed in Sections 2 and 4, approaches that control the test error via uniform convergence fail for overparametrized interpolating models. So far, the most complete generalization results for such models have been obtained in the linear regime, namely under the assumption that we can approximate $f(\boldsymbol{\theta})$ by its first order Taylor approximation $f_{\mathrm{lin}}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_0) + \boldsymbol{D}f(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. While Theorem 5.1 provides a set of sufficient conditions for this approximation to be accurate, in this section we leave aside the question of whether or when this is indeed the case, and review what we know about the generalization properties of these linearized models. We begin in Section 6.1 by discussing the inductive bias induced by gradient descent on

wide two-layer networks. Section 6.2 describes a general setup. Section 6.3 reviews random features models: two-layer neural networks in which the first layer is not trained and entirely random. While these are simpler than neural networks in the linear regime, their generalization behavior is in many ways similar. Finally, in Section 6.4 we review progress on the generalization error of linearized two-layer networks.

## 6.1 The implicit regularization of gradient-based training

As emphasized in previous sections, in an overparametrized setting, convergence to global minima is not sufficient to characterize the generalization properties of neural networks. It is equally important to understand which global minima are selected by the training algorithm, in particular by gradient-based training. As shown in Section 3, in linear models gradient descent converges to the minimum $\ell_2$-norm interpolator. Under the assumption that training takes place in the linear regime (see Section 5.1), we can apply this observation to neural networks. Namely, the neural network trained by gradient descent will be well approximated by the model[6] $f_{\rm lin}(\widehat{\boldsymbol{a}}) = f(\boldsymbol{\theta}_0) + \boldsymbol{D}f(\boldsymbol{\theta}_0)\widehat{\boldsymbol{a}}$ where $\widehat{\boldsymbol{a}}$ minimizes $\|\boldsymbol{a}\|_2$ among empirical risk minimizers

$$\widehat{\boldsymbol{a}} := \operatorname*{argmin}_{\boldsymbol{a}\in\mathbb{R}^p} \Big\{ \|\boldsymbol{a}\|_2 : \ y_i = f_{\rm lin}(\boldsymbol{x}_i;\boldsymbol{a}) \text{ for all } i \le n \Big\}. \tag{117}$$

For simplicity, we will set $f(\boldsymbol{x};\boldsymbol{\theta}_0) = 0$. This can be achieved either by properly constructing the initialization $\boldsymbol{\theta}_0$ (as in the initialization $\boldsymbol{\theta}_0^{(2)}$ in Section 5.1) or by redefining the response vector $\boldsymbol{y}' = \boldsymbol{y} - f_n(\boldsymbol{\theta}_0)$. If $f(\boldsymbol{x};\boldsymbol{\theta}_0) = 0$, the interpolation constraint $y_i = f_{\rm lin}(\boldsymbol{x}_i;\boldsymbol{a})$ for all $i \le n$ can be written as $\boldsymbol{D}f_n(\boldsymbol{\theta}_0)\boldsymbol{a} = \boldsymbol{y}$.

Consider the case of two-layer neural networks in which only first-layer weights are trained. Recalling the form of the Jacobian (93), we can rewrite (117) as

$$\widehat{\boldsymbol{a}} := \operatorname*{argmin}_{\boldsymbol{a}\in\mathbb{R}^{md}} \Big\{ \|\boldsymbol{a}\|_2 : \ y_i = \sum_{j=1}^m \langle \boldsymbol{a}_j, \boldsymbol{x}_i\rangle \sigma'(\langle \boldsymbol{w}_j, \boldsymbol{x}_i\rangle) \Big\}, \tag{118}$$

where we write $\boldsymbol{a} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_m)$, $\boldsymbol{a}_i \in \mathbb{R}^d$. In this section we will study the generalization properties of this *neural tangent* (NT) model and some of its close relatives. Before formally defining our setup, it is instructive to rewrite the norm that we are minimizing in function space:

$$\|f\|_{\mathsf{NT},m} := \inf \Big\{ \frac{1}{\sqrt{m}} \|\boldsymbol{a}\|_2 : \ f(\boldsymbol{x}) = \frac{1}{m} \sum_{j=1}^m \langle \boldsymbol{a}_j, \boldsymbol{x}_i\rangle \sigma'(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle) \text{ a.e.} \Big\}. \tag{119}$$

This is an RKHS norm defining a finite-dimensional subspace of $L^2(\mathbb{R}^d, \mathbb{P})$. We can also think of it as a finite approximation to the norm

$$\|f\|_{\mathsf{NT}} := \inf \Big\{ \|\boldsymbol{a}\|_{L^2(\rho_0)} : \ f(\boldsymbol{x}) = \int \langle \boldsymbol{a}(\boldsymbol{w}), \boldsymbol{x}_i\rangle \sigma'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\, \rho_0(\mathrm{d}\boldsymbol{w}) \Big\}. \tag{120}$$

Here $\boldsymbol{a} : \mathbb{R}^d \to \mathbb{R}^d$ is a measurable function with

$$\|\boldsymbol{a}\|_{L^2(\rho_0)}^2 := \int \|\boldsymbol{a}(\boldsymbol{w})\|^2 \rho_0(\mathrm{d}\boldsymbol{w}) < \infty,$$

and we are assuming that the weights $\boldsymbol{w}_j$ in (119) are initialized as

$$(\boldsymbol{w}_j)_{j\le m} \sim_{i.i.d.} \rho_0.$$

This is also an RKHS norm whose kernel $K_{\mathsf{NT}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ will be described below; see (129).

Let us emphasize that moving out of the linear regime leads to different—and possibly more interesting—inductive biases than those described in (119) or (120). As an example, [CB20] analyze the mean field limit

[6] With a slight abuse of notation, in this section we parametrize the linearized model by the shift with respect to the initialization $\boldsymbol{\theta}_0$.

of two-layer networks, trained with logistic loss, for activation functions that have Lipschitz gradient and are positively 2-homogeneous. For instance, the square ReLU $\sigma(x) = (x_+)^2$ with fixed second-layer coefficients fits this framework. The usual ReLU with trained second-layer coefficients $b_j\sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle) = b_j(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle)_+$ is 2-homogeneous but not differentiable. In this setting, and under a convergence assumption, they show that gradient flow minimizes the following norm among interpolators:

$$\|f\|_\sigma := \inf\left\{\|\nu\|_{\text{TV}} : \ f(\boldsymbol{x}) = \int \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\, \nu(\mathrm{d}\boldsymbol{w}) \text{ a.e.}\right\}. \tag{121}$$

Here, minimization is over the finite signed measure $\nu$ with Hahn decomposition $\nu = \nu_+ - \nu_-$, and $\|\nu\|_{\text{TV}} := \nu_+(\mathbb{R}^d) + \nu_-(\mathbb{R}^d)$ is the associated total variation. The norm $\|f\|_\sigma$ is a special example of the variation norms introduced in [Kur97] and further studied in [KS01, KS02].

This norm differs in two ways from the RKHS norm of (120). Each is defined in terms of a different integral operator,

$$\boldsymbol{a} \mapsto \int \langle \boldsymbol{a}(\boldsymbol{w}), \boldsymbol{x}\rangle \sigma'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)\, \rho_0(\mathrm{d}\boldsymbol{w})$$

for (120) and

$$\nu \mapsto \int \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\, \nu(\mathrm{d}\boldsymbol{w})$$

for (121). However, more importantly, the norms are very different: in (120) it is a Euclidean norm while in (121) it is a total variation norm. Intuitively, the total variation norm $\|\nu\|_{\text{TV}}$ promotes 'sparse' measures $\nu$, and hence the functional norm $\|f\|_\sigma$ promotes functions that depend primarily on a small number of directions in $\mathbb{R}^d$ [Bac17].

## 6.2 Ridge regression in the linear regime

We generalize the min-norm procedure of (117) to consider the ridge regression estimator:

$$\widehat{\boldsymbol{a}}(\lambda) := \operatorname*{argmin}_{\boldsymbol{a}\in\mathbb{R}^p} \left\{\frac{1}{n}\sum_{i=1}^n \big(y_i - f_{\text{lin}}(\boldsymbol{x}_i; \boldsymbol{a})\big)^2 + \lambda\|\boldsymbol{a}\|_2^2\right\}, \tag{122}$$

$$f_{\text{lin}}(\boldsymbol{x}_i; \boldsymbol{a}) := \langle \boldsymbol{a}, \boldsymbol{D}f(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\rangle\,. \tag{123}$$

The min-norm estimator can be recovered by taking the limit of vanishing regularization $\lim_{\lambda\to 0}\widehat{\boldsymbol{a}}(\lambda) = \widehat{\boldsymbol{a}}(0^+)$ (with a slight abuse of notation, we will identify $\lambda = 0$ with this limit). Apart from being intrinsically interesting, the behavior of $\widehat{\boldsymbol{a}}(\lambda)$ for $\lambda > 0$ is a good approximation of the behavior of the estimator produced by gradient flow with early stopping [AKT19]. More precisely, letting $(\widehat{\boldsymbol{a}}_{\text{GF}}(t))_{t\ge 0}$ denote the path of gradient flow initialized at $\widehat{\boldsymbol{a}}_{\text{GF}}(0) = 0$, there exists a parametrization $t \mapsto \lambda(t)$, such that the test error at $\widehat{\boldsymbol{a}}_{\text{GF}}(t)$ is well approximated by the test error at $\widehat{\boldsymbol{a}}(\lambda(t))$.

Note that the function class $\{f_{\text{lin}}(\boldsymbol{x}_i; \boldsymbol{a}) := \langle \boldsymbol{a}, \boldsymbol{D}f(\boldsymbol{x}_i; \boldsymbol{\theta}_0)\rangle : \ \boldsymbol{a} \in \mathbb{R}^p\}$ is a linear space, which is linearly parametrized by $\boldsymbol{a}$. We consider two specific examples which are obtained by linearizing two-layer neural networks (see (90)):

$$\mathcal{F}_{\mathsf{RF}}^m := \left\{f_{\text{lin}}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{i=1}^m a_i\sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) : \ \ a_i \in \mathbb{R}\right\}, \tag{124}$$

$$\mathcal{F}_{\mathsf{NT}}^m := \left\{f_{\text{lin}}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{i=1}^m \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) : \ \ \boldsymbol{a}_i \in \mathbb{R}^d\right\}. \tag{125}$$

Namely, $\mathcal{F}_{\mathsf{RF}}^m$ (RF stands for 'random features') is the class of functions obtained by linearizing a two-layer network with respect to second-layer weights and keeping the first layer fixed, and $\mathcal{F}_{\mathsf{NT}}^m$ (NT stands for 'neural tangent') is the class obtained by linearizing a two-layer network with respect to the first layer and keeping the second fixed. The first example was introduced by [BBV06] and [RR07] and can be viewed as

a linearization of the two-layer neural networks in which only second-layer weights are trained. Of course, since the network is linear in the second-layer weights, it coincides with its linearization. The second example is the linearization of a neural network in which only the first-layer weights are trained. In both cases, we draw $(\boldsymbol{w}_i)_{i\le m} \sim_{i.i.d.} \mathsf{Unif}(\mathbb{S}^{d-1})$ (the Gaussian initialization $\boldsymbol{w}_i \sim \mathsf{N}(0, \mathbf{I}_d/d)$ behaves very similarly).

Ridge regression (122) within either model $\mathcal{F}_{\mathsf{RF}}$ or $\mathcal{F}_{\mathsf{NT}}$ can be viewed as kernel ridge regression (KRR) with respect to the kernels

$$K_{\mathsf{RF},m}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \frac{1}{m}\sum_{i=1}^{m} \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}_1\rangle)\sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}_2\rangle)\,, \tag{126}$$

$$K_{\mathsf{NT},m}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \frac{1}{m}\sum_{i=1}^{m} \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}_1\rangle)\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}_2\rangle)\,. \tag{127}$$

These kernels are random (because the weights $\boldsymbol{w}_i$ are) and have finite rank, namely rank at most $p$, where $p = m$ in the first case and $p = md$ in the second. The last property is equivalent to the fact that the RKHS is at most $p$-dimensional. As the number of neurons diverge, these kernels converge to their expectations $K_{\mathsf{RF}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $K_{\mathsf{NT}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Since the distribution of $\boldsymbol{w}_i$ is invariant under rotations in $\mathbb{R}^d$, so are these kernels. The kernels $K_{\mathsf{RF}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $K_{\mathsf{NT}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ can therefore be written as functions of $\|\boldsymbol{x}_1\|_2$, $\|\boldsymbol{x}_2\|_2$ and $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle$. In particular, if we assume that data are normalized, say $\|\boldsymbol{x}_1\|_2 = \|\boldsymbol{x}_2\|_2 = \sqrt{d}$, then we have the particularly simple form

$$K_{\mathsf{RF}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = H_{\mathsf{RF},d}(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d)\,, \tag{128}$$

$$K_{\mathsf{NT}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = d\, H_{\mathsf{NT},d}(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d)\,, \tag{129}$$

where

$$H_{\mathsf{RF},d}(q) := \mathbb{E}_{\boldsymbol{w}}\{\sigma(\sqrt{d}\langle \boldsymbol{w}, \boldsymbol{e}_1\rangle)\sigma(\sqrt{d}\langle \boldsymbol{w}, q\boldsymbol{e}_1 + \bar{q}\boldsymbol{e}_2\rangle)\}\,, \tag{130}$$

$$H_{\mathsf{NT},d}(q) := q\mathbb{E}_{\boldsymbol{w}}\{\sigma'(\sqrt{d}\langle \boldsymbol{w}, \boldsymbol{e}_1\rangle)\sigma'(\sqrt{d}\langle \boldsymbol{w}, q\boldsymbol{e}_1 + \bar{q}\boldsymbol{e}_2\rangle)\}\,, \tag{131}$$

with $\bar{q} := \sqrt{1-q^2}$.

The convergence $K_{\mathsf{RF},m} \to K_{\mathsf{RF}}$, $K_{\mathsf{NT},m} \to K_{\mathsf{NT},m}$ takes place under suitable assumptions, pointwise [RR07]. However, we would like to understand the qualitative behavior of the generalization error in the above linearized models.

(*i*) Does the procedure (122) share qualitative behavior with KRR, as discussed in Section 4? In particular, can min-norm interpolation be (nearly) optimal in the RF or NT models as well?

(*ii*) How large should $m$ be for the generalization properties of RF or NT ridge regression to match those of the associated kernel?

(*iii*) What discrepancies between KRR and RF or NT regression can we observe when $m$ is not sufficiently large?

(*iv*) Is there any advantage of one of the three methods (KRR, RF, NT) over the others?

Throughout this section we assume an isotropic model for the distribution of the covariates $\boldsymbol{x}_i$, namely we assume $\{(\boldsymbol{x}_i, y_i)\}_{i\le n}$ to be i.i.d., with

$$y_i = f^*(\boldsymbol{x}_i) + \varepsilon_i\,, \quad \boldsymbol{x}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))\,, \tag{132}$$

where $f^* \in L^2(\mathbb{S}^{d-1})$ is a square-integrable function on the sphere and $\varepsilon_i$ is noise independent of $\boldsymbol{x}_i$, with $\mathbb{E}\{\varepsilon_i\} = 0$, $\mathbb{E}\{\varepsilon_i^2\} = \tau^2$. We will also consider a modification of this model in which $\boldsymbol{x}_i \sim \mathsf{N}(0, \mathbf{I}_d)$; the two settings are very close to each other in high dimension. Let us emphasize that we do not make any regularity assumption about the target function beyond square integrability, which is the bare minimum for the risk

to be well defined. On the other hand, the covariates have a simple isotropic distribution and the noise has variance independent of $\boldsymbol{x}_i$ (it is homoscedastic).

While homoscedasticity is not hard to relax to an upper bound on the noise variance, it is useful to comment on the isotropicity assumption. The main content of this assumption is that the ambient dimension $d$ of the covariate vectors does coincide with the intrinsic dimension of the data. If, for instance, the $\boldsymbol{x}_i$ lie on a $d_0$-dimensional subspace in $\mathbb{R}^d$, $d_0 \ll d$, then it is intuitively clear that $d$ would have to be replaced by $d_0$ below. Indeed this is a special case of a generalization studied in [GMMM20b]. An even more general setting is considered in [MMM21], where $\boldsymbol{x}_i$ belongs to an abstract space. The key assumption there is that leading eigenfunctions of the associated kernel are delocalized.

We evaluate the quality of method (122) using the square loss

$$L(\lambda) := \mathbb{E}_{\boldsymbol{x}}\big\{(f^*(\boldsymbol{x}) - f_{\rm lin}(\boldsymbol{x}; \widehat{\boldsymbol{a}}(\lambda))^2\big\}\,. \tag{133}$$

The expectation is with respect to the test point $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$; note that the risk is random because $\widehat{\boldsymbol{a}}(\lambda)$ depends on the training data. However, in all the results below, it concentrates around a non-random value. We add subscripts, and write $L_{\mathsf{RF}}(\lambda)$ or $L_{\mathsf{NT}}(\lambda)$ to refer to the two classes of models above.

## 6.3 Random features model

We begin by considering the random features model $\mathcal{F}_{\mathsf{RF}}$. A number of authors have established upper bounds on its minimax generalization error for suitably chosen positive values of the regularization [RR17, RR09]. Besides the connection to neural networks, $\mathcal{F}_{\mathsf{RF}}$ can be viewed as a randomized approximation for the RKHS associated with $K_{\mathsf{RF}}$. A closely related approach in this context is provided by randomized subset selection, also known as Nyström's method [WS01, Bac13, EAM15, RCR15].

The classical random features model $\mathcal{F}_{\mathsf{RF}}$ is mathematically easier to analyze than the neural tangent model $\mathcal{F}_{\mathsf{NT}}$, and a precise picture can be established that covers the interpolation limit. Several elements of this picture have been proved to generalize to the $\mathsf{NT}$ model as well, as discussed in the next subsection.

We focus on the high-dimensional regime, $m, n, d \to \infty$; as discussed in Section 4, interpolation methods have appealing properties in high dimension. Complementary asymptotic descriptions are obtained depending on how $m, n, d$ diverge. In Section 6.3.1 we discuss the behavior at a coarser scale, namely when $m$ and $n$ scale polynomially in $d$: this type of analysis provides a simple quantitative answer to the question of how large $m$ should be to approach the $m = \infty$ limit. Next, in Section 6.3.2, we consider the proportional regime $m \asymp n \asymp d$. This allows us to explore more precisely what happens in the transition from underparametrized to overparametrized.

### 6.3.1 Polynomial scaling

The following characterization was proved in [MMM21] (earlier work by [GMMM20a] established this result for the two limiting cases $m = \infty$ and $n = \infty$). In what follows, we let $L^2(\gamma)$ denote the space of square integrable functions on $\mathbb{R}$, with respect to the standard Gaussian measure $\gamma(\mathrm{d}x) = (2\pi)^{-1/2} e^{-x^2/2}\mathrm{d}x$, and we write $\langle\,\cdot\,,\,\cdot\,\rangle_{L^2(\gamma)}$, $\|\,\cdot\,\|_{L^2(\gamma)}$ for the associated scalar product and norm.

**Theorem 6.1.** *Fix an integer $\ell > 0$. Let the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ be independent of $d$ and such that: (i) $|\sigma(x)| \le c_0 \exp(|x|^{c_1})$ for some constants $c_0 > 0$ and $c_1 < 1$, and (ii) $\langle \sigma, q\rangle_{L^2(\gamma)} \neq 0$ for any non-vanishing polynomial $q$, with $\deg(q) \le \ell$. Assume $\max((n/m), (m/n)) \ge d^{\delta}$ and $d^{\ell+\delta} \le \min(m, n) \le d^{\ell+1-\delta}$ for some constant $\delta > 0$. Then for any $\lambda = O_d((m/n) \vee 1)$, and all $\eta > 0$,*

$$L_{\mathsf{RF}}(\lambda) = \|\mathsf{P}_{>\ell} f^*\|_{L^2}^2 + o_d(1)\big(\|f^*\|_{L^2}^2 + \|\mathsf{P}_{>\ell} f^*\|_{L^{2+\eta}}^2 + \tau^2\big)\,. \tag{134}$$

In words, as long as the number of parameters $m$ and the number of samples $n$ are well separated, the test error is determined by the minimum of $m$ and $n$:

- For $m \ll n$, the approximation error dominates. If $d^\ell \ll m \ll d^{\ell+1}$, the model fits the projection of $f$ onto degree-$\ell$ polynomials perfectly but does not fit the higher degree components at all: $\widehat{f}_\lambda \approx \mathsf{P}_{\le \ell} f$. This is consistent with a parameter-counting heuristic: degree-$\ell$ polynomials form a subspace of dimension $\Theta(d^\ell)$ and in order to approximate them we need a network with $\Omega(d^\ell)$ parameters. Surprisingly, this transition is sharp.
- For $n \ll m$, the statistical error dominates. If $d^\ell \ll n \ll d^{\ell+1}$, $\widehat{f}_\lambda \approx \mathsf{P}_{\le \ell} f$. This is again consistent with a parameter-counting heuristic: to learn degree-$\ell$ polynomials we need roughly as many samples as parameters.
- Both of the above are achieved for any sufficiently small value of the regularization parameter $\lambda$. In particular, they apply to min-norm interpolation (corresponding to the case $\lambda = 0^+$).

From a practical perspective, if the sample size $n$ is given, we might be interested in choosing the number of neurons $m$. The above result indicates that the test error roughly decreases until the overparametrization threshold $m \approx n$, and that there is limited improvement from increasing the network size beyond $m \ge nd^\delta$. At this point, RF ridge regression achieves the same error as the corresponding kernel method. Indeed the statement of Theorem 6.1 holds for the case of KRR as well, by identifying it with the limit $m = \infty$ [GMMM20a].

Note that the infinite width (kernel) limit $m = \infty$ corresponds to the setting already investigated in Theorem 4.10. Indeed, the staircase phenomenon in the $m = \infty$ case of Theorem 6.1 corresponds to the multiple descent behavior seen in Theorem 4.10. The two results do not imply each other because Theorem 4.10 assumes $f^*$ to have bounded RKHS norm; Theorem 6.1 does not make this assumption, but is not as sharp for functions with bounded RKHS norm.

The significance of polynomials in Theorem 6.1 is related to the fact that the kernel $K_{\mathsf{RF}}$ is invariant under rotations (see (128)). As a consequence, the eigenfunctions of $K_{\mathsf{RF}}$ are spherical harmonics, that is, restrictions of homogeneous harmonic polynomials in $\mathbb{R}^d$ to the sphere $\mathbb{S}^{d-1}(\sqrt{d})$, with eigenvalues given by their degrees. [MMM21] have obtained analogous results for more general probability spaces $(\mathcal{X}, \mathbb{P})$ for the covariates, and more general random features models. The role of low-degree polynomials is played by the top eigenfunctions of the associated kernel.

The mathematical phenomenon underlying Theorem 6.1 can be understood by considering the feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n \times m}$:

$$\boldsymbol{\Phi} := \begin{bmatrix} \sigma(\langle \boldsymbol{x}_1, \boldsymbol{w}_1 \rangle) & \sigma(\langle \boldsymbol{x}_1, \boldsymbol{w}_2 \rangle) & \cdots & \sigma(\langle \boldsymbol{x}_1, \boldsymbol{w}_m \rangle) \\ \sigma(\langle \boldsymbol{x}_2, \boldsymbol{w}_1 \rangle) & \sigma(\langle \boldsymbol{x}_2, \boldsymbol{w}_2 \rangle) & \cdots & \sigma(\langle \boldsymbol{x}_2, \boldsymbol{w}_m \rangle) \\ \vdots & \vdots & & \vdots \\ \sigma(\langle \boldsymbol{x}_n, \boldsymbol{w}_1 \rangle) & \sigma(\langle \boldsymbol{x}_n, \boldsymbol{w}_2 \rangle) & \cdots & \sigma(\langle \boldsymbol{x}_n, \boldsymbol{w}_m \rangle) \end{bmatrix}. \tag{135}$$

The $i$th row of this matrix is the feature vector associated with the $i$th sample. We can decompose $\boldsymbol{\Phi}$ according to the eigenvalue decomposition of $\sigma$, seen as an integral operator from $L^2(\mathbb{S}^{d-1}(1))$ to $L^2(\mathbb{S}^{d-1}(\sqrt{d}))$:

$$a(\boldsymbol{w}) \mapsto \int \sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)\, a(\boldsymbol{w})\, \tau_d(\mathrm{d}\boldsymbol{w})$$

(where $\tau_d$ is the uniform measure on $\mathbb{S}^{d-1}(1)$). This takes the form

$$\sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle) = \sum_{k=0}^{\infty} s_k \psi_k(\boldsymbol{x}) \phi_k(\boldsymbol{w}), \tag{136}$$

where $(\psi_j)_{j \ge 1}$ and $(\phi_j)_{j \ge 1}$ are two orthonormal systems in $L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ and $L^2(\mathbb{S}^{d-1}(1))$ respectively, and the $s_j$ are singular values $s_0 \ge s_1 \ge \cdots \ge 0$. (In the present example, $\sigma$ can be regarded as a self-adjoint operator on $L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ after rescaling the $\boldsymbol{w}_j$, and hence the $\phi_j$ and $\psi_j$ can be taken to coincide up to a rescaling, but this is not crucial.)

The eigenvectors are grouped into eigenspaces $\mathcal{V}_\ell$ indexed by $\ell \in \mathbb{Z}_{\ge 0}$, where $\mathcal{V}_\ell$ consists of the degree-$\ell$ polynomials, and

$$\dim(\mathcal{V}_\ell) =: B(d,\ell) = \frac{d-2+2\ell}{d-2}\binom{d-3+\ell}{\ell}, \ B(d,\ell) \asymp d^\ell/\ell!.$$

We write $s^{(\ell)}$ for the eigenvalue associated with eigenspace $\mathcal{V}_\ell$: it turns out that $s^{(\ell)} \asymp d^{-\ell/2}$, for a generic $\sigma$; $(s^{(\ell)})^2 B(d,\ell) \le C$ since $\sigma$ is square integrable. Let $\boldsymbol{\psi}_k = (\psi_k(\boldsymbol{x}_1), \dots, \psi_k(\boldsymbol{x}_n))^\mathsf{T}$ be the evaluation of the $k$th left eigenfunction at the $n$ data points, and let $\boldsymbol{\phi}_k = (\phi_k(\boldsymbol{w}_1), \dots, \phi_k(\boldsymbol{w}_m))^\mathsf{T}$ be the evaluation of the $k$th right eigenfunction at the $m$ neuron parameters. Further, let $k(\ell) := \sum_{\ell' \le \ell} B(d,\ell')$. Following our approach in Section 4, we decompose $\boldsymbol{\Phi}$ into a 'low-frequency' and a 'high-frequency' component,

$$\boldsymbol{\Phi} = \boldsymbol{\Phi}_{\le \ell} + \boldsymbol{\Phi}_{>\ell}\,, \tag{137}$$

$$\boldsymbol{\Phi}_{\le \ell} = \sum_{j=0}^{k(\ell)} s_j \boldsymbol{\psi}_j \boldsymbol{\phi}_j^\top = \boldsymbol{\psi}_{\le \ell} \boldsymbol{S}_{\le \ell} \boldsymbol{\phi}_{\le \ell}^\top\,, \tag{138}$$

where $\boldsymbol{S}_{\le \ell} = \mathrm{diag}(s_1, \dots, s_{k(\ell)})$, $\boldsymbol{\psi}_{\le \ell} \in \mathbb{R}^{n \times k(\ell)}$ is the matrix whose $j$th column is $\boldsymbol{\psi}_j$, and $\boldsymbol{\phi}_{\le \ell} \in \mathbb{R}^{m \times k(\ell)}$ is the matrix whose $j$th column is $\boldsymbol{\phi}_j$.

Consider, to be definite, the overparametrized case $m \ge n^{1+\delta}$, and assume $d^{\ell+\delta} \le n$. Then we can think of $\boldsymbol{\phi}_j$, $\boldsymbol{\psi}_j$, $j \le k(\ell)$ as densely sampled eigenfunctions. This intuition is accurate in the sense that $\boldsymbol{\psi}_{\le \ell}^\mathsf{T} \boldsymbol{\psi}_{\le \ell} \approx n \mathbf{I}_{k(\ell)}$ and $\boldsymbol{\phi}_{\le \ell}^\mathsf{T} \boldsymbol{\phi}_{\le \ell} \approx m \mathbf{I}_{k(\ell)}$ [MMM21]. Further, if $n \le d^{\ell+1-\delta}$, the 'high-frequency' part of the decomposition (137) behaves similarly to noise along directions orthogonal to the previous ones. Namely, $(i)$ $\boldsymbol{\Phi}_{>\ell} \boldsymbol{\phi}_{\le \ell} \approx 0$, $\boldsymbol{\psi}_{\le \ell}^\mathsf{T} \boldsymbol{\Phi}_{>\ell} \approx 0$, and $(ii)$ its singular values (except those along the low-frequency components) concentrate: for any $\delta' > 0$,

$$\kappa_\ell^{1/2} n^{-\delta'} \le \sigma_{n-k(\ell)}(\boldsymbol{\Phi}_{>\ell})/m^{1/2} \le \sigma_1(\boldsymbol{\Phi}_{>\ell})/m^{1/2} \le \kappa_\ell^{1/2} n^{\delta'}\,,$$

where $\kappa_\ell := \sum_{j \ge k(\ell)+1} s_j^2$.

In summary, regression with respect to the random features $\sigma(\langle \boldsymbol{w}_j, \cdot \rangle)$ turns out to be essentially equivalent to kernel ridge regression with respect to a polynomial kernel of degree $\ell$, where $\ell$ depends on the smaller of the sample size and the network size. Higher degree parts in the activation function effectively behave as noise in the regressors. We will next see that this picture can become even more precise in the proportional regime $m \asymp n$.

### 6.3.2 Proportional scaling

Theorem 6.1 requires that $m$ and $n$ are well separated. When $m, n$ are close to each other, the feature matrix (135) is nearly square and we might expect its condition number to be large. When this is the case, the variance component of the risk can also be large.

Theorem 6.1 also requires the smaller of $m$ and $n$ to be well separated from $d^\ell$, with $\ell$ any integer. For $d^\ell \ll m \ll d^{\ell+1}$ the model has enough degrees of freedom to represent (at least in principle) all polynomials of degree at most $\ell$ and not enough to represent even a vanishing fraction of all polynomials of degree $\ell+1$. Hence it behaves in a particularly simple way. On the other hand, when $m$ is comparable to $d^\ell$, the model can partially represent degree-$\ell$ polynomials, and its behavior will be more complex. Similar considerations apply to the sample size $n$.

What happens when $m$ is comparable to $n$, and both are comparable to an integer power of $d$? Figure 1 reports simulations within the data model introduced above. We performed ridge regression as per (122), with a small value of the regularization parameter, $\lambda = 10^{-3}(m/d)$. We report test error and train error for several network widths $m$, plotting them as a function of the overparametrization ratio $m/n$.

We observe that the train error decreases with the overparametrization ratio, and becomes very small for $m/n \ge 1$: it is not exactly 0 because we are using $\lambda > 0$, but for $m/n > 1$ it vanishes as $\lambda \to 0$. On the other hand, the test error displays a peak at the interpolation threshold $m/n = 1$. For $\lambda = 0^+$ the

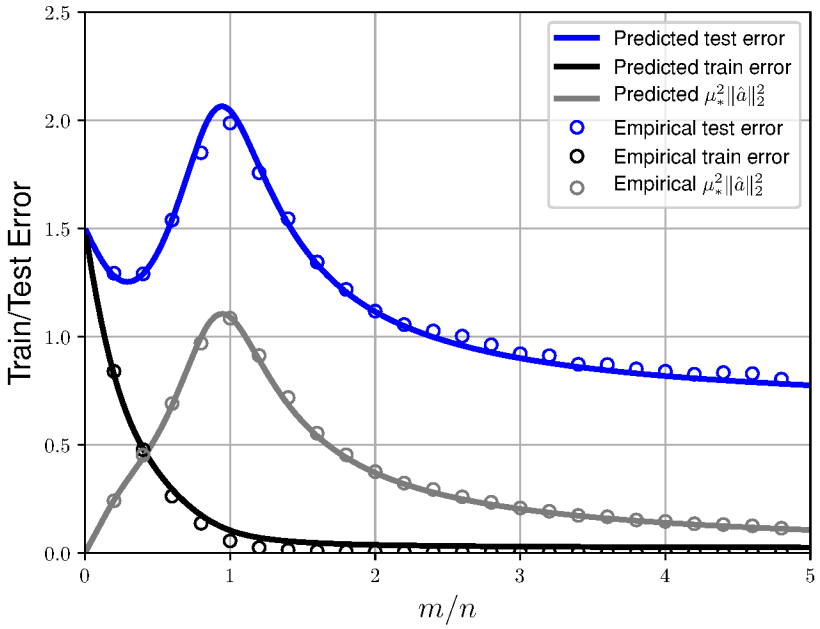


Figure 1: Train and test error of a random features model (two-layer neural net with random first layer) as a function of the overparametrization ratio $m/n$. Here $d = 100$, $n = 400$, $\tau^2 = 0.5$, and the target function is $f^* = \langle \boldsymbol{\beta}_0, \boldsymbol{x} \rangle$, $\|\boldsymbol{\beta}_0\|_2 = 1$. The model is fitted using ridge regression with a small regularization parameter $\lambda = 10^{-3}(m/d)$. Circles report the results of numerical simulations (averaged over 20 realizations), while lines are theoretical predictions for the $m, n, d \to \infty$ asymptotics.

error actually diverges at this threshold. It then decreases and converges rapidly to an asymptotic value as $m/n \gg 1$. If both $n/d \gg 1$, and $m/n \gg 1$, the asymptotic value of the test error is given by $\|\mathsf{P}_{>1} f^*\|_{L^2}$: the model is fitting the degree-one polynomial component of the target function perfectly and behaves trivially on higher degree components. This matches the picture obtained under polynomial scalings, in Theorem 6.1, and actually indicates that a far smaller separation between $m$ and $n$ is required than assumed in that theorem. Namely, $m/n \gg 1$ instead of $m/n \geq d^{\delta}$ appears to be sufficient for the risk to be dominated by the statistical error.

The peculiar behavior illustrated in Figure 1 was first observed empirically in neural networks and then shown to be ubiquitous for numerous over-parametrized models [GSd+19, SGd+19, BHMM19]. It is commonly referred to as the 'double descent phenomenon', after [BHMM19].

Figure 1 displays curves that are exact asymptotic predictions in the limit $m, n, d \to \infty$, with $m/d \to \psi_{\mathrm{w}}$, $n/d \to \psi_{\mathrm{s}}$. Explicit formulas for these asymptotics were originally established in [MM19] using an approach from random matrix theory, which we will briefly outline. The first step is to write the risk as an explicit function of the matrices $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ (the matrix whose $i$th row is the sample $\boldsymbol{x}_i$), $\boldsymbol{\Theta} \in \mathbb{R}^{m \times d}$ (the matrix whose $j$th row is the sample $\boldsymbol{\theta}_j = \sqrt{d}\boldsymbol{w}_j$), and $\boldsymbol{\Phi} = \sigma(\boldsymbol{X}\boldsymbol{\Theta}^{\mathsf{T}}/\sqrt{d})$ (the feature matrix in (135)). After a straightforward calculation, one obtains

$$\begin{aligned} L_{\mathsf{RF}}(\lambda) =& \mathbb{E}_{\boldsymbol{x}}[f^*(\boldsymbol{x})^2] - \frac{2}{n}\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}/n + \lambda \mathbf{I}_m)^{-1}\boldsymbol{V} \\ &+ \frac{1}{n^2}\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}/n + \lambda \mathbf{I}_m)^{-1}\boldsymbol{U}(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}/n + \lambda \mathbf{I}_m)^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y}\,, \end{aligned} \tag{139}$$

where $\boldsymbol{V} \in \mathbb{R}^m$, $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ are matrices with entries

$$V_i := \mathbb{E}_{\boldsymbol{x}}\{\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d})\, f^*(\boldsymbol{x})\}\,, \tag{140}$$

$$U_{ij} := \mathbb{E}_{\boldsymbol{x}}\{\sigma(\langle \boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d})\, \sigma(\langle \boldsymbol{\theta}_j, \boldsymbol{x}\rangle/\sqrt{d})\}\,. \tag{141}$$

Note that the matrix $\boldsymbol{U}$ takes the form of an empirical kernel matrix, although expectation is taken over the covariates $\boldsymbol{x}$ and the kernel is evaluated at the neuron parameters $(\boldsymbol{\theta}_i)_{i\le m}$. Namely, we have $U_{ij} = H_{\mathsf{RF},d}(\langle \boldsymbol{\theta}_i, \boldsymbol{\theta}_j\rangle/d)$, where the kernel $H_{\mathsf{RF},d}$ is defined exactly[7] as in (130). Estimates similar to those of Section 4 apply here (see also [EK10]): since $m \asymp d$ we can approximate the kernel $H_{\mathsf{RF},d}$ by a linear kernel in operator norm. Namely, if we decompose $\sigma(x) = \mu_0 + \mu_1 x + \sigma_\perp(x)$, where $\mathbb{E}\{\sigma_\perp(G)\} = \mathbb{E}\{G\sigma_\perp(G)\} = 0$, and $\mathbb{E}\{\sigma_\perp(G)^2\} = \mu_*^2$, we have

$$\boldsymbol{U} = \mu_0^2 \mathbf{1}\mathbf{1}^{\mathsf{T}} + \mu_1^2 \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}} + \mu_* \mathbf{I}_m + \boldsymbol{\Delta}\,, \tag{142}$$

where $\boldsymbol{\Delta}$ is an error term that vanishes asymptotically in operator norm. Analogously, $\boldsymbol{V}$ can be approximated as $\boldsymbol{V} \approx a\mathbf{1} + \boldsymbol{\Theta}\boldsymbol{b}$ for suitable coefficients $a \in \mathbb{R}$, $\boldsymbol{b} \in \mathbb{R}^d$.

Substituting these approximations for $\boldsymbol{U}$ and $\boldsymbol{V}$ in (139) yields an expression of the risk in terms of the three (correlated) random matrices $\boldsymbol{X}$, $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$. Standard random matrix theory does not apply directly to compute the asymptotics of this expression. The main difficulty is that the matrix $\boldsymbol{\Phi}$ does not have independent or nearly independent entries. It is instead obtained by applying a nonlinear function to a product of matrices with (nearly) independent entries; see (135). The name 'nonlinear random matrix theory' has been coined to refer to this setting [PW17]. Techniques from random matrix theory have been adapted to this new class of random matrices. In particular, the leave-one-out method can be used to derive a recursion for the resolvent, as first shown for this type of matrices in [CS13], and the moments method was first used in [FM19] (both of these papers consider symmetric random matrices, but these techniques extend to the asymmetric case). Further results on kernel random matrices can be found in [DV13, LLC18] and [PW18].

Using these approaches, the exact asymptotics of $L_{\mathsf{RF}}(\lambda)$ was determined in the proportional asymptotics $m, n, d \to \infty$ with $m/d \to \psi_{\mathrm{w}}$ ( $\psi_{\mathrm{w}}$ represents the number of neurons per dimension), $n/d \to \psi_{\mathrm{s}}$ ($\psi_{\mathrm{s}}$ represents the number of samples per dimension). The target function $f^*$ is assumed to be square integrable and such that $\mathsf{P}_{>1} f^*$ is a Gaussian isotropic function.[8] In this setting, the risk takes the form

$$\begin{aligned} L_{\mathsf{RF}}(\lambda) =& \|\mathsf{P}_1 f^*\|_{L^2}^2 \mathscr{B}(\zeta, \psi_{\mathrm{w}}, \psi_{\mathrm{s}}, \lambda/\mu_*^2) \\ &+ (\tau^2 + \|\mathsf{P}_{>1} f^*\|_{L^2}^2)\mathscr{V}(\zeta, \psi_{\mathrm{w}}, \psi_{\mathrm{s}}, \lambda/\mu_*^2) + \|\mathsf{P}_{>1} f^*\|_{L^2}^2 + o_d(1)\,, \end{aligned} \tag{143}$$

where $\zeta := |\mu_1|/\mu_*$. The functions $\mathscr{B}$, $\mathscr{V} \ge 0$ are explicit and correspond to an effective bias term and an effective variance term. Note the additive term $\|\mathsf{P}_{>1} f^*\|_{L^2}^2$: in agreement with Theorem 6.1, the nonlinear component of $f^*$ cannot be learnt at all (recall that $m, n = O(d)$ here). Further $\|\mathsf{P}_{>1} f^*\|_{L^2}^2$ is added to the noise strength in the 'variance' term: high degree components of $f^*$ are equivalent to white noise at small sample/network size.

The expressions for $\mathscr{B}$, $\mathscr{V}$ can be used to plot curves such as those in Figure 1: we refer to [MMM21] for explicit formulas. As an interesting conceptual consequence, these results establish a universality phenomenon: the risk under the random features model is asymptotically the same as the risk of a mathematically simpler model. This simpler model can be analyzed by a direct application of standard random matrix theory [HMRT20].

We refer to the simpler equivalent model as the 'noisy features model.' In order to motivate it, recall the decomposition $\sigma(x) = \mu_0 + \mu_1 x + \sigma_\perp(x)$ (with the three components being orthogonal in $L^2(\gamma)$). Accordingly,

[7] The two kernels coincide because we are using the same distribution for $\boldsymbol{x}_i$ and $\boldsymbol{\theta}_j$: while this symmetry simplifies some calculations, it is not really crucial.

[8] Concretely, for each $\ell \ge 2$, let $\boldsymbol{f}_\ell = (f_{k,\ell})_{k\le B(d,\ell)}$ be the coefficients of $f^*$ in a basis of degree-$\ell$ spherical harmonics. Then $\boldsymbol{f}_\ell \sim \mathsf{N}(0, F_\ell^2 \mathbf{I}_{B(d,\ell)})$ independently across $\ell$.

we decompose the feature matrix as

$$\begin{aligned}\boldsymbol{\Phi} &= \boldsymbol{\Phi}_{\le 1} + \boldsymbol{\Phi}_{>1} \\ &= \mu_0 \mathbf{1}\mathbf{1}^{\mathsf{T}} + \frac{\mu_1}{\sqrt{d}} \boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}} + \mu_* \tilde{\boldsymbol{Z}}\,,\end{aligned}$$

where $\tilde{Z}_{ij} = \sigma_\perp(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j\rangle/\sqrt{d})/\mu_*$. Note that the entries of $\tilde{\boldsymbol{Z}}$ have zero mean and are asymptotically uncorrelated. Further they are asymptotically uncorrelated with the entries[9] of $\boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}}/\sqrt{d}$.

As we have seen in Section 6.3.1, the matrix $\tilde{\boldsymbol{Z}}$ behaves in many ways as a matrix with independent entries, independent of $\boldsymbol{\Theta}, \boldsymbol{X}$. In particular, if $\max(m,n) \ll d^2$ and either $m \gg n$ or $m \ll n$, its eigenvalues concentrate around a deterministic value (see discussion below (137)).

The noisy features model is obtained by replacing $\tilde{\boldsymbol{Z}}$ with a matrix $\boldsymbol{Z}$, with independent entries, independent of $\boldsymbol{\Theta}$, $\boldsymbol{X}$. Accordingly, we replace the target function with a linear function with additional noise. In summary:

$$\boldsymbol{\Phi}^{\rm NF} = \mu_0 \mathbf{1}\mathbf{1}^{\mathsf{T}} + \frac{\mu_1}{\sqrt{d}} \boldsymbol{\Theta}\boldsymbol{X}^{\mathsf{T}} + \mu_* \boldsymbol{Z}, \quad (Z_{ij})_{i\le n, j\le m} \sim \mathsf{N}(0,1)\,, \tag{144}$$

$$\boldsymbol{y} = b_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \tau_+ \tilde{\boldsymbol{g}}, \quad (\tilde{g}_i)_{i\le n} \sim \mathsf{N}(0,1)\,. \tag{145}$$

Here the random variables $(\tilde{g}_i)_{i\le n}, (Z_{ij})_{i\le n, j\le m}$ are mutually independent, and independent of all the others, and the parameters $b_0, \boldsymbol{\beta}, \tau_+$ are fixed by the conditions $\mathsf{P}_{\le 1} f^*(\boldsymbol{x}) = b_0 + \langle \boldsymbol{\beta}, \boldsymbol{x}\rangle$ and $\tau_+^2 = \tau^2 + \|\mathsf{P}_{>1} f^*\|_{L^2}^2$.

The next statement establishes asymptotic equivalence of the noisy and random features model.

**Theorem 6.2.** *Under the data distribution introduced above, let $L_{\mathsf{RF}}(\lambda)$ denote the risk of ridge regression in the random features model with regularization $\lambda$, and let $L_{\rm NF}(\lambda)$ be the risk in the noisy features model. Then we have, in $n, m, d \to \infty$ with $m/d \to \psi_{\rm w}$, $n/d \to \psi_{\rm s}$,*

$$L_{\mathsf{RF}}(\lambda) = L_{\rm NF}(\lambda) \cdot \big(1 + o_n(1)\big)\,. \tag{146}$$

Knowing the exact asymptotics of the risk allows us to identify phenomena that otherwise would be out of reach. A particularly interesting one is the optimality of interpolation at high signal-to-noise ratio.

**Corollary 6.3.** *Define the signal-to-noise ratio of the random features model as $\mathsf{SNR}_d := \|\mathsf{P}_1 f^*\|_{L^2}^2/(\|\mathsf{P}_{>1} f^*\|_{L^2}^2 + \tau^2)$, and let $L_{\mathsf{RF}}(\lambda)$ be the risk of ridge regression with regularization $\lambda$. Then there exists a critical value $\mathsf{SNR}_* > 0$ such that the following hold.*

- *(i) If $\lim_{d\to\infty} \mathsf{SNR}_d = \mathsf{SNR}_\infty > \mathsf{SNR}_*$, then the optimal regularization parameter is $\lambda = 0^+$, in the sense that $L_{\mathsf{RF},\infty}(\lambda) := \lim_{d\to\infty} L_{\mathsf{RF}}(\lambda)$ is monotone increasing for $\lambda \in (0,\infty)$.*
- *(ii) If $\lim_{d\to\infty} \mathsf{SNR}_d = \mathsf{SNR}_\infty < \mathsf{SNR}_*$, then the optimal regularization parameter is $\lambda > 0$, in the sense that $L_{\mathsf{RF},\infty}(\lambda) := \lim_{d\to\infty} L_{\mathsf{RF}}(\lambda)$ is monotone decreasing for $\lambda \in (0, \lambda_0)$ with $\lambda_0 > 0$.*

In other words, above a certain threshold in $\mathsf{SNR}$, (near) interpolation is required in order to achieve optimal risk, not just optimal rates.

The universality phenomenon of Theorem 6.2 first emerged in random matrix theory studies of (symmetric) kernel inner product random matrices. In that case, the spectrum of such a random matrix was shown in [CS13] to behave asymptotically as the one of the sum of independent Wishart and Wigner matrices, which correspond respectively to the linear and nonlinear parts of the kernel (see also [FM19] where this remark is made more explicit). In the context of random features ridge regression, this type of universality was first pointed out in [HMRT20], which proved a special case of Theorem 6.2. In [GMKZ19] and [GRM+20], a universality conjecture was put forward on the basis of statistical physics arguments and proved to hold in online learning schemes (that is, if each sample is visited only once).

[9] Uncorrelatedness holds only asymptotically, because the distribution of $\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j\rangle/\sqrt{d}$ is not exactly Gaussian, but only asymptotically so, while the decomposition $\sigma(x) = \sigma_0 + \sigma_1 x + \sigma_\perp(x)$ is taken in $L^2(\gamma)$.

Universality is conjectured to hold in significantly broader settings than ridge-regularized least-squares. This is interesting because analysing the noisy feature models is often significantly easier than the original random features model. For instance [MRSY19] studied max margin classification under the universality hypothesis, and derived an asymptotic characterization of the test error using Gaussian comparison inequalities. Related results were obtained by [TPT20] and [KT20], among others.

Finally, a direct proof of universality for general strongly convex smooth losses was recently proposed in [HL20] using the Lindeberg interpolation method.

## 6.4 Neural tangent model

The neural tangent model $\mathcal{F}_{\mathsf{NT}}$ —recall (125)— has not (yet) been studied in as much detail as the random features model. The fundamental difficulty is related to the fact that the features matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n\times md}$ no longer has independent columns:

$$\boldsymbol{\Phi} := \begin{bmatrix} \sigma'(\langle \boldsymbol{x}_1, \boldsymbol{w}_1\rangle)\boldsymbol{x}_1^{\mathsf{T}} & \sigma'(\langle \boldsymbol{x}_1, \boldsymbol{w}_2\rangle)\boldsymbol{x}_1^{\mathsf{T}} & \cdots & \sigma(\langle \boldsymbol{x}_1, \boldsymbol{w}_m\rangle)\boldsymbol{x}_1^{\mathsf{T}} \\ \sigma'(\langle \boldsymbol{x}_2, \boldsymbol{w}_1\rangle)\boldsymbol{x}_2^{\mathsf{T}} & \sigma(\langle \boldsymbol{x}_2, \boldsymbol{w}_2\rangle)\boldsymbol{x}_2^{\mathsf{T}} & \cdots & \sigma(\langle \boldsymbol{x}_2, \boldsymbol{w}_m\rangle)\boldsymbol{x}_2^{\mathsf{T}} \\ \vdots & \vdots & & \vdots \\ \sigma'(\langle \boldsymbol{x}_n, \boldsymbol{w}_1\rangle)\boldsymbol{x}_n^{\mathsf{T}} & \sigma(\langle \boldsymbol{x}_n, \boldsymbol{w}_2\rangle)\boldsymbol{x}_n^{\mathsf{T}} & \cdots & \sigma(\langle \boldsymbol{x}_n, \boldsymbol{w}_m\rangle)\boldsymbol{x}_n^{\mathsf{T}} \end{bmatrix}. \tag{147}$$

Nevertheless, several results are available and point to a common conclusion: the generalization properties of $\mathsf{NT}$ are very similar to those of $\mathsf{RF}$, provided we keep the number of parameters constant, which amounts to reducing the number of neurons according to $m_{\mathsf{NT}}d = p_{\mathsf{NT}} = p_{\mathsf{RF}} = m_{\mathsf{RF}}$.

Before discussing rigorous results pointing in this direction, it is important to emphasize that, even if the two models are statistically equivalent, they can differ from other points of view. In particular, at prediction time both models have complexity $O(md)$. Indeed, in the case of $\mathsf{RF}$ the most complex operation is the matrix vector multiplication $\boldsymbol{x} \mapsto \boldsymbol{W}\boldsymbol{x}$, while for $\mathsf{NT}$ two such multiplications are needed $\boldsymbol{x} \mapsto \boldsymbol{W}\boldsymbol{x}$ and $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$ (here $\boldsymbol{A} \in \mathbb{R}^{m\times d}$ is the matrix with rows $(\boldsymbol{a}_i)_{i\le m}$. If we keep the same number of parameters (which we can regard as a proxy for expressivity of the model), we obtain complexity $O(pd)$ for $\mathsf{RF}$ and $O(p)$ for $\mathsf{NT}$. Similar considerations apply at training time. In other words, if we are constrained by computational complexity, in high dimension $\mathsf{NT}$ allows significantly better expressivity.

A first element confirming this picture is provided by the following result, which partially generalizes Theorem 6.1. In order to state this theorem, we introduce a useful notation. Given a function $f : \mathbb{R} \to \mathbb{R}$, such that $\mathbb{E}\{f(G)^2\} < \infty$, we let $\mu_k(f) := \mathbb{E}\{\mathrm{He}_k(G)f(G)\}$ denote the $k$th coefficient of $f$ in the basis of Hermite polynomials.

**Theorem 6.4.** *Fix an integer $\ell > 0$. Let the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ be weakly differentiable, independent of $d$, and such that: (i) $|\sigma'(x)| \le c_0 \exp(c_1 x^2/4)$ for some constants $c_0 > 0$, and $c_1 < 1$, (ii) there exist $k_1$, $k_2 \ge 2\ell + 7$ such that $\mu_{k_1}(\sigma'), \mu_{k_2}(\sigma') \ne 0$, and $\mu_{k_1}(x^2\sigma')/\mu_{k_1}(\sigma') \ne \mu_{k_1}(x^2\sigma')/\mu_{k_1}(\sigma')$, and (iii) $\mu_k(\sigma) \ne 0$ for all $k \le \ell + 1$. Then the following holds.*

*Assume either $n = \infty$ (in which case we are considering pure approximation error) or $m = \infty$ (that is, the test error of kernel ridge regression) and $d^{\ell+\delta} \le \min(md; n) \le d^{\ell+1-\delta}$ for some constant $\delta > 0$. Then, for any $\lambda = o_d(1)$ and all $\eta > 0$,*

$$L_{\mathsf{NT}}(\lambda) = \|\mathsf{P}_{>\ell} f^*\|_{L^2}^2 + o_d(1)\big(\|f^*\|_{L^2}^2 + \tau^2\big)\,. \tag{148}$$

In this statement we abused notation in letting $m = \infty$ denote the case of KRR, and letting $n = \infty$ refer to the approximation error:

$$\lim_{n\to\infty} L_{\mathsf{NT}}(\lambda) = \inf_{\widehat{f}\in\mathcal{F}_{\mathsf{NT}}^m} \mathbb{E}\big\{[f^*(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x})]^2\big\}\,. \tag{149}$$

Note that here the $\mathsf{NT}$ kernel is a rotationally invariant kernel on $\mathbb{S}^{d-1}(\sqrt{d})$ and hence takes the same form as the $\mathsf{RF}$ kernel, namely $K_{\mathsf{NT}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = d\, H_{\mathsf{NT},d}(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d)$ (see (128)). Hence the $m = \infty$ case of the last theorem is not new: it can be regarded as a special case of Theorem 6.1.

On the other hand, the $n = \infty$ portion of the last theorem is new. In words, if $d^{\ell+\delta} \le md \le d^{\ell+1-\delta}$, then $\mathcal{F}_{\mathsf{NT}}^m$ can approximate degree-$\ell$ polynomials to an arbitrarily good relative accuracy, but is roughly orthogonal to polynomials of higher degree (more precisely, to polynomials that have vanishing projection onto degree-$\ell$ ones). Apart from the technical assumptions, this result is identical to the $n = \infty$ case of Theorem 6.1, with the caveat that, as mentioned above, the two models should be compared by keeping the number of parameters (not the number of neurons) constant.

How do NT models behave when both $m$ and $n$ are finite? By analogy with the RF model, we would expect that the model undergoes an 'interpolation' phase transition at $md \approx n$: the test error is bounded away from 0 for $md \lesssim n$ and can instead vanish for $md \gtrsim n$. Note that finding an interpolating function $f \in \mathcal{F}_{\mathsf{NT}}^m$ amounts to solving the system of linear equations $\boldsymbol{\Phi}\boldsymbol{a} = \boldsymbol{y}$, and hence a solution exists for generic $\boldsymbol{y}$ if and only if $\mathrm{rank}(\boldsymbol{\Phi}) = n$. Lemma 5.3 implies[10] that this is indeed the case for $md \ge C_0 n \log n$ and $n \le d^{\ell_0}$ for some constant $\ell_0$ (see (96)).

In order to study the test error, it is not sufficient to lower-bound the minimum singular value of $\boldsymbol{\Phi}$, but we need to understand the structure of this matrix: results in this direction were obtained in [MZ20], for $m \le C_0 d$, for some constant $C_0$. Following the same strategy of previous sections, we decompose

$$\boldsymbol{\Phi} = \boldsymbol{\Phi}_0 + \boldsymbol{\Phi}_{\ge 1}, \tag{150}$$

$$\boldsymbol{\Phi}_0 = \mu_1 \begin{bmatrix} \boldsymbol{x}_1^{\mathsf{T}} & \boldsymbol{x}_1^{\mathsf{T}} & \cdots & \boldsymbol{x}_1^{\mathsf{T}} \\ \boldsymbol{x}_2^{\mathsf{T}} & \boldsymbol{x}_2^{\mathsf{T}} & \cdots & \boldsymbol{x}_2^{\mathsf{T}} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{x}_n^{\mathsf{T}} & \boldsymbol{x}_n^{\mathsf{T}} & \cdots & \boldsymbol{x}_n^{\mathsf{T}} \end{bmatrix}, \tag{151}$$

where $\mu_1 := \mathbb{E}\{\sigma'(G)\}$ for $G \sim \mathsf{N}(0,1)$. The empirical kernel matrix $\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}}/m$ then reads

$$\boldsymbol{K} = \frac{1}{m}\boldsymbol{\Phi}_0\boldsymbol{\Phi}_0^{\mathsf{T}} + \frac{1}{m}\boldsymbol{\Phi}_0\boldsymbol{\Phi}_{\ge 1}^{\mathsf{T}} + \frac{1}{m}\boldsymbol{\Phi}_{\ge 1}\boldsymbol{\Phi}_0^{\mathsf{T}} + \frac{1}{m}\boldsymbol{\Phi}_{\ge 1}\boldsymbol{\Phi}_{\ge 1}^{\mathsf{T}} \tag{152}$$

$$= \mu_1^2 \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}} + \frac{1}{m}\boldsymbol{\Phi}_{\ge 1}\boldsymbol{P}^{\perp}\boldsymbol{\Phi}_{\ge 1}^{\mathsf{T}} + \boldsymbol{\Delta}\,. \tag{153}$$

Here $\boldsymbol{P} \in \mathbb{R}^{md \times md}$ is a block-diagonal projector, with $m$ blocks of dimension $d \times d$, with $\ell$th block given by

$$\boldsymbol{P}_\ell := \boldsymbol{w}_\ell \boldsymbol{w}_\ell^{\mathsf{T}},\ \boldsymbol{P}^{\perp} = \mathbf{I}_{md} - \boldsymbol{P} \text{ and } \boldsymbol{\Delta} := (\boldsymbol{\Phi}_0\boldsymbol{\Phi}_{\ge 1}^{\mathsf{T}} + \boldsymbol{\Phi}_{\ge 1}\boldsymbol{\Phi}_0^{\mathsf{T}} + \boldsymbol{\Phi}_{\ge 1}\boldsymbol{P}\boldsymbol{\Phi}_{\ge 1}^{\mathsf{T}})/m.$$

For the diagonal entries we have (assuming for simplicity $\boldsymbol{x}_i \sim \mathsf{N}(0, \mathbf{I}_d)$),

$$\begin{aligned}
\mathbb{E}\Big\{\frac{1}{m}\big(\boldsymbol{\Phi}_{\ge 1}\boldsymbol{P}^{\perp}\boldsymbol{\Phi}_{\ge 1}^{\mathsf{T}}\big)_{ii}\Big\} &= \mathbb{E}\big\{\langle \boldsymbol{x}_i, (\mathbf{I}_d - \boldsymbol{P}_\ell)\boldsymbol{x}_i\rangle(\sigma'(\langle \boldsymbol{w}_\ell, \boldsymbol{x}_i\rangle) - \mu_1)^2\big\} \\
&= \mathbb{E}\big\{\langle \boldsymbol{x}_i, (\mathbf{I}_d - \boldsymbol{P}_\ell)\boldsymbol{x}_i\rangle\big\}\,\mathbb{E}\big\{\sigma'(\langle \boldsymbol{w}_\ell, \boldsymbol{x}_i\rangle) - \mu_1)^2\big\} \\
&= (d-1)\mathbb{E}\{(\sigma'(G) - \mathbb{E}\sigma'(G))^2\} =: (d-1)v(\sigma),
\end{aligned}$$

where the second equality follows because $(\mathbf{I}_d - \boldsymbol{P}_\ell)\boldsymbol{x}_i$ and $\langle \boldsymbol{w}_\ell, \boldsymbol{x}_i\rangle$ are independent for $\boldsymbol{x}_i \sim \mathsf{N}(0, \mathbf{I}_d)$, and the last expectation is with respect to $G \sim \mathsf{N}(0,1)$. As proved in [MZ20] the matrix $\boldsymbol{\Phi}_{\ge 1}\boldsymbol{P}^{\perp}\boldsymbol{\Phi}_{\ge 1}^{\mathsf{T}}$ is well approximated by this diagonal expectation. Namely, under the model above, there exists a constant $C$ such that, with high probability:

$$\Big\|\frac{1}{md}\boldsymbol{\Phi}_{\ge 1}\boldsymbol{P}^{\perp}\boldsymbol{\Phi}_{\ge 1}^{\mathsf{T}} - v(\sigma)\,\mathbf{I}_n\Big\| \le \sqrt{\frac{n(\log d)^C}{md}}\,. \tag{154}$$

Equations (153) and (154) suggest that for $m = O(d)$, ridge regression in the NT model can be approximated by ridge regression in the raw covariates, as long as the regularization parameter is suitably modified. The next theorem confirms this intuition [MZ20]. We define ridge regression with respect to the raw covariates as per

$$\widehat{\boldsymbol{\beta}}(\gamma) := \mathrm{argmin}\,\Big\{\frac{1}{d}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \gamma\|\boldsymbol{\beta}\|_2^2\Big\}. \tag{155}$$

[10] To be precise, Lemma 5.3 assumes the covariate vectors $\boldsymbol{x}_i \sim \mathsf{N}(0, \mathbf{I}_d)$.

**Theorem 6.5.** *Assume $d^{1/C_0} \le m \le C_0 d$, $n \ge d/C_0$ and $md \gg n$. Then with high probability there exists an interpolator. Further assume $\boldsymbol{x}_i \sim \mathsf{N}(0, \mathbf{I}_d)$ and $f^*(\boldsymbol{x}) = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle$. Let*

$$L_{\rm lin}(\gamma) := \mathbb{E}\{(f^*(\boldsymbol{x}) - \langle \widehat{\boldsymbol{\beta}}(\gamma), \boldsymbol{x} \rangle)^2\}$$

*denote the risk of ridge regression with respect to the raw features.*

*Set $\lambda = \lambda_0(md/n)$ for some $\lambda_0 \ge 0$. Then there exists a constant $C > 0$ such that, with high probability,*

$$L_{\mathsf{NT}}(\lambda) = L_{\rm lin}(\gamma_{\rm eff}(\lambda_0, \sigma)) + O\left(\sqrt{\frac{n(\log d)^C}{md}}\right), \tag{156}$$

*where $\gamma_{\rm eff}(\lambda_0, \sigma) := (\lambda_0 + v(\sigma))/\mathbb{E}\{\sigma'(G)\}^2$.*

Notice that the shift in regularization parameter matches the heuristics given above (the scaling in $\lambda = \lambda_0(md/n)$ is introduced to match the typical scale of $\boldsymbol{\Phi}$).

# 7 Conclusions and future directions

Classical statistical learning theory establishes guarantees on the performance of a statistical estimator $\widehat{f}$, by bounding the generalization error $L(\widehat{f}) - \widehat{L}(\widehat{f})$. This is often thought of as a small quantity compared to the training error $L(\widehat{f}) - \widehat{L}(\widehat{f}) \ll \widehat{L}(\widehat{f})$. Regularization methods are designed precisely with the aim of keeping the generalization error $L(\widehat{f}) - \widehat{L}(\widehat{f})$ small.

The effort to understand deep learning has recently led to the discovery of a different learning scenario, in which the test error $L(\widehat{f})$ is optimal or nearly optimal, despite being much larger than the training error. Indeed in deep learning the training error often vanishes or is extremely small. The model is so rich that it overfits the data, that is, $\widehat{L}(\widehat{f}) \ll \inf_f L(f)$. When pushed, gradient-based training leads to interpolation or near-interpolation $\widehat{L}(\widehat{f}) \approx 0$ [ZBH+17]. We regard this as a particularly illuminating limit case.

This behavior is especially puzzling from a statistical point of view, that is, if we view data $(\boldsymbol{x}_i, y_i)$ as inherently noisy. In this case $y_i - f^*(\boldsymbol{x}_i)$ is of the order of the noise level and therefore, for a model that interpolates, $\widehat{f}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)$ is also large. Despite this, near-optimal test error means that $\widehat{f}(\boldsymbol{x}_{\rm test}) - f^*(\boldsymbol{x}_{\rm test})$ must be small at 'most' test points $\boldsymbol{x}_{\rm test} \sim \mathbb{P}$.

As pointed out in Section 2, interpolation poses less of a conceptual problem if data are noiseless. Indeed, unlike the noisy case, we can exhibit at least one interpolating solution that has vanishing test error, for any sample size: the true function $f^*$. Stronger results can also be established in the noiseless case: [Fel20] proved that interpolation is necessary to achieve optimal error rates when the data distribution is heavy-tailed in a suitable sense.

In this review we have focused on understanding when and why interpolation can be optimal or nearly optimal even with noisy data. Rigorous work has largely focused on models that are linear in a certain feature space, with the featurization map being independent of the data. Examples are RKHSs, the features produced by random network layers, or the neural tangent features defined by the Jacobian of the network at initialization. Mathematical work has established that interpolation can indeed be optimal and has described the underlying mechanism in a number of settings. While the scope of this analysis might appear to be limited (neural networks are notoriously nonlinear in their parameters), it is relevant to deep learning in two ways. First, in a direct way: as explained in Section 5, there are training regimes in which an overparametrized neural network is well approximated by a linear model that corresponds to the first-order Taylor expansion of the network around its initialization (the 'neural tangent' model). Second, in an indirect way: insights and hypotheses arising from the analysis of linear models can provide useful guidance for studying more complex settings.

Based on the work presented in this review, we can distill a few insights worthy of exploration in broader contexts.

**Simple-plus-spiky decomposition.** The function learnt in the overfitting (interpolating) regime takes the form

$$\widehat{f}(\boldsymbol{x}) = \widehat{f}_0(\boldsymbol{x}) + \Delta(\boldsymbol{x})\,. \tag{157}$$

Here $\widehat{f}_0$ is simple in a suitable sense (for instance, it is smooth) and hence is far from interpolating the data, while $\Delta$ is spiky: it has large complexity and allows interpolation of the data, but it is small, in the sense that it has negligible effect on the test error, *i.e.* $L(\widehat{f}_0 + \Delta) \approx L(\widehat{f}_0)$.

In the case of linear models, the decomposition (157) corresponds to a decomposition of $\widehat{f}$ into two orthogonal subspaces that do not depend on the data. Namely, $\widehat{f}_0$ is the projection of $\widehat{f}$ onto the top eigenvectors of the associated kernel and $\Delta$ is its orthogonal complement. In nonlinear models, the two components need not be orthogonal and the associated subspaces are likely to be data-dependent.

Understanding whether such a decomposition is possible, and what is its nature is a wide-open problem, which could be investigated both empirically and mathematically. A related question is whether the decomposition (157) is related to the widely observed 'compressibility' of neural network models. This is the observation that the test error of deep learning models does not change significantly if —after training— the model is simplified by a suitable compression operation [HMD15].

**Implicit regularization.** Not all interpolating models generalize equally well. This is easily seen in the case of linear models, where the set of interpolating models forms an affine space of dimension $p-n$ (where $p$ is the number of parameters). Among these, we can find models of arbitrarily large norm, that are arbitrarily far from the target regression function. Gradient-based training selects a specific model in this subspace, which is the closest in $\ell_2$ norm to the initialization.

The mechanism by which the training algorithm selects a specific empirical risk minimizer is understood in only a handful of cases: we refer to Section 3 for pointers to this literature. It would be important to understand how the model nonlinearity interacts with gradient flow dynamics. This in turn impacts the decomposition (157), namely which part of the function $\widehat{f}$ is to be considered 'simple' and which one is 'spiky'. Finally, the examples of kernel machines, random features and neural tangent models show that—in certain regimes—the simple component $\widehat{f}_0$ is also regularized in a non-trivial way, a phenomenon that we called self-induced regularization. Understanding these mechanisms in a more general setting is an outstanding challenge.

**Role of dimension.** As pointed out in Section 4, interpolation is sub-optimal in a fixed dimension in the presence of noise, for certain kernel methods [RZ19]. The underlying mechanism is as described above: for an interpolating model, $\widehat{f}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)$ is of the order of the noise level. If $\widehat{f}$ and $f^*$ are sufficiently regular (for instance, uniformly continuous, both in $\boldsymbol{x}$ and in $n$) $\widehat{f}(\boldsymbol{x}_{\text{test}}) - f^*(\boldsymbol{x}_{\text{test}})$ is expected to be of the same order when $\boldsymbol{x}_{\text{test}}$ is close to the training set. This happens with constant probability in fixed dimension. However, this probability decays rapidly with the dimension.

Typical data in deep learning applications are high-dimensional (images, text, and so on). On the other hand, it is reasonable to believe that deep learning methods are not affected by the ambient dimension (the number of pixels in an image), but rather by an effective or intrinsic dimension. This is the case for random feature models [GMMM20b]. This raises the question of how deep learning methods escape the intrinsic limitations of interpolators in low dimension. Is it because they construct a (near) interpolant $\widehat{f}$ that is highly irregular (not uniformly continuous)? Or perhaps because the effective dimension is at least moderately large? (After all the lower bounds mentioned above decrease rapidly with dimension.) What is the proper mathematical definition of effective dimension?

**Adaptive model complexity.** As mentioned above, in the case of linear models, the terms $\widehat{f}_0$ and $\Delta$ in the decomposition (157) correspond to the projections of $\widehat{f}$ onto $\mathcal{V}_k$ and $\mathcal{V}_k^{\perp}$. Here $\mathcal{V}_k$ is the space spanned by the top $k$ eigenfunctions of the kernel associated with the linear regression problem. Note that this is the case also for the random features and neural tangent models of Section 6. In this case the relevant kernel is the expectation of the finite-network kernel $\boldsymbol{D}f(\boldsymbol{\theta}_0)^{\mathsf{T}}\boldsymbol{D}f(\boldsymbol{\theta}_0)$ with respect to the choice of random weights at initialization.

A crucial element of this behavior is the dependence of $k$ (the dimension of the eigenspace $\mathcal{V}_k$) on various features of the problem at hand: indeed $k$ governs the complexity of the 'simple' part of the model $\widehat{f}_0$, which is the one actually relevant for prediction. As discussed in Section 4, in kernel methods $k$ increases with the sample size $n$: as more data are used, the model $\widehat{f}_0$ becomes more complex. In random features and neural tangent models (see Section 6), $k$ depends on the minimum of $n$ and the number of network parameters (which is proportional to the width for two-layer networks). The model complexity increases with sample size, but saturates when it reaches the number of network parameters.

This suggests a general hypothesis that would be interesting to investigate beyond linear models. Namely, if a decomposition of the type (157) is possible, then the complexity of the simple part $\widehat{f}_0$ increases with the sample size and the network size.

**Computational role of overparametrization.** We largely focused on the surprising discovery that overparametrization and interpolation do not necessarily hurt generalization, even in the presence of noise. However, we should emphasize once more that the real motivation for working with overparametrized models is not statistical but computational. The empirical risk minimization problem for neural networks is computationally hard, and in general we cannot hope to be able to find a global minimizer using gradient-based algorithms. However, empirical evidence indicates that global optimization becomes tractable when the model is sufficiently overparametrized.

The linearized and mean field theories of Section 5 provide general arguments to confirm this empirical finding. However, we are far from understanding precisely what amount of overparametrization is necessary, even in simple neural network models.

## Acknowledgements

PB, AM and AR acknowledge support from the NSF through award DMS-2031883 and from the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning. For insightful discussions on these topics, the authors also thank the other members of that Collaboration and many other collaborators and colleagues, including Emmanuel Abbe, Misha Belkin, Niladri Chatterji, Amit Daniely, Tengyuan Liang, Philip Long, Gábor Lugosi, Song Mei, Theodor Misiakiewicz, Hossein Mobahi, Elchanan Mossel, Phan-Minh Nguyen, Nati Srebro, Nike Sun, Alexander Tsigler, Roman Vershynin, and Bin Yu. We thank Tengyuan Liang and Song Mei for insightful comments on the draft. PB acknowledges support from the NSF through grant DMS-2023505. AM acknowledges support from the ONR through grant N00014-18-1-2729. AR acknowledges support from the NSF through grant DMS-1953181, and support from the MIT-IBM Watson AI Lab and the NSF AI Institute for Artificial Intelligence and Fundamental Interactions.

## References

[AB99] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[ABR64] MA Aizerman, E M Braverman, and LI Rozonoer. Theoretical foundations of the potential function method in pattern recognition. *Avtomat. i Telemeh*, 25(6):917–936, 1964.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

[AKT19] Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1370–1378. PMLR, 2019.

[AM97] Dimitris Achlioptas and Michael Molloy. The analysis of a list-coloring algorithm on a random graph. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 204–212. IEEE, 1997.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 2019.

[Bac13] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 185–209. PMLR, 2013.

[Bac17] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[Bar98] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[Bar08] Peter L. Bartlett. Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory*, 24(2):545–552, April 2008.

[BBD02] P. L. Bartlett and S. Ben-David. Hardness results for neural network approximation problems. *Theoretical Computer Science*, 284(1):53–66, 2002.

[BBL02] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[BBM05] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[BBV06] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65(1):79–94, 2006.

[BD07] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Pearson Prentice Hall, 2007.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[BEHW89] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[BFT17] Peter L. Bartlett, Dylan Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc., 2017.

[BH89] Eric B. Baum and David Haussler. What size net gives valid generalization? *Neural Computation*, 1(1):151–160, 1989.

[BHLM19] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[BHM18] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 2300–2311. Curran Associates, Inc., 2018.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BJM06] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BL99] P. L. Bartlett and G. Lugosi. An inequality for uniform deviations of sample averages from their means. *Statistics and Probability Letters*, 44(1):55–62, 1999.

[BL20a] Yu Bai and Jason D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*, 2020. arXiv:1910.01619.

[BL20b] Peter L. Bartlett and Philip M. Long. Failures of model-dependent generalization bounds for least-norm interpolation. *arXiv preprint arXiv:2010.08479*, 2020.

[BLLT20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BM02] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[BMM98] P. L. Bartlett, V. Maiorov, and R. Meir. Almost linear VC dimension bounds for piecewise polynomial networks. *Neural Computation*, 10(8):2159–2173, 1998.

[BMM18] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549, 2018.

[BR92] Avrim Blum and Ronald L. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5(1):117–127, 1992.

[Bre98] Leo Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–849, 1998.

[BRT19] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

[CB18] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 3036–3046. Curran Associates, Inc., 2018.

[CB20] Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1305–1338. PMLR, 2020. arXiv:2002.04486.

[CDV07] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[CH67] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[CLG01] Rich Caruana, Steve Lawrence, and C. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.

[CO10] Amin Coja-Oghlan. A better algorithm for random k-SAT. *SIAM Journal on Computing*, 39(7):2823–2864, 2010.

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

[CS13] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[CX21] Lin Chen and Sheng Xu. Deep neural tangent kernel and Laplace kernel have the same RKHS. In *International Conference on Learning Representations*, 2021. arXiv:2009.10683.

[DC95] Harris Drucker and Corinna Cortes. Boosting decision trees. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, page 479–485, Cambridge, MA, USA, 1995. MIT Press.

[DFKU13] Paramveer S. Dhillon, Dean P. Foster, Sham M. Kakade, and Lyle H. Ungar. A risk comparison of ordinary least squares vs ridge regression. *Journal of Machine Learning Research*, 14(10):1505–1511, 2013.

[DGA20] Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from Feynman diagrams. In *International Conference on Learning Representations*, 2020. arXiv:1909.11304.

[DGK98] Luc Devroye, Laszlo Györfi, and Adam Krzyżak. The Hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998.

[DLL+19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019.

[DSS95] Bhaskar DasGupta, Hava T. Siegelmann, and Eduardo D. Sontag. On the complexity of training neural networks with continuous activation functions. *IEEE Transactions on Neural Networks*, 6(6):1490–1504, 1995.

[DV13] Yen Do and Van Vu. The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 2(03):1350005, 2013.

[DW79] Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

[DZPS19] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. arXiv:1810.02054.

[EAM15] Ahmed El Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

[EHKV89] A. Ehrenfeucht, David Haussler, Michael J. Kearns, and Leslie G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989.

[EK10] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

[Fel20] Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer, 2001.

[FM19] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1-2):27–85, 2019.

[Fri01] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 10 2001.

[FS96] Alan Frieze and Stephen Suen. Analysis of two simple heuristics on a random instance of $k$-SAT. *Journal of Algorithms*, 20(2):312–355, 1996.

[FS97] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[GJ79] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[GLSS18a] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

[GLSS18b] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

[GMKZ19] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modelling the influence of data structure on learning in neural networks. *arXiv:1909.11500*, 2019.

[GMMM20a] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv:1904.12191. Annals of Statistics (To appear).*, 2020.

[GMMM20b] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14820–14830. Curran Associates, Inc., 2020.

[GRM+20] Sebastian Goldt, Galen Reeves, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with two-layer neural networks. *arXiv preprint arXiv:2006.14709*, 2020.

[GRS18] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 297–299. PMLR, 06–09 Jul 2018.

[GSd+19] Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

[GWB+17] Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6152–6160, 2017.

[GYK+20] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the Laplace and neural tangent kernels. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1451–1461. Curran Associates, Inc., 2020. arXiv:2007.01580.

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[HL20] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv:2009.07669*, 2020.

[HMD15] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv:1510.00149*, 2015.

[HMRT20] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560v5*, 2020.

[HN20] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020. arXiv:1909.05989.

[HY20] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4542–4551. PMLR, 13–18 Jul 2020.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580. Curran Associates, Inc., 2018.

[Joh19] Iain M. Johnstone. *Gaussian Estimation: Sequence and Wavelet Models*. 2019. Manuscript, available at `http://statweb.stanford.edu/~imj/`.

[JP78] David S. Johnson and F. P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6:93–107, 1978.

[JT18] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[JT19] Ziwei Ji and Matus Telgarsky. A refined primal-dual analysis of the implicit bias. *arXiv preprint arXiv:1906.04540*, 2019.

[Jud90] J. S. Judd. *Neural Network Design and the Complexity of Learning*. MIT Press, 1990.

[KL17] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

[KM97] Marek Karpinski and Angus J. Macintyre. Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *Journal of Computer and System Sciences*, 54:169–176, 1997.

[KM15] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[Kol01] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, July 2001.

[Kol06] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34:2593–2656, 2006.

[KP00] V. I. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In Evarist Giné, David M. Mason, and Jon A. Wellner, editors, *High Dimensional Probability II*, volume 47, pages 443–459. Birkhäuser, 2000.

[KS01] Vera Kurková and Marcello Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6):2659–2665, 2001.

[KS02] Vera Kurková and Marcello Sanguineti. Comparison of worst case errors in linear and neural network approximation. *IEEE Transactions on Information Theory*, 48(1):264–275, 2002.

[KT20] Ganesh Ramachandra Kini and Christos Thrampoulidis. Analytic study of double descent in binary classification: The impact of loss. In *IEEE International Symposium on Information Theory, ISIT 2020, Los Angeles, CA, USA, June 21-26, 2020*, pages 2527–2532. IEEE, 2020. arXiv:2001.11572.

[Kur97] Věra Kurková. Dimension-independent rates of approximation by neural networks. In *Computer Intensive Methods in Control and Signal Processing*, pages 261–270. Springer, 1997.

[KY17] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1-2):257–352, 2017.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[LBW96] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Society, 2001.

[LGT97] Steve Lawrence, C. Lee Giles, and Ah Chung Tsoi. Lessons in neural network training: Overfitting may be harder than expected. In *In Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI-97*, pages 540–545. AAAI Press, 1997.

[Lia20] Tengyuan Liang, 2020. Personal communication.

[Lin04] Y. Lin. A note on margin-based loss functions in classification. *Statistics and Probability Letters*, 68:73–82, 2004.

[LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[LMZ18] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[LR20] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

[LRS15] Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset Rademacher complexity. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1260–1285, Paris, France, 03–06 Jul 2015. PMLR.

[LRZ20] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2683–2711. PMLR, 2020. arXiv:1908.10292.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

[LV04] G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 32:30–55, 2004.

[LZB20] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15954–15964. Curran Associates, Inc., 2020.

[Men02] Shahar Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48:1977–1991, 2002.

[Men20] Shahar Mendelson. Extending the scope of the small-ball method. *Studia Mathematica*, pages 1–21, 2020.

[MM19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *Communications in Pure and Applied Mathematics (To appear)*, 2019. arXiv:1908.05355.

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464, 2019.

[MMM21] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. *arXiv preprint arXiv:2101.10588*, 2021.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[MP90] Gale Martin and James Pittman. Recognizing hand-printed letters and digits. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1990.

[MRSY19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[MZ20] Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *arXiv:2007.12826*, 2020.

[Nad64] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[NK19] Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *NeurIPS*, pages 11611–11622, 2019.

[NLG+19] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[NP20] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv:2001.11443*, 2020.

[NS17] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv:1712.05438*, 2017.

[NTS15] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1376–1401. PMLR, 2015.

[NTSS17] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

[OS19] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

[OS20] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

[Pol90] David Pollard. *Empirical Processes: Theory and Applications*, volume 2. Institute of Mathematical Statistics, 1990.

[Pol95] David Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22:271–278, 1995.

[PW17] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2017.

[PW18] Jeffrey Pennington and Pratik Worah. The spectrum of the Fisher information matrix of a single-hidden-layer neural network. *Advances in Neural Information Processing Systems*, 31:5410–5419, 2018.

[Qui96] J. R. Quinlan. Bagging, boosting, and C4.5. In *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 725–730, 1996.

[RCR15] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, volume 28, pages 1657–1665, 2015.

[RR07] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184, 2007.

[RR09] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 1313–1320, 2009.

[RR17] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3215–3225, 2017.

[RST17] Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

[RV06] M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, 164(2):603–648, 2006.

[RVE18] Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. arXiv:1805.00915.

[RZ19] Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pages 2595–2623, 2019.

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87. Birkhäuser, 2015.

[SFBL98] Robert E Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[SGd+19] Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, 2019.

[SHN+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[SS20] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[SST10] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv:1009.3896*, 2010.

[Tal94] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.

[TB20] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

[Tel13] Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315, 2013.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[TPT20] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. *arXiv:2006.08917*, 2020.

[Tsy08] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

[VC71] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[VC74] V. N. Vapnik and A. Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, 1974.

[vdG90] Sara van de Geer. Estimating a regression function. *Annals of Statistics*, 18:907–924, 1990.

[Ver18] Roman Vershynin. *High-Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[Vu98] Van H. Vu. On the infeasibility of training neural networks with small squared errors. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 371–377. MIT Press, 1998.

[Was13] Larry Wasserman. *All of Statistics: a Concise Course in Statistical Inference*. Springer Science & Business Media, 2013.

[Wat64] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[WOBM17] Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of AdaBoost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017.

[WS01] Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2001.

[ZBH+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. arXiv:1611.03530.

[ZCZG20] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109(3):467–492, 2020.

[Zha04] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–85, 2004.

[ZY05] Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

# A Kernels on $\mathbb{R}^d$ with $d \asymp n$

## A.1 Bound on the variance of the minimum-norm interpolant

**Lemma A.1.** *For any $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and any positive semidefinite $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, for $n \lesssim d$ and any $k < d$,*

$$\mathsf{tr}\left((\boldsymbol{X}\boldsymbol{X}^\top + d\gamma \mathbf{I}_n)^{-2}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\right) \lesssim \frac{1}{\gamma}\left(\frac{\lambda_1 k}{n} + \lambda_{k+1}\right), \tag{158}$$

*where $\lambda_1 \geq \ldots \geq \lambda_d$ are the eigenvalues of $\boldsymbol{\Sigma}$.*

*Proof.* This deterministic argument is due to T. Liang [Lia20]. We write $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\leq k} + \boldsymbol{\Sigma}_{>k}$, with $\boldsymbol{\Sigma}_{\leq k} = \sum_{i\leq k}\lambda_i \boldsymbol{u}_i\boldsymbol{u}_i^\top$. Then by the argument in [LR20, Remark 5.1],

$$\mathsf{tr}\left((\boldsymbol{X}\boldsymbol{X}^\top + d\gamma \mathbf{I}_n)^{-2}\boldsymbol{X}\boldsymbol{\Sigma}_{>k}\boldsymbol{X}^\top\right) \leq \lambda_{k+1}\sum_{i=1}^{n}\frac{\widehat{\lambda}_i}{(d\gamma+\widehat{\lambda}_i)^2} \leq \lambda_{k+1}\frac{n}{4d\gamma} \lesssim \frac{\lambda_{k+1}}{\gamma} \tag{159}$$

where $\widehat{\lambda}_i$ are the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$. Here we use the fact that $\frac{t}{(r+t)^2} \leq \frac{1}{4r}$ for all $t, r > 0$. On the other hand,

$$\mathsf{tr}\left((\boldsymbol{X}\boldsymbol{X}^\top + d\gamma \mathbf{I}_n)^{-2}\boldsymbol{X}\boldsymbol{\Sigma}_{\leq k}\boldsymbol{X}^\top\right) \leq \sum_{i\leq k}\lambda_i \left\|(d\gamma \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{u}_i\right\|^2. \tag{160}$$

Now, using the argument similar to that in [BLLT20], we define $A_{-i} = d\gamma \boldsymbol{I}_n + \boldsymbol{X}(\boldsymbol{I}_n - \boldsymbol{u}_i\boldsymbol{u}_i^\top)\boldsymbol{X}^\top$, $\boldsymbol{v} = \boldsymbol{X}\boldsymbol{u}_i$ and write

$$\left\|(d\gamma \boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{u}_i\right\|^2 = \left\|(A_{-i} + \boldsymbol{v}\boldsymbol{v}^\top)^{-1}\boldsymbol{v}\right\|^2 = \frac{\boldsymbol{v}^\top A_{-i}^{-2}\boldsymbol{v}}{(1+\boldsymbol{v}^\top A_{-i}^{-1}\boldsymbol{v})^2} \tag{161}$$

by the Sherman-Morrison formula. The last quantity is upper bounded by

$$\frac{1}{d\gamma}\frac{\boldsymbol{v}^\top A_{-i}^{-1}\boldsymbol{v}}{(1+\boldsymbol{v}^\top A_{-i}^{-1}\boldsymbol{v})^2} \leq \frac{1}{4\gamma d}. \tag{162}$$

Substituting in (160), we obtain an upper bound of

$$\frac{1}{4\gamma d}\sum_{i\leq k}\lambda_i \lesssim \frac{\lambda_1 k}{\gamma n},$$

assuming $n \lesssim d$. □

## A.2 Exact characterization in the proportional asymptotics

We will denote by $\boldsymbol{K} = (h(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle/d))_{i,j\leq n}$ the kernel matrix. We will also denote by $\boldsymbol{K}_1$ the linearized kernel

$$\boldsymbol{K}_1 = \beta\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \beta\gamma\mathbf{I}_n + \alpha\mathbf{1}\mathbf{1}^\top, \tag{163}$$

$$\alpha := h(0) + h''(0)\frac{\mathsf{tr}(\boldsymbol{\Sigma}^2)}{2d^2},\ \beta := h'(0), \tag{164}$$

$$\gamma := \frac{1}{h'(0)}\big[h(\mathsf{tr}(\boldsymbol{\Sigma})/d) - h(0) - h'(0)\mathsf{tr}(\boldsymbol{\Sigma}/d)\big]. \tag{165}$$

**Assumption 4.12.** *We assume that the coordinates of $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ are independent, with zero mean and unit variance, so that $\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}$. Further assume there are constants $0 < \eta, M < \infty$, such that the following hold.*

(a) *For all $i \le d$, $\mathbb{E}[|\boldsymbol{z}_i|^{8+\eta}] \le M$.*

(b) *$\|\boldsymbol{\Sigma}\| \le M$, $d^{-1}\sum_{i=1}^{d}\lambda_i^{-1} \le M$, where $\lambda_1, \dots, \lambda_d$ are the eigenvalues of $\boldsymbol{\Sigma}$.*

**Theorem 4.13.** *Let $0 < M, \eta < \infty$ be fixed constants and suppose that Assumption 4.12 holds with $M^{-1} \le d/n \le M$. Further assume that $h$ is continuous on $\mathbb{R}$ and smooth in a neighborhood of $0$ with $h(0), h'(0) > 0$, that $\|f^*\|_{L^{4+\eta}(\mathbb{P})} \le M$ and that the $z_i$'s are $M$-sub-Gaussian. Let $y_i = f^*(\boldsymbol{x}_i) + \xi_i$, $\mathbb{E}(\xi_i^2) = \sigma_\xi^2$, and $\boldsymbol{\beta}_0 := \boldsymbol{\Sigma}^{-1}\mathbb{E}[\boldsymbol{x}f^*(\boldsymbol{x})]$. Let $\lambda_* > 0$ be the unique positive solution of*

$$n\Big(1 - \frac{\gamma}{\lambda_*}\Big) = \mathsf{tr}\Big(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_*\mathbf{I})^{-1}\Big)\,. \tag{166}$$

*Define $\mathscr{B}(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0)$ and $\mathscr{V}(\boldsymbol{\Sigma})$ by*

$$\mathscr{V}(\boldsymbol{\Sigma}) := \frac{\mathsf{tr}\big(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_*\mathbf{I})^{-2}\big)}{n - \mathsf{tr}\big(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_*\mathbf{I})^{-2}\big)}\,, \tag{167}$$

$$\mathscr{B}(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0) := \frac{\lambda_*^2\langle\boldsymbol{\beta}_0, (\boldsymbol{\Sigma} + \lambda_*\mathbf{I})^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}_0\rangle}{1 - n^{-1}\mathsf{tr}\big(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_*\mathbf{I})^{-2}\big)}\,. \tag{168}$$

*Finally, let $\widehat{\mathrm{BIAS}}^2$ and $\widehat{\mathrm{VAR}}$ denote the squared bias and variance for the minimum-norm interpolant. Then there exist $C, c_0 > 0$ (depending also on the constants in Assumption 4.12) such that the following holds with probability at least $1 - Cn^{-1/4}$ (here $\mathsf{P}_{>1}$ denotes the projector orthogonal to affine functions in $L^2(\mathbb{P})$):*

$$\Big|\widehat{\mathrm{BIAS}}^2 - \mathscr{B}(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0) - \|\mathsf{P}_{>1}f^*\|_{L^2}^2(1 + \mathscr{V}(\Sigma))\Big| \le Cn^{-c_0}\,, \tag{169}$$

$$\Big|\widehat{\mathrm{VAR}} - \sigma_\xi^2\mathscr{V}(\boldsymbol{\Sigma})\Big| \le Cn^{-c_0}\,. \tag{170}$$

**Remark A.1.** The result for the variance will be proved under weaker assumptions and in a stronger form than stated. In particular, it does not require any assumption on the target function $f_*$, and it holds with smaller error terms than stated.

**Remark A.2.** Notice that by positive definiteness of the kernel, we have $h'(0), h''(0) \ge 0$. Hence the conditions that these are strictly positive is essentially a non-degeneracy requirement.

We note for future reference that the target function $f^*$ is decomposed as

$$f^*(\boldsymbol{x}) = b_0 + \langle\boldsymbol{\beta}_0, \boldsymbol{x}\rangle + \mathsf{P}_{>1}f^*(\boldsymbol{x})\,, \tag{171}$$

where $b_0 := \mathbb{E}\{f^*(\boldsymbol{x})\}$, $\boldsymbol{\beta}_0 := \boldsymbol{\Sigma}^{-1}\mathbb{E}[\boldsymbol{x}f^*(\boldsymbol{x})]$ as defined above and $\mathbb{E}\{\mathsf{P}_{>1}f^*(\boldsymbol{x})\}$, $\mathbb{E}\{\boldsymbol{x}\mathsf{P}_{>1}f^*(\boldsymbol{x})\} = \mathbf{0}$.

### A.2.1 Preliminaries

Throughout the proof, we will use $C$ for constants that depend uniquely on the constants in Assumption 4.12 and Theorem 4.13. We also write that an inequality holds *with very high probability* if, for any $A > 0$, we can choose the constants $C$ in the inequality such that this holds with probability at least $1 - n^{-A}$ for all $A$ large enough.

We will repeatedly use the following bound, see e.g. [EK10].

**Lemma A.2.** *Under the assumptions of Theorem 4.13, we have, with very high probability*

$$\boldsymbol{K} = \boldsymbol{K}_1 + \boldsymbol{\Delta}\,, \qquad \|\boldsymbol{\Delta}\| \le n^{-c_0}\,. \tag{172}$$

*In particular, as long as $h$ is non-linear, we have $\boldsymbol{K} \succeq c_*\mathbf{I}_n$, $c_* = \beta\gamma > 0$ with probability at least $1 - Cn^{-D}$.*

Define the matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, and the vector $\boldsymbol{v} \in \mathbb{R}^n$ by

$$M_{ij} := \mathbb{E}_{\boldsymbol{x}}\left\{h\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big)h\Big(\frac{1}{d}\langle \boldsymbol{x}_j, \boldsymbol{x}\rangle\Big)\right\}, \tag{173}$$

$$v_i := \mathbb{E}_{\boldsymbol{x}}\left\{h\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big)f^*(\boldsymbol{x})\right\}. \tag{174}$$

Our first lemma provides useful approximations of these quantities.

**Lemma A.3.** *Define (here expectations are over $G \sim \mathsf{N}(0,1)$):*

$$\boldsymbol{v}_0 := \boldsymbol{a}_0 b_0 + \frac{1}{d}h'(0)\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{\beta}_0\,, \tag{175}$$

$$a_{i,0} := \mathbb{E}\Big\{h\Big(\sqrt{\frac{Q_{ii}}{d}}\,G\Big)\Big\}, \quad Q_{ij} := \frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{\Sigma}\boldsymbol{x}_j\rangle\,. \tag{176}$$

*and*

$$\boldsymbol{M}_0 := \boldsymbol{a}\boldsymbol{a}^{\mathsf{T}} + \boldsymbol{B}\,, \qquad \boldsymbol{B} := \frac{1}{d}\,\boldsymbol{D}\boldsymbol{Q}\boldsymbol{D}\,, \tag{177}$$

$$a_i := a_{i,0} + a_{i,1}\,, \quad a_{i,1} = \frac{1}{6}\Big(\frac{Q_{ii}}{d}\Big)^{3/2}h^{(3)}(0)\sum_{j=1}^{d}\frac{(\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_i)_j^3}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_i\|_2^3}\mathbb{E}(z_j^3)\,, \tag{178}$$

$$\boldsymbol{D} := \mathrm{diag}(D_1, \dots, D_n)\,, \qquad D_i := \mathbb{E}\Big\{h'\Big(\sqrt{\frac{Q_{ii}}{d}}\,G\Big)\Big\}. \tag{179}$$

*Then the following hold with very high probability (in other words, for any $A > 0$ there exists $C$ such that the following hold with probability at least $1 - n^{-A}$ for all $n$ large enough)*

$$\max_{i\le n}\big|v_i - v_{0,i}\big| \le C\frac{\sqrt{\log d}}{d^{3/2}}\,, \tag{180}$$

$$\max_{i\ne j\le n}\big|M_{ij} - M_{0,ij}\big| \le C\frac{\log d}{d^{5/2}}\,, \tag{181}$$

$$\max_{i\le n}\big|M_{ii} - M_{0,ii}\big| \le C\frac{\log d}{d^{2}}\,. \tag{182}$$

*In particular, this implies $\|\boldsymbol{v} - \boldsymbol{v}_0\|_2 \le Cd^{-1}\sqrt{\log d}$, $\|\boldsymbol{M} - \boldsymbol{M}_0\|_F \le Cd^{-3/2}\log d$.*

*Proof.* Throughout the proof we will work on the intersection $\mathcal{E}_1 \cap \mathcal{E}_2$ of following events, which hold with very high probability by standard concentration arguments. These events are defined by

$$\mathcal{E}_1 := \Big\{C^{-1} \le \frac{1}{\sqrt{d}}\|\boldsymbol{\Sigma}\boldsymbol{z}_i\|_2 \le C;\ \ \frac{1}{\sqrt{d}}\|\boldsymbol{\Sigma}\boldsymbol{z}_i\|_\infty \le C\sqrt{\frac{\log d}{d}}\ \ \forall i \le n\Big\} \tag{183}$$

$$= \Big\{C^{-2} \le \frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{\Sigma}\boldsymbol{x}_i\rangle \le C^2;\ \ \frac{1}{\sqrt{d}}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_i\|_\infty \le C\sqrt{\frac{\log d}{d}}\ \ \forall i \le n\Big\}, \tag{184}$$

and

$$\mathcal{E}_2 := \Big\{\frac{1}{d}\sum_{\ell=1}^{d}(\boldsymbol{\Sigma}\boldsymbol{z}_i)_\ell(\boldsymbol{\Sigma}\boldsymbol{z}_j)_\ell^2 \le \frac{\log d}{d^{1/2}}\,;\ \ \frac{1}{d}|\langle \boldsymbol{z}_i, \boldsymbol{\Sigma}\boldsymbol{z}_j\rangle| \le C\sqrt{\frac{\log d}{d}}\,;\ \ \frac{1}{d}|\langle \boldsymbol{z}_i, \boldsymbol{\Sigma}^2\boldsymbol{z}_j\rangle| \le C\sqrt{\frac{\log d}{d}}\ \ \forall i \ne j \le n\Big\} \tag{185}$$

$$= \Big\{\frac{1}{d}\sum_{\ell=1}^{d}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_i)_\ell(\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_j)_\ell^2 \le \frac{\log d}{d^{1/2}}\,;\ \ \frac{1}{d}|\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle| \le C\sqrt{\frac{\log d}{d}}\ \ \frac{1}{d}|\langle \boldsymbol{x}_i, \boldsymbol{\Sigma}\boldsymbol{x}_j\rangle| \le C\sqrt{\frac{\log d}{d}}\ \ \forall i \ne j \le n\Big\}.$$

Recall that, by assumption, $h$ is smooth on an interval $[-t_0, t_0]$, $t_0 > 0$. On the event $\mathcal{E}_2$, we have $\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle/d \in [-t_0, t_0]$ for all $i \neq j$. If $h$ is not smooth everywhere, we can always modify it outside $[-t_0/2, t_0/2]$ to obtain a kernel $\tilde{h}$ that is smooth everywhere. Since $\boldsymbol{x}$ is sub-Gaussian, as long as $\|\boldsymbol{x}_i\|/\sqrt{d} \le C$ for all $i \le n$ (this happens on $\mathcal{E}_1$) we have (for $\boldsymbol{x} \sim \mathbb{P}$), $\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle/d \in [-t_0/2, t_0/2]$ with probability at least $1 - e^{-d/C}$. Further using the fact that $f$ is bounded in Eqs. (173), (174), we get,

$$M_{ij} := \mathbb{E}_{\boldsymbol{x}}\left\{\tilde{h}\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big)\tilde{h}\Big(\frac{1}{d}\langle \boldsymbol{x}_j, \boldsymbol{x}\rangle\Big)\right\} + O(e^{-d/C})\,, \tag{186}$$

$$v_i := \mathbb{E}_{\boldsymbol{x}}\left\{\tilde{h}\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big) f^*(\boldsymbol{x})\right\} + O(e^{-d/C})\,, \tag{187}$$

where the term $O(e^{-d/C})$ is uniform over $i, j \le n$. Analogously, in the definition of $\boldsymbol{v}_0$, $\boldsymbol{M}_0$ (more precisely, in defining $\boldsymbol{a}_0$, $\boldsymbol{D}$), we can replace $h$ by $\tilde{h}$ at the price of an $O(e^{-d/C})$ error. Since these error terms are negligible as compared to the ones in the statement, we shall hereafter neglect them and set $\tilde{h} = h$ (which corresponds to defining arbitrarily the derivatives of $h$ outside a neighborhood of 0).

We denote by $h_{i,k}$ the $k$-th coefficient of $h((Q_{ii}/d)^{1/2}x)$ in the basis of Hermite polynomials. Namely:

$$h_{i,k} = \mathbb{E}\Big\{h\Big(\sqrt{\frac{Q_{ii}}{d}}\,G\Big)\,\mathrm{He}_k(G)\Big\} = \Big(\frac{Q_{ii}}{d}\Big)^{k/2}\mathbb{E}\Big\{h^{(k)}\Big(\sqrt{\frac{Q_{ii}}{d}}\,G\Big)\Big\}\,. \tag{188}$$

Here $h^{(k)}$ denotes the $k$-th derivative of $h$ (recall that by the argument above we can assume, without loss of generality, that $h$ is $k$-times differentiable).

We write $h_{i,>k}$ for the remainder after the first $k$ terms of the Hermite expansion have been removed:

$$\begin{aligned} h_{i,>k}\Big(\sqrt{\frac{Q_{ii}}{d}}\,x\Big) &:= h\Big(\sqrt{\frac{Q_{ii}}{d}}\,x\Big) - \sum_{\ell=0}^{k}\frac{1}{\ell!}h_{i,\ell}\,\mathrm{He}_\ell(x) \\ &= h\Big(\sqrt{\frac{Q_{ii}}{d}}\,x\Big) - \sum_{\ell=0}^{k}\frac{1}{\ell!}\Big(\frac{Q_{ii}}{d}\Big)^{\ell/2}\mathbb{E}\Big\{h^{(k)}\Big(\sqrt{\frac{Q_{ii}}{d}}\,G\Big)\Big\}\,\mathrm{He}_\ell(x)\,. \end{aligned} \tag{189}$$

Finally, we denote by $\hat{h}_{i,>k}(x)$ the remainder after the first $k$ terms in the Taylor expansion have been subtracted:

$$\hat{h}_{>k}(x) := h(x) - \sum_{\ell=0}^{k}\frac{1}{\ell!}h^{(\ell)}(0)\,x^\ell\,. \tag{190}$$

Of course $h - \hat{h}_{>k}$ is a polynomial of degree $k$, and therefore its projection orthogonal to the first $k$ Hermite polynomials vanishes, whence

$$h_{i,>k}\Big(\sqrt{\frac{Q_{ii}}{d}}\,x\Big) = \hat{h}_{>k}\Big(\sqrt{\frac{Q_{ii}}{d}}\,x\Big) - \sum_{\ell=0}^{k}\frac{1}{\ell!}\Big(\frac{Q_{ii}}{d}\Big)^{\ell/2}\mathbb{E}\Big\{\hat{h}^{(\ell)}_{>k}\Big(\sqrt{\frac{Q_{ii}}{d}}\,G\Big)\Big\}\,\mathrm{He}_\ell(x)\,. \tag{191}$$

Note that, by smoothness of $h$, we have $|\hat{h}^{(\ell)}_{>k}(t)| \le C\min(|t|^{k+1-\ell}, 1)$, and therefore

$$\left|\frac{1}{\ell!}\Big(\frac{Q_{ii}}{d}\Big)^{\ell/2}\mathbb{E}\Big\{\hat{h}^{(\ell)}_{>k}\Big(\sqrt{\frac{Q_{ii}}{d}}\,G\Big)\Big\}\right| \le Cd^{-(k+1)/2}\,. \tag{192}$$

We also have that $|\hat{h}_{>k}(t)| \le C\min(1, |t|^{k+1})$. Define $\boldsymbol{v}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_i/\sqrt{d}$, $\|\boldsymbol{v}_i\|_2^2 = Q_{ii}$. For any fixed $m \ge 2$, by Eq. (191) and the triangle inequality,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{z}}\Big\{\Big|h_{i,>k}\Big(\frac{1}{\sqrt{d}}\langle \boldsymbol{v}_i, \boldsymbol{z}\rangle\Big)\Big|^m\Big\}^{1/m} &\overset{(a)}{\le} \mathbb{E}\Big\{\Big|\hat{h}_{>k}\Big(\frac{1}{\sqrt{d}}\langle \boldsymbol{v}_i, \boldsymbol{z}\rangle\Big)\Big|^m\Big\}^{1/m} + Cd^{-(k+1)/2}\sum_{\ell=0}^{k}\mathbb{E}\Big\{\Big|\mathrm{He}_\ell\Big(\frac{\langle \boldsymbol{v}_i, \boldsymbol{z}\rangle}{\|\boldsymbol{v}_i\|_2}\Big)\Big|^m\Big\}^{1/m} \\ &\le C\Big(\frac{Q_{ii}}{d}\Big)^{(k+1)/2} + C\,d^{-(k+1)/2} \le C\,d^{-(k+1)/2}\,, \end{aligned} \tag{193}$$

where the inequality $(a)$ follows since $\langle \boldsymbol{v}_i, \boldsymbol{z}\rangle$ is $C$-sub-Gaussian. Note that Eqs. (189), (193) can also be rewritten as

$$h\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big) = \sum_{\ell=0}^{k} \frac{1}{\ell!} h_{i,\ell}\, \mathrm{He}_\ell\Big(\frac{1}{\sqrt{dQ_{ii}}}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big) + h_{i,>k}\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big)\,, \tag{194}$$

$$\mathbb{E}\Big\{\Big|h_{i,>k}\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big)\Big|^m\Big\}^{1/m} \le C\, d^{-(k+1)/2}\,. \tag{195}$$

We next prove Eq. (180). Using Eq. (194) with $k=2$ and recalling $\mathrm{He}_0(x)=1$, $\mathrm{He}_1(x)=x$, $\mathrm{He}_2(x)=x^2-1$, we get

$$\begin{aligned}
v_i &= \mathbb{E}_{\boldsymbol{x}}\left\{h\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big) f^*(\boldsymbol{x})\right\} \\
&= h_{i,0}\mathbb{E}_{\boldsymbol{x}}\left\{f^*(\boldsymbol{x})\right\} + \frac{h_{i,1}}{\sqrt{dQ_{ii}}}\langle \boldsymbol{x}_i, \mathbb{E}_{\boldsymbol{x}}\left\{\boldsymbol{x} f^*(\boldsymbol{x})\right\}\rangle \\
&\quad + \frac{h_{i,2}}{2dQ_{ii}}\mathbb{E}_{\boldsymbol{x}}\left\{f^*(\boldsymbol{x})(\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle^2 - dQ_{ii})\right\} + \mathbb{E}_{\boldsymbol{x}}\left\{h_{i,>2}\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big) f^*(\boldsymbol{x})\right\} \\
&= h_{i,0} b_0 + \frac{h_{i,1}}{\sqrt{dQ_{ii}}}\langle \boldsymbol{\Sigma}\boldsymbol{\beta}_0, \boldsymbol{x}_i\rangle + \frac{h_{i,2}}{2dQ_{ii}}\langle \boldsymbol{x}_i, \boldsymbol{F}_2 \boldsymbol{x}_i\rangle + \mathbb{E}_{\boldsymbol{x}}\left\{h_{i,>2}\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big) f^*(\boldsymbol{x})\right\}\,.
\end{aligned}$$

Here we defined the $d\times d$ matrix $\boldsymbol{F}_2 = \mathbb{E}\{[f^*(\boldsymbol{x})-b_0]\boldsymbol{x}\boldsymbol{x}^\intercal\}$. Recalling the definitions of $h_{i,k}$, in Eq. (188), we get $h_{i,0}=a_{i,0}$. Comparing other terms we obtain that the following holds with very high probability,

$$\begin{aligned}
|v_i - v_{0,i}| \le& \frac{1}{d}\Big|\mathbb{E}\Big\{h'\Big(\sqrt{\frac{Q_{ii}}{d}}\, G\Big)\Big\} - h'(0)\Big| \cdot |\langle \boldsymbol{\Sigma}\boldsymbol{\beta}_0, \boldsymbol{x}_i\rangle| + \frac{1}{d^2}\Big|\mathbb{E}\Big\{h''\Big(\sqrt{\frac{Q_{ii}}{d}}\, G\Big)\Big\}\Big| \cdot \Big|\langle \boldsymbol{x}_i, \boldsymbol{F}_2\boldsymbol{x}_i\rangle\Big| \\
&+ \Big|\mathbb{E}_{\boldsymbol{x}}\left\{h_{i,>2}\Big(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle\Big) f^*(\boldsymbol{x})\right\}\Big| \\
\overset{(a)}{\le}& \frac{1}{d}\times\frac{C}{d}\times C\log d + \frac{C}{d^2}\Big|\langle \boldsymbol{x}_i, \boldsymbol{F}_2\boldsymbol{x}_i\rangle\Big| + C\, d^{-3/2} \\
\le& \frac{C}{d^2}\Big|\langle \boldsymbol{x}_i, \boldsymbol{F}_2\boldsymbol{x}_i\rangle\Big| + C\, d^{-3/2}.
\end{aligned}$$

Here the inequality $(a)$ follows since $|\mathbb{E}\{h'(Z)-h'(0)\}| \le C\mathbb{E}\{Z^2\}$ by smoothness of $h$ and Taylor expansion, $\max_{i\le n}|\langle \boldsymbol{\Sigma}\boldsymbol{\beta}_0, \boldsymbol{x}_i\rangle| \le C\sqrt{\log n}$ by sub-Gaussian tail bounds, and we used Eq. (195) for the last term.

The proof of Eq. (180) is completed by showing that, with very high probability, $\max_{i\le n}|\langle \boldsymbol{x}_i, \boldsymbol{F}_2\boldsymbol{x}_i\rangle| \le C\|\mathsf{P}_{>1}f^*\|_{L^2}\sqrt{d\log d}$. Without loss of generality, we assume here $\|\mathsf{P}_{>1}f^*\|_{L^2}=1$. In order to show this claim, note that (defining $\mathsf{P}_{>0}f^*(\boldsymbol{x}) := f^*(\boldsymbol{x}) - \mathbb{E}f^*(\boldsymbol{x})$)

$$\mathbb{E}\langle \boldsymbol{x}_i, \boldsymbol{F}_2\boldsymbol{x}_i\rangle = \mathsf{tr}(\boldsymbol{\Sigma}\boldsymbol{F}_2) \le C\mathbb{E}\{\mathsf{P}_{>0}f^*(\boldsymbol{x})\|\boldsymbol{x}\|_2^2\} \le \mathrm{Var}(\|\boldsymbol{x}\|_2^2)^{1/2} \le C\sqrt{d}\,. \tag{196}$$

Further notice that

$$\begin{aligned}
\|\boldsymbol{F}_2\| &= \max_{\|\boldsymbol{v}\|_2=1} |\langle \boldsymbol{v}, \boldsymbol{F}_2\boldsymbol{v}\rangle| && (197)\\
&= \max_{\|\boldsymbol{v}\|_2=1} \big|\mathbb{E}\big\{\mathsf{P}_{>0}f^*(\boldsymbol{x})\langle \boldsymbol{v}, \boldsymbol{x}\rangle^2\big\}\big| && (198)\\
&\le \max_{\|\boldsymbol{v}\|_2=1} \mathbb{E}\big\{\langle \boldsymbol{v}, \boldsymbol{x}\rangle^4\big\}^{1/2} \le C\,. && (199)
\end{aligned}$$

By the above and the Hanson-Wright inequality

$$\mathbb{P}\big(\langle \boldsymbol{x}_i, \boldsymbol{F}_2\boldsymbol{x}_i\rangle \ge C\sqrt{d}+t\big) \le 2\,\exp\Big(-c\Big(\frac{t^2}{\|\boldsymbol{F}_2\|_F^2} \wedge \frac{t}{\|\boldsymbol{F}_2\|}\Big)\Big) \le 2\, e^{-c((t^2/d)\wedge t)}\,, \tag{200}$$

and similarly for the lower tail. By taking a union bound over $i \le n$, we obtain $\max_{i\le n} |\langle \boldsymbol{x}_i, \boldsymbol{F}_2 \boldsymbol{x}_i\rangle| \le C\sqrt{d\log d}$ as claimed, thus completing the proof of Eq. (180).

We next prove Eq. (181). We claim that this bound holds for any realization in $\mathcal{E}_1 \cap \mathcal{E}_2$. Therefore we can fix without loss of generality $i = 1$, $j = 2$. We use Eq. (194) with $k = 4$. Using Cauchy-Schwarz and Eqs. (194), (195), we get

$$M_{12} = \sum_{\ell_1,\ell_2=0}^{4} \frac{1}{\ell_1!\ell_2!} h_{1,\ell_1} h_{2,\ell_2} M_{1,2}(\ell_1,\ell_2) + \Delta_{12}\,, \tag{201}$$

$$M_{1,2}(\ell_1,\ell_2) := \mathbb{E}_{\boldsymbol{x}}\left\{ \mathrm{He}_{\ell_1}\Big(\frac{1}{\sqrt{dQ_{11}}}\langle \boldsymbol{x}_1, \boldsymbol{x}\rangle\Big)\mathrm{He}_{\ell_2}\Big(\frac{1}{\sqrt{dQ_{22}}}\langle \boldsymbol{x}_2, \boldsymbol{x}\rangle\Big)\right\}, \quad |\Delta_{12}| \le Cd^{-5/2}\,. \tag{202}$$

Note that, by Eq. (188), $|h_{ik}| \le Cd^{-k/2}$, and $M_{1,2}(\ell_1,\ell_2)$ is bounded on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, by the sub-Gaussianity of $\boldsymbol{z}$. Comparing with Eqs. (177), (179), we get

$$|M_{12} - M_{0,12}| \le \Big| \sum_{\ell_1,\ell_2=0}^{4} \frac{1}{\ell_1!\ell_2!} h_{1,\ell_1} h_{2,\ell_2} M_{1,2}(\ell_1,\ell_2) - M_{0,12}\Big| + Cd^{-5/2} \tag{203}$$

$$+ 2\sum_{(\ell_1,\ell_2)\in\mathcal{S}} \big|h_{1,\ell_1} h_{2,\ell_2} M_{1,2}(\ell_1,\ell_2)\big| \tag{204}$$

$$+ 2\Big|\frac{1}{6} h_{1,0} h_{2,3} M_{1,2}(0,3) - a_{1,0}a_{2,1}\Big| + |a_{1,1}a_{2,1}| + Cd^{-5/2}\,,$$

$$\mathcal{S} := \big\{(0,1),(0,2),(0,4),(1,2),(1,3),(2,2)\big\}\,, \tag{205}$$

where in the inequality we used the identities $h_{1,0}h_{2,0}M_{1,2}(0,0) = h_{1,0}h_{2,0} = a_{1,0}a_{2,0}$, and

$$h_{1,1}h_{2,1}M_{1,2}(1,1) = \frac{1}{d^2}\langle \boldsymbol{x}_1, \boldsymbol{\Sigma}\boldsymbol{x}_2\rangle \mathbb{E}h'\Big(\sqrt{\frac{Q_{11}}{d}}G\Big)\mathbb{E}h'\Big(\sqrt{\frac{Q_{22}}{d}}G\Big) = B_{12}\,.$$

We next bound each of the terms above separately.

We begin with the terms $(\ell_1,\ell_2) \in \mathcal{S}$. Since by Eq. (188), $|h_{ik}| \le Cd^{-k/2}$, for each of these pairs, we need to show $|M_{1,2}(\ell_1,\ell_2)| \le Cd^{(\ell_1+\ell_2-5)/2}\log d$. Consider $(\ell_1,\ell_2) = (0,k)$, $k \in \{1,2,4\}$. Set $\boldsymbol{w} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_2/\sqrt{dQ_{22}}$, $\|\boldsymbol{w}\|_2 = 1$, and write $\mathrm{He}_k(x) = \sum_{m=0}^{k} c_{k,\ell}x^\ell$. If $\boldsymbol{g}$ is a standard Gaussian vector, we have $\mathbb{E}_{\boldsymbol{g}}\mathrm{He}_k\big(\langle \boldsymbol{w}, \boldsymbol{g}\rangle\big) = 0$ and therefore

$$M_{1,2}(0,k) = \mathbb{E}_{\boldsymbol{z}}\big\{\mathrm{He}_k\big(\langle \boldsymbol{w}, \boldsymbol{z}\rangle\big)\big\} - \mathbb{E}_{\boldsymbol{g}}\big\{\mathrm{He}_k\big(\langle \boldsymbol{w}, \boldsymbol{g}\rangle\big)\big\} \tag{206}$$

$$= \sum_{\ell=0}^{k} c_{k,\ell} \sum_{i_1,\dots,i_\ell\le n} w_{i_1}\cdots w_{i_\ell}\big\{\mathbb{E}(z_{i_1}\cdots z_{i_\ell}) - \mathbb{E}(g_{i_1}\cdots g_{i_\ell})\big\}\,. \tag{207}$$

Note that the only non-vanishing terms in the above sum are those in which all of the indices appearing in $(i_1,\dots,i_\ell)$ appear at least twice, and at least one of the indices appears at least 3 times (because otherwise the two expectations are equal). This immediately implies $M_{1,2}(0,1) = M_{1,2}(0,2) = 0$. Analogously, all terms $\ell \le 2$ vanish in the above sum.

As for $k = 4$, we have (recalling $\mathrm{He}_4(x) = x^4 - 3x^2$):

$$M_{1,2}(0,4) = \Big|\sum_{i_1,\dots,i_4\le n} w_{i_1}\cdots w_{i_4}\big\{\mathbb{E}(z_{i_1}\cdots z_{i_4}) - \mathbb{E}(g_{i_1}\cdots g_{i_4})\big\}\Big| \tag{208}$$

$$\le \sum_{i\le n} w_i^4\big|\mathbb{E}(z_i^4) - 3\big| \le C\|\boldsymbol{w}\|_\infty^2\|\boldsymbol{w}\|_2^2 \le \frac{C\log d}{d}\,, \tag{209}$$

where the last inequality follows since $\|\boldsymbol{w}\|_2 = 1$ by construction and $\|\boldsymbol{w}\|_\infty \le C\sqrt{(\log d)/d}$ on $\mathcal{E}_1 \cap \mathcal{E}_2$.

Next consider $(\ell_1, \ell_2) = (1, 2)$. Setting $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_i/\sqrt{dQ_{ii}}$, $i \in \{1, 2\}$, we get

$$M_{1,2}(1,2) = \mathbb{E}_{\boldsymbol{z}}\left\{\mathrm{He}_1\big(\langle \boldsymbol{w}_1, \boldsymbol{z}\rangle\big)\mathrm{He}_2\big(\langle \boldsymbol{w}_2, \boldsymbol{z}\rangle\big)\right\} - \mathbb{E}_{\boldsymbol{g}}\left\{\mathrm{He}_1\big(\langle \boldsymbol{w}_1, \boldsymbol{g}\rangle\big)\mathrm{He}_2\big(\langle \boldsymbol{w}_2, \boldsymbol{g}\rangle\big)\right\} \tag{210}$$

$$= \mathbb{E}_{\boldsymbol{z}}\left\{\big(\langle \boldsymbol{w}_1, \boldsymbol{z}\rangle\big)\big(\langle \boldsymbol{w}_2, \boldsymbol{z}\rangle\big)^2\right\} - \mathbb{E}_{\boldsymbol{g}}\left\{\big(\langle \boldsymbol{w}_1, \boldsymbol{g}\rangle\big)\big(\langle \boldsymbol{w}_2, \boldsymbol{g}\rangle\big)^2\right\} \tag{211}$$

$$= \sum_{i_1,i_2,i_3 \le n} w_{1,i_1}w_{2,i_2}w_{2,i_3}\big\{\mathbb{E}(z_{i_1}z_{i_2}z_{i_2}) - \mathbb{E}(g_{i_1}g_{i_2}g_{i_3})\big\} \tag{212}$$

$$= \sum_{i=1}^{n} w_{1,i}w_{2,i}\mathbb{E}(z_i^3)\,. \tag{213}$$

Therefore, on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\big|M_{1,2}(1,2)\big| \le C\Big|\sum_{i=1}^{n} w_{1,i}w_{2,i}^2\Big| \le \frac{C\log d}{d}\,. \tag{214}$$

Next consider $(\ell_1, \ell_2) = (1, 3)$. Proceeding as above (and noting that the degree-one term in $\mathrm{He}_3$ does not contribute), we get

$$M_{1,2}(1,3) = \mathbb{E}_{\boldsymbol{z}}\left\{\mathrm{He}_1\big(\langle \boldsymbol{w}_1, \boldsymbol{z}\rangle\big)\mathrm{He}_3\big(\langle \boldsymbol{w}_2, \boldsymbol{z}\rangle\big)\right\} - \mathbb{E}_{\boldsymbol{g}}\left\{\mathrm{He}_1\big(\langle \boldsymbol{w}_1, \boldsymbol{g}\rangle\big)\mathrm{He}_3\big(\langle \boldsymbol{w}_2, \boldsymbol{g}\rangle\big)\right\} \tag{215}$$

$$= \mathbb{E}_{\boldsymbol{z}}\left\{\big(\langle \boldsymbol{w}_1, \boldsymbol{z}\rangle\big)\big(\langle \boldsymbol{w}_2, \boldsymbol{z}\rangle\big)^3\right\} - \mathbb{E}_{\boldsymbol{g}}\left\{\big(\langle \boldsymbol{w}_1, \boldsymbol{g}\rangle\big)\big(\langle \boldsymbol{w}_2, \boldsymbol{g}\rangle\big)^3\right\} \tag{216}$$

$$= \sum_{i_1,\dots,i_4 \le d} w_{1,i_1}w_{2,i_2}w_{2,i_3}w_{2,i_4}\big\{\mathbb{E}(z_{i_1}z_{i_2}z_{i_2}z_{i_4}) - \mathbb{E}(g_{i_1}g_{i_2}g_{i_3}g_{i_4})\big\} \tag{217}$$

$$= \sum_{i=1}^{d} w_{1,i}w_{2,i}^3(\mathbb{E}(z_i^4) - 3)\,. \tag{218}$$

Therefore, on $\mathcal{E}_1 \cap \mathcal{E}_2$

$$\big|M_{1,2}(1,3)\big| \le C\Big|\sum_{i=1}^{d} w_{1,i}w_{2,i}^3\Big| \le C\|\boldsymbol{w}_1\|_\infty\|\boldsymbol{w}_2\|_\infty\|\boldsymbol{w}_2\|_2^2 \le \frac{C\log d}{d}\,. \tag{219}$$

Finally, for $(\ell_1, \ell_2) = (2, 2)$, proceeding as above we get

$$M_{1,2}(2,2) = \left|\sum_{i=1}^{d} w_{1,i}^2w_{2,i}^2(\mathbb{E}(z_i^4) - 3)\right| \le C\|\boldsymbol{w}_1\|_\infty^2\|\boldsymbol{w}_2\|_2^2 \le \frac{C\log d}{d}\,. \tag{220}$$

Next consider the term $|h_{1,0}h_{2,3}M_{1,2}(0,3)/6 - a_{1,0}a_{2,1}|$ in Eq. (204). Using the fact that $h_{1,0} = a_{1,0}$ is bounded, we get

$$\Big|\frac{1}{6}h_{1,0}h_{2,3}M_{1,2}(0,3) - a_{1,0}a_{2,1}\Big| \le C\big|h_{2,3}M_{1,2}(0,3) - 6a_{2,1}\big|\,. \tag{221}$$

Recalling $\mathrm{He}_3(x) = x^3 - 3x$, and letting $\boldsymbol{w} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_2/\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_2\|_2$:

$$M_{1,2}(0,3) = \sum_{i_1,\dots,i_3 \le d} w_{i_1}w_{i_2}w_{i_3}\big\{\mathbb{E}(z_{i_1}z_{i_2}z_{i_3}) - \mathbb{E}(g_{i_1}g_{i_2}g_{i_3})\big\} \tag{222}$$

$$= \sum_{i \le d} w_i^3\mathbb{E}(z_i^3)\,. \tag{223}$$

In particular, on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, $|M_{1,2}(0,3)| \le C\sqrt{(\log d)/d}$. Comparing the definitions of $a_{2,1}$ and $h_{2,3}$, we get

$$\left|h_{1,0}h_{2,3}M_{1,2}(0,3) - a_{1,0}a_{2,1}\right| \le C|M_{1,2}(0,3)| \times \left(\frac{Q_{22}}{d}\right)^{3/2} \left|\mathbb{E}\Big\{h^{(3)}\Big(\sqrt{\frac{Q_{ii}}{d}}\, G\Big)\Big\} - h^{(3)}(0)\right| \tag{224}$$

$$\le C\sqrt{\frac{\log d}{d}} \times \frac{1}{d^{3/2}} \times \frac{1}{d^{1/2}} \le \frac{C(\log d)^{1/2}}{d^{5/2}}\,. \tag{225}$$

Finally, consider term $|a_{1,1}a_{2,1}|$ in Eq. (204). By the above estimates, we get $|a_{2,1}| \le Cd^{-2}(\log d)^{1/2}$, and hence this term is negligible as well. This completes the proof of Eq. (181).

Equation (182) follows by a similar argument, which we omit. □

### A.2.2 An estimate on the entries of the resolvent

**Lemma A.4.** *Let $\boldsymbol{Z} = (z_{ij})_{i\le n, j\le d}$ be a random matrix with iid rows $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n \in \mathbb{R}^d$ that are zero mean and $C$-sub-Gaussian. Further assume $C^{-1} \le n/d \le C$. Let $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ be a symmetric matrix such that $\mathbf{0} \preceq \boldsymbol{S} \preceq C\mathbf{I}_d$ for some finite constant $C > 1$. Finally, let $g : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that $\mathbb{E}\{g(\boldsymbol{z}_1)\} = \mathbb{E}\{\boldsymbol{z}_1 g(\boldsymbol{z}_1)\} = 0$, and $\mathbb{E}\{g(\boldsymbol{z}_1)^2\} = 1$.*

*Then, for any $\lambda > 0$ there exists a finite constant $C$ such that, for any $i \ne j$,*

$$\left|\mathbb{E}\Big\{\big(\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}^{\mathsf{T}}/d + \lambda\mathbf{I}_n\big)^{-1}_{i,j} g(\boldsymbol{z}_i)g(\boldsymbol{z}_j)\Big\}\right| \le C\, d^{-3/2}\,. \tag{226}$$

*Proof.* Without loss of generality, we can consider $i = 1$, $j = 2$. Further, we let $\boldsymbol{Z}_0 \in \mathbb{R}^{(n-2)\times d}$ be the matrix comprising the last $n-2$ rows of $\boldsymbol{Z}$, and $\boldsymbol{U} \in \mathbb{R}^{d\times 2}$ be the matrix with columns $\boldsymbol{U}\boldsymbol{e}_1 = \boldsymbol{z}_1$, $\boldsymbol{U}\boldsymbol{e}_2 = \boldsymbol{z}_2$. We finally define the matrices $\boldsymbol{R}_0 \in \mathbb{R}^{d\times d}$ and $\boldsymbol{Y} = (Y_{ij})_{i,j\le 2}$:

$$\boldsymbol{R}_0 := \lambda\boldsymbol{S}^{1/2}\big(\boldsymbol{S}^{1/2}\boldsymbol{Z}_0^{\mathsf{T}}\boldsymbol{Z}_0\boldsymbol{S}^{1/2}/d + \lambda\mathbf{I}_d\big)^{-1}\boldsymbol{S}^{1/2}\,, \tag{227}$$

$$\boldsymbol{Y} := \big(\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}^{\mathsf{T}}/d + \lambda\mathbf{I}_n\big)^{-1}\,. \tag{228}$$

Then, by a simple linear algebra calculation, we have

$$\boldsymbol{Y} = \Big(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{R}_0\boldsymbol{U}/d + \lambda\mathbf{I}_2\Big)^{-1}\,, \tag{229}$$

$$Y_{12} = -\frac{\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle/d}{(\lambda + \langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle/d)(\lambda + \langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle/d) - \langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle^2/d^2}\,. \tag{230}$$

Note that since $\boldsymbol{R}_0 \succeq 0$, we have $\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle^2 \le \langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_1\rangle\langle \boldsymbol{z}_2, \boldsymbol{R}_0\boldsymbol{z}_2\rangle$, and therefore

$$Y_{12} = Y_{12}^{(1)} + Y_{12}^{(2)}\,, \tag{231}$$

$$Y_{12}^{(1)} := -\frac{\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle/d}{(\lambda + \langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_1\rangle/d)(\lambda + \langle \boldsymbol{z}_2, \boldsymbol{R}_0\boldsymbol{z}_2\rangle/d)}\,, \tag{232}$$

$$|Y_{12}^{(2)}| \le \frac{1}{\lambda^4 d^3}|\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle|^3\,. \tag{233}$$

Denote by $\mathbb{E}_+$ expectation with respect to $\boldsymbol{z}_1, \boldsymbol{z}_2$ (conditional on $(\boldsymbol{z}_i)_{2<i\le n}$). We have

$$\begin{aligned}
\left|\mathbb{E}_+\{Y_{12}\, g(\boldsymbol{z}_1)\, g(\boldsymbol{z}_2)\}\right| &\le \left|\mathbb{E}_+\{Y_{12}^{(1)}\, g(\boldsymbol{z}_1)\, g(\boldsymbol{z}_2)\}\right| + \mathbb{E}_+\{(Y_{12}^{(2)})^2\}^{1/2}\, \mathbb{E}_+\{g(\boldsymbol{z}_1)^2\, g(\boldsymbol{z}_2)^2\}^{1/2} \\
&\le \left|\mathbb{E}_+\{Y_{12}^{(1)}\, g(\boldsymbol{z}_1)\, g(\boldsymbol{z}_2)\}\right| + \mathbb{E}_+\{(Y_{12}^{(2)})^2\}^{1/2} \\
&\le \left|\mathbb{E}_+\{Y_{12}^{(1)}\, g(\boldsymbol{z}_1)\, g(\boldsymbol{z}_2)\}\right| + C\, d^{-3/2}\,.
\end{aligned}$$

Here the last step follows by the Hanson-Wright inequality. We therefore only have to bound the first term. Defining $q_j := \lambda + \langle \boldsymbol{z}_j, \boldsymbol{R}_0 \boldsymbol{z}_j\rangle/d$, $\overline{q}_j = \mathbb{E}_+ q_j$, $g_j = g(\boldsymbol{z}_j)$, $j \in \{1,2\}$,

$$\begin{aligned}
\big|\mathbb{E}_+\{Y_{12}^{(1)}\, g_1\, g_2\}\big| &\le \Big|\mathbb{E}_+\Big\{\overline{q}^{-2}\frac{\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle}{d} g_1 g_2\Big\}\Big| \\
&\quad + 2\Big|\mathbb{E}_+\Big\{\Big(q_1^{-1} - \overline{q}^{-1}\Big)\overline{q}^{-2}\frac{\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle}{d} g_1 g_2\Big\}\Big| \\
&\quad + \Big|\mathbb{E}\Big\{\Big(q_1^{-1} - \overline{q}^{-1}\Big)\Big(q_2^{-1} - \overline{q}^{-1}\Big)\frac{\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle}{d} g_1 g_2\Big\}\Big| \\
&\overset{(a)}{\le} \Big|\mathbb{E}_+\Big\{\Big(q_1^{-1} - \overline{q}^{-1}\Big)\Big(q_2^{-1} - \overline{q}^{-1}\Big)\frac{\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle}{d} g_1 g_2\Big\}\Big| \\
&\le \frac{1}{\lambda^4}\mathbb{E}\Big\{|q_1 - \overline{q}||q_2 - \overline{q}|\Big|\frac{\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle}{d}\Big||g_1 g_2|\Big\}\,.
\end{aligned}$$

Here $(a)$ follows from the orthogonality of $g(\boldsymbol{z})$ to linear functions.

We then conclude

$$\begin{aligned}
\big|\mathbb{E}_+\{Y_{12}^{(1)}\, g_1\, g_2\}\big| &\overset{(a)}{\le} C\mathbb{E}_+\big\{|q_1 - \overline{q}|^8\big\}^{1/4}\mathbb{E}_+\big\{(\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle/d)^4\big\}^{1/4} \\
&\le C\mathbb{E}_+\big\{|\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_1\rangle/d - \mathbb{E}_+\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_1\rangle/d|^8\big\}^{1/4}\mathbb{E}_+\big\{(\langle \boldsymbol{z}_1, \boldsymbol{R}_0\boldsymbol{z}_2\rangle/d)^4\big\}^{1/4} \\
&\overset{(b)}{\le} C(d^{-1/2})^2 \times Cd^{-1/2} \le Cd^{-3/2}\,.
\end{aligned}$$

Here $(a)$ follows from Hölder's inequality and $(b)$ from the Hanson-Wright inequality using the fact that $\|\boldsymbol{R}_0\|$ is bounded. The proof is completed by taking expectation over $(\boldsymbol{z}_i)_{2<i\le n}$. □

**Lemma A.5.** *Under the definitions and assumptions of Lemma A.4, let $Y_{ij} := (\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}^\intercal/d + \lambda \mathbf{I}_n)^{-1}_{i,j}$. Then, for any tuple of four distinct indices $i,j,k,l$, we have*

$$\big|\mathbb{E}\{Y_{ij}Y_{kl}g(\boldsymbol{z}_i)g(\boldsymbol{z}_j)g(\boldsymbol{z}_k)g(\boldsymbol{z}_l)\}\big| \le Cd^{-5/2}\,. \tag{234}$$

*Proof.* The proof is analogous to the one of Lemma A.4. Without loss of generality, we set $(i,j,k,l) = (1,2,3,4)$, denote by $\boldsymbol{Z}_0 \in \mathbb{R}^{(n-4)\times d}$ the matrix with rows $(\boldsymbol{z}_\ell)_{\ell\ge 5}$, and define the $d\times d$ matrix

$$\boldsymbol{R}_0 := \lambda \boldsymbol{S}^{1/2}\big(\boldsymbol{S}^{1/2}\boldsymbol{Z}_0^\intercal\boldsymbol{Z}_0\boldsymbol{S}^{1/2}/d + \lambda \mathbf{I}_{n-2}\big)^{-1}\boldsymbol{S}^{1/2}\,. \tag{235}$$

We then have that $\boldsymbol{Y} = (Y_{ij})_{i,j\le 4}$ is given by

$$\boldsymbol{Y} = (\mathrm{diag}(\boldsymbol{q}) + \boldsymbol{A})^{-1}\,, \tag{236}$$

$$q_i := \overline{q} + Q_i\,,\quad \overline{q} := \lambda + \mathsf{tr}(\boldsymbol{R}_0)/d\,,\quad Q_i = (\langle \boldsymbol{z}_i, \boldsymbol{R}_0\boldsymbol{z}_i\rangle - \mathbb{E}\langle \boldsymbol{z}_i, \boldsymbol{R}_0\boldsymbol{z}_i\rangle)/d\,, \tag{237}$$

$$A_{ij} := \begin{cases} \langle \boldsymbol{z}_i, \boldsymbol{R}_0\boldsymbol{z}_j\rangle/d & \text{if } i\ne j,\\ 0 & \text{if } i = j.\end{cases} \tag{238}$$

In what follows we denote by $\mathbb{E}_+$ expectation with respect to $(\boldsymbol{z}_i)_{i\le 4}$, with $\boldsymbol{Z}_0$ fixed. Note that, by the Hanson-Wright inequality, $\mathbb{E}_+\{|A_{ij}|^k\}^{1/k} \le c_k\, d^{-1/2}$, $\mathbb{E}_+\{|Q_i|^k\}^{1/k} \le c_k\, d^{-1/2}$ for each $k \ge 1$. We next compute the Taylor expansion of $Y_{12}$ and $Y_{3,4}$ in powers of $\boldsymbol{A}$ to get

$$Y_{12} = Y_{12}^{(1)} + Y_{12}^{(2)} + Y_{12}^{(3)} + Y_{12}^{(4)}\,, \tag{239}$$

$$Y_{12}^{(1)} := -q_1^{-1}A_{12}q_2\,, \tag{240}$$

$$Y_{12}^{(2)} := q_1^{-1}A_{13}q_3^{-1}A_{32}q_2^{-1} + q_1^{-1}A_{14}q_4^{-1}A_{41}q_2^{-1}\,, \tag{241}$$

$$Y_{12}^{(3)} := -\sum_{i_1\ne i_2, i_1\ne 1 i_2\ne 2} q_1^{-1}A_{1i_1}q_{i_1}^{-1}A_{i_1i_2}q_{i_2}^{-1}A_{i_2 2}q_2^{-1}\,, \tag{242}$$

and similarly for $Y_{34}$. It is easy to show that $\mathbb{E}_+\{|Y_{ab}^{(\ell)}|^k\}^{1/k} \le c_k d^{-\ell/2}$, for all $k \ge 1$. Therefore, using $\mathbb{E}\{g(\boldsymbol{z}_i)^2\} \le C$ and Cauchy-Schwarz inequality, and writing $g_i = g(\boldsymbol{z}_i)$:

$$\big|\mathbb{E}\{Y_{12}Y_{34}g_1g_2g_3g_4\}\big| = \sum_{\ell_1+\ell_2\le 4} \big|\mathbb{E}\{Y_{12}^{(\ell_1)}Y_{34}^{(\ell_2)}g_1g_2g_3g_4\}\big| + Cd^{-5/2}\,. \tag{243}$$

The proof is completed by bounding each of the terms above, which we now do. By symmetry it is sufficient to consider $\ell_1 \le \ell_2$ and therefore we are left with the 4 pairs $(\ell_1,\ell_2) \in \{(1,1),(1,2),(1,3),(2,2)\}$.

**Term $(\ell_1,\ell_2) = (1,1)$.** By the same argument as in the proof of Lemma A.4, we have $|\mathbb{E}\{A_{ij}q_i^{-1}q_j^{-1}g_ig_j\}| \le Cd^{-3/2}$ and therefore

$$\big|\mathbb{E}_+\{Y_{12}^{(1)}Y_{34}^{(1)}g_1g_2g_3g_4\}\big| = \big|\mathbb{E}_+\{A_{12}q_1^{-1}q_2^{-1}g_1g_2\}\big|\cdot\big|\mathbb{E}\{A_{34}q_3^{-1}q_4^{-1}g_3g_4\}\big| \le Cd^{-3}\,. \tag{244}$$

**Term $(\ell_1,\ell_2) = (1,2)$.** Note that each of the two terms in the definition of $Y_{34}^{(2)}$ contributes a summand with the same structure. Hence we can consider just the one resulting in the largest expectation, say $q_3^{-1}A_{31}q_1^{-1}A_{14}q_4^{-1}$

$$\begin{aligned}
\big|\mathbb{E}_+\{Y_{12}^{(1)}Y_{34}^{(2)}g_1g_2g_3g_4\}\big| &= 2\big|\mathbb{E}_+\{q_1^{-1}A_{12}q_2^{-1}q_3^{-1}A_{31}q_1^{-1}A_{14}q_4^{-1}g_1g_2g_3g_4\}\big| \\
&\overset{(a)}{=} 2\big|\mathbb{E}_+\{q_1^{-2}A_{12}(q_2^{-1}-\overline{q}^{-1})(q_3^{-1}-\overline{q}^{-1})A_{31}A_{14}(q_4^{-1}-\overline{q}^{-1})g_1g_2g_3g_4\}\big| \\
&\overset{(b)}{\le} C\mathbb{E}_+\{|A_{12}|^p\}^{1/p}\mathbb{E}\{|A_{13}|^p\}^{1/p}\mathbb{E}\{|A_{13}|^p\}^{1/p}\mathbb{E}_+\{|q_2^{-1}-\overline{q}^{-1}|^p\}^{1/p}\mathbb{E}\{|q_3^{-1}-\overline{q}^{-1}|^p\}^{1/p} \\
&\qquad\cdot\mathbb{E}\{|q_4^{-1}-\overline{q}^{-1}|^p\}^{1/p}\|g\|_{L^2}^4 \\
&\overset{(c)}{\le} Cd^{-3}\,.
\end{aligned}$$

Here $(a)$ holds because $g_i$ is orthogonal to $\boldsymbol{z}_i$ for $i \in \{2,3,4\}$ and hence the terms $\overline{q}^{-1}$ have vanishing contribution; $(b)$ by Hölder for $p = 12$, and using the fact that $q_i^{-1}$ is bounded; $(c)$ by the above bounds on the moments of $A_{ij}$, $Q_i$, plus $|q_i^{-1}-\overline{q}^{-1}| \le C|Q_i|$.

**Term $(\ell_1,\ell_2) = (1,3)$.** Taking into account symmetries, there are only two distinct terms to consider in the sum defining $Y_{34}^{(3)}$, which we can identify with the following ones:

$$\begin{aligned}
\big|\mathbb{E}_+\{Y_{12}^{(1)}Y_{34}^{(3)}g_1g_2g_3g_4\}\big| &\le C\big|\mathbb{E}_+\big\{q_1^{-1}A_{12}q_2^{-1}q_3^{-1}A_{31}q_1^{-1}A_{12}q_2^{-1}A_{24}q_4^{-1}g_1g_2g_3g_4\big\}\big| \\
&\quad + C\big|\mathbb{E}_+\big\{q_1^{-1}A_{12}q_2^{-1}q_3^{-1}A_{31}q_1^{-1}A_{13}q_3^{-1}A_{34}q_4^{-1}g_1g_2g_3g_4\big\}\big| =: C\cdot T_1 + C\cdot T_2\,.
\end{aligned}$$

Notice that in the first term $\boldsymbol{z}_3$ only appears in $q_3$, $A_{31}$, and $g_3$, and similarly $\boldsymbol{z}_4$ only appears in $q_4$, $A_{24}$, and $g_4$. Hence

$$T_1 = \big|\mathbb{E}_+\big\{q_1^{-1}A_{12}q_2^{-1}(q_3^{-1}-\overline{q}^{-1})A_{31}q_1^{-1}A_{12}q_2^{-1}A_{24}(q_4^{-1}-\overline{q}^{-1})g_1g_2g_3g_4\big\}\big| \le Cd^{-3}\,,$$

where the last inequality follows again by Hölder. Analogously, for the second term we have

$$T_2 = \big|\mathbb{E}_+\big\{q_1^{-1}A_{12}(q_2^{-1}-\overline{q}^{-1})q_3^{-1}A_{31}q_1^{-1}A_{32}q_3^{-1}A_{24}(q_4^{-1}-\overline{q}^{-1})g_1g_2g_3g_4\big\}\big| \le Cd^{-3}\,,$$

This proves the desired bound for $(\ell_1,\ell_2) = (1,3)$.

**Term $(\ell_1,\ell_2) = (2,2)$.** There are four terms that arise from the sum in the definition of $Y_{ij}^{(2)}$. By symmetry, these are equivalent by pairs

$$\begin{aligned}
\big|\mathbb{E}_+\{Y_{12}^{(2)}Y_{34}^{(2)}g_1g_2g_3g_4\}\big| &\le 2\big|\mathbb{E}_+\big\{q_1^{-1}A_{13}q_3^{-1}A_{32}q_2^{-1}q_3^{-1}A_{31}q_1^{-1}A_{14}q_4^{-1}g_1g_2g_3g_4\big\}\big| \\
&\quad + 2\big|\mathbb{E}_+\big\{q_1^{-1}A_{13}q_3^{-1}A_{32}q_2^{-1}q_3^{-1}A_{32}q_2^{-1}A_{24}q_4^{-1}g_1g_2g_3g_4\big\}\big| \\
&\le 2\big|\mathbb{E}_+\big\{q_1^{-1}A_{13}q_3^{-1}A_{32}(q_2^{-1}-\overline{q}^{-1})q_3^{-1}A_{31}q_1^{-1}A_{14}(q_4^{-1}-\overline{q}^{-1})g_1g_2g_3g_4\big\}\big| \\
&\quad + 2\big|\mathbb{E}_+\big\{(q_1^{-1}-\overline{q}^{-1})A_{13}q_3^{-1}A_{32}q_2^{-1}q_3^{-1}A_{32}q_2^{-1}A_{24}(q_4^{-1}-\overline{q}^{-1})g_1g_2g_3g_4\big\}\big| \\
&\le Cd^{-3}\,.
\end{aligned}$$

This completes the proof of this lemma. □

**Lemma A.6.** *Under the definitions and assumptions of Lemma A.5, further assume* $\mathbb{E}\{|g(\boldsymbol{z})|^{2+\eta}\} \le C$ *for some constants* $0 < C, \eta < \infty$. *for any triple of four distinct indices* $i, j, k$, *we have*

$$\left|\mathbb{E}\{Y_{ij}Y_{jk}g(\boldsymbol{z}_i)g(\boldsymbol{z}_j)^2g(\boldsymbol{z}_k)\}\right| \le Cd^{-3/2}\,, \tag{245}$$

$$\left|\mathbb{E}\{Y_{ij}^2g(\boldsymbol{z}_i)^2g(\boldsymbol{z}_l)^2\}\right| \le Cd^{-1}\,. \tag{246}$$

*Proof.* This proof is very similar to the one of Lemma A.5, and we will follow the same notation introduced there.

Consider Eq. (245). Without loss of generality, we take $(i, j, k) = (1, 2, 3)$. Since $\mathbb{E}\{|Y_{ij}^{(\ell)}|^k\} \le c_k d^{-\ell/2}$, we have

$$\left|\mathbb{E}_+\{Y_{12}Y_{23}g_1g_2^2g_3\}\right| \le \left|\mathbb{E}_+\{Y_{12}^{(1)}Y_{23}^{(1)}g_1g_2^2g_3\}\right| + Cd^{-3/2}\,. \tag{247}$$

Further

$$\begin{aligned}\left|\mathbb{E}_+\{Y_{12}^{(1)}Y_{23}^{(1)}g_1g_2^2g_3\}\right| &= \left|\mathbb{E}_+\{q_1^{-1}A_{12}q_2^{-2}A_{23}q_3^{-1}g_1g_2^2g_3\}\right| \\ &= \left|\mathbb{E}_+\{(q_1^{-1}-\overline{q}^{-1})A_{12}q_2^{-2}A_{23}(q_3^{-1}-\overline{q}^{-1})g_1g_2^2g_3\}\right| \\ &\le Cd^{-2}\,,\end{aligned}$$

where the last bound follows from Hölder inequality.

Finally, Eq. (246) follows immediately by Hölder inequality since $\mathbb{E}\{|Y_{ij}|^k\}^{1/k} \le C_k d^{-1/2}$ for all $k$. □

**Theorem A.7.** *Let* $\boldsymbol{Z} = (z_{ij})_{i\le n, j\le d}$ *be a random matrix with iid rows* $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n \in \mathbb{R}^d$, *with zero mean C-sub-Gaussian. Let* $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ *be a symmetric matrix such that* $\mathbf{0} \preceq \boldsymbol{S} \preceq C\mathbf{I}_d$ *for some finite constant* $C > 1$. *Finally, let* $g : \mathbb{R}^d \to \mathbb{R}$ *be a measurable function such that* $\mathbb{E}\{g(\boldsymbol{z}_1)\} = \mathbb{E}\{\boldsymbol{z}_1 g(\boldsymbol{z}_1)\} = 0$, *and* $\mathbb{E}\{|g(\boldsymbol{z}_1)|^{4+\eta}\} \le C$.

*Then, for any* $\lambda > 0$, *with probability at least* $1 - Cd^{-1/4}$, *we have*

$$\left|\frac{1}{d}\sum_{i<j\le n}\left(\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}^\intercal/d + \lambda\mathbf{I}_n\right)^{-1}_{i,j} g(\boldsymbol{z}_i)g(\boldsymbol{z}_j)\right| \le C\, d^{-1/8}\,. \tag{248}$$

*Proof.* Denote by $X$ the sum on the left-hand side of Eq. (248), and define $Y_{ij} := (\boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}^\intercal/d + \lambda\mathbf{I}_n)^{-1}_{i,j}$, $g_i = g(\boldsymbol{z}_i)$. Further, let $\mathcal{I}_m := \{(i,j,k,l) :\ i < j \le n, k < l \le n, |\{i,j\} \cap \{k,j\}| = m\}$, $m \in \{0, 1\}$. Then we have

$$\begin{aligned}\mathbb{E}\{X^2\} &= \frac{1}{d^2}\sum_{i<j}\sum_{k<l}\mathbb{E}\{Y_{ij}Y_{kl}g_ig_jg_kg_l\} \\ &\le \frac{1}{d^2}\sum_{(i,j,k,l)\in\mathcal{I}_0}\mathbb{E}\{Y_{ij}Y_{kl}g_ig_jg_kg_l\} + \frac{1}{d^2}\sum_{(i,j,k,l)\in\mathcal{I}_1}\mathbb{E}\{Y_{ij}Y_{kl}g_ig_jg_kg_l\} + + \frac{1}{d^2}\sum_{i<j}\mathbb{E}\{Y_{ij}^2g_i^2g_j^2\} \\ &\le Cd^2\left|\mathbb{E}\{Y_{12}Y_{34}g_1g_2g_3g_4\}\right| + Cd\left|\mathbb{E}\{Y_{12}Y_{23}g_1g_2^2g_3\}\right| + C\left|\mathbb{E}\{Y_{12}^2g_1^2g_2^2\}\right| \\ &\le Cd^{-1/2}\,.\end{aligned}$$

The proof is completed by Chebyshev inequality. □

### A.2.3 Proof of Theorem 4.13: Variance term

Throughout this section we will refer to the events $\mathcal{E}_1$, $\mathcal{E}_2$ defined in Eqs. (183), (185). The variance is given by

$$\widehat{\mathrm{VAR}} = \sigma_\xi^2\mathbb{E}_{\boldsymbol{x}}\left\{K(\boldsymbol{x},\boldsymbol{X})^\intercal K(\boldsymbol{X},\boldsymbol{X})^{-2}K(\boldsymbol{x},\boldsymbol{X})\right\}. \tag{249}$$

The following lemma allows us to take the expectation with respect to $\boldsymbol{x}$.

**Lemma A.8.** *Under the assumptions of Theorem 4.13, define $\boldsymbol{M}_0 \in \mathbb{R}^{n\times n}$ as in the statement of Lemma A.3. Then, with very high probability, we have*

$$\left|\frac{1}{\sigma_\xi^2}\widehat{\mathrm{VAR}} - \langle \boldsymbol{M}_0, \boldsymbol{K}^{-2}\rangle\right| \le \frac{C\log d}{d}\,. \tag{250}$$

*Proof.* First notice that, defining $\boldsymbol{M}$ as in Eq. (173), we have

$$\frac{1}{\sigma_\xi^2}\widehat{\mathrm{VAR}} = \langle \boldsymbol{M}, \boldsymbol{K}^{-2}\rangle\,. \tag{251}$$

We then have, with very high probability,

$$\begin{aligned}\left|\frac{1}{\sigma^2}\widehat{\mathrm{VAR}} - \langle \boldsymbol{M}_0, \boldsymbol{K}^{-2}\rangle\right| &\le \left|\langle \boldsymbol{M}-\boldsymbol{M}_0, \boldsymbol{K}^{-2}\rangle\right| &(252)\\ &\le \|\boldsymbol{M}-\boldsymbol{M}_0\|_F\sqrt{n}\|\boldsymbol{K}^{-2}\| &(253)\\ &\overset{(a)}{\le} \frac{C\log d}{d^{3/2}}\times\sqrt{d}\times\|\boldsymbol{K}^{-1}\|^2 &(254)\\ &\overset{(b)}{\le} \frac{C\log d}{d}\,, &(255)\end{aligned}$$

where $(a)$ follows from Lemma A.3 and $(b)$ from Lemma A.2. □

In the following we define $\boldsymbol{B}_0 \in \mathbb{R}^{n\times n}$ via

$$\boldsymbol{B}_0 := \frac{h'(0)}{d^2}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\intercal\,. \tag{256}$$

The next lemma shows that $\boldsymbol{B}_0$ is a good approximation for $\boldsymbol{B}$, defined in Eq. (177).

**Lemma A.9.** *Let $\boldsymbol{B}$ be defined as per Eq. (177). With very high probability, we have $\|\boldsymbol{B}-\boldsymbol{B}_0\| \le Cd^{-3/2}$ and $\|\boldsymbol{B}-\boldsymbol{B}_0\|_* \le Cd^{-1/2}$.*

*Proof.* Notice that $\boldsymbol{B} = \boldsymbol{D}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\intercal\boldsymbol{D}/d^2$ and, on $\mathcal{E}_1\cap\mathcal{E}_2$,

$$\big\|\boldsymbol{D}-h'(0)\mathbf{I}\big\| = \max_{i\le n}\left|\mathbb{E}h'\Big(\sqrt{\frac{Q_{ii}}{d}}G\Big) - h'(0)\right| \le \frac{C}{\sqrt{d}}\,. \tag{257}$$

We then have

$$\big\|\boldsymbol{B}-\boldsymbol{B}_0\big\| \le \frac{C}{\sqrt{d}}\Big\|\frac{1}{d^2}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\intercal\Big\| \le \frac{C}{d^{5/2}}\|\boldsymbol{X}\|^2 \le \frac{C}{d^{3/2}}\,. \tag{258}$$

This immediately implies $\|\boldsymbol{B}-\boldsymbol{B}_0\|_* \le n\|\boldsymbol{B}-\boldsymbol{B}_0\| \le C/\sqrt{d}$. □

**Lemma A.10.** *Under the assumptions of Theorem 4.13, let $\boldsymbol{B}$ be defined as per Eq. (177) and $\boldsymbol{B}_0$ as per Eq. (256). Also, recall the definition of $\boldsymbol{K}_1$ in Eq. (163). Then, with very high probability, we have*

$$\big|\langle \boldsymbol{B}, \boldsymbol{K}^{-2}\rangle - \langle \boldsymbol{B}_0, \boldsymbol{K}_1^{-2}\rangle\big| \le C\,n^{-c_0}\,. \tag{259}$$

*Proof.* Throughout this proof, we work under events $\mathcal{E}_1\cap\mathcal{E}_2$ defined in the proof of Lemma A.3. Recall that $\max_{i\le n}|D_i|$ is bounded (see, e.g., Eq. (257)), whence

$$|B_{ij}| \le \frac{C}{d^2}\big|\langle \boldsymbol{x}_i, \boldsymbol{\Sigma}\boldsymbol{x}_j\rangle\big| \le \begin{cases} C/d & \text{if } i=j,\\ C(\log d)^{1/2}/d^{3/2} & \text{if } i\neq j,\end{cases} \tag{260}$$

whence $\|\boldsymbol{B}\|_F \le C\sqrt{(\log d)/d}$. Using Lemma A.2, we have

$$\begin{aligned}\left|\langle \boldsymbol{B}, \boldsymbol{K}^{-2}\rangle - \langle \boldsymbol{B}, \boldsymbol{K}_1^{-2}\rangle\right| &\le \|\boldsymbol{B}\|_F n^{1/2}\|\boldsymbol{K}^{-2} - \boldsymbol{K}_1^{-2}\| \\ &\le C\sqrt{(\log d)/d} \times n^{1/2}[\lambda_{\min}(\boldsymbol{K}) \wedge \lambda_{\min}(\boldsymbol{K}_1)]^{-3}\|\boldsymbol{K} - \boldsymbol{K}_1\| \\ &\le C\sqrt{\log d}\|\boldsymbol{K} - \boldsymbol{K}_1\| \le C\, n^{-c_0}\,.\end{aligned} \tag{261}$$

Using again Lemma A.2 together with Lemma A.9, we obtain that the following holds with very high probability:

$$\begin{aligned}\left|\langle \boldsymbol{B}, \boldsymbol{K}_1^{-2}\rangle - \langle \boldsymbol{B}_0, \boldsymbol{K}_1^{-2}\rangle\right| &\le \lambda_{\min}(\boldsymbol{K}_1)^{-2}\left\|\boldsymbol{B} - \boldsymbol{B}_0\right\|_* \\ &\le \frac{C}{d^{1/2}}\,.\end{aligned}$$

The desired claim follows from this display alongside Eq. (261). □

**Lemma A.11.** *Under the assumptions of Theorem 4.13, let $\boldsymbol{a}$ be defined as in Lemma A.3. Then, with very high probability we have*

$$0 \le \langle \boldsymbol{a}, \boldsymbol{K}^{-2}\boldsymbol{a}\rangle \le \frac{C}{n}\,. \tag{262}$$

*Proof.* Notice that the lower bound is trivial since $\boldsymbol{K}$ is positive semidefinite. We will write

$$\boldsymbol{K} = \alpha\, \mathbf{1}\mathbf{1}^{\mathsf{T}} + \boldsymbol{K}_*\,, \tag{263}$$

$$\boldsymbol{a} = h(0)\mathbf{1} + \tilde{\boldsymbol{a}}\,. \tag{264}$$

By standard bounds on the norm of matrices with i.i.d. rows (and using $\|\boldsymbol{\Sigma}\| \le C$), we have $0 \preceq \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}/d \preceq C\,\mathbf{I}$, with probability at least $1 - C\exp(-n/C)$. Therefore, by Lemma A.2, and since $\beta\gamma > 0$ is bounded away from zero by assumption, with very high probability we have $C^{-1}\mathbf{I} \preceq \boldsymbol{K}_* \preceq C\mathbf{I}$, for a suitable constant $C$. Note that $\tilde{\boldsymbol{a}} = (\boldsymbol{a}_0 - h(0)\mathbf{1}) + \boldsymbol{a}_1$. Under event $\mathcal{E}_1 \cap \mathcal{E}_2$, the following holds by smoothness of $h$:

$$\|\boldsymbol{a}_0 - h(0)\mathbf{1}\|_\infty = \max_{i\le d}\left|\mathbb{E}\left\{h\Big(\sqrt{\frac{Q_{ii}}{d}}G\Big) - h(0)\right\}\right| \le \frac{C}{d}\,. \tag{265}$$

On the other hand, recalling the definition of $\boldsymbol{a}_1$ in Eq. (178), we have, always on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\|\boldsymbol{a}_1\|_\infty \le C\frac{1}{d^{3/2}}\max_{i\le d} Q_{ii}^{3/2} \times d \times \max_{i\le n}\frac{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_i\|_\infty^3}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_i\|_2^3} \tag{266}$$

$$\le C\frac{1}{d^{3/2}} \times d \times \Big(\frac{\log d}{d}\Big)^{3/2} \le C\frac{(\log d)^{3/2}}{d^2}\,. \tag{267}$$

Therefore we conclude that $\|\tilde{\boldsymbol{a}}\|_\infty \le C/d$, whence $\|\tilde{\boldsymbol{a}}\|_2 \le C/\sqrt{d}$.

We therefore obtain, again using Lemma A.2,

$$\left|\langle \boldsymbol{a}, \boldsymbol{K}^{-2}\boldsymbol{a}\rangle - h(0)^2\langle \mathbf{1}, \boldsymbol{K}^{-2}\mathbf{1}\rangle - 2h(0)\langle \mathbf{1}, \boldsymbol{K}^{-2}\tilde{\boldsymbol{a}}\rangle\right| = \langle \tilde{\boldsymbol{a}}, \boldsymbol{K}^{-2}\tilde{\boldsymbol{a}}\rangle \le \lambda_{\min}(\boldsymbol{K})^{-2}\|\tilde{\boldsymbol{a}}\|_2^2 \le \frac{C}{d}\,. \tag{268}$$

We are therefore left with the task of controlling the two terms $\langle \mathbf{1}, \boldsymbol{K}^{-2}\mathbf{1}\rangle$ and $\langle \tilde{\boldsymbol{a}}, \boldsymbol{K}^{-2}\mathbf{1}\rangle$. We will assume $h(0) \neq 0$ because otherwise there is nothing to control. Since $h$ is a positive semidefinite kernel, this also implies $h(0) > 0$ and $\alpha \ge h(0) > 0$. By an application of the Sherman-Morrison formula, we get

$$\langle \mathbf{1}, \boldsymbol{K}^{-2}\mathbf{1}\rangle = \langle \mathbf{1}, (\boldsymbol{K}_* + \alpha\mathbf{1}\mathbf{1}^{\mathsf{T}})^{-2}\mathbf{1}\rangle \tag{269}$$

$$= \frac{\langle \mathbf{1}, \boldsymbol{K}_*^{-2}\mathbf{1}\rangle}{(1 + \alpha\langle \mathbf{1}, \boldsymbol{K}_*^{-1}\mathbf{1}\rangle)^2} \tag{270}$$

$$\le \frac{1}{\alpha^2}\frac{\langle \mathbf{1}, \boldsymbol{K}_*^{-2}\mathbf{1}\rangle}{\langle \mathbf{1}, \boldsymbol{K}_*^{-1}\mathbf{1}\rangle^2} \le \frac{C}{\alpha^2}\frac{1}{\|\mathbf{1}\|^2} \le \frac{C}{d}\,, \tag{271}$$

where we used the above remark $C^{-1}\mathbf{I} \preceq \boldsymbol{K}_* \preceq C\mathbf{I}$.

Using again Sherman-Morrison formula,

$$\langle \mathbf{1}, \boldsymbol{K}^{-2}\tilde{\boldsymbol{a}}\rangle = \frac{\langle \tilde{\boldsymbol{a}}, \boldsymbol{K}_*^{-2}\mathbf{1}\rangle}{1+\alpha\langle \mathbf{1}, \boldsymbol{K}_*^{-1}\mathbf{1}\rangle} - \frac{\alpha\langle \mathbf{1}, \boldsymbol{K}_*^{-2}\mathbf{1}\rangle\langle \tilde{\boldsymbol{a}}, \boldsymbol{K}_*^{-1}\mathbf{1}\rangle}{(1+\alpha\langle \mathbf{1}, \boldsymbol{K}_*^{-1}\mathbf{1}\rangle)^2}\,, \tag{272}$$

$$\left|\langle \mathbf{1}, \boldsymbol{K}^{-2}\tilde{\boldsymbol{a}}\rangle\right| \le C\frac{\|\tilde{\boldsymbol{a}}\|_2\|\mathbf{1}\|_2}{\alpha\|\mathbf{1}\|_2^2} + \frac{\|\mathbf{1}\|_2^3\|\tilde{\boldsymbol{a}}\|_2}{\alpha\|\mathbf{1}\|_2^4} \tag{273}$$

$$\le \frac{C}{d}\,. \tag{274}$$

Using the last two displays in Eq. (268) yields the desired claim. □

*Proof of Theorem 4.13: Variance term.* By virtue of Lemmas A.8, A.10, A.11, we have

$$\frac{1}{\sigma_\xi^2}\widehat{\mathrm{VAR}} = \langle \boldsymbol{B}_0, \boldsymbol{K}_1^{-2}\rangle + \mathrm{Err}(n) \tag{275}$$

$$= \langle \boldsymbol{B}_0, (\boldsymbol{K}_0 + \alpha\mathbf{1}\mathbf{1}^\mathsf{T})^{-2}\rangle + \mathrm{Err}(n)\,. \tag{276}$$

Here and below we denote by $\mathrm{Err}(n)$ an error term bounded as $|\mathrm{Err}(n)| \le Cn^{-c_0}$ with very high probability, and we defined

$$\boldsymbol{K}_0 := \beta\frac{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}}{d} + \beta\gamma\mathbf{I}_n\,. \tag{277}$$

By an application of the Sherman-Morrison formula, and recalling that $\beta\gamma > 0$ is bounded away from zero, we get

$$\frac{1}{\sigma_\xi^2}\widehat{\mathrm{VAR}} = \mathsf{tr}(\boldsymbol{B}_0\boldsymbol{K}_0^{-2}) - \frac{2\alpha}{1+\alpha A_1}\,\mathsf{tr}(\boldsymbol{B}_0\boldsymbol{K}_0^{-2}\mathbf{1}\mathbf{1}^\mathsf{T}\boldsymbol{K}_0^{-1}) \tag{278}$$

$$+ \frac{\alpha^2 A_2}{(1+\alpha A_1)^2}\,\mathsf{tr}(\boldsymbol{B}_0\boldsymbol{K}_0^{-1}\mathbf{1}\mathbf{1}^\mathsf{T}\boldsymbol{K}_0^{-1}) + \mathrm{Err}(n)\,, \tag{279}$$

where $A_\ell := \langle \mathbf{1}, \boldsymbol{K}_0^{-\ell}\mathbf{1}\rangle$, $\ell \in \{1,2\}$. By standard bounds on the norm of matrices with i.i.d. rows (and using $\|\boldsymbol{\Sigma}\| \le C$), we have $0 \preceq \boldsymbol{X}\boldsymbol{X}^\mathsf{T}/d \preceq C\,\mathbf{I}$. Therefore $C^{-1}\mathbf{I} \preceq \boldsymbol{K}_0 \preceq C\mathbf{I}$, for a suitable constant $C$, with very high probability. This implies $d/C \le A_\ell \le Cd$ for $\ell \in \{1,2\}$ and some constant $C > 0$. Further $\|\boldsymbol{B}_0\| \le C\|\boldsymbol{X}\|^2/d^2 \le C/d$. Therefore, (since $\alpha > 0$):

$$\left|\frac{1}{\sigma_\xi^2}\widehat{\mathrm{VAR}} - \mathsf{tr}(\boldsymbol{B}_0\boldsymbol{K}_0^{-2})\right| \le \frac{C}{d}\left|\langle \mathbf{1}, \boldsymbol{K}_0^{-1}\boldsymbol{B}_0\boldsymbol{K}_0^{-2}\mathbf{1}\rangle\right| + \frac{C}{d}\,\langle \mathbf{1}, \boldsymbol{K}_0^{-1}\boldsymbol{B}_0\boldsymbol{K}_0^{-1}\mathbf{1}\rangle + \mathrm{Err}(n) \tag{280}$$

$$\le \frac{C}{d} + \mathrm{Err}(n)\,. \tag{281}$$

We are therefore left with the task of evaluating the asymptotics of

$$\mathsf{tr}(\boldsymbol{B}_0\boldsymbol{K}_0^{-2}) = \mathsf{tr}\big(\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T} + \gamma d\mathbf{I}_n)^2\big)\,. \tag{282}$$

However, this is just the variance of ridge regression with respect to the simple features $\boldsymbol{X}$, with ridge regularization proportional to $\gamma$. We apply the results of [HMRT20] to obtain the claim. □

### A.2.4 Proof of Theorem 4.13: Bias term

We recall the decomposition

$$f^*(\boldsymbol{x}) = b_0 + \langle \boldsymbol{\beta}_0, \boldsymbol{x}\rangle + f^*_{\mathrm{NL}}(\boldsymbol{x}) =: f^*_{\mathrm{L}}(\boldsymbol{x}) + f^*_{\mathrm{NL}}(\boldsymbol{x})\,, \tag{283}$$

where $b_0$, $\boldsymbol{\beta}_0$ are defined by the orthogonality conditions $\mathbb{E}\{f^*_{\rm NL}(\boldsymbol{x})\} = \mathbb{E}\{\boldsymbol{x} f^*_{\rm NL}(\boldsymbol{x})\} = 0$. This yields $b_0 = \mathbb{E}\{f^*(\boldsymbol{x})\}$ and $\boldsymbol{\beta}_0 = \boldsymbol{\Sigma}^{-1}\mathbb{E}\{f^*(\boldsymbol{x})\boldsymbol{x}\}$. We denote by $\boldsymbol{f}^* = (f^*(\boldsymbol{x}_1), \dots, f^*(\boldsymbol{x}_n))^{\mathsf{T}}$ the vector of noiseless responses, which we correspondingly decompose as $\boldsymbol{f}^* = \boldsymbol{f}^*_{\rm L} + \boldsymbol{f}^*_{\rm NL}$. Recalling the definition of $\boldsymbol{M}$, $\boldsymbol{v}$ in Eqs. (173), (174), the bias reads

$$\widehat{\rm BIAS}^2 = \langle \boldsymbol{f}^*, \boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{K}^{-1}\boldsymbol{f}^*\rangle - 2\langle \boldsymbol{v}, \boldsymbol{K}^{-1}\boldsymbol{f}^*\rangle + \|f^*\|^2_{L^2}\,. \tag{284}$$

We begin with an elementary lemma on the norm of $\boldsymbol{f}^*$.

**Lemma A.12.** *Assume $\mathbb{E}\{f^*(\boldsymbol{x})^4\} \le C_0$ for a constant $C_0$ (in particular, this is the case if $\mathbb{E}\{|f^*(\boldsymbol{x})|^{4+\eta}\} \le C_0$). Then, there exists a constant $C$ depending uniquely on $C_0$ such that the following hold:*

(a) $|b_0| \le C$, $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_0\|_2 \le C$, $\mathbb{E}\{f^*_{\rm NL}(\boldsymbol{x})^2\} \le C$.

(b) *With probability at least $1 - Cn^{-1/4}$, we have $|\|\boldsymbol{f}^*\|_2^2/n - \|f^*\|^2_{L^2}| \le n^{-3/8}$.*

(c) *With probability at least $1 - Cn^{-1/4}$, we have $|\|\boldsymbol{f}^*_{\rm NL}\|_2^2/n - \|f^*_{\rm NL}\|^2_{L^2}| \le n^{-3/8}$.*

*Proof.* By Jensen's inequality we have $\mathbb{E}\{f^*(\boldsymbol{x})^2\} \le C$. By orthogonality of $f^*_{\rm NL}$ to linear and constant functions, we also have $\mathbb{E}\{f^*(\boldsymbol{x})^2\} = b_0^2 + \mathbb{E}\{\langle\boldsymbol{\beta}_0, \boldsymbol{x}\rangle^2\} + \mathbb{E}\{f^*_{\rm NL}(\boldsymbol{x})^2\} = b_0^2 + \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_0\|_2^2 + \mathbb{E}\{f^*_{\rm NL}(\boldsymbol{x})^2\}$, which proves claim $(a)$.

To prove $(b)$, simply call $Z = \|\boldsymbol{f}^*\|_2^2/n - \|f^*\|^2_{L^2}$, and note that $\mathbb{E}\{Z^2\} = (\mathbb{E}\{f^*(\boldsymbol{x})^4\} - \mathbb{E}\{f^*(\boldsymbol{x})^2\}^2)/n \le C/n$. The claim follows by Chebyshev inequality.

Finally, $(c)$ follows by the same argument as for claim $(b)$, once we bound $\|f^*_{\rm NL}\|_{L^4}$. In order to show this, notice that, by triangle inequality, $\|f^*_{\rm NL}\|_{L^4} \le \|f^*\|_{L^4} + \|f_0\|_{L^4} + \|f_1\|_{L^4}$, where $f_0(\boldsymbol{x}) = b_0$, $f_1(\boldsymbol{x}) = \langle\boldsymbol{\beta}_0, \boldsymbol{x}\rangle$. Since $\boldsymbol{x} = \boldsymbol{\Sigma}\boldsymbol{z}$, with $\boldsymbol{z}$ $C$-sub-Gaussian, $\|f^*_{\rm NL}\|_{L^4} \le \|f^*\|_{L^4} + b_0 + C\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_0\|_2 \le C$. □

**Lemma A.13.** *Under the assumptions of Theorem 4.13, let $\boldsymbol{M}_0$, $\boldsymbol{v}_0$ be defined as in the statement of Lemma A.3. Then, with probability at least $1 - Cn^{-1/4}$, we have*

$$\big|\widehat{\rm BIAS}^2 - \widehat{\rm BIAS}_0^2\big| \le \frac{C\log d}{\sqrt{d}}\,, \tag{285}$$

$$\widehat{\rm BIAS}_0^2 := \langle \boldsymbol{f}^*, \boldsymbol{K}^{-1}\boldsymbol{M}_0\boldsymbol{K}^{-1}\boldsymbol{f}^*\rangle - 2\langle \boldsymbol{v}_0, \boldsymbol{K}^{-1}\boldsymbol{f}^*\rangle + \|f^*\|^2_{L^2} \tag{286}$$

*Proof.* We have

$$\begin{aligned}
\big|\widehat{\rm BIAS}^2 - \widehat{\rm BIAS}_0^2\big| &\le \big|\langle \boldsymbol{f}^*, \boldsymbol{K}^{-1}(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{K}^{-1}\boldsymbol{f}^*\rangle\big| + 2\big|\langle \boldsymbol{v} - \boldsymbol{v}_0, \boldsymbol{K}^{-1}\boldsymbol{f}^*\rangle\big| && (287)\\
&\le \|\boldsymbol{M} - \boldsymbol{M}_0\|_F\|\boldsymbol{K}^{-1}\boldsymbol{f}^*\|_2^2 + 2\|\boldsymbol{v} - \boldsymbol{v}_0\|_2\|\boldsymbol{K}^{-1}\boldsymbol{f}^*\|_2 && (288)\\
&\le \|\boldsymbol{M} - \boldsymbol{M}_0\|_F\|\boldsymbol{K}^{-1}\|^2\|\boldsymbol{f}^*\|_2^2 + 2\|\boldsymbol{v} - \boldsymbol{v}_0\|_2\|\boldsymbol{K}^{-1}\|\|\boldsymbol{f}^*\|_2 && (289)\\
&\le C\frac{\log d}{d^{3/2}} \times n + C\frac{\log d}{d} \times \sqrt{n} \le \frac{C\log d}{\sqrt{d}}\,. && (290)
\end{aligned}$$

Here, in the last line, we used Lemmas A.2, A.3 and the fact that $\|\boldsymbol{f}^*\|^2 \le Cn$ by Lemma A.12. □

In view of the last lemma, it is sufficient to work with $\widehat{\rm BIAS}_0^2$. We decompose it as

$$\widehat{\rm BIAS}_0^2 = \widehat{\rm BIAS}_{\rm L}^2 + \widehat{\rm BIAS}_{\rm NL}^2 + \widehat{\rm BIAS}_{\rm mix}^2 + \|f^*_{\rm NL}\|^2_{L^2}\,, \tag{291}$$

$$\widehat{\rm BIAS}_{\rm L}^2 := \langle \boldsymbol{f}^*_{\rm L}, \boldsymbol{K}^{-1}\boldsymbol{M}_0\boldsymbol{K}^{-1}\boldsymbol{f}^*_{\rm L}\rangle - 2\langle \boldsymbol{v}_0, \boldsymbol{K}^{-1}\boldsymbol{f}^*_{\rm L}\rangle + \|f^*_{\rm L}\|^2_{L^2}\,, \tag{292}$$

$$\widehat{\rm BIAS}_{\rm NL}^2 := \langle \boldsymbol{f}^*_{\rm NL}, \boldsymbol{K}^{-1}\boldsymbol{M}_0\boldsymbol{K}^{-1}\boldsymbol{f}^*_{\rm NL}\rangle\,, \tag{293}$$

$$\widehat{\rm BIAS}_{\rm mix}^2 := 2\langle \boldsymbol{f}^*_{\rm L}, \boldsymbol{K}^{-1}\boldsymbol{M}_0\boldsymbol{K}^{-1}\boldsymbol{f}^*_{\rm NL}\rangle - 2\langle \boldsymbol{v}_0, \boldsymbol{K}^{-1}\boldsymbol{f}^*_{\rm NL}\rangle\,. \tag{294}$$

We next show that the contribution of the constant term in $f^*_{\rm L}(\boldsymbol{x})$ and $\boldsymbol{M}_0$ is negligible.

**Lemma A.14.** *Under the assumptions of Theorem 4.13, let $\boldsymbol{M}_0$, $\boldsymbol{B}$, $\boldsymbol{v}_0$ be defined as in the statement of Lemma A.3. Further define*

$$R_{\rm L} := \langle \boldsymbol{X}\boldsymbol{\beta}_0, \boldsymbol{K}^{-1}\boldsymbol{B}\boldsymbol{K}^{-1}\boldsymbol{X}\boldsymbol{\beta}_0\rangle - \frac{2h'(0)}{d}\langle \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{\beta}_0, \boldsymbol{K}^{-1}\boldsymbol{X}\boldsymbol{\beta}_0\rangle + \langle \boldsymbol{\beta}_0, \boldsymbol{\Sigma}\boldsymbol{\beta}_0\rangle\,, \tag{295}$$

$$R_{\rm NL} := \langle \boldsymbol{f}^*_{\rm NL}, \boldsymbol{K}^{-1}\boldsymbol{B}\boldsymbol{K}^{-1}\boldsymbol{f}^*_{\rm NL}\rangle\,, \tag{296}$$

$$R_{\rm mix} := 2\langle \boldsymbol{X}\boldsymbol{\beta}_0, \boldsymbol{K}^{-1}\boldsymbol{B}\boldsymbol{K}^{-1}\boldsymbol{f}^*_{\rm NL}\rangle - \frac{2h'(0)}{d}\langle \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{\beta}_0, \boldsymbol{K}^{-1}\boldsymbol{f}^*_{\rm NL}\rangle\,. \tag{297}$$

*Then, with very high probability we have*

$$\big|\widehat{\rm BIAS}^2_{\rm L} - R_{\rm L}\big| \le \frac{C}{n}\,, \tag{298}$$

$$\big|\widehat{\rm BIAS}^2_{\rm NL} - R_{\rm NL}\big| \le \frac{C}{n}\,, \tag{299}$$

$$\big|\widehat{\rm BIAS}^2_{\rm mix} - R_{\rm mix}\big| \le \frac{C}{n}\,. \tag{300}$$

*Proof.* The proof of this lemma is very similar to the one of Lemma A.11, and we omit it. □

**Lemma A.15.** *Under the assumptions of Theorem 4.13, let $\mathscr{B}(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0)$ be defined as in Eq. (168), and $R_{\rm L}$ be defined as in the statement of Lemma A.14. Let $a \in (0, 1/2)$. Then we have, with very high probability*

$$\big|R_{\rm L} - \mathscr{B}(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0)\big| \le C\, n^{-a}\,. \tag{301}$$

*Proof.* Recall the definition of $\boldsymbol{K}_1$ in Eq. (163). and define $\tilde{R}_{\rm L}$ as $R_{\rm L}$ (cf. Eq. (295)) except with $\boldsymbol{B}$ replaced by $\boldsymbol{B}_0$ defined in Eq. (256), and $\boldsymbol{K}$ replaced by $\boldsymbol{K}_1$ defined in Eq. (163). Namely:

$$\tilde{R}_{\rm L} := \langle \boldsymbol{X}\boldsymbol{\beta}_0, \boldsymbol{K}_1^{-1}\boldsymbol{B}_0\boldsymbol{K}_1^{-1}\boldsymbol{X}\boldsymbol{\beta}_0\rangle - \frac{2h'(0)}{d}\langle \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{\beta}_0, \boldsymbol{K}_1^{-1}\boldsymbol{X}\boldsymbol{\beta}_0\rangle + \langle \boldsymbol{\beta}_0, \boldsymbol{\Sigma}\boldsymbol{\beta}_0\rangle\,. \tag{302}$$

Letting $\boldsymbol{u} = \boldsymbol{X}\boldsymbol{\beta}_0 = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_0$, note that $\|\boldsymbol{u}\|_2 \le \|\boldsymbol{Z}\|\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_0\|_2 \le C\sqrt{n}$ with very high probability (using Lemma A.12). We then have

$$\begin{aligned}
\big|R_{\rm L} - \tilde{R}_{\rm L}\big| &\le \big|\langle \boldsymbol{u}, \boldsymbol{K}^{-1}\boldsymbol{B}\boldsymbol{K}^{-1}\boldsymbol{u}\rangle - \langle \boldsymbol{u}, \boldsymbol{K}^{-1}\boldsymbol{B}_0\boldsymbol{K}^{-1}\boldsymbol{u}\rangle\big| \\
&\quad + \big|\langle \boldsymbol{u}, \boldsymbol{K}^{-1}\boldsymbol{B}_0\boldsymbol{K}^{-1}\boldsymbol{u}\rangle - \langle \boldsymbol{u}, \boldsymbol{K}_1^{-1}\boldsymbol{B}_0\boldsymbol{K}_1^{-1}\boldsymbol{u}\rangle\big| + \frac{C}{d}\big|\langle \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{\beta}_0, \boldsymbol{K}^{-1}\boldsymbol{u}\rangle - \langle \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{\beta}_0, \boldsymbol{K}_1^{-1}\boldsymbol{u}\rangle\big| \\
&=: E_1 + E_2 + E_3\,.
\end{aligned}$$

We bound each of the three terms with very high probability:

$$E_1 \le \|\boldsymbol{B} - \boldsymbol{B}_0\| \cdot \|\boldsymbol{K}^{-1}\|^2 \cdot \|\boldsymbol{u}\|_2^2 \le \frac{C}{d^{3/2}} \times C \times Cn \le \frac{C}{n^{1/2}}\,, \tag{303}$$

$$\begin{aligned}
E_2 &\le \big(\|\boldsymbol{B}_0\boldsymbol{K}^{-1}\boldsymbol{u}\|_2 + \|\boldsymbol{B}_0\boldsymbol{K}_1^{-1}\boldsymbol{u}\|_2\big)\|\boldsymbol{u}\|_2\|\boldsymbol{K}^{-1} - \boldsymbol{K}_1^{-1}\| \\
&\le \|\boldsymbol{B}_0\|\big(\|\boldsymbol{K}^{-1}\| + \|\boldsymbol{K}_1^{-1}\|\big)\|\boldsymbol{u}\|_2^2\|\boldsymbol{K}^{-1} - \boldsymbol{K}_1^{-1}\| \qquad (304) \\
&\le \frac{C}{d} \times C \times Cn \times n^{-c_0} \le C\, n^{-c_0}\,,
\end{aligned}$$

$$\begin{aligned}
E_3 &\le \frac{C}{d}\|\boldsymbol{X}\|\|\boldsymbol{\Sigma}\boldsymbol{\beta}_0\|_2\|\boldsymbol{u}\|_2\|\boldsymbol{K}^{-1} - \boldsymbol{K}_1^{-1}\| \qquad (305) \\
&\le \frac{C}{d} \times C\sqrt{n} \times C \times C\sqrt{n} \times Cn^{-c_0} \le Cn^{-c_0}\,.
\end{aligned}$$

Here in Eq. (303) we used Lemma A.2 and Lemma A.9; in Eq. (304) Lemma A.2 and the fact that $\|\boldsymbol{B}_0\| \le C/d$; in Eq. (305), Lemma A.2 and $\|\boldsymbol{X}\| \le C\sqrt{d}$. Hence we conclude that

$$\big|R_{\rm L} - \tilde{R}_{\rm L}\big| \le Cn^{-c_0}\,. \tag{306}$$

Finally define $\tilde{\tilde{R}}_{\text{L}}$ as $\tilde{R}_{\text{L}}$, with $\boldsymbol{K}_1$ replaced by $\boldsymbol{K}_0 = \beta\frac{\boldsymbol{X}\boldsymbol{X}^{\text{T}}}{d} + \beta\gamma\mathbf{I}_n$.

$$\begin{aligned}\big|\tilde{R}_{\text{L}} - \tilde{\tilde{R}}_{\text{L}}\big| &\le \big|\langle \boldsymbol{u}, (\boldsymbol{K}_1^{-1} + \boldsymbol{K}_0^{-1})\boldsymbol{B}_0(\boldsymbol{K}_1^{-1} - \boldsymbol{K}_0^{-1})\boldsymbol{u}\rangle\big| + \frac{C}{d}\big|\langle \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{\beta}_0, (\boldsymbol{K}_1^{-1} - \boldsymbol{K}_0^{-1})\boldsymbol{u}\rangle\big| \\ &=: G_1 + G_2\,.\end{aligned}$$

By the Sherman-Morrison formula, for any two vectors $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^n$, we have

$$\big|\langle \boldsymbol{w}_1, (\boldsymbol{K}_1^{-1} - \boldsymbol{K}_0^{-1})\boldsymbol{w}_2\rangle\big| = \alpha\frac{\big|\langle \mathbf{1}, \boldsymbol{K}_0^{-1}\boldsymbol{w}_1\rangle\langle \mathbf{1}, \boldsymbol{K}_0^{-1}\boldsymbol{w}_2\rangle\big|}{1 + \alpha\langle \mathbf{1}, \boldsymbol{K}_0^{-1}\mathbf{1}\rangle} \tag{307}$$

$$\le \frac{C}{d}\big|\langle \mathbf{1}, \boldsymbol{K}_0^{-1}\boldsymbol{w}_1\rangle\big| \cdot \big|\langle \mathbf{1}, \boldsymbol{K}_0^{-1}\boldsymbol{w}_2\rangle\big|\,. \tag{308}$$

Further notice that

$$|\langle \boldsymbol{u}, \boldsymbol{K}_0^{-1}, \mathbf{1}\rangle| = |\langle \boldsymbol{\beta}_0, \boldsymbol{X}^{\text{T}}(\beta\boldsymbol{X}\boldsymbol{X}^{\text{T}}/d + \beta\gamma\mathbf{I}_n)^{-1}\mathbf{1}\rangle| \le C\sqrt{d\log d}\,,$$

where the last inequality holds with very high probability by [KY17, Theorem 3.16] (cf. also Lemma 4.4 in the same paper). We therefore have

$$G_1 \le \frac{C}{d}\big|\langle \boldsymbol{u}, (\boldsymbol{K}_1^{-1} + \boldsymbol{K}_0^{-1})\boldsymbol{B}_0\boldsymbol{K}_0^{-1}\mathbf{1}\rangle\big| \cdot \big|\langle \boldsymbol{u}, \boldsymbol{K}_0^{-1}\mathbf{1}\rangle\big| \tag{309}$$

$$\le \frac{C}{d}\|\boldsymbol{B}_0\|\|\boldsymbol{u}\|_2\|\mathbf{1}\|_2\big|\langle \boldsymbol{u}, \boldsymbol{K}_0^{-1}\mathbf{1}\rangle\big| \tag{310}$$

$$\le \frac{C}{d} \times \frac{1}{d} \times \sqrt{d} \times \sqrt{d} \times \sqrt{d\log d} \le C\sqrt{\frac{\log d}{d}}\,. \tag{311}$$

Analogously

$$G_2 \le \frac{C}{d^2}\big|\langle \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{\beta}_0, \boldsymbol{K}_0^{-1}\mathbf{1}\rangle\big| \cdot \big|\langle \boldsymbol{u}, \boldsymbol{K}_0^{-1}\mathbf{1}\rangle\big| \tag{312}$$

$$\le \frac{C}{d^2}\|\boldsymbol{X}\|\|\boldsymbol{\Sigma}\boldsymbol{\beta}_0\|_2\|\boldsymbol{K}_0^{-1}\|\|\mathbf{1}\|_2\big|\langle \boldsymbol{u}, \boldsymbol{K}_0^{-1}\mathbf{1}\rangle\big| \tag{313}$$

$$\le \frac{C}{d^2} \times C\sqrt{d} \times C \times C \times C\sqrt{n} \times C\sqrt{d\log d} \le C\sqrt{\frac{\log d}{d}}\,. \tag{314}$$

Summarizing

$$\big|\tilde{R}_{\text{L}} - \tilde{\tilde{R}}_{\text{L}}\big| \le C\sqrt{\frac{\log d}{d}}\,. \tag{315}$$

We are left with the task of estimating $\tilde{\tilde{R}}_{\text{L}}$ which we rewrite explicitly as

$$\tilde{\tilde{R}}_{\text{L}} = \gamma^2\big\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{X}\boldsymbol{X}^{\text{T}} + \gamma\mathbf{I}_n)^{-1}\boldsymbol{\beta}_0\big\|_2^2\,. \tag{316}$$

We recognize in this the bias of ridge regression with respect to the linear features $\boldsymbol{x}_i$, when the responses are also linear $\langle \boldsymbol{\beta}_0, \boldsymbol{x}_i\rangle$. Using the results of [HMRT20], we obtain that, for any $a \in (0, 1/2)$, the following holds with very high probability.

$$\big|\tilde{\tilde{R}}_{\text{L}} - \mathscr{B}(\boldsymbol{\Sigma}, \boldsymbol{\beta}_0)\big| \le C\,n^{-c_0}\,. \tag{317}$$

The proof is completed by using Eqs. (306), (315), (317). □

We next consider the nonlinear term $R_{\text{NL}}$, cf. Eq. (296).

**Lemma A.16.** *Under the assumptions of Theorem 4.13, let $\mathscr{V}(\boldsymbol{\Sigma})$ be defined as in Eq. (167), and $R_{\rm NL}$ be defined as in the statement of Lemma A.14. Then there exists $c_0 > 0$ such that, with probability at least $1 - Cn^{-1/4}$,*

$$\big|R_{\rm NL} - \mathscr{V}(\boldsymbol{\Sigma})\|\mathsf{P}_{>1} f^*\|_{L^2}^2\big| \le C\, n^{-c_0}\,. \tag{318}$$

*Proof.* Define

$$\begin{aligned}
\widetilde{\widetilde{R}}_{\rm NL} &:= \langle \boldsymbol{f}^*_{\rm NL}, \boldsymbol{K}_0^{-1}\boldsymbol{B}_0\boldsymbol{K}_0^{-1}\boldsymbol{f}^*_{\rm NL}\rangle && (319)\\
&= \frac{1}{d^2}\langle \boldsymbol{f}^*_{\rm NL}, (\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}/d + \gamma \mathbf{I}_n)^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}/d + \gamma \mathbf{I}_n)^{-1}\boldsymbol{f}^*_{\rm NL}\rangle && (320)\\
&= \frac{1}{d^2}\langle \boldsymbol{f}^*_{\rm NL}, (\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^{\mathsf{T}}/d + \gamma \mathbf{I}_n)^{-1}\boldsymbol{Z}\boldsymbol{\Sigma}^2\boldsymbol{Z}^{\mathsf{T}}(\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^{\mathsf{T}}/d + \gamma \mathbf{I}_n)^{-1}\boldsymbol{f}^*_{\rm NL}\rangle\,. && (321)
\end{aligned}$$

By the same argument as in the proof of Lemma A.15, we have, with very high probability,

$$\big|R_{\rm NL} - \widetilde{\widetilde{R}}_{\rm NL}\big| \le C\sqrt{\frac{\log d}{d}}\,. \tag{322}$$

We next use the following identity, which holds for any two symmetric matrices $\boldsymbol{A}$, $\boldsymbol{M}$, and any $t \neq 0$,

$$\boldsymbol{A}^{-1}\boldsymbol{M}\boldsymbol{A}^{-1} = \frac{1}{t}\big[\boldsymbol{A}^{-1} - (\boldsymbol{A}+t\boldsymbol{M})^{-1}\big] + t\boldsymbol{A}^{-1}\boldsymbol{M}\boldsymbol{A}^{-1}\boldsymbol{M}(\boldsymbol{A}+t\boldsymbol{M})^{-1}\,. \tag{323}$$

Therefore, for any matrix $\boldsymbol{U}$ and any $t > 0$, we have

$$\big|\langle \boldsymbol{A}^{-1}\boldsymbol{M}\boldsymbol{A}^{-1}, \boldsymbol{U}\rangle\big| \le \frac{1}{t}\big|\langle \boldsymbol{A}^{-1}, \boldsymbol{U}\rangle\big| + \frac{1}{t}\big|\langle (\boldsymbol{A}+t\boldsymbol{M})^{-1}, \boldsymbol{U}\rangle\big| + t\|\boldsymbol{A}^{-1}\|^2\|\boldsymbol{M}\|^2\|(\boldsymbol{A}+t\boldsymbol{M})^{-1}\|\,\|\boldsymbol{U}\|_*\,. \tag{324}$$

We apply this inequality to $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^{\mathsf{T}}/d + \gamma\mathbf{I}_n$, $\boldsymbol{M} = \boldsymbol{Z}\boldsymbol{\Sigma}^2\boldsymbol{Z}^{\mathsf{T}}/d$ and $U_{ij} = f^*_{\rm NL}(\boldsymbol{x}_i)f^*_{\rm NL}(\boldsymbol{x}_i)\mathbf{1}_{i\neq j}$. Note that $\|\boldsymbol{A}^{-1}\|, \|\boldsymbol{M}\|, \|(\boldsymbol{A}+t\boldsymbol{M})^{-1}\| \le C$. Further $\|\boldsymbol{U}\|_* \le 2\|\boldsymbol{f}^*_{\rm NL}\|_2^2 \le Cn$ with probability at least $1 - Cn^{-1/4}$ by Lemma A.12. Finally for any $t \in (0,1)$, by Theorem A.7, the following hold with probability at least $1 - Cd^{-1/4}$:

$$\frac{1}{d}\big|\langle \boldsymbol{A}^{-1}, \boldsymbol{U}\rangle\big| \le C\, d^{-1/8}\,, \qquad \frac{1}{d}\big|\langle (\boldsymbol{A}+t\boldsymbol{M})^{-1}, \boldsymbol{U}\rangle\big| \le C\, d^{-1/8}\,. \tag{325}$$

Therefore, applying Eq. (324) we obtain

$$\frac{1}{d}\big|\langle \boldsymbol{A}^{-1}\boldsymbol{M}\boldsymbol{A}^{-1}, \boldsymbol{U}\rangle\big| \le \frac{1}{t}C\, d^{-1/8} + Ct \le Cd^{-1/16}\,, \tag{326}$$

where in the last step we selected $t = d^{-1/16}$. Recalling the definitions of $\boldsymbol{A}, \boldsymbol{M}, \boldsymbol{U}$, we have proved:

$$\left|\widetilde{\widetilde{R}}_{\rm NL} - \frac{1}{d^2}\sum_{i=1}^n [\boldsymbol{A}^{-1}\boldsymbol{M}\boldsymbol{A}^{-1}]_{ii} f^*_{\rm NL}(\boldsymbol{x}_i)^2\right| \le Cd^{-1/16}\,. \tag{327}$$

We are therefore left with the task of controlling the diagonal terms. Using the results of [KY17], we get

$$\max_{i\le n}\left|[\boldsymbol{A}^{-1}\boldsymbol{M}\boldsymbol{A}^{-1}]_{ii} - \frac{1}{n}\mathsf{tr}(\boldsymbol{A}^{-1}\boldsymbol{M}\boldsymbol{A}^{-1})\right| \le Cn^{-1/8}\,. \tag{328}$$

Further $|\|\boldsymbol{f}^*_{\rm NL}\|_2^2/n - \|f^*_{\rm NL}\|_{L^2}^2| \le Cn^{-1/2}$ with probability at least $1 - Cn^{-1/4}$ by Lemma A.12. Therefore, with probability at least $1 - Cd^{-1/4}$,

$$\big|\widetilde{\widetilde{R}}_{\rm NL} - V_{\rm RR}\|f^*_{\rm NL}\|_{L^2}^2\big| \le Cd^{-1/16}\,, \tag{329}$$

$$V_{\rm RR} := \frac{1}{d^2}\big\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}/d + \gamma\mathbf{I}_n)^{-1}\big\|_F^2\,. \tag{330}$$

We finally recognize that the term $V_{\rm RR}$ is just the variance of ridge regression with respect to the linear features $\boldsymbol{x}_i$, and using [HMRT20], we obtain

$$\left|\widetilde{\widetilde{R}}_{\rm NL} - \mathscr{V}(\boldsymbol{\Sigma})\|f^*_{\rm NL}\|^2_{L^2}\right| \le C d^{-1/16}\,. \tag{331}$$

The proof of the lemma is concluded by using the last equation together with Eq. (322). □

**Lemma A.17.** *Under the assumptions of Theorem 4.13, $R_{\rm mix}$ be defined as in the statement of Lemma A.14. Then we have, with probability at least $1 - Cd^{-1/4}$,*

$$\left|R_{\rm mix}\right| \le C\, n^{-1/16}\,. \tag{332}$$

*Proof.* The proof of this lemma is analogous to the one of Lemma A.16 and we omit it. □

We are now in a position to prove Theorem 4.13.

*Proof of Theorem 4.13: Bias term.* Using Lemma A.13, Eq. (291) and Lemma A.14, we obtain that, with very high probability,

$$\left|\widehat{\rm BIAS}^2 - (R_{\rm L} + R_{\rm NL} + R_{\rm mix} + \|f^*_{\rm NL}\|^2_{L^2})\right| \le C\sqrt{\frac{\log n}{n}}\,. \tag{333}$$

Hence the proof is completed by using Lemmas A.15, A.16, A.17. □

### A.2.5 Consequences: Proof of Corollary 4.14

We denote by $\lambda_1 \ge \cdots \ge \lambda_d$ the eigenvalues of $\boldsymbol{\Sigma}$ in decreasing order.

First note that the left hand side of Eq. (166) is strictly increasing in $\lambda_*$, while the right hand side is strictly decreasing. By considering the limits as $\lambda_* \to 0$ and $\lambda_* \to \infty$, it is easy to see that this equation admits indeed a unique solution.

Next denoting by $F(x) := \mathsf{tr}\Big(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + x\mathbf{I})^{-1}\Big)$ the function appearing on the right hand side of Eq. (166), we have, for $x \ge c_*\lambda_{k+1}$,

$$F(x) = \sum_{i=1}^{d} \frac{\lambda_i}{x+\lambda_i} \ge \sum_{i=k+1}^{d} \frac{\lambda_i}{x+\lambda_i} \tag{334}$$

$$\ge \frac{c_*}{(1+c_*)x} \sum_{i=k+1}^{d} \lambda_i =: \underline{F}(x)\,. \tag{335}$$

Let $\underline{\lambda}_*$ be the unique non-negative solution of $n(1-(\gamma/\underline{\lambda}_*)) = \underline{F}(\underline{\lambda}_*)$. Then, the above inequality implies that whenever $\underline{\lambda}_* \ge c_*\lambda_{k+1}$ we have $\lambda_* \ge \underline{\lambda}_*$. Solving explicitly for $\underline{\lambda}_*$, we get

$$\frac{(1+c_*)\gamma}{c_*\lambda_{k+1}} + \frac{r_k(\boldsymbol{\Sigma})}{n} \ge (1+c_*) \quad \Rightarrow \quad \lambda_* \ge \gamma + \frac{c_*}{1+c_*}\frac{1}{n}\sum_{i=k+1}^{d}\lambda_i\,. \tag{336}$$

Next, we upper bound

$$\mathsf{tr}\big(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*\mathbf{I})^{-2}\big) = \sum_{i=1}^{d}\frac{\lambda_i^2}{(\lambda_i+\lambda_*)^2} \tag{337}$$

$$\le k + \frac{1}{\lambda_*^2}\sum_{i=k+1}^{d}\lambda_i^2 \tag{338}$$

$$\le k + (1+c_*^{-1})^2 n^2 \frac{\sum_{i=k+1}^{d}\lambda_i^2}{(n\gamma/c_* + \sum_{i=k+1}^{d}\lambda_i)^2}\,. \tag{339}$$

If we assume that the right-hand side is less than $1/2$, using Theorem 4.13, we obtain that, with high probability,

$$\frac{1}{\sigma_\xi^2}\widehat{\mathrm{VAR}} \le k + (1+c_*^{-1})^2 n^2 \frac{\sum_{i=k+1}^d \lambda_i^2}{(n\gamma/c_* + \sum_{i=k+1}^d \lambda_i)^2} + n^{-c_0}\,. \tag{340}$$

Next, considering again Eq. (166) and upper bounding the right-hand side, we get

$$n\Big(1-\frac{\gamma}{\lambda_*}\Big) \le k + \frac{1}{\lambda_*}\sum_{i=k+1}^d \lambda_i\,. \tag{341}$$

Hence, using the assumption that the right hand side of Eq. (339) is upper bounded by $1/2$, which implies $k \le n/2$, we get

$$\lambda_* \le 2\gamma + \frac{2}{n}\sum_{i=k+1}^d \lambda_i\,. \tag{342}$$

Next consider the formula for the bias term, Eq. (168). Denoting by $(\beta_{0,i})_{i\le p}$ the coordinates of $\boldsymbol{\beta}_0$ in the basis of the eigenvectors of $\boldsymbol{\Sigma}$, we get

$$\lambda_*^2\langle \boldsymbol{\beta}_0, (\boldsymbol{\Sigma}+\lambda_*\mathbf{I})^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}_0\rangle = \sum_{i=1}^d \frac{\lambda_*^2\lambda_i\beta_{0,i}^2}{(\lambda_i+\lambda_*)^2} \tag{343}$$

$$\le \lambda_*^2\sum_{i=1}^k \lambda_i^{-1}\beta_{0,i}^2 + \sum_{i=1}^d \lambda_i\beta_{0,i}^2 \tag{344}$$

$$\le 4\Big(\gamma + \frac{1}{n}\sum_{i=k+1}^d \lambda_i\Big)^2 \|\boldsymbol{\beta}_{0,\le k}\|_{\boldsymbol{\Sigma}^{-1}}^2 + \|\boldsymbol{\beta}_{0,>k}\|_{\boldsymbol{\Sigma}}^2\,. \tag{345}$$

Together with Theorem 4.13, this implies the desired bound on the bias.

# B Optimization in the linear regime

**Theorem 5.1.** *Assume*

$$\mathrm{Lip}(\boldsymbol{D}f_n)\,\|\boldsymbol{y}-f_n(\boldsymbol{\theta}_0)\|_2 < \frac{1}{4}\sigma_{\min}^2(\boldsymbol{D}f_n(\boldsymbol{\theta}_0))\,. \tag{346}$$

*Further define*

$$\sigma_{\max} := \sigma_{\max}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)), \sigma_{\min} := \sigma_{\min}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)).$$

*Then the following hold for all $t>0$:*

1. *The empirical risk decreases exponentially fast to 0, with rate $\lambda_0 = \sigma_{\min}^2/(2n)$:*

$$\widehat{L}(\boldsymbol{\theta}_t) \le \widehat{L}(\boldsymbol{\theta}_0)\,e^{-\lambda_0 t}\,. \tag{347}$$

2. *The parameters stay close to the initialization and are closely tracked by those of the linearized flow. Specifically, letting $L_n := \mathrm{Lip}(\boldsymbol{D}f_n)$,*

$$\|\boldsymbol{\theta}_t-\boldsymbol{\theta}_0\|_2 \le \frac{2}{\sigma_{\min}}\|\boldsymbol{y}-f_n(\boldsymbol{\theta}_0)\|_2\,, \tag{348}$$

$$\|\boldsymbol{\theta}_t-\overline{\boldsymbol{\theta}}_t\|_2 \le \Big\{\frac{32\sigma_{\max}}{\sigma_{\min}^2}\|\boldsymbol{y}-f_n(\boldsymbol{\theta}_0)\|_2 + \frac{16L_n}{\sigma_{\min}^3}\|\boldsymbol{y}-f_n(\boldsymbol{\theta}_0)\|_2^2\Big\} \wedge \frac{180L_n\sigma_{\max}^2}{\sigma_{\min}^5}\|\boldsymbol{y}-f_n(\boldsymbol{\theta}_0)\|_2^2\,. \tag{349}$$

3. *The models constructed by gradient flow and by the linearized flow are similar on test data. Specifically, writing* $f^{\rm lin}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_0) + \boldsymbol{D}f(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$*, we have*

$$\begin{aligned}&\|f(\boldsymbol{\theta}_t) - f^{\rm lin}(\overline{\boldsymbol{\theta}}_t)\|_{L^2(\mathbb{P})}\\ &\le \Big\{4\,\mathrm{Lip}(\boldsymbol{D}f)\frac{1}{\sigma_{\min}^2} + 180\|\boldsymbol{D}f(\boldsymbol{\theta}_0)\|\frac{L_n\sigma_{\max}^2}{\sigma_{\min}^5}\Big\}\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0)\|_2^2\,.\end{aligned} \tag{350}$$

*Proof.* Throughout the proof we let $L_n := \mathrm{Lip}(\boldsymbol{D}f_n)$, and we use $\dot{\boldsymbol{a}}_t$ to denote the derivative of quantity $\boldsymbol{a}_t$ with respect to time.

Let $\boldsymbol{y}_t = f_n(\boldsymbol{\theta}_t)$. By the gradient flow equation,

$$\dot{\boldsymbol{y}}_t = \boldsymbol{D}f_n(\boldsymbol{\theta}_t)\,\dot{\boldsymbol{\theta}}_t = -\frac{1}{n}\boldsymbol{D}f_n(\boldsymbol{\theta}_t)\boldsymbol{D}f_n(\boldsymbol{\theta}_t)^{\mathsf{T}}(\boldsymbol{y}_t - \boldsymbol{y})\,. \tag{351}$$

Defining the empirical kernel at time $t$, $\boldsymbol{K}_t := \boldsymbol{D}f_n(\boldsymbol{\theta}_t)\boldsymbol{D}f_n(\boldsymbol{\theta}_t)^{\mathsf{T}}$, we thus have

$$\dot{\boldsymbol{y}}_t = -\frac{1}{n}\boldsymbol{K}_t(\boldsymbol{y}_t - \boldsymbol{y})\,, \tag{352}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{y}_t - \boldsymbol{y}\|_2^2 = -\frac{2}{n}\langle \boldsymbol{y}_t - \boldsymbol{y}, \boldsymbol{K}_t(\boldsymbol{y}_t - \boldsymbol{y})\rangle\,. \tag{353}$$

Letting $r_* := \sigma_{\min}/(2L_n)$ and $t_* := \inf\{t:\ \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 > r_*\}$, we have $\lambda_{\min}(\boldsymbol{K}_t) \ge (\sigma_{\min}/2)^2$ for all $t \le t_*$, whence

$$t \le t_* \ \ \Rightarrow\ \ \|\boldsymbol{y}_t - \boldsymbol{y}\|_2^2 \le \|\boldsymbol{y}_0 - \boldsymbol{y}\|_2^2 e^{-\lambda_0 t}\,, \tag{354}$$

with $\lambda_0 = \sigma_{\min}^2/(2n)$.

Note that, for any $t \le t_*$, $\sigma_{\min}(\boldsymbol{D}f_n(\boldsymbol{\theta}_t)) \ge \sigma_{\min}/2$. Therefore, by the gradient flow equations, for any $t \le t_*$,

$$\|\dot{\boldsymbol{\theta}}_t\|_2 = \frac{1}{n}\big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_t)^{\mathsf{T}}(\boldsymbol{y}_t - \boldsymbol{y})\big\|_2\,, \tag{355}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{y}_t - \boldsymbol{y}\|_2 = -\frac{1}{n}\cdot\frac{\|\boldsymbol{D}f_n(\boldsymbol{\theta}_t)^{\mathsf{T}}(\boldsymbol{y}_t - \boldsymbol{y})\|_2^2}{\|\boldsymbol{y}_y - \boldsymbol{y}\|_2} \tag{356}$$

$$\le -\frac{\sigma_{\min}}{2n}\|\boldsymbol{D}f_n(\boldsymbol{\theta}_t)^{\mathsf{T}}(\boldsymbol{y}_t - \boldsymbol{y})\big\|_2\,. \tag{357}$$

Therefore, by Cauchy-Schwartz,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\|\boldsymbol{y}_t - \boldsymbol{y}\|_2 + \frac{\sigma_{\min}}{2}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2\Big) \le \frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{y}_t - \boldsymbol{y}\|_2 + \frac{\sigma_{\min}}{2}\|\dot{\boldsymbol{\theta}}_t\|_2 \le 0\,. \tag{358}$$

This implies, for all $t \le t_*$,

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \le \frac{2}{\sigma_{\min}}\|\boldsymbol{y} - \boldsymbol{y}_0\|_2\,. \tag{359}$$

Assume by contradiction $t_* < \infty$. The last equation together with the assumption (346) implies $\|\boldsymbol{\theta}_{t_*} - \boldsymbol{\theta}_0\|_2 < r_*$, which contradicts the definition of $t_*$. We conclude that $t_* = \infty$, and Eq. (347) follows from Eq. (354).

Equation (348) follows from Eq. (359).

In order to prove Eq. (349), let $\overline{\boldsymbol{y}}_t := f_n(\boldsymbol{\theta}_0) + \boldsymbol{D}f_n(\boldsymbol{\theta}_0)(\overline{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0)$. Note that this satisfies an equation similar to (352), namely

$$\dot{\overline{\boldsymbol{y}}}_t = -\frac{1}{n}\boldsymbol{K}_0(\overline{\boldsymbol{y}}_t - \boldsymbol{y})\,. \tag{360}$$

Define the difference $\boldsymbol{r}_t := \boldsymbol{y}_t - \overline{\boldsymbol{y}}_t$. We then have $\dot{\boldsymbol{r}}_t = -(\boldsymbol{K}_t/n)\boldsymbol{r}_t - ((\boldsymbol{K}_t - \boldsymbol{K}_0)/n)(\overline{\boldsymbol{y}}_t - \boldsymbol{y})$, whence

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{r}_t\|_2^2 = -\frac{2}{n}\langle \boldsymbol{r}_t, \boldsymbol{K}_t\boldsymbol{r}_t\rangle - \frac{2}{n}\langle \boldsymbol{r}_t, (\boldsymbol{K}_t - \boldsymbol{K}_0)(\overline{\boldsymbol{y}}_t - \boldsymbol{y})\rangle \tag{361}$$

$$\leq -\frac{2}{n}\lambda_{\min}(\boldsymbol{K}_t)\|\boldsymbol{r}_t\|_2^2 + \frac{2}{n}\|\boldsymbol{r}_t\|_2\big\|\boldsymbol{K}_t - \boldsymbol{K}_0\big\|\,\|\overline{\boldsymbol{y}}_t - \boldsymbol{y}\|_2\,. \tag{362}$$

Using $2\lambda_{\min}(\boldsymbol{K}_t)/n \geq \lambda_0$ and $\|\overline{\boldsymbol{y}}_t - \boldsymbol{y}_t\|_2 \leq \|\boldsymbol{y}_0 - \boldsymbol{y}\|_2 e^{-\lambda_0 t/2}$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{r}_t\|_2 = -\frac{\lambda_0}{2}\|\boldsymbol{r}_t\|_2 + \frac{1}{n}\big\|\boldsymbol{K}_t - \boldsymbol{K}_0\big\|\,\|\boldsymbol{y}_0 - \boldsymbol{y}\|_2\, e^{-\lambda_0 t/2}\,. \tag{363}$$

Note that

$$\big\|\boldsymbol{K}_t - \boldsymbol{K}_0\big\| = \big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_t)\boldsymbol{D}f_n(\boldsymbol{\theta}_t)^{\mathsf{T}} - \boldsymbol{D}f_n(\boldsymbol{\theta}_0)\boldsymbol{D}f_n(\boldsymbol{\theta}_0)^{\mathsf{T}}\big\| \tag{364}$$

$$\leq 2\big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_0)\big\|\,\big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_t) - \boldsymbol{D}f_n(\boldsymbol{\theta}_0)\big\| + \big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_t) - \boldsymbol{D}f_n(\boldsymbol{\theta}_0)\big\|^2 \tag{365}$$

$$\leq 2\sigma_{\max}L_n\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\| + L_n^2\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|^2 \tag{366}$$

$$\leq \frac{5}{2}\sigma_{\max}L_n\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|\,. \tag{367}$$

(In the last inequality, we used the fact that $L_n\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\| \leq \sigma_{\min}/2$ by definition of $r_*$.) Applying Grönwall's inequality, and using $\boldsymbol{r}_0 = 0$, we obtain

$$\|\boldsymbol{r}_t\|_2 \leq e^{-\lambda_0 t/2}\|\boldsymbol{y}_0 - \boldsymbol{y}\|_2 \int_0^t \frac{1}{n}\big\|\boldsymbol{K}_s - \boldsymbol{K}_0\big\|\mathrm{d}s \tag{368}$$

$$\leq e^{-\lambda_0 t/2}t\|\boldsymbol{y}_0 - \boldsymbol{y}\|_2 \sup_{s\in[0,t]}\frac{1}{n}\big\|\boldsymbol{K}_s - \boldsymbol{K}_0\big\| \tag{369}$$

$$\leq e^{-\lambda_0 t/4}\frac{2}{\lambda_0}\|\boldsymbol{y}_0 - \boldsymbol{y}\|_2 \sup_{s\geq 0}\frac{1}{n}\big\|\boldsymbol{K}_s - \boldsymbol{K}_0\big\| \tag{370}$$

$$\overset{(a)}{\leq} e^{-\lambda_0 t/4}\frac{2}{\lambda_0}\|\boldsymbol{y}_0 - \boldsymbol{y}\|_2 \frac{5}{2n}L_n\sigma_{\max}\sup_{s\geq 0}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_0\|_2 \tag{371}$$

$$\overset{(b)}{\leq} e^{-\lambda_0 t/4}\frac{2}{\lambda_0}\|\boldsymbol{y}_0 - \boldsymbol{y}\|_2 \frac{5}{2n}L_n\sigma_{\max}\cdot\frac{2}{\sigma_{\min}}\|\boldsymbol{y}_0 - \boldsymbol{y}\|_2 \tag{372}$$

$$\leq 20\, e^{-\lambda_0 t/4}\frac{\sigma_{\max}}{\sigma_{\min}^3}L_n\|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2\,. \tag{373}$$

Here in $(a)$ we used Eq. (367) and in $(b)$ Eq. (359). Further using $\|\boldsymbol{r}_t\|_2 \leq \|\boldsymbol{y}_t - \boldsymbol{y}\|_2 + \|\overline{\boldsymbol{y}}_t - \boldsymbol{y}\|_2 \leq 2\|\boldsymbol{y}_0 - \boldsymbol{y}\|\exp(-\lambda_0 t/2)$, we get

$$\|\boldsymbol{y}_t - \overline{\boldsymbol{y}}_t\|_2 \leq 2e^{-\lambda_0 t/4}\|\boldsymbol{y} - \boldsymbol{y}_0\|_2\Big\{1 \wedge \frac{10\sigma_{\max}}{\sigma_{\min}^3}L_n\|\boldsymbol{y} - \boldsymbol{y}_0\|_2\Big\}\,. \tag{374}$$

Recall the gradient flow equations for $\boldsymbol{\theta}_t$ and $\overline{\boldsymbol{\theta}}_t$:

$$\dot{\boldsymbol{\theta}}_t = \frac{1}{n}\boldsymbol{D}f_n(\boldsymbol{\theta}_t)^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{y}_t)\,, \tag{375}$$

$$\dot{\overline{\boldsymbol{\theta}}}_t = \frac{1}{n}\boldsymbol{D}f_n(\boldsymbol{\theta}_0)^{\mathsf{T}}(\boldsymbol{y} - \overline{\boldsymbol{y}}_t)\,. \tag{376}$$

Taking the difference of these equations, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{\theta}_t - \overline{\boldsymbol{\theta}}_t\|_2 \leq \frac{1}{n}\big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_t) - \boldsymbol{D}f_n(\boldsymbol{\theta}_0)\big\|\,\|\boldsymbol{y}_t - \boldsymbol{y}\|_2 + \frac{1}{n}\big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_0)\big\|\,\|\boldsymbol{y}_t - \overline{\boldsymbol{y}}_t\|_2 \tag{377}$$

$$\leq \frac{L_n}{n}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2\|\boldsymbol{y}_t - \boldsymbol{y}\|_2 + \frac{\sigma_{\max}}{n}\|\boldsymbol{y}_t - \overline{\boldsymbol{y}}_t\|_2 \tag{378}$$

$$\overset{(a)}{\leq} \frac{L_n}{n}\cdot\frac{2}{\sigma_{\min}}\|\boldsymbol{y} - \boldsymbol{y}_0\|_2^2 e^{-\lambda_0 t/2} + \frac{\sigma_{\max}}{n}\cdot 2e^{-\lambda_0 t/4}\|\boldsymbol{y} - \boldsymbol{y}_0\|_2\Big\{1 \wedge \frac{10\sigma_{\max}}{\sigma_{\min}^3}L_n\|\boldsymbol{y} - \boldsymbol{y}_0\|_2\Big\} \tag{379}$$

where in $(a)$ we used Eqs. (348), (354) and (374). Integrating the last expression (thanks to $\overline{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$), we get

$$\|\boldsymbol{\theta}_t - \overline{\boldsymbol{\theta}}_t\|_2 \le \frac{8L_n}{\sigma_{\min}^3}\|\boldsymbol{y}-\boldsymbol{y}_0\|_2^2 + \Big\{\frac{16\sigma_{\max}}{\sigma_{\min}^2}\|\boldsymbol{y}-\boldsymbol{y}_0\|_2 \wedge \frac{160\sigma_{\max}^2}{\sigma_{\min}^5}L_n\|\boldsymbol{y}-\boldsymbol{y}_0\|_2^2\Big\}. \tag{380}$$

Simplifying, we get Eq. (349).

Finally, to prove Eq. (350), write

$$\|f(\boldsymbol{\theta}_t) - f_{\rm lin}(\overline{\boldsymbol{\theta}}_t)\|_{L^2} \le \underbrace{\|f(\boldsymbol{\theta}_t) - f_{\rm lin}(\boldsymbol{\theta}_t)\|_{L^2}}_{E_1} + \underbrace{\|f_{\rm lin}(\boldsymbol{\theta}_t) - f_{\rm lin}(\overline{\boldsymbol{\theta}}_t)\|_{L^2}}_{E_2}. \tag{381}$$

By writing $f(\boldsymbol{\theta}_t) - f_{\rm lin}(\boldsymbol{\theta}_t) = \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}[f(\boldsymbol{\theta}_s) - f_{\rm lin}(\boldsymbol{\theta}_s)]\mathrm{d}s$, we get

$$E_1 = \Big\|\int_0^t [\boldsymbol{D}f(\boldsymbol{\theta}_s) - \boldsymbol{D}f(\boldsymbol{\theta}_0)]\dot{\boldsymbol{\theta}}_s \mathrm{d}s\Big\|_{L^2} \tag{382}$$

$$\le \mathrm{Lip}(\boldsymbol{D}f)\sup_{s\ge 0}\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_0\|_2 \int_0^t \|\dot{\boldsymbol{\theta}}_s\|_2 \mathrm{d}s \tag{383}$$

$$\le \mathrm{Lip}(\boldsymbol{D}f)\cdot\frac{4\|\boldsymbol{y}-\boldsymbol{y}_0\|_2^2}{\sigma_{\min}^2}. \tag{384}$$

In the last step we used Eq. (348) and noted that the same argument to prove the latter indeed also bounds the integral $\int_0^t \|\dot{\boldsymbol{\theta}}_s\|_2 \mathrm{d}s$ (see Eq. (358)).

Finally, to bound term $E_2$, note that $f_{\rm lin}(\boldsymbol{\theta}_t) - f_{\rm lin}(\overline{\boldsymbol{\theta}}_t) = \boldsymbol{D}f(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_t - \overline{\boldsymbol{\theta}}_t)$, and using Eq. (349), we get

$$E_2 \le 180\|\boldsymbol{D}f(\boldsymbol{\theta}_0)\|\frac{L_n\sigma_{\max}^2}{\sigma_{\min}^5}\|\boldsymbol{y}-\boldsymbol{y}_0\|_2^2. \tag{385}$$

Equation (350) follows by putting together the above bounds for $E_1$ and $E_2$. □

We next pass to the case of two-layers networks:

$$f(\boldsymbol{x};\boldsymbol{\theta}) := \frac{\alpha}{\sqrt{m}}\sum_{j=1}^m b_j\sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle), \quad \boldsymbol{\theta} = (\boldsymbol{w}_1,\dots,\boldsymbol{w}_m). \tag{386}$$

**Lemma 5.3.** *Under Assumption 5.2, further assume $\{(y_i,\boldsymbol{x}_i)\}_{i\le n}$ to be i.i.d. with $\boldsymbol{x}_i \sim_{iid} \mathsf{N}(0,\mathbf{I}_d)$, and $y_i$ $B^2$-sub-Gaussian. Then there exist constants $C_i$, depending uniquely on $\sigma$, such that the following hold with probability at least $1-2\exp\{-n/C_0\}$, provided $md \ge C_0 n\log n$, $n \le d^{\ell_0}$ (whenever not specified, these hold for both $\boldsymbol{\theta}_0 \in \{\boldsymbol{\theta}_0^{(1)}, \boldsymbol{\theta}_0^{(2)}\}$):*

$$\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0^{(1)})\|_2 \le C_1(B+\alpha)\sqrt{n} \tag{387}$$

$$\|\boldsymbol{y} - f_n(\boldsymbol{\theta}_0^{(2)})\|_2 \le C_1 B\sqrt{n}, \tag{388}$$

$$\sigma_{\min}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)) \ge C_2\alpha\sqrt{d}, \tag{389}$$

$$\sigma_{\max}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)) \le C_3\alpha(\sqrt{n}+\sqrt{d}), \tag{390}$$

$$\mathrm{Lip}(\boldsymbol{D}f_n) \le C_4\alpha\sqrt{\frac{d}{m}}(\sqrt{n}+\sqrt{d}). \tag{391}$$

*Further*

$$\|\boldsymbol{D}f(\boldsymbol{\theta}_0)\| \le C_1'\alpha, \tag{392}$$

$$\mathrm{Lip}(\boldsymbol{D}f) \le C_4'\alpha\sqrt{\frac{d}{m}}. \tag{393}$$

*Proof.* Since the $y_i$ are $B^2$ sub-Gaussian, we have $\|\boldsymbol{y}\|_2 \le C_1 B\sqrt{n}$ with the stated probability. Equation (388) follows since by construction $f_n(\boldsymbol{\theta}_0^{(2)}) = 0$.

For Eq. (387) we claim that $\|f_n(\boldsymbol{\theta}_0^{(1)})\|_2 \le C_1\alpha\sqrt{n}$ with the claimed probability. To show this, it is sufficient of course to consider $\alpha = 1$. Let $F(\boldsymbol{X}, \boldsymbol{W}) := \|f_n(\boldsymbol{\theta}_0^{(1)})\|_2$, where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ contains as rows the vectors $\boldsymbol{x}_i$, and $\boldsymbol{W}$ the vectors $\boldsymbol{w}_i$. We also write $\boldsymbol{\theta}_0^{(1)} = \boldsymbol{\theta}_0$ for simplicity. We have

$$\mathbb{E}\{F(\boldsymbol{X}, \boldsymbol{W})\}^2 \le \mathbb{E}\{\|f_n(\boldsymbol{\theta}_0)\|_2^2\} = n\mathbb{E}\{f(\boldsymbol{x}_1; \boldsymbol{\theta}_0)^2\} \tag{394}$$

$$= n\mathrm{Var}\{\sigma(\langle \boldsymbol{w}_1, \boldsymbol{x}_1\rangle)\} \le Cn\,. \tag{395}$$

Next, proceeding as in the proof of [OS20, Lemma 7] (letting $\boldsymbol{b} = (b_j)_{j\le m}$)

$$\begin{aligned}
\big|F(\boldsymbol{X}, \boldsymbol{W}_1) - F(\boldsymbol{X}, \boldsymbol{W}_2)\big| &\le \frac{1}{\sqrt{m}}\big\|\sigma(\boldsymbol{X}\boldsymbol{W}_1^{\mathsf{T}})\boldsymbol{b} - \sigma(\boldsymbol{X}\boldsymbol{W}_2^{\mathsf{T}})\boldsymbol{b}\big\|_2 \\
&\le \big\|\sigma(\boldsymbol{X}\boldsymbol{W}_1^{\mathsf{T}}) - \sigma(\boldsymbol{X}\boldsymbol{W}_2^{\mathsf{T}})\big\|_F \\
&\le C\big\|\boldsymbol{X}\boldsymbol{W}_1^{\mathsf{T}} - \boldsymbol{X}\boldsymbol{W}_2^{\mathsf{T}}\big\|_F \\
&\le C\|\boldsymbol{X}\|\big\|\boldsymbol{W}_1^{\mathsf{T}} - \boldsymbol{W}_2^{\mathsf{T}}\big\|_F\,.
\end{aligned}$$

We have $\|\boldsymbol{X}\| \le 2(\sqrt{n} + \sqrt{d})$ with the probability at least $1 - 2\exp\{-(n \vee d)/C)$ [Ver18]. On this event, $F(\boldsymbol{X}, \cdot\,)$ is $2(\sqrt{n} + \sqrt{d})$-Lipschitz with respect to $\boldsymbol{W}$. Recall that the uniform measure on the sphere of radius $\sqrt{d}$ satisfies a log-Sobolev inequality with $\Theta(1)$ constant, [Led01, Chapter 5], that the log-Sobolev constant for a product measure is the same as the worst constant of each of the terms. We then have

$$\mathbb{P}\big(F(\boldsymbol{X}, \boldsymbol{W}) \ge \mathbb{E}F(\boldsymbol{X}, \boldsymbol{W}) + t\big) \le e^{-dt^2/C(n+d)} + 2e^{-(n\vee d)/C}\,. \tag{396}$$

Taking $t = C_1\sqrt{n}$ for a sufficiently large constant $C_1$ implies that the right-hand side is at most $2\exp(-(n \vee d)/C)$, which proves the claim.

Notice that all the following inequalities are homogeneous in $\alpha > 0$. Hence, we will assume—without loss of generality—that $\alpha = 1$. Equation (389) follows from [OS20, Lemma 4]. Indeed this lemma implies

$$m \ge \frac{C(n+d)\log n}{d\lambda_{\min}(\boldsymbol{K})} \quad \Rightarrow \quad \sigma_{\min}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)) \ge c_0\sqrt{d\lambda_{\min}(\boldsymbol{K})}\,, \tag{397}$$

where $\boldsymbol{K}$ is the empirical NT kernel

$$\boldsymbol{K} = \frac{1}{d}\mathbb{E}\big\{\boldsymbol{D}f_n(\boldsymbol{\theta}_0)\boldsymbol{D}f_n(\boldsymbol{\theta}_0)\big\} = \big(K_{\mathsf{NT}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\big)_{i,j\le n}\,. \tag{398}$$

Under Assumption 5.2 (in particular $\sigma'$ having non-vanishing Hermite coefficients $\mu_\ell(\sigma)$ for all $\ell \le \ell_0$), and $n \le d^{\ell_0}$, we have $\lambda_{\min}(\boldsymbol{K}) \ge c_0$ with the stated probability, see for instance [EK10]. This implies the claim.

For Eq. (390), note that, for any vector $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|_2 = 1$ we have

$$\big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_0)^{\mathsf{T}}\boldsymbol{v}\big\|_2^2 = \frac{1}{m}\sum_{i,j\le n}\sum_{\ell=1}^m v_i\sigma'(\langle \boldsymbol{w}_\ell, \boldsymbol{x}_i\rangle)v_j\sigma'(\langle \boldsymbol{w}_\ell, \boldsymbol{x}_j\rangle)\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \tag{399}$$

$$= \langle \boldsymbol{M}, \boldsymbol{X}, \boldsymbol{X}^{\mathsf{T}}\rangle\,, \tag{400}$$

$$M_{ij} := \frac{1}{m}\sum_{\ell=1}^m v_i\sigma'(\langle \boldsymbol{w}_\ell, \boldsymbol{x}_i\rangle)v_j\sigma'(\langle \boldsymbol{w}_\ell, \boldsymbol{x}_j\rangle)\,. \tag{401}$$

Since $\boldsymbol{M} \succeq 0$, we have

$$\big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_0)^{\mathsf{T}}\boldsymbol{v}\big\|_2^2 \le \mathsf{tr}(\boldsymbol{M})\|\boldsymbol{X}\|^2 \tag{402}$$

$$= \frac{1}{m}\sum_{\ell=1}^m\sum_{i=1}^n v_i^2\sigma'(\langle \boldsymbol{w}_\ell, \boldsymbol{x}_i\rangle)^2 \cdot \|\boldsymbol{X}\|^2 \tag{403}$$

$$\le B^2\|\boldsymbol{v}\|_2^2\|\boldsymbol{X}\|^2\,. \tag{404}$$

Hence $\sigma_{\max}(\boldsymbol{D}f_n(\boldsymbol{\theta}_0)) \le B\|\boldsymbol{X}\|$ and the claim follows from standard estimates of operator norms of random matrices with independent entries.

Equation (391) follows from [OS20, Lemma 5], which yields (after adapting to the different normalization of the $\boldsymbol{x}_i$, and using the fact that $\max_{i\le n}\|\boldsymbol{x}_i\|_2 \le C\sqrt{d}$ with probability at least $1-2\exp(-d/C)$):

$$\big\|\boldsymbol{D}f_n(\boldsymbol{\theta}_1) - \boldsymbol{D}f_n(\boldsymbol{\theta}_2)\big\| \le C\sqrt{\frac{d}{m}}\|\boldsymbol{X}\|\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2\,.$$

(Here $\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2 = \|\boldsymbol{W}_1-\boldsymbol{W}_2\|_F$, where $\boldsymbol{W}_i \in \mathbb{R}^{m\times d}$ is the matrix whose rows are the weight vectors.) The claim follows once more by using $\|\boldsymbol{X}\| \le 2(\sqrt{n}+\sqrt{d})$ with probability at least $1-2\exp\{-(n\vee d)/C\}$.

In order to prove Eq. (392), note that, for $h \in L^2(\mathbb{R}^d,\mathbb{P})$,

$$\big\|\boldsymbol{D}f(\boldsymbol{\theta}_0)^*h\big\|_2 = \mathbb{E}\{Q_h(\boldsymbol{x}_1,\boldsymbol{x}_2)\,P(\boldsymbol{x}_1,\boldsymbol{x}_2)\}\,, \tag{405}$$

$$Q_h(\boldsymbol{x}_1,\boldsymbol{x}_2) := \frac{1}{m}\sum_{\ell=1}^m \sigma'(\langle\boldsymbol{w}_\ell,\boldsymbol{x}_1\rangle)h(\boldsymbol{x}_1)\sigma'(\langle\boldsymbol{w}_\ell,\boldsymbol{x}_2\rangle)h(\boldsymbol{x}_2)\,, \tag{406}$$

$$P(\boldsymbol{x}_1,\boldsymbol{x}_2) := \langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle\,. \tag{407}$$

Here expectation is with respect to independent random vectors $\boldsymbol{x}_1,\boldsymbol{x}_2\sim\mathbb{P}$. Denote by $\boldsymbol{Q}_h$ and $\boldsymbol{P}$ the integral operators in $L^2(\mathbb{R}^d,\mathbb{P})$ with kernels $Q_h$ and $P$. It is easy to see that $\boldsymbol{P}$ is the projector onto the subspace of linear functions, and $\boldsymbol{Q}_h$ is positive semidefinite. Therefore

$$\big\|\boldsymbol{D}f(\boldsymbol{\theta}_0)^*h\big\|_2 \le \mathsf{tr}(\boldsymbol{Q}_h) = \frac{1}{m}\sum_{\ell=1}^m \mathbb{E}\big\{\sigma'(\langle\boldsymbol{w}_\ell,\boldsymbol{x}_1\rangle)^2h(\boldsymbol{x}_1)^2\big\} \tag{408}$$

$$\le B^2\|h\|_{L^2}^2\,. \tag{409}$$

This implies $\|\boldsymbol{D}f(\boldsymbol{\theta}_0)\| \le B$.

In order to prove Eq. (393), define $\Delta_\ell(\boldsymbol{x}) := \sigma'(\langle\boldsymbol{w}_{1,\ell},\boldsymbol{x}\rangle) - \sigma'(\langle\boldsymbol{w}_{2,\ell},\boldsymbol{x}\rangle)$. Let $h\in L^2(\mathbb{R}^d,\mathbb{P})$ and note that

$$\big\|\boldsymbol{D}f(\boldsymbol{\theta}_0)^*h\big\|_2^2 = \frac{1}{m}\sum_{\ell=1}^m \Big\|\mathbb{E}\big\{\boldsymbol{x}h(\boldsymbol{x})\Delta_\ell(\boldsymbol{x})\big\}\Big\|_2^2 \tag{410}$$

$$\le \frac{1}{m}\sum_{\ell=1}^m \mathbb{E}\big\{\|\boldsymbol{x}\|\,|h(\boldsymbol{x})\Delta_\ell(\boldsymbol{x})|\big\}^2 \tag{411}$$

$$\le \frac{1}{m}\sum_{\ell=1}^m \mathbb{E}\big\{\|\boldsymbol{x}\|^2\,\Delta_\ell(\boldsymbol{x})^2\big\}\,\|h\|_{L^2}\,. \tag{412}$$

Note that $|\Delta_\ell(\boldsymbol{x})| \le B\,|\langle\boldsymbol{w}_{1,\ell}-\boldsymbol{w}_{2,\ell},\boldsymbol{x}\rangle|$. Using this and the last expression above, we get

$$\big\|\boldsymbol{D}f(\boldsymbol{\theta}_0)\big\|^2 \le \frac{B^2}{m}\sum_{\ell=1}^m \mathbb{E}\big\{\|\boldsymbol{x}\|^2\,\langle\boldsymbol{x},\boldsymbol{w}_{1,\ell}-\boldsymbol{w}_{2,\ell}\rangle^2\big\} \tag{413}$$

$$\le \frac{B^2}{m}(d+2)\sum_{\ell=1}^m \|\boldsymbol{w}_{1,\ell}-\boldsymbol{w}_{2,\ell}\|_2^2 = \frac{B^2}{m}(d+2)\|\boldsymbol{W}_1-\boldsymbol{W}_2\|_F^2\,, \tag{414}$$

where the second inequality follows from the Gaussian identity $\mathbb{E}\{\|\boldsymbol{x}\|^2\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\} = (d+2)\mathbf{I}_d$. This proves Eq. (393). □

**Theorem 5.4.** *Consider the two layer neural network of (386) under the assumptions of Lemma 5.3. Further let $\overline{\alpha} := \alpha/(1+\alpha)$ for initialization $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^{(1)}$ and $\overline{\alpha} := \alpha$ for $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^{(2)}$. Then there exist constants $C_i$, depending uniquely on $\sigma$, such that if $md \ge C_0 n\log n$, $d\le n\le d^{\ell_0}$ and*

$$\overline{\alpha} \ge C_0\sqrt{\frac{n^2}{md}}\,, \tag{415}$$

*then, with probability at least* $1 - 2\exp\{-n/C_0\}$*, the following hold for all* $t \geq 0$*.*

1. *Gradient flow converges exponentially fast to a global minimizer. Specifically, letting* $\lambda_* = C_1\alpha^2 d/n$*, we have*

$$\widehat{L}(\boldsymbol{\theta}_t) \leq \widehat{L}(\boldsymbol{\theta}_0)\, e^{-\lambda_* t}\,. \tag{416}$$

2. *The model constructed by gradient flow and linearized flow are similar on test data, namely*

$$\|f(\boldsymbol{\theta}_t) - f_{\rm lin}(\overline{\boldsymbol{\theta}}_t)\|_{L^2(\mathbb{P})} \leq C_1 \left\{ \frac{\alpha}{\overline{\alpha}^2}\sqrt{\frac{n^2}{md}} + \frac{1}{\overline{\alpha}^2}\sqrt{\frac{n^5}{md^4}} \right\}. \tag{417}$$

*Proof.* Throughout the proof, we use $C$ to denote constants depending only on $\sigma$, that might change from line to line. Using Lemma 5.3, the condition (346) reads

$$\alpha\sqrt{\frac{dn}{m}} \cdot \frac{\alpha}{\overline{\alpha}}\sqrt{n} \leq C\big(\alpha\sqrt{d}\big)^2\,. \tag{418}$$

which is equivalent to Eq. (415). We can therefore apply Theorem 5.1.

Equation (416) follows from Theorem 5.1, point 1, using the lower bound on $\sigma_{\min}$ given in Eq. (389).

Equation (417) follows from Theorem 5.1, point 3, using the estimates in Lemma 5.3. □