Published in final edited form as:
Ann Stat. 2022 April ; 50(2): 949–986. doi:10.1214/21-aos2133.

SURPRISES IN HIGH-DIMENSIONAL RIDGELESS LEAST SQUARES INTERPOLATION

Trevor Hastie[1,a], **Andrea Montanari**[2,b], **Saharon Rosset**[3,c], **Ryan J. Tibshirani**[4,d]

[1]Department of Statistics and Department of Biomedical Data Science, Stanford University

[2]Department of Statistics and Department of Electrical Engineering, Stanford University

[3]School of Mathematical Sciences, Tel Aviv University

[4]Department of Statistics and Department of Machine Learning, Carnegie Mellon University

Abstract

Interpolators—estimators that achieve zero training error—have attracted growing attention in machine learning, mainly because state-of-the art neural networks appear to be models of this type. In this paper, we study minimum ℓ_2 norm ("ridgeless") interpolation least squares regression, focusing on the high-dimensional regime in which the number of unknown parameters p is of the same order as the number of samples n. We consider two different models for the feature distribution: a linear model, where the feature vectors $x_i \in \mathbb{R}^p$ are obtained by applying a linear transform to a vector of i.i.d. entries, $x_i = \Sigma^{1/2} z_i$ (with $z_i \in \mathbb{R}^p$); and a nonlinear model, where the feature vectors are obtained by passing the input through a random one-layer neural network, $x_i = \varphi(W z_i)$ (with $z_i \in \mathbb{R}^d$, $W \in \mathbb{R}^{p \times d}$ a matrix of i.i.d. entries, and φ an activation function acting componentwise on $W z_i$). We recover—in a precise quantitative way—several phenomena that have been observed in large-scale neural networks and kernel machines, including the "double descent" behavior of the prediction risk, and the potential benefits of overparametrization.

Key words and phrases.

Regression; interpolation; overparametrization; ridge regression; random matrix theory

MSC2020 subject classifications.

Primary 62J05, 62J07; secondary 62J02, 62F12

[a] hastie@stanford.edu . [b] montanar@stanford.edu . [c] saharon@post.tau.ac.il . [d] ryantibs@cmu.edu .

SUPPLEMENTARY MATERIAL
Supplement to "Surprises in high-dimensional ridgeless least squares interpolation" (DOI: 10.1214/21-AOS2133SUPP; .pdf). Supplementary information.

1. Introduction.

Modern deep learning models involve a huge number of parameters. In many applications, current practice suggests that we should design the network to be sufficiently complex so that the model (as trained, typically, by gradient descent) interpolates the data, that is, achieves zero training error. Indeed, in a thought-provoking experiment, Zhang et al. [71] showed that state-of-the-art deep neural network architectures are complex enough that they can be trained to interpolate the data even when the actual labels are replaced by entirely random ones.

Despite their enormous complexity, deep neural networks are frequently observed to generalize well in practice. At first sight, this seems to defy conventional statistical wisdom: interpolation (vanishing training error) is commonly taken to be a proxy for poor generalization (large gap between training and test error), and hence large test error. In an insightful series of papers, Belkin et al. [8, 10, 11] pointed out that these concepts are in general distinct, and interpolation does not contradict generalization. For example, recent work has investigated interpolation—via kernel ridge regression—in reproducing kernel Hilbert spaces [30, 47]. While in low dimension a positive regularization is needed to achieve good interpolation, in certain high-dimensional settings interpolation can be nearly optimal.

In this paper, we investigate these phenomena in the context of simple linear models. We assume to be given i.i.d. data (y_i, x_i), $i \le n$, with $x_i \in \mathbb{R}^p$ a feature vector and $y_i \in \mathbb{R}$ a response variable. These are distributed according to the model (see Section 2 for further definitions)

$$(x_i, \epsilon_i) \sim P_x \times P_\epsilon, \quad i = 1, \dots, n, \tag{1}$$

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \dots, n, \tag{2}$$

where P_x is a distribution on $\mathbb{R}^p$ such that $\mathbb{E}(x_i) = 0$, Cov(x_i) Σ, and P_ϵ is a distribution on $\mathbb{R}$ such that $\mathbb{E}(\epsilon_i) = 0$, Var$(\epsilon_i)$ σ^2.

We estimate β by linear regression. Since our focus is on the overparametrized regime $p > n$, the usual least square objective does not have a unique minimizer, and needs to be regularized. We consider two approaches: min-norm regression, which estimates β by the least squares solution with minimum ℓ_2 norm; and ridge regression, which penalizes a coefficients vector β by its ℓ_2 norm square $\|\beta\|_2^2$. We denote these estimates by $\hat{\beta}$ and $\hat{\beta}_\lambda$ (λ being the regularization parameter), and note that $\lim_{\lambda \to 0} \hat{\beta}_\lambda = \hat{\beta}$. If the design matrix has full row rank, which is generically the case for $p > n$, the min-norm estimator is an interpolator, namely $x_i^T \hat{\beta} = y_i$ for all $i \le n$. In order to evaluate these methods, we will study the prediction risk at a new (unseen) test point (y_0, x_0).

We study the model (2) in the proportional regime $p \asymp n$, with a special focus on the overparametrized case $p > n$. Our main contribution is to show that, by considering different choices of the features distribution P_x, we can reproduce a number of statistically interesting phenomena that have emerged in the context of deep learning.

From a technical perspective, our main results are: Theorems 2 and 5, which assume the linear model $x_i \Sigma^{1/2} z_i$ with z_i a vector with independent coordinates and Theorem 8, which assumes a nonlinear model $x_i = \varphi(Wz_i)$ with $z_i \sim N(0, I_d)$. While the linear model has already a attracted significant amount of work (see Section 1.3 for an overview), Theorems 2 and 5 provide a more accurate approximation of the prediction risk in the proportional regime $n \asymp p$, as compared to available results in the literature, and hold in a more general setting.

The prediction risk depends on the geometry of the pair (Σ, β). We consider a few different choices for this geometry, which are broadly motivated by our objective to understand overparametrized models, and specialize our formulas to these special cases:

1. *Isotropic features*. This is the simplest case, in which $\Sigma = I_p$ and, therefore, as we will see the asymptotic risk depends on β only through its norm $\|\beta\|_2$. This simple model captures some interesting features of overparametrization, but misses others.

   We first consider a well-specified case in which $x_i \in \mathbb{R}^p$ and we regress against x_i. We then pass to a misspecified case, in which the model (2) holds for covariates $x_i \in \mathbb{R}^{p+q}$, but we regress only against the first p covariates.

2. *Latent space features*. In the overparametrized regime, it is natural to assume that both the covariates x_i, and the coefficients vector β lie close to a low-dimensional subspace. In order to model this property, we assume $\Sigma = WW^T + I$, with $W \in \mathbb{R}^{p \times d}$, $d \ll p$ and β lies in the span of the columns of W. Interestingly, this model reproduces many phenomena observed in more complex nonlinear models, and has a more direct connection to neural networks.

3. *Nonlinear model*. In all of the previous cases, the distribution of x_i is of the form $x_i = \Sigma^{1/2} z_i$ where z_i is a vector with independent coordinates. In order to test the generality of our results, we consider a model in which x_i is obtained by passing $z_i \sim N(0, I_d)$ through a one-layer neural net with random first layer weights, namely $x_i = \varphi(Wz_i)$, for $W \in \mathbb{R}^{p \times d}$.

We will summarize our results for these four examples in the next subsection.

A skeptical reader might ask what linear models have to do with neural networks. We emphasize that linear models provide more than a simple analogy, and a recent line of work outlines a concrete connection between the two settings [4, 19, 24, 25, 39]. We will discuss this connection in Section 1.2.

1.1. Summary of results.

As mentioned above, we analyze the out-of-sample prediction risk of the minimum ℓ_2 norm (or min-norm, for short) least squares estimator, and of ridge-regularized least squares.

We denote by $\gamma_n := p/n \in (0, \infty)$ the overparametrization ratio. When $\gamma < 1$, we call the problem *underparametrized*, and when $\gamma > 1$, we call it *overparametrized*. Our most general results for the linear model (Theorem 2 and 5) apply to a nonasymptotic setting in which n, p are finite, and provide a deterministic approximation of the risk with error bounds that are uniform in the distribution of the data. We use these general results to derive asymptotic formulas in the limit in which both p and n diverge with $\gamma_n = p/n \to \gamma$. (We will drop the subscript from γ_n whenever this is not cause for confusion.)

We assume the model (2) and denote by $\mathrm{SNR} = \|\beta\|_2^2/\sigma^2$ the signal-to-noise ratio. We refer to Figure 1 for supporting plots of the asymptotic risk curves for different cases of interest.

Our main results are twofold: (i) We show that by suitable choices of β, Σ, we can easily construct scenarios in which the minimum of the risk is achieved in the overparamertized regime $p > n$; (ii) We show that these findings are robust to the details of the distribution of (y_i, x_i).

As a preliminary remark, note that in the underparametrized regime ($\gamma < 1$), the min-norm estimator coincides with the standard least squares estimator. Its risk is purely variance (there is no bias), and does not depend on β, Σ (see Proposition 2). Interestingly, the asymptotic risk diverges as we approach the interpolation boundary (as $\gamma \to 1$).

In contrast, in the overparametrized regime ($\gamma > 1$), the risk is composed of both bias and variance,[1] and generally depends on β, Σ (see Theorem 2).

We next highlight some concrete results for the four models discussed in the previous section (unless explicitly said, we refer to the min-norm estimator).

Isotropic features.—The asymptotic risk depends on the coefficients vector only through its norm $\|\beta\|_2^2$ or, up to a scaling, on $\mathrm{SNR} = \|\beta\|_2^2/\sigma^2$.

1. If the model is *well specified*, we observe two different behaviors. For SNR ≤ 1, the risk is decreasing for $\gamma \in (1, \infty)$. For SNR > 1, the risk has a *local* minimum on $\gamma \in (1, \infty)$.

   In either case, the risk approaches the null risk as $\gamma \to \infty$, and achieves its global minimum in the underparametrized regime (see Section 3.2).

2. If the model is *misspecified*, when SNR > 1, the risk can attain its *global* minimum in the overparametrized regime $\gamma \in (1, \infty)$ (when there is strong

[1] Note that in the overparametrized regime the bias is nonvanishing even in the interpolation limit $\lambda \to 0$. The reason is that the set of interpolators is an affine space of dimension $p - n$, and the min-norm criterion selects one specific interpolator, whose mean has—in general—norm smaller than β.

enough approximation bias, see Section 5.1.3). However, the risk is again increasing for γ large enough.

3. Optimally-tuned ridge regression uses a nonvanishing regularization $\lambda > 0$, and dominates the min-norm least squares estimator in risk, across all values of γ and SNR, both the well-specified and misspecified settings. For a misspecified model, optimally-tuned ridge regression attains its global minimum around $\gamma = 1$ (see Section 6).

4. Optimal tuning of the ridge penalty can be achieved by leave-one-out cross-validation (see Theorem 7).

Anisotropic features.—In this case, $\Sigma \neq I$ and the risk depends on the geometry of (Σ, β), and in particular on how β aligns with the eigenvectors of Σ.

1. If the coefficients vector is equidistributed along the eigenvectors of Σ, the behavior is qualitatively similar to the isotropic case. This situation arises, for instance, if β is itself random with a spherical prior.

2. If β is aligned with the top eigenvectors of Σ, the situation is qualitatively different. As an example we obtain an explicit formula for the asymptotic risk in the latent space model discussed above; see the red line of Figure 1 (and Figure 5) for an illustration. We find that, for natural choices of the model parameters, the risk is monotone decreasing in the overparametrized regime, and reaches its global minimum as $\gamma \to \infty$. This qualitative behavior matches the one observed for neural networks (see Section 5.2).

3. For the latent space model, we observe that, at large overparametrization, the minimum error is achieved as $\lambda \to 0$, that is, by min-norm interpolators (see Section 6.2, and Section 1.3 for related work).

Nonlinear model.—Finally, we consider a nonlinear model in which $x_i = \varphi(Wz_i)$ where φ is a nonlinear activation function applied componentwise, $W \in \mathbb{R}^{p \times d}$ and $z_i \sim N(0, I_d)$.

1. We first consider the case of a purely nonlinear activation. We compute the limiting risk of min-norm regression, and show that this matches the one for Gaussian $x_i \sim N(0, I_d)$ (see Theorem 8). This is illustrated by the “x” symbols in Figure 1.

2. We then compute the limit of the variance component of the risk for more general activations φ. We show that this depends on the activation function only through the size of its linear component. Further, the resulting variance turns out to coincide asymptotically with the variance in the linear model $x_i = \Sigma^{1/2}\tilde{z}_i$, if we take $\Sigma = (1 - c_1)I_p + c_1 WW^T$ for a certain constant c_1. (see Theorem 9).

These results confirm that the results established for the case $x_i \Sigma^{1/2} z_i$ with z_i an i.i.d. vector hold in greater generality

From a technical viewpoint, analysis of the isotropic covariates model is straightforward and relies on standard random matrix theory results. However, we believe it provides useful insights.

In contrast, the results for general covariance and coefficients structure (Σ, β) is technically novel. We discuss related work in Section 1.3. Our results for the nonlinear model are also technically novel. In this setting, we derive a new asymptotic result on resolvents of certain block matrices, which may be of independent interest (see Lemma 3).

We next discuss the intuitions that emerge from our results as well as earlier literature.

Bias and variance.—The shape of the asymptotic risk curve for min-norm least squares is, of course, controlled by its components: bias and variance. For fully specified models, the bias increases with γ in the overparametrized regime, which is intuitive. When $p > n$, the min-norm least squares estimate of β is constrained to lie the row space of X, the training feature matrix. This is a subspace of dimension n lying in a feature space of dimension p. Thus as p increases, so does the bias, since this row space accounts for less and less of the ambient p-dimensional feature space.

Meanwhile, we find that, in the overparametrized regime, the variance *decreases* with γ. This may seem counterintuitive at first, because it says, in a sense, that the min-norm least squares estimator becomes *more* regularized as p grows. However, this too can be explained intuitively, as follows. As p grows, the minimum ℓ_2 norm least squares solution—that is, the minimum ℓ_2 norm solution to the linear system $Xb = y$, for a training feature matrix X and response vector y—will generally have decreasing ℓ_2 norm. Why? Compare two such linear systems: in each, we are asking for the min-norm solution to a linear system with the same y, but in one instance we are given more columns in X, so we can generally decrease the components of b (by distributing them over more columns), and achieve a smaller ℓ_2 norm. This can in fact be formalized asymptotically; see Corollaries 1 and 3.

Double descent.—Recently, Belkin et al. [8] pointed out a fascinating empirical trend where, for popular methods like neural networks and random forests, we can see a *second* bias-variance tradeoff in the out-of-sample prediction risk beyond the interpolation limit. The risk curve here resembles a traditional U-shape curve before the interpolation limit, and then descends again beyond the interpolation limit, which these authors call "double descent." A closely related phenomenon was found earlier by Spigler et al. [63], who studied the "jamming transition" from underparametrized to overparametrized neural networks. Our results formally verify that this double descent phenomenon occurs even in the simple and fundamental case of least squares regression. The appearance of the second descent in the risk, past the interpolation boundary ($\gamma = 1$), is explained by the fact that the variance decreases as γ grows, as discussed above.

In the misspecified case, the variance still decreases with γ (for the same reasons), but interestingly, the bias can now also decrease with γ, provided γ is not too large (not too far past the interpolation boundary). The intuition here is that in a misspecified model, some part of the true regression function is always unaccounted for, and adding features generally improves our approximation capacity. As a consequence, the double descent phenomenon can be even more pronounced in the misspecified case (depending on the strength of the approximation bias), in that the risk can attain its global minimum past the interpolation limit.

Finally, in the latent space model, we observe that the overall risk can be monotone decreasing in the overparametrized regime, and attain its global minimum for large overparametrization $\gamma \to \infty$(after p, $n \to \infty$). In this case, we can write the design matrix as $X ZW^T + U$, where U is noise, and Z is the $n \times d$ matrix of latent covariates. Equivalently, the ith column of X (the ith feature) takes the form $\tilde{x}_i = Zw_i + \tilde{u}_i$, where w_i is the ith column of W^T and $\tilde{u}_i$ is the ith column of U. Therefore, each new feature provides new information about the underlying low-dimensional latent variables Z. As p gets large, ridge regression with respect to the feature matrix X approximates increasingly well a ridge regression with respect to the latent variables Z.

Interpolation versus regularization.—The min-norm least squares estimator can be seen as the limit of ridge regression as the tuning parameter tends to zero. A natural and important question is whether (or when) letting the regularization to 0 is optimal. Min-norm least squares is also the convergence point of gradient descent run on the least squares loss. Early-stopped gradient descent is known to be closely connected to ridge regularization; see, for example, Ali et al. [3], which proves a tight coupling between the two (see Section 1.3 for further related work). The question of whether letting the regularization vanish is optimal is closely related to the question of whether running gradient descent until convergence is optimal or early stopping provides some advantage.

Closely related questions have been investigated in the context of classification. For instance, it is common to run boosting until the training error is zero, and the boosting path is tied to ℓ_1 regularization [37, 58, 65]. It is empirically observed that, for noisy labels, early stopping (treating the number of boosting iterations as a tuning parameter) can be beneficial.

We would not expect the best-predicting ridge solution to be always at the end of its regularization path. Our results, comparing min-norm least squares to optimally-tuned ridge regression, show that (asymptotically) this is never the case, when β is incoherent with respect to the eigenvectors of Σ. This is for instance the case when $\Sigma = I_p$, or β is distributed according to a spherically symmetric prior. In contrast, [42] recently pointed out that—when β is aligned with the leading eigenvectors of Σ—min-norm regression can have optimal risk (i.e., the optimal regularization vanishes). We show that this is indeed the case in the latent space model mentioned above: this provides indeed an extremely simple example of a phenomenon that has been observed in the past for kernel methods [47]. Notice that the results we obtain for the latent space model are asymptotically sharper than the one of [47], in that they imply optimality up to subleading terms (not just up to constant factors).

In practice, of course, we would not have access to the optimal tuning parameter for ridge (optimal stopping for gradient descent), and we would rely on, for example, cross-validation (CV). Our theory shows that for ridge regression, CV tuning is asymptotically equivalent to optimal tuning (and we would expect the same results to carry over to gradient descent, but have not pursued this formally).

1.2. Connection to neural networks.

As mentioned above, recent literature has established a direct connection between linear models and more complex models such as neural networks, in a certain training regime [4, 19, 24, 25, 39]. Here, we will briefly outline this connection, referring the reader to the literature for a more detailed exposition.

For the discussion in this section, it is convenient to consider a more general setting, in which we are given data (y_i, z_i), $i \le n$, $y_i \in \mathbb{R}$, $z_i \in \mathbb{R}^d$, which are i.i.d. from an arbitrary distribution $(y_i, z_i) \sim P_{y,z}$. Imagine training a neural network with parameters (weights) $\theta \in \mathbb{R}^p$, $f(\cdot\,; \theta) : \mathbb{R}^d \to \mathbb{R}$, $z \longrightarrow f(z; \theta)$. The specific form or architecture of the network is not important for our discussion. However, it is important to distinguish two conceptually different functions. One the one hand, we have the true regression function $f_*(z_i) = \mathbb{E}\{y_i \mid z_i\}$; this is unknown to the statistician. For theoretical purposes, f_* can be assumed to belong so some function class, but for this section we will not specify this choice. On the other hand, we have a parametric model $f(z_i; \theta)$, which is determined by the specific network architecture. A specific network with the given architecture is determined by assigning the network weights $\theta \in \mathbb{R}^p$.

From the point of view of optimization, the central role is played by the parametric model $f(z; \theta)$. In modern machine learning, the number of parameters p is so large that—under certain training schemes—θ only changes by a small amount with respect to a random initialization $\theta_0 \in \mathbb{R}^p$. It thus makes sense to linearize the model around θ_0. Supposing that the initialization is such that $f(z; \theta_0) \approx 0$, and letting $\theta = \theta_0 + \beta$, we can approximate the statistical model $z \mapsto f(z; \theta)$ by

$$z \mapsto \nabla_\theta f(z; \theta_0)^T \beta \,. \tag{3}$$

This model is still nonlinear in the input z, but is linear in the parameters β. In other words, if the linear approximation is accurate, learning reduces to computing feature vectors $x_i = \nabla_\theta f(z_i; \theta_0)$, $i = 1, \ldots, n$, from the data, and then fitting a linear model in the x_i's. Notice that these feature vectors have high dimension ($p > n$) since the network is overparametrized, and that the "featurization map" $z_i \mapsto \nabla f(z_i; \theta_0)$ is random because the initialization θ_0 is. Further, since $p > n$, many vectors β give rise to a model that interpolates the data.

The above scenario was made rigorous in a number of papers [4, 19, 24, 25, 39]. In particular, [19] shows—under some technical conditions—that the linearization (3) can be accurate if the model is overparametrized ($p > n$), and closed under scalings (if $f(\cdot)$ is encoded by a neural network, then $sf(\cdot)$ is also a neural network for any $s \in \mathbb{R}$). Under these conditions, there exists a scaling of the network's parameters such that gradient-based training converges to a model that can be approximated arbitrarily well by (3). Further, under the linearization (3), gradient descent converges to the interpolator that minimizes[2] the ℓ_2 norm $\|\beta\|_2$ (see Proposition 1 below).

What are the *statistical consequences* of these linearization results? In principle, we could consider $\{(y_i, z_i)\}_{i \le n}$ to be i.i.d. samples with a certain population distribution $P_{y,z}$, and then study the behavior of minimum ℓ_2 norm interpolator of the form (3) under this data model. Denoting by $\phi(z) := \nabla_\theta f(z_i; \theta_0)$ the featurization map, this would amount to study

$$\hat{\beta} = \arg\min\left\{\|b\|_2 \text{ subject to } x_i^T b = y_i,\ x_i = \phi(z_i) \forall i \le n\right\}. \tag{4}$$

Even starting from a simple joint distribution $P_{y,z}$ for the data (y_i, z_i), the resulting joint distribution for (y_i, x_i) (induced by the map $z_i \rightarrow \phi(z_i) = \nabla_\theta f(z_i; \theta_0)$) can be very complicated.

From this point of view, the present paper establishes results for two types of featurization maps: in the linear model $\phi(z_i)\ \Sigma^{1/2} z_i$, and in the nonlinear model $\phi(\bar{z}_i) = \varphi(W\bar{z}_i)$ where $W \in \mathbb{R}^{p \times d}$ and $\varphi: \mathbb{R} \to \mathbb{R}$ is applied componentwise. While both models are significantly simpler than the featurization map $\phi(z) = \nabla_\theta f(z; \theta_0)$ for a multilayer neural network, our results imply that certain universality phenomena hold in the proportional asymptotics n, p, $d \rightarrow \infty$, with $n \asymp p \asymp d$. Namely, under the assumption of z_i with independent coordinates in the linear model, and $\bar{z}_i \sim N(0, I_d)$ in the nonlinear model, we prove that:

1. If the activation function φ is "purely nonlinear" (in a sense to be made precise below), then the risk of the nonlinear model is asymptotically equal to the one of the linear model with $\Sigma = I_p$.

2. For more general activations φ, we compute explicitly the asymptotics of the variance of the nonlinear model. This does not coincide with the variance in the isotropic model, but depends on φ only through the size of the linear component of φ (to be defined below). Once more, the details of the activation function do not matter.

After the present work appeared, the analysis of the nonlinear model was generalized in [49], which obtained the asymptotics of the risk for general activations, under a nonparametric model for the responses y_i. This required computing the bias term beyond purely nonlinear activations, and hence solving several technical challenges. The results of [49] confirmed that universality extends beyond purely nonlinear activations. For general activations, the nonlinear model with isotropic $\bar{z}_i \in \mathbb{R}^d$ is asymptotically equivalent to a linear model with anisotropic $x_i \in \mathbb{R}^p$ (analogous to the latent space model of Section 5.2). It is currently an open problem to which extent universality applies beyond the proportional regime.

Let us emphasize that universality is not expected to hold for any distribution of the data (y_i, z_i), and for any function f. In particular, we not expect it to hold when z_i is low-dimensional. This is quite obvious from the proof of Theorem 8, and consistent with the findings of [56], which point at a qualitatively different behavior for interpolating methods in low dimension.

[2] Understanding the bias induced by gradient-based algorithms on fully nonlinear models is a broadly open problem, which has attracted considerable attention recently; see, for example, [34, 35].

Finally, the correspondence outlined above only holds in a certain “lazy training” regime, in which network weights do not change much during training, More generally, in a neural network, the feature representation and the regression function or classifier are learned *simultaneously*. In terms of the first-order Taylor expansion (3) this means that θ_0 depends itself on the data, and hence the feature vectors $x_i = \nabla_\theta f(z_i; \theta_0)$ are not merely observed but trained. Learning the feature map could significantly change some aspects of the behavior of an interpolator. (See, for instance, Chapter 9 of Goodfellow et al. [33], and also Chizat and Bach [19], Zhang et al. [72], which emphasize the importance of learning the representation.)

1.3. Related work.

The present work connects to and is motivated by the recent interest in interpolators in machine learning [8, 10, 11, 28, 47, 48]. Several authors have argued that minimum ℓ_2 norm least squares regression captures the basic behavior of deep neural networks, at least in early (lazy) training [4, 19, 24, 25, 39, 44, 73]. The connection between neural networks and kernel ridge regression arises when the number of hidden units diverges. The infinite width limit was also studied (beyond the linearized regime) in [18, 50, 59, 62].

Interpolation has a long history in signal processing, where it is a method of choice to reconstruct a subsampled signal. The overparametrized regime corresponds to the use of over-complete dictionaries, and the minimum-ℓ_2 norm criterion was used for selecting a specific interpolator [21]. It was subsequently recognized that sparsity promoting interpolators provide better data representations [16].

Ridge regression with random designs has been studied in the past. Dicker [22] considers a model in which the covariates are isotropic Gaussian $x_i \sim N(0, I_p)$ and computes the asymptotic risk of ridge regression in the proportional asymptotics $p, n \to \infty$, with $p/n \to \gamma \in (0, \infty)$. Dobriban and Wager [23] generalize these results to $x_i = \Sigma^{1/2} z_i$, where z_i has independent entries with bounded 12th moment.

Recently, Advani and Saxe [2] study the effect of early stopping and ridge regularization in a model with isotropic Gaussian covariates $x_i \sim N(0, I_p)$, again focusing on the proportional asymptotics p, n, with $p/n \to \gamma \in (0, \infty)$. They show that this simple model reproduces several phenomena observed in neural networks training. The same model is reconsidered in concurrent work by Belkin et al. [9], who obtain exact results for the expected risk of min-norm regression, relying on the jointly Gaussian distribution of (y_i, x_i). We contribute to this line of work by extending the analysis to general covariance structures, non-Gaussian covariates and to misspecified models. As we will see, these generalizations allow to produce examples for which the global minimum of the risk is achieved in the overparametrized regime $\gamma > 1$.

The importance of the relation between the coefficient vector β and the eigenvectors of Σ was emphasized by Kobak et al. [42] and Bartlett et al. [6]. These papers point out—under different asymptotic settings—that $\lambda = 0+$ (i.e., min-norm regression) can be optimal or nearly optimal. After a preprint of this paper appeared, Wu and Xu [69] and Richards et al. [57] generalized our earlier results to cover the case in which β is potentially aligned with

Σ. We review in further detail these important generalizations in Section 4. We contribute to this line of work by obtaining nonasymptotic approximations for the risk, with explicit and nearly optimal error bounds. These hold under weaker assumptions on the geometry of (Σ, β) than the results of [57, 69].

High-dimensional regression under factor models for the covariates was recently studied by Bunea et al. [14], Bing et al. [12]. These models are related to the latent space model of Section 5.2 and present results complementary to ours.

For the nonlinear model, the random matrix theory literature is much sparser, and focuses on the related model of kernel random matrices, namely, symmetric matrices of the form $K_{ij} = \varphi\left(z_i^T z_j\right)$. El Karoui [26] studied the spectrum of such matrices in a regime in which φ can be approximated by a linear function (for $i \neq j$), and hence the spectrum converges to a rescaled Marchenko–Pastur law. This approximation does not hold for the regime of interest here, which was studied instead by Cheng and Singer [17] (who determined the limiting spectral distribution) and Fan and Montanari [27] (who characterized the extreme eigenvalues). The resulting eigenvalue distribution is the free convolution of a semicircle law and a Marchenko–Pastur law. In the current paper, we must consider asymmetric (rectangular) matrices $x_{ij} = \varphi\left(w_j^T z_i\right)$, whose singular value distribution was recently computed by Pennington and Worah [55], using the moment method. Unfortunately, the prediction variance depends on both the singular values and vectors of this matrix. In order to address this issue, we apply the leave-one out method of Cheng and Singer [17] to compute the asymptotics of the resolvent of a suitably extended matrix. We then extract the information of interest from this matrix. After appearance of a preprint of this paper, Mei and Montanari [49] extended the results presented here, to obtain a complete characterization of the risk for the nonlinear random features model.

Let us finally mention that the universality (or "invariance") phenomenon is quite common in random matrix theory [64]. In the context of kernel inner product random matrices, it appears (somewhat implicitly) in [17] and (more explicitly) in [27]. After a first appearance of this manuscript, universality has been investigated in the context of neural networks in several papers [1, 29, 31, 38, 49, 52].

1.4. Outline.

Section 2 provides important background. Sections 3–7 consider the linear model, focusing on isotropic features, correlated features, misspecified models, ridge regularization and cross-validation, respectively. Section 8 covers the nonlinear model case. Nearly all proofs are deferred until the Appendix.

2. Preliminaries.

We describe our setup and gather a number of important preliminary results.

2.1. Data model and risk.

Assume we observe training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ from the model of equations (1), (2). We collect the responses in a vector $y \in \mathbb{R}^n$, and the features in a matrix $X \in \mathbb{R}^{n \times p}$ (with rows $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$).

Consider a test point $x_0 \sim P_x$, independent of the training data. For an estimator $\hat{\beta}$ (a function of the training data X, y), we define its out-of-sample prediction risk (or simply, risk) as

$$R_X(\hat{\beta}; \beta) = \mathbb{E}\left[\left(x_0^T\hat{\beta} - x_0^T\beta\right)^2 \mid X\right] = \mathbb{E}\left[\|\hat{\beta} - \beta\|_\Sigma^2 \mid X\right],$$

where $\|x\|_\Sigma^2 = x^T \Sigma x$. Note that our definition of risk is conditional on X (as emphasized by our notation R_X). Note also that we have the bias-variance decomposition

$$R_X(\hat{\beta}; \beta) = \underbrace{\|\mathbb{E}(\hat{\beta} \mid X) - \beta\|_\Sigma^2}_{B_X(\hat{\beta}; \beta)} + \underbrace{\mathrm{Tr}[\mathrm{Cov}(\hat{\beta} \mid X)\Sigma]}_{V_X(\hat{\beta}; \beta)}. \tag{5}$$

2.2. Ridgeless least squares.

We consider the minimum ℓ_2 norm (min-norm) least squares regression estimator, of y on X, defined by

$$\hat{\beta} = \arg\min\left\{\|b\|_2 : b \text{ minimizes } \|y - Xb\|_2^2\right\}. \tag{6}$$

This can be equivalently written as $\hat{\beta} = \left(X^T X\right)^+ X^T y$, where $(X^T X)^+$ is the pseudoinverse of $X^T X$. An alternative name for (6) is the "ridgeless" least squares estimator, motivated by the fact that $\hat{\beta} = \lim_{\lambda \to 0^+} \hat{\beta}_\lambda$, where $\hat{\beta}_\lambda$ denotes the ridge regression estimator:

$$\hat{\beta}_\lambda = \arg \min_{b \in \mathbb{R}^p} \left\{\frac{1}{n}\left\|y - Xb\right\|_2^2 + \lambda\left\|b\right\|_2^2\right\}, \tag{7}$$

or, equivalently, $\hat{\beta}_\lambda = \left(X^T X + n\lambda I\right)^{-1} X^T y$.

When X has full column rank the min-norm least squares estimator reduces to $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, the usual least squares estimator. When X has rank n, importantly, this estimator interpolates the training data: $y_i = x_i^T \hat{\beta}$, for $i = 1, \ldots, n$.

Lastly, the following is a well-known fact that connects the min-norm least squares solution to gradient descent (as referenced in the Introduction).

Proposition 1. *Initialize* $\beta^{(0)}$ 0, *and consider running gradient descent on the least squares loss, yielding iterates*

$$\beta^{(k)} = \beta^{(k-1)} + tX^T\left(y - X\beta^{(k-1)}\right), \quad k = 1, 2, 3, \ldots,$$

where we take $0 < t \le 1/\lambda_{\max}(X^T X)$ (*and* $\lambda_{\max}(X^T X)$ *is the largest eigenvalue of* $X^T X$). *Then* $\lim_{k \to \infty}\beta^{(k)} = \hat{\beta}$, *the min-norm least squares solution in* (6).

Proof. The choice of step size guarantees that $\beta^{(k)}$ converges to a least squares solution as $k \to \infty$, call it $\tilde{\beta}$. Note that $\beta^{(k)}$, $k = 1, 2, 3, \ldots$ all lie in the row space of X; therefore, $\tilde{\beta}$ must also lie in the row space of X; and the min-norm least squares solution $\hat{\beta}$ is the unique least squares solution with this property. □

2.3. Bias and variance.

We recall expressions for the bias and variance of the min-norm least squares estimator, which are standard.

Lemma 1. *Under the model* (1), (2), *the min-norm least squares estimator* (6) *has bias and variance*

$$B_X(\hat{\beta}; \beta) = \beta^T \Pi \Sigma \Pi \beta \quad \text{and} \quad V_X(\hat{\beta}; \beta) = \frac{\sigma^2}{n}\mathrm{Tr}\left(\hat{\Sigma}^+ \Sigma\right),$$

where $\hat{\Sigma} = X^T X/n$ *is the* (*uncentered*) *sample covariance of* X, *and* $\Pi = I - \hat{\Sigma}^+ \hat{\Sigma}$ *is the projection onto the null space of* X.

Proof. As $\mathbb{E}(\hat{\beta} \mid X) = \left(X^T X\right)^+ X^T X\beta = \hat{\Sigma}^+ \hat{\Sigma}\beta$ and $\mathrm{Cov}(\hat{\beta} \mid X) = \sigma^2\left(X^T X\right)^+ X^T \times X\left(X^T X\right)^+ = \sigma^2 \hat{\Sigma}^+/n$, the bias and variance expressions follow from plugging these into their respective definitions. □

2.4. Underparametrized asymptotics.

We consider an asymptotic setup where $n, p \to \infty$, in such a way that $p/n \to \gamma \in (0, \infty)$. Recall that when $\gamma < 1$, we call the problem underparametrized; when $\gamma > 1$, we call it overparametrized. Here, we recall the risk of the min-norm least squares estimator in the underparametrized case. The rest of this paper focuses on the overparametrized case.

The following is a known result in random matrix theory, and can be found in Chapter 6 of Serdobolskii [61]. It can also be found in the wireless communications literature; see Chapter 4 of Tulino and Verdu [67].

Proposition 2. *Assume the model* (1), (2), *and assume* $x \sim P_x$ *is of the form* $x = \Sigma^{1/2}z$, *where* z *is a random vector with i.i.d. entries that have zero mean, unit variance and a finite* 4*th moment, and* Σ *is a* (*sequence of*) *deterministic positive definite matrix, such that* $\lambda_{\min}(\Sigma) \ge c > 0$, *for all* n, p *and a constant* c (*here* $\lambda_{\min}(\Sigma)$ *is the smallest eigenvalue of* Σ). *Then as* $n, p \to \infty$, *such that* $p/n \to \gamma < 1$, *the risk of the least squares estimator* (6) *satisfies, almost surely*

$$\lim_{n \to \infty} R_X(\hat{\beta}; \beta) = \sigma^2 \frac{\gamma}{1 - \gamma} .$$

As it can be seen from the last proposition, in the underparametrized case the risk is just variance. In contrast, in the overparametrized case, the bias $B_X(\hat{\beta}; \beta) = \beta^T \Pi \Sigma \Pi \beta$ is nonzero, because $\prod$ is. This will be the focus of the next sections.

3. Isotropic features.

We begin by considering the simpler case in which $\Sigma = I$. In this case, the limiting bias is relatively straightforward to compute and depends on β only through $\|\beta\|_2^2$. In Section 4, we generalize our analysis and study the dependence of the prediction risk on the geometry of Σ and β.

3.1. Limiting bias.

As mentioned above, in the isotropic case the risk depends β only on through $r^2 = \|\beta\|_2^2$. To give some intuition as to why this is true, consider the special case where X has i.i.d. entries from $N(0, 1)$. By rotational invariance, for any orthogonal $U \in \mathbb{R}^{p \times p}$, the distribution of X and XU is the same. Thus

$$\begin{aligned} B_X(\hat{\beta}; \beta) &= \beta^T \left(I - \left(X^T X\right)^+ X^T X\right) \beta \\ &\overset{d}{=} \beta^T \left(I - U^T \left(X^T X\right)^+ U U^T X^T X U\right) \beta \\ &= r^2 - (U\beta)^T \left(X^T X\right)^+ X^T X (U\beta) . \end{aligned}$$

Choosing U so that $U\beta = re_i$, the ith standard basis vector, then averaging over $i = 1, \ldots, p$, yields

$$\mathbb{E} B_X(\hat{\beta}; \beta) = r^2 \mathbb{E}\left[1 - \mathrm{Tr}\left(\left(X^T X\right)^+ X^T X\right)/p\right] = r^2 (1 - n/p) .$$

It is possible to show that, $B_X(\hat{\beta}; \beta)$ concentrates around its expectation and, therefore, $B_X(\hat{\beta}; \beta) \to r^2(1 - 1/\gamma)$, almost surely. This is stated formally in the next section.

3.2. Limiting risk.

As the next result shows, the independence of the risk on β is still true outside of the Gaussian case, provided the features are isotropic. The next result can be proved as a corollary of the more general Theorem 3 below. We give a simpler self-contained proof using a theorem of Rubio and Mestre [60] in Appendix A.4.2.

THEOREM 1. *Assume the model* (1), (2), *where* $x_i \sim P_x$ *has independent entries with zero mean, unit variance. Further assume that either of these conditions hold for* $x \sim P_x$: (i) *the entries* $(x_j)_{j \le p}$ *have uniformly bounded moments of all order* $\mathbb{E}\left[\left|x_j\right|^k\right] \le C_k$ *for all k and some* *constants* C_k; (ii) *entries* $(x_j)_{j\le p}$ *are identically distributed and have finite moment of order* $4+\delta$, $\mathbb{E}\left\{\left|x_j\right|^{4+\delta}\right\} \le C$, *for some* C, $\delta > 0$. *Also assume that* $\|\beta\|_2^2 = r^2$ *for all n, p. Then for the min-norm least squares estimator* $\hat{\beta}$ *in* (6), *as* $n, p \to \infty$, *such that* $p/n \to \gamma \in (1, \infty)$, *it holds almost surely that*

$$B_X(\hat{\beta}; \beta) \to r^2\left(1 - \frac{1}{\gamma}\right), \tag{8}$$

$$V_X(\hat{\beta}; \beta) \to \sigma^2 \frac{1}{\gamma - 1}\,. \tag{9}$$

Hence, summarizing with Proposition 2, *we have*

$$R_X(\hat{\beta}; \beta) \to \begin{cases} \sigma^2 \dfrac{\gamma}{1-\gamma} & for\ \gamma < 1, \\ r^2\left(1 - \dfrac{1}{\gamma}\right) + \sigma^2 \dfrac{1}{\gamma - 1} & for\ \gamma > 1. \end{cases} \tag{10}$$

For $\gamma \in (0, 1)$, there is no bias, and the variance increases with γ. For $\gamma \in (1, \infty)$, the bias increases with γ, and the variance decreases with γ. Let SNR = r^2/σ^2. Observe that the risk of the null estimateor $\tilde{\beta} = 0$ is r^2, which we hence call the null risk. The following facts are immediate from the form of the risk curve in (10). See Figure 2 for an accompanying plot when SNR varies from 1 to 5.

1. On (0, 1), the least squares risk $R(\gamma)$ is better than the null risk if and only if $\gamma < \frac{\mathrm{SNR}}{\mathrm{SNR}+1}$.

2. On $(1, \infty)$, when SNR ≤ 1, the min-norm least squares risk $R(\gamma)$ is always worse than the null risk. Moreover, it is monotonically decreasing, and approaches the null risk (from above) as $\gamma \to \infty$.

3. On $(1, \infty)$, when SNR > 1, the min-norm least squares risk $R(\gamma)$ beats the null risk if and only if $\gamma > \frac{\mathrm{SNR}}{\mathrm{SNR}-1}$. Further, it has a local minimum at $\gamma = \frac{\sqrt{\mathrm{SNR}}}{\sqrt{\mathrm{SNR}} - 1}$, and approaches the null risk (from below) as $\gamma \to \infty$.

3.3. Limiting ℓ_2 norm.

Calculation of the limiting ℓ_2 norm of the min-norm least squares estimator is quite similar to the study of the limiting risk in Theorem 1 and, therefore, we state the next result without proof.

Corollary 1. *Assume the conditions of Theorem* 1. *Then as* $n, p \to \infty$, *such that* $p/n \to \gamma$, *the squared* ℓ_2 *norm of the min-norm least squares estimator* (6) *satisfies*, *almost surely*

$$\mathbb{E}\left[\|\hat{\beta}\|_2^2 \mid X\right] \to \begin{cases} r^2 + \sigma^2 \frac{\gamma}{1-\gamma} & for\ \gamma < 1, \\ r^2 \frac{1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1} & for\ \gamma > 1. \end{cases}$$

We can see that the limiting norm, as a function of γ, has a somewhat similar profile to the limiting risk in (10): it is monotonically increasing for γ ε (0, 1), diverges at the interpolation boundary and is monotonically decreasing for (1, ∞). These findings confirm the intuition given in the Introduction: as γ grows above the interpolation threshold, the minimum norm interpolator becomes increasingly simpler, in the sense of having smaller ℓ_2 norm.

4. Correlated features.

We broaden the scope of our analysis from the last section, where we examined isotropic features. In this section, we take $x \sim P_x$ to be of the form $x = \Sigma^{1/2}z$, where z is a random vector with independent entries that have zero mean and unit variance, and Σ is arbitrary (but still deterministic and positive definite).

The risk of min-norm regression depends on the geometry of Σ and β. Denote by $\Sigma = \sum_{i=1}^{p} s_i v_i v_i^T$ he eigenvalue decomposition of Σ with $s_1 \geq s_2 \geq \cdots \geq s_p \geq 0$. The geometry of the problem is captured by the sequence of eigenvalues $(s_1, \ldots, s_p)$, and by the coefficients of β in the basis of eigenvectors $(\langle v_1, \beta\rangle, \ldots, \langle v_p, \beta\rangle)$. We encode these via two probability distributions on $\mathbb{R}_{\geq 0}$:

$$\widehat{H}_n(s) := \frac{1}{p} \sum_{i=1}^{p} 1_{\{s \geq s_i\}}, \quad \widehat{G}_n(s) = \frac{1}{\|\beta\|_2^2} \sum_{i=1}^{p} \langle \beta, v_i \rangle^2 1_{\{s \geq s_i\}}. \tag{11}$$

We next state our assumptions about the data distribution: our results will be uniform with respect to the (large) constants M, $\{C_k\}_{k \geq 2}$ appearing in this assumption.

Assumption 1. The covariates vector $x \sim P_x$ is of the form $x = \Sigma^{1/2}z$, where defining $\widehat{H}_n$ as per equation (11), we have:

- **a.** The vector $z = (z_1, \ldots, z_p)$ has independent (not necessarily identically distributed) entries with $\mathbb{E}\{z_i\} = 0$, $\mathbb{E}\left\{z_i^2\right\} = 1$, and $\mathbb{E}\left\{|z_i|^k\right\} \leq C_k < \infty$ for all $i \leq p$, $k \geq 2$.
- **b.** $s_1 = \|\Sigma\|_{op} \leq M, \int s^{-1} d\widehat{H}_n(s) < M$.
- **c.** $|1 - (p/n)| \geq 1/M$, $1/M \leq p/n \leq M$.

Condition (a) bounds the tail probabilities on the covariates. Requiring finite moment of all orders is useful to get strong bounds on the deviations of the risk from its predicted value. As discussed below, bounds on the first few moments are sufficient if we are satisfied with weaker probability bounds.

Conditions (b) requires the eigenvalues of Σ to be bounded, and not to accumulate[3] near 0. For the analysis of min-norm interpolation, we will add the additional assumption that the minimum eigenvalue of Σ is bounded away from zero. However, condition (b) is sufficient for the analysis of ridge regression in Section 6.

Finally, as our statements are nonasymptotic, we do not assume p/n to converge to a value. However, condition (c) requires p/n to be bounded and bounded away from the interpolation threshold $p/n = 1$.

4.1. Prediction risk.

Definition 1 (Predicted bias and variance: min-norm regression). Let $\widehat{H}_n$ be the empirical distribution of eigenvalues of Σ, and $\widehat{G}_n$ the reweighted distribution as per equation (11). For $\gamma \in \mathbb{R}_{>0}$, define $c_0 = c_0\left(\gamma, \widehat{H}_n\right) \in \mathbb{R}_{>0}$ to be the unique nonnegative solution of

$$1 - \frac{1}{\gamma} = \int \frac{1}{1 + c_0\gamma s} d\widehat{H}_n(s)\,. \qquad (12)$$

We then define the predicted bias and variance by

$$\mathscr{B}\left(\widehat{H}_n, \widehat{G}_n, \gamma\right) := \left\|\beta\right\|_2^2 \left\{1 + \gamma c_0 \frac{\int \frac{s^2}{(1 + c_0\gamma s)^2} d\widehat{H}_n(s)}{\int \frac{s}{(1 + c_0\gamma s)^2} d\widehat{H}_n(s)}\right\} \cdot \int \frac{s}{(1 + c_0\gamma s)^2} d\widehat{G}_n(s), \qquad (13)$$

$$\mathscr{V}\left(\widehat{H}_n, \gamma\right) := \sigma^2 \gamma c_0 \frac{\int \frac{s^2}{(1 + c_0\gamma s)^2} d\widehat{H}_n(s)}{\int \frac{s}{(1 + c_0\gamma s)^2} d\widehat{H}_n(s)}\,. \qquad (14)$$

Note that evaluating $\mathscr{B}(H, G, \gamma)$, $\mathscr{V}(H, \gamma)$ numerically is relatively straightforward, with the most complex part being the solution of equation (12). The next theorem establishes that—under suitable technical conditions—the functions $\mathscr{B}$, $\mathscr{V}$ characterize the test error. Similar theorems were proved in [57, 69], which generalized an earlier version of this manuscript to account for the geometry of (Σ, β).

Theorem 2. *Assume the data model* (1), (2) *and that the covariates distribution satisfies Assumption* 1. *Further assume* $s_p = \lambda_{\min}(\Sigma) > 1/M$. *Define* $\gamma = p/n$ *and let* $\hat{\beta}$ *be the min-norm least squares estimator in* equation (6).

Then for any constant $D > 0$ (*arbitrarily large*) *there exist* $C = C(M, D)$ *such that, with probability at least* $1 - Cn^{-D}$ *the following hold*:

[3] The latter assumption could have been further relaxed, by requiring only p_+ of the p eigenvalues to be nonvanishing and to satisfy the other conditions. This would require to redefine γ as p_+/n.

$$R_X(\hat{\beta};\beta) = B_X(\hat{\beta};\beta) + V_X(\hat{\beta};\beta), \tag{15}$$

$$\left|B_X(\hat{\beta};\beta) - \mathscr{B}\left(\widehat{H}_n, \widehat{G}_n, \gamma\right)\right| \le \frac{C\|\beta\|_2^2}{n^{1/7}}, \tag{16}$$

$$\left|V_X(\hat{\beta};\beta) - \mathscr{V}\left(\widehat{H}_n, \gamma\right)\right| \le \frac{C}{n^{1/7}}, \tag{17}$$

where $\mathscr{B}$ *and* $\mathscr{V}$ *are given in Definition* 2, *and the first identity is just the general bias-variance decomposition of* equation (5).

The proof of this theorem is deferred to Section A.2.

Remark 1. The order of the error bound in equations (16), (17) is not optimal: a central limit theorem heuristics suggests the deterministic approximation to be accurate up to an error of order $n^{-1/2}$. Indeed, we are able to establish the optimal order in the case of ridge regression; see Theorem 5.

Let us emphasize that, while suboptimal, the $O(n^{-1/7})$ terms in equation (16), (17) are often negligible as compared to the leading terms $\mathscr{B}\left(\widehat{H}_n, \widehat{G}_n, \gamma\right)$, $\mathscr{V}\left(\widehat{H}_n, \gamma\right)$. In particular, as stated in Theorem 3 below, whenever $p, n \to \infty$ with p/n $\gamma \in (0, \infty)$ and the two probability measures $\widehat{H}_n$, $\widehat{G}_n$ converge weakly to finite limits H, G, $\mathscr{B}\left(\widehat{H}_n, \widehat{G}_n, \gamma\right)$, $\mathscr{V}\left(\widehat{H}_n, \gamma\right)$ remain bounded away from zero, and hence dominate the $O(n^{-1/7})$ errors.

Notice that this is in particular the case for isotropic features, and Theorem 1 (under the stronger moment assumption on the z_i's) is recovered as a corollary of Theorem 2.

Remark 2. Note that Theorem 2 establishes deterministic approximations for the bias and variance, that are valid at finite n, p. The overparametrization ratio $\gamma = p/n$ is a nonasymptotic quantity, and the error bounds are uniform, that is, depend only on the constant M. This is to be contrasted with the asymptotic setting of [57, 69]. Both of these papers assume a sequence of regression problems with $n, p \to \infty$, $p/n \to \gamma$, and obtain an asymptotically exact expression for the risk.

In order for the asymptotics to make sense, additional assumptions are required by [57, 69]. In [57], this is achieved by assuming β to be random with $\mathbb{E}\left[\beta\beta^T\right] = r^2\Phi(\Sigma)/d$ for a certain (deterministic) function $\Phi: \mathbb{R} \to \mathbb{R}$ (promoted to a function on matrices in the usual way). In addition, the empirical spectral distribution of Σ is assumed to converge. To state the assumptions in [69], recall that $(s_i)_{i \le p}$ are the eigenvalues of Σ, and denote by $b_i = p \cdot \left(v_i^T\beta\right)^2$ the projection of β onto the eigenvectors of Σ. Then [69] assumes that the joint empirical distribution $p^{-1}\sum_{i=1}^{p}\delta_{s_i, b_i}$ converges weakly as $n, p \to \infty$.

Technically, [57, 69] apply asymptotic random matrix theory results, such as [43], while we have to take a longer detour to exploit nonasymptotic results established in [41]. We believe that the nonasymptotic approach provides more concrete and accurate statements.

Theorem 2 also implies asymptotic predictions under minimal assumptions. In particular, if the two probability measures $\widehat{H}_n$, $\widehat{G}_n$ converge weakly to probability measures H, G on $[0, \infty)$, then we obtainn[4] $B_X(\hat{\beta};\beta)/\|\beta\|^2 \to \mathscr{B}(H,G,\gamma)$, $V_X(\hat{\beta};\beta) \to \mathscr{V}(H,\gamma)$. Hence, the first part of the following asymptotic statement follows immediately from Theorem 2 by taking the limit $n, p \to \infty$ in equations (16), (17) (and using Borel–Cantelli to obtain almost sure convergence).

Theorem 3. *Consider the setting of Theorem* 2. *Further assume* $n, p \to \infty$, $p/n \to \gamma \in (0, \infty)$, $\widehat{H}_n \Rightarrow H$, $\widehat{G}_n \Rightarrow G$. *Define* $\mathscr{B}_1(H, G, \gamma)$ *as in* equation (13), *with* $\|\beta\|_2^2$ *replaced by* 1. *Then, almost surely*

$$\frac{1}{\|\beta\|_2^2} B_X(\hat{\beta};\beta) \to \mathscr{B}_1(H,G,\gamma), \quad V_X(\hat{\beta};\beta) \to \mathscr{V}(H,\gamma)\,. \tag{18}$$

with $\mathscr{B}_1(H, G, \gamma), \mathscr{V}(H, \gamma) > 0$ *strictly.*

The same conclusion holds if instead of Assumption 1(*a*), *the coordinates of z*, $(z_i)_{i \le p}$ *are i.i.d. and satisfy the conditions* $\mathbb{E}z_i = 0$, $\mathbb{E}\left(z_i^2\right) = 1$, $\mathbb{E}\left(|z_i|^{4+\delta}\right) \le C < \infty$.

The last part of this theorem (under the weaker moment condition $\mathbb{E}\left(|z_i|^{4+\delta}\right) < C$) is proved via a truncation argument in Appendix A.1.4. For carrying out this argument, we make use of estimates on the norm of random matrices that are available only for the case of identically distributed entries $(z_i)_{i \le p}$.

As pointed out above the condition $\widehat{H}_n \Rightarrow H$, $\widehat{G}_n \Rightarrow G$ (here ⇒ denotes weak convergence) is strictly weaker than the condition assumed in [69] to establish asymptotic results. Further, we require weaker moment conditions.

In the next sections, we illustrate the role of the geometry of β, Σ by considering two models for which $\widehat{H}_n \Rightarrow H$, $\widehat{G}_n \Rightarrow G$ as $n, p \to \infty$. First, we consider the case $G = H$, which we refer to as “equidistributed”: the components of β are roughly equally distributed along the eigenvectors of Σ. In this case, there is no special relation between β and Σ.

As a further application, we consider a latent space model in which β is aligned with the top eigenvectors of Σ. This can be regarded as a misspecified model, and is therefore presented in Section 5.2 below.

[4] Indeed all the expressions in equations (12), (13), (14) are continuous in $\widehat{H}_n$, $\widehat{G}_n$ (with respect to the weak topology) since they are expectations of bounded continuous functions.

4.2. Equidistributed coefficients.

In this section, we assume $G = H$, $\|\beta\|_2 \to r$, and $p/n \to \gamma$. One way to generate β satisfying this condition is to draw it uniformly at random on the p-dimensional sphere of radius $\|\beta\|_2 = r$. In this case, the conditions of Theorem 2 (or Theorem 3), hold with $\widehat{G}_n \to G = H$.

Corollary 2. *Under the assumptions of Theorem* 3, *further assume $G = H$, $\|\beta\|^2 \to r^2$. Then for $n, p \to \infty$, with $p/n \to \gamma > 1$, almost surely*

$$B_X(\hat\beta;\beta) \to \mathscr{B}_{\text{equi}}(H,\gamma) := \frac{r^2}{c_0(H,\gamma)\gamma^2}, \tag{19}$$

$$V_X(\hat\beta;\beta) \to \mathscr{V}_{\text{equi}}(H,\gamma) := \mathscr{V}(H,\gamma)\,. \tag{20}$$

As a special case, we can revisit the isotropic case $\Sigma = I$, which results in $dH = \delta_1$. In this case, $c_0(H, \gamma) = \gamma(\gamma - 1)$ yielding immediately $\mathscr{B}_{\text{equi}}(H, \gamma) = 1 - \gamma^{-1}$ and $\mathscr{V}_{\text{equi}}(H, \gamma) = 1/(\gamma - 1)$.

4.3. Limiting ℓ_2 norm.

Again, as in the isotropic case, analysis of the limiting ℓ_2 norm is similar to analysis of the risk in Theorem 2. We give the next result without proof, as it is an immediate generalization of previous results.

Corollary 3. *Under the assumptions of Theorem* 3, *further assume $\|\beta\|^2 \to r^2$, and let $c_0 = c_0(H, \gamma)$ be defined as there. Then as $n, p \to \infty$, such that $p/n \to \gamma$, the min-norm least squares estimator* (6) *satisfies, almost surely*

$$\left\|\hat\beta\right\|_2^2 \to \begin{cases} r^2 + \sigma^2\dfrac{\gamma}{1-\gamma}\displaystyle\int \frac{1}{s}\,dH(s) & for\ \gamma < 1, \\ \displaystyle\int \frac{c_0\gamma s}{1+c_0\gamma s}\,dG(s) + c_0\gamma\sigma^2 & for\ \gamma > 1, \end{cases} \tag{21}$$

4.4. Benign overfitting.

Theorem 2 (and its generalization to nonzero ridge regularization, Theorem 5) can be used to delineate regimes in which interpolation is statistically optimal or nearly optimal. "Statistical optimality" can be given different meanings in this context. Section 6.2 explores optimality in the context of the latent space model and shows that (in certain regimes) $R_X(\hat\beta_0;\beta) \le (1 + o_n(1))R_X(\hat\beta_\lambda;\beta)$ for any $\lambda > 0$: min-norm interpolation is optimal up to subleading factors.

A different notion of optimality was explored (in concurrent work) in [6] and [66]. In these works, optimality is understood to hold up to constant multiplicative factors. A weaker notion is also considered whereby the interpolator is only required to be consistent. The term "benign overfitting" was proposed in [6] for such phenomena.

Theorem 2 can be used to establish upper bounds that are closely related to the ones of [6, 66] and in particular imply benign overfitting. As an example, the following bound was proven in joint work with Peter Bartlett and Alexander Rakhlin [7]. Recall that $s_1 \geq s_2 \geq \cdots \geq s_p$ denote the eigenvalues of the covariance Σ in decreasing order, and we define the effective rank $r_k(\Sigma) = \sum_{i=k+1}^{p} \lambda_i / \lambda_{k+1}$. We also denote by $\beta_{\leq k}$ the projection of β onto the top k eigenvectors of Σ and $\beta_{>k} = \beta - \beta_{\leq k}$.

Corollary 4 ([7]). *Under the assumptions of Theorem* 2, *further assume that there exists an integer k and a constant* $c_* > 0$ *such that* $r_k(\Sigma) \geq (1 + c_*)n$. *Then there exists a constant* $C = C(M, D)$ *such that, with probability at least* $1 - Cn^{-D}$,

$$B_X(\hat{\beta}; \beta) \leq 4\left(\frac{1}{n}\sum_{i=k+1}^{p} \lambda_i\right)^2 \|\beta_{\leq k}\|_{\Sigma^{-1}}^2 + \|\beta_{>k}\|_{\Sigma}^2 + Cn^{-1/7}, \tag{22}$$

$$V_X(\hat{\beta}; \beta) \leq \frac{2k\sigma^2}{n} + \frac{4n\sigma^2}{c_*}\frac{\sum_{i=k+1}^{p} \lambda_i^2}{\left(\sum_{i=k+1}^{p} \lambda_i\right)^2} + Cn^{-1/7}. \tag{23}$$

This corollary is an immediate consequence of Theorem 2. It upper bounds the excess risk over an “ideal” (underparametrized) estimator that only fits the projection of β onto the top eigenvector of Σ. This excess risk will be small when β is well aligned to the top eigenvectors of Σ and the ratio $\sum_{i=k+1}^{p} \lambda_i^2 / \left(\sum_{i=k+1}^{p} \lambda_i\right)^2$ is small.

This result is analogous to the ones of [6, 66] although not precisely comparable. The upper bound on the variance term in [6] is sharp up to universal multiplicative constants. The upper bound on the bias in [66] depends on the (random) condition number of the component of the bias along less important directions. On the other hand, both of [6, 66] apply to cases in which $\Sigma^{-1/2}x$ does not have independent coordinates. Corollary 4 has a larger additive slack $Cn^{-1/7}$ (which can be improved to $Cn^{-1/2}$ for ridge regression), but a more precise prefactor.

5. Misspecified models.

5.1. Regression with respect to a subset of features.

In this section, we consider a misspecified model, in which the regression function is still linear, but we observe only a subset of the features. Such a setting provides another potential motivation for interpolation: in many problems, we do not know the form of the regression function, and we generate features in order to improve our approximation capacity. Increasing the number of features past the point of interpolation (increasing γ past 1) can now decrease *both* bias and variance.

As such, the misspecified model setting also yields further interesting asymptotic comparisons between the $\gamma < 1$ and $\gamma > 1$ regimes. Recall the isotropic features model of Section 3.2: the risk function in (10) is globally minimized at $\gamma = 0$. This is a consequence of the fact that, in a well-specified linear model at $\gamma = 0$ there is no bias and no variance, and hence no risk. In a misspecified model, we will see that the story can be quite different, and the asymptotic risk can actually attain its *global* minimum on (1, ∞).

5.1.1. Data model and risk.

—Consider, instead of (1), (2), a data model

$$((x_i, w_i), \epsilon_i) \sim P_{x,w} \times P_\epsilon, \quad i = 1, \ldots, n, \tag{24}$$

$$y_i = x_i^T \beta + w_i^T \theta + \epsilon_i, \quad i = 1, \ldots, n, \tag{25}$$

where as before the random draws across $i = 1, \ldots, n$ are independent. Here, we partition the features according to $(x_i, w_i) \in \mathbb{R}^{p+q}$, $i = 1, \ldots, n$, where the joint distribution $P_{x,w}$ is such that $\mathbb{E}((x_i, w_i)) = 0$ and

$$\mathrm{Cov}((x_i, w_i)) = \Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xw} \\ \Sigma_{xw}^T & \Sigma_w \end{bmatrix}.$$

We collect the features in a block matrix $[XW] \in \mathbb{R}^{n \times (p+q)}$ (which has rows $(x_i, w_i) \in \mathbb{R}^{p+q}$, i 1, …, n). We presume that X is observed but W is unobserved, and focus on the min-norm least squares estimator exactly as before in (6), from the regression of y on X (not the full feature matrix $[XW]$).

Given a test point $(x_0, w_0) \sim P_{x,w}$, and an estimator $\hat{\beta}$ (fit using X, y only, and not W), we define its out-of-sample prediction risk as

$$R_X(\hat{\beta}; \beta, \theta) = \mathbb{E}\left[\left(x_0^T \hat{\beta} - \mathbb{E}(y_0 \mid x_0, w_0)\right)^2 \mid X\right] = \mathbb{E}\left[\left(x_0^T \hat{\beta} - x_0^T \beta - w_0^T \theta\right)^2 \mid X\right].$$

Note that this definition is conditional on X, and we are integrating over the randomness not only in ϵ (the training errors), but in the unobserved features W, as well. The next lemma decomposes this notion of risk in a useful way.

Lemma 2. *Under the misspecified model* (24), (25), *for any estimator* $\hat{\beta}$, *we have*

$$R_X(\hat{\beta}; \beta, \theta) = \underbrace{\mathbb{E}\left[\left(x_0^T \hat{\beta} - \mathbb{E}(y_0 \mid x_0)\right)^2 \mid X\right]}_{R_X^*(\hat{\beta}; \beta, \theta)} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}(y_0 \mid x_0) - \mathbb{E}(y_0 \mid x_0, w_0)\right)^2\right]}_{M(\beta, \theta)}.$$

Proof. Simply add an subtract $\mathbb{E}(y_0 \mid x_0)$ inside the square in the definition of $R_X(\hat{\beta}; \beta, \theta)$, then expand, and note that the cross term vanishes because $\mathbb{E}[(\mathbb{E}(y_0 \mid x_0) - \mathbb{E}(y_0 \mid x_0, w_0)) \mid x_0] = 0$. □

The first term $R_X^*(\hat{\beta}; \beta, \theta)$ in the decomposition in Lemma 2 is precisely the risk that we studied previously in the well-specified case, except that the response distribution has changed (due to the presence of the middle term in (25)). We call the second term $M(\beta, \theta)$ in Lemma 2 the *misspecification bias*.

Remark 3. If (x_i, w_i) are jointly Gaussian, then the above expressions simplify and Theorem 2 can be used to characterize the risk $R_X(\hat{\beta}; \beta, \theta)$. In particular, the conditional distribution of w given x is $P_{w \mid x} = N\left(\Sigma_{wx}\Sigma_x^{-1}x, \Sigma_{w \mid x}\right)$ where $\Sigma_{wx} = \Sigma_{xw}^T$, and $\Sigma_{w \mid x} = \Sigma_w - \Sigma_{wx}\Sigma_x^{-1}\Sigma_{wx}^T$. Further, $y = \tilde{\beta}^T x + \tilde{\varepsilon}$, where $\tilde{\beta} = \beta + \Sigma_x^{-1}\Sigma_{xw}\theta$ and $\tilde{\varepsilon} \sim N\left(0, \tilde{\sigma}^2\right)$, $\tilde{\sigma}^2 = \sigma^2 + \theta^T \Sigma_{w \mid x}\theta$. It is then easy to show that the misspecification bias is $M(\beta, \theta) = \theta^T \Sigma_{w|x}\theta$ and the term $R_X^*(\hat{\beta}; \beta, \theta)$ can be approximated using Theorem 2.

In order to discuss some qualitative features, we focus on the simplest possible model by assuming independent covariates.

5.1.2. Isotropic features.

Here, we make the additional simplifying assumption that $(x, w) \sim P_{x,w}$ has i.i.d. entries with unit variance, which implies that $\Sigma = I$. (The case of independent features but general covariances Σ_x, Σ_w is similar, and we omit the details.) Therefore, we may write the response distribution in (25) as

$$y_i = x_i^T \beta + \delta_i, \quad i = 1, \ldots, n,$$

where δ_i is independent of x_i, having mean zero and variance $\sigma^2 + \|\theta\|_2^2$, for $i = 1, \ldots, n$. Denote the total signal by $r^2 = \|\beta\|_2^2 + \|\theta\|_2^2$, and the fraction of the signal captured by the observed features by $\kappa = \|\beta\|_2^2/r^2$. Then $R_X^*(\hat{\beta}; \beta, \theta)$ behaves exactly as we computed previously, for isotropic features in the well-specified setting (Theorem 2 for $\gamma < 1$, and Theorem 1 for $\gamma > 1$), after we make the substitutions:

$$r^2 \mapsto r^2\kappa \quad \text{and} \quad \sigma^2 \mapsto \sigma^2 + r^2(1 - \kappa)\,. \tag{26}$$

Furthermore, we can easily calculate the misspecification bias:

$$M(\beta, \theta) = \mathbb{E}\left(w_0^T \theta\right)^2 = r^2(1 - \kappa)\,.$$

Putting these results together leads to the next conclusion.

Theorem 4. *Assume the misspecified model* (24), (25) *and assume* $(x, w) \sim P_{x,w}$ *has i.i.d. entries with zero mean, unit variance and a finite moment of order* $4 + \delta$, *for some* $\delta > 0$. *Also assume that* $\|\beta\|_2^2 + \|\theta\|_2^2 = r^2$ *and* $\|\beta\|_2^2/r^2 = \kappa$ *for all n, p. Then for the min-norm least squares estimator* $\hat{\beta}$ *in* (6), *as n, p* $\to \infty$, *with p/n* $\to \gamma$, *it holds almost surely that*

$$R_X(\hat{\beta}; \beta, \theta) \to \begin{cases} r^2(1 - \kappa) + \left(r^2(1 - \kappa) + \sigma^2\right)\dfrac{\gamma}{1 - \gamma} & for\ \gamma < 1, \\ r^2(1 - \kappa) + r^2\kappa\left(1 - \dfrac{1}{\gamma}\right) + \left(r^2(1 - \kappa) + \sigma^2\right)\dfrac{1}{\gamma - 1} & for\ \gamma > 1. \end{cases}$$

We remark that, in the independence setting considered in Theorem 4, the dimension q of the unobserved feature space does not play any role: we may equally well take $q = \infty$ for all n, p (i.e., infinitely many unobserved features).

The components of the limiting risk from Theorem 4 are intuitive and can be interpreted as follows. The first term $r^2(1 - \kappa)$ is the misspecification bias (irreducible). The second term, which we deem as 0 for $\gamma < 1$ and $r^2\kappa(1 - 1/\gamma)$ for $\gamma > 1$, is the bias. The third term, $r^2(1 - \kappa)\gamma/(1 - \gamma)$ for $\gamma < 1$ and $r^2(1 - \kappa)/(\gamma - 1)$ for $\gamma > 1$, is what we call the *misspecification bias*: the inflation in risk due to unobserved features, when we take $\mathbb{E}(y_0 \mid x_0)$ to be the target of estimation. The last term, $\sigma^2\gamma/(1 - \gamma)$ for $\gamma < 1$ and $\sigma^2/(\gamma 1)$ for $\gamma > 1$, is the variance itself.

5.1.3. Polynomial approximation bias.—

Since adding features should generally improve our approximation capacity, it is reasonable to model $\kappa = \kappa(\gamma)$ as an increasing function of γ. To get an idea of the possible shapes taken by the asymptotic risk curve from Theorem 4, we consider the example of a *polynomial decay* for the approximation bias,

$$1 - \kappa(\gamma) = (1 + \gamma)^{-a}, \tag{27}$$

for some $a > 0$. In this case, the limiting risk in the isotropic setting, from Theorem 4, becomes

$$R_a(\gamma) = \begin{cases} r^2(1+\gamma)^{-a} + \left(r^2(1+\gamma)^{-a} + \sigma^2\right)\dfrac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2(1+\gamma)^{-a} + r^2\left(1 - (1+\gamma)^{-a}\right)\left(1 - \dfrac{1}{\gamma}\right) + \left(r^2(1+\gamma)^{-a} + \sigma^2\right)\dfrac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases} \tag{28}$$

We next summarize some interesting features of these risk curves, and Figures 3 and 4 give accompanying plots for SNR = 1 and 5, respectively. Recall that the null risk is r^2, which comes from predicting with the null estimator $\tilde{\beta} = 0$.

1. On (0, 1), the least squares risk $R_a(\gamma)$ can only be better than the null risk if $a > 1 + \frac{1}{\mathrm{SNR}}$. Further, in this case, we have $R_a(\gamma) < r^2$ if and only if $\gamma < \gamma_0$, where γ_0 is the unique zero of the function

$$(1+x)^{-a} + \left(1 + \frac{1}{\mathrm{SNR}}\right)x - 1$$

that lies in $\left(0, \frac{\mathrm{SNR}}{\mathrm{SNR}+1}\right)$. Finally, on $\left(\frac{\mathrm{SNR}}{\mathrm{SNR}+1}, 1\right)$, the least squares risk $R_a(\gamma)$ is always worse than the null risk, regardless of $a > 0$, and it is monotonically increasing.

2. On $(1, \infty)$, when SNR 1, the min-norm least squares risk $R_a(\gamma)$ is always worse than the null risk. Moreover, it is monotonically decreasing, and approaches the null risk (from above) as $\gamma \to \infty$.

3. On $(1, \infty)$, when SNR > 1, the min-norm least squares risk $R_a(\gamma)$ can be better than the null risk for any $a > 0$, and in particular we have $R_a(\gamma) < r^2$ if and only if $\gamma < \gamma_0$, where γ_0 is the unique zero of the function

$$(1+x)^{-a}(2x-1)+1-\left(1-\frac{1}{\mathrm{SNR}}\right)x$$

   lying in $\left(\frac{\mathrm{SNR}}{\mathrm{SNR}-1}, \infty\right)$. Indeed, on $\left(1, \frac{\mathrm{SNR}}{\mathrm{SNR}-1}\right)$, the min-norm least squares risk $R_a(\gamma)$ is always worse than the null risk (regardless of $a > 0$), and it is monotonically decreasing.

4. When SNR > 1, for small enough $a > 0$, the global minimum of the min-norm least squares risk $R_a(\gamma)$ occurs after $\gamma = 1$. A sufficient but not necessary condition is $a \le 1 + \frac{1}{\mathrm{SNR}}$ (because, due to points 1 and 3 above, we see that in this case $R_a(\gamma)$ is always worse than null risk for $\gamma < 1$, but will be better than the null risk at some $\gamma > 1$).

5.2. Latent space model.

We next consider an example in which β is aligned with the top eigenvectors of Σ. To motivate this example, assume that the responses y_i are linear in the latent features vectors $z_i \in \mathbb{R}^d$. We do not observe this latent vector, but rather observe $p \ge d$ covariates $x_i := (x_{i1}, \ldots, x_{ip})$ that are also linear in the latent vector z_i:

$$y_i = \theta^T z_i + \xi_i, \quad x_{ij} = w_j^T z_i + u_{ij}\,. \tag{29}$$

Here, $(\xi_i)_{i \le n}$, $(u_{ij})_{i \le n, j \le p}$ are noise variables that are mutually independent, and independent of z_i, with $\xi_i \sim N\left(0, \sigma_\xi^2\right)$, $u_{ij} \sim N(0, 1)$. The features matrix takes the form $X = ZW^T + U$ and, therefore, for $p > n$, min-norm regression amounts to

$$\hat{\beta} = \arg\min\left\{\|b\|_2 : ZW^T b + Ub = y\right\}. \tag{30}$$

Apart from its intrinsic interest, this latent-space model is directly connected to nonlinear random features models, as the ones studied in Section 8. Indeed, in nonlinear random features models we have $x_{ij} = \varphi\left(w_j^T z_i\right)$. We can decompose this as $x_{ij} = a_0 + a_1 w_j^T z_i + \tilde{\varphi}\left(w_j^T z_i\right)$, where $\tilde{\varphi}$ is such that $\tilde{\varphi}\left(w_j^T z_i\right)$ has zero mean and is uncorrelated with $w_j^T z_i$, conditional on w_j. Equation (29) then corresponds to replacing the uncorrelated random variable $\tilde{\varphi}\left(w_j^T z_i\right)$ by the independent Gaussians u_{ij}. This connection was discussed in [49, 52], after a first appearance of the present paper. Recent studies of high-dimensional linear discriminant analysis [15, 40, 53] share important elements of the model studied here: anisotropic covariance and signal aligned with its top eigenvectors.

We consider a variant of model (29), which is a special case of the model studied in Section 4. Namely we assume $x_i = \Sigma^{1/2}\tilde{z}_i$, $y_i\,\beta^T x_i + \varepsilon_i$, where $\tilde{z}_i$ is a vector with independent coordinates satisfying Assumption 1, and we set

$$\Sigma = I_p + WW^T, \quad \beta = W\left(I + W^T W\right)^{-1}\theta, \tag{31}$$

$$\mathbb{E}(\varepsilon_i) = 0, \quad \mathbb{E}\left(\varepsilon_i^2\right) = \sigma^2, \quad \sigma^2 = \sigma_\xi^2 + \theta^T\left(I + W^T W\right)^{-1}\theta\,. \tag{32}$$

Here, $W \in \mathbb{R}^{p \times d}$ is the matrix with rows $(w_i)_{i \le p}$. In what follows, $r_\theta^2 := \|\theta\|_2^2$, $\psi = d/p$. As anticipated, the coefficients vector is aligned with the top eigenspace of Σ (the span of the columns of W).

The connection between the last formulation and the model of equation (29) is easy to see if the latent vector $z_i \sim \mathcal{N}(0, I_d)$. In this case, the two models coincide because $(y_i, x_i) \in \mathbb{R}^{p+1}$ is a centered Gaussian vector with the same covariance structure.

In order to simplify our calculations, we assume all the nonzero singular values of W to be equal, whence $W^T W = (p\mu/d)I_d$, for $\mu > 0$ a constant. The factor p/d is justified by the remark that the average norm of the vectors w_j is given by

$$\frac{1}{p}\sum_{j=1}^{p}\|w_j\|_2^2 = \frac{1}{p}\mathrm{tr}\left(W^T W\right) = \mu\,.$$

Hence, μ is the signal-to-noise ratio in the features $x_{ij} = w_j^T z_i + u_{ij}$, and keeping μ constant corresponds to keeping this signal-to-noise ratio constant. This is also motivated by the nonlinear random features model $x_{ij} = \varphi\left(w_j^T z_i\right)$ (if we identify the nonlinear component with the noise); see Section 8 and [49].

Note that with this setting, the eigenvalues of Σ are

$$s_1 = s_2 = \cdots = s_d = 1 + \mu\psi^{-1} > s_{d+1} = s_2 = \cdots = s_p = 1.$$

If $p, d, n \to \infty$, with $p/n \to \gamma$, $d/p \to \psi$, then this model satisfies the assumptions of Theorem 2, with

$$H(s) = (1 - \psi)\mathbf{1}(s \ge 1) + \psi\mathbf{1}\left(s \ge 1 + \psi^{-1}\right), \tag{33}$$

$$G(s) = \mathbf{1}\left(s \ge 1 + \psi^{-1}\right), \quad \left\|\beta\right\|_2^2 = \frac{\mu\psi^{-1} r_\theta^2}{\left(1 + \mu\psi^{-1}\right)^2} \tag{34}$$

Using Theorem 2, we get the following explicit expressions.

Corollary 5. *Consider the latent space model described above, namely* $x_i = \Sigma^{1/2}\tilde{z}_i$, $y_i = \beta^T x_i + \varepsilon_i$, *where* $\tilde{z}_i$ *is a vector with independent coordinates satisfying Assumption* 1. *Further assume* equations (31), (32) *and* $d/p \to \psi \in (1, \infty)$, $p/n \to \gamma\,(1, \infty)$ *to hold.* (*The case* $\gamma \in (0, 1)$ *being covered by Proposition* 2.)

Then almost surely

$$R_X(\hat{\beta}; \beta) \to \mathscr{B}_{\text{lat}}(\psi, \gamma) + \mathscr{V}_{\text{lat}}(\psi, \gamma), \tag{35}$$

$$\mathscr{B}_{\text{lat}}(\psi, \gamma) := \left\{1 + \gamma c_0 \frac{\mathscr{E}_1(\psi, \gamma)}{\mathscr{E}_2(\psi, \gamma)}\right\} \cdot \frac{\mu\psi^{-1} r_\theta^2}{\left(1 + \mu\psi^{-1}\right)\left(1 + c_0\gamma\left(1 + \mu\psi^{-1}\right)\right)^2}, \tag{36}$$

$$\mathscr{V}_{\text{lat}}(\psi, \gamma) := \sigma^2 \gamma c_0 \frac{\mathscr{E}_1(\psi, \gamma)}{\mathscr{E}_2(\psi, \gamma)}, \tag{37}$$

$$\mathscr{E}_1(\psi, \gamma) := \frac{1 - \psi}{(1 + c_0\gamma)^2} + \frac{\psi\left(1 + \mu\psi^{-1}\right)^2}{\left(1 + c_0\left(1 + \mu\psi^{-1}\right)\gamma\right)^2}, \tag{38}$$

$$\mathscr{E}_2(\psi, \gamma) := \frac{1 - \psi}{(1 + c_0\gamma)^2} + \frac{\psi\left(1 + \mu\psi^{-1}\right)}{\left(1 + c_0\left(1 + \mu\psi^{-1}\right)\gamma\right)^2}. \tag{39}$$

where $\sigma^2 = \sigma_\xi^2 + r_\theta^2/\left(1 + \mu\psi^{-1}\right)$, *and* $c_0 = c_0(\psi, \gamma) \ge 0$ *is the unique nonnegative solution of the following second-order equation*:

$$1 - \frac{1}{\gamma} = \frac{1 - \psi}{1 + c_0\gamma} + \frac{\psi}{1 + c_0\left(1 + \mu\psi^{-1}\right)\gamma}. \tag{40}$$

Remark 4. The proof of Theorem 2 holds almost unchanged for the case in which $x_i = \Sigma^{1/2}\tilde{z}_i$ with $\Sigma^{1/2}$ a nonsymmetric square root of Σ and $\tilde{z}_i$ a vector with independent entries satisfying Assumption 2. This case includes the general model equation (29) for z_i with independent entries as a special case. It is sufficient to set $\tilde{z}_i = (z_i, u_i)$ and $\Sigma^{1/2} = (W, I_p)$.

Figures 5 and 6 illustrate this corollary by comparing analytical predictions to numerical simulations. We observe that the prediction risk is monotone decreasing in the over-parametrization ratio for $\gamma > 1$, and reaches its global minimum asymptotically as $\gamma \to$

∞ (after p, n, $d \to \infty$). To understand why this happens, notice that each feature vector x_i can be viewed as a noisy measurement of the latent covariates z_i. If the noise u_{ij} was absent, then performing min-norm regression with respect to $(x_i)_{i \le n}$ would be equivalent to min-norm regression with respect to $(z_i)_{i \le n}$. To see this, consider again equation (30). If we drop the noise U, we are minimizing $\|b\|_2$ subject to $Z(W^T b) = 0$, and the regression function is $\hat{f}(z) = x^T \hat{\beta} = z^T \left(W^T \hat{\beta}\right)$. Since W is orthogonal, this is equivalent to computing $\hat{\theta} = \arg\min\{\|t\|_2 \colon \text{subject to } Zt = y\}$, with $y = Z\theta + \xi$. In other words, we are back to the underparametrized model.

In presence of noise u_{ij}, the latent features cannot be estimated exactly. However, as p gets larger, the noise is effectively "averaged out" and we approach the idealized situation in which the z_i 's are observed.

All of the simulations in Figures 5, 6 are carried out with $\mu = 1$. In Appendix A.3, we explore the dependence on μ, and show that the generalization curves are insensitive over a broad range of choices of this parameter.

6. Ridge regularization.

We generalize the formulas of Section 4 to nonvanishing ridge regularization. We work under the same assumptions of that section. In particular, recall that $\widehat{H}_n(s) = p^{-1} \sum_{i=1}^{p} 1_{\{s \ge s_i\}}$ is the empirical distribution of the eigenvalues of Σ, and $\widehat{G}_n(s) = \sum_{i=1}^{p} \langle \beta, v_i \rangle^2 1_{\{s \ge s_i\}} / \|\beta\|^2$ the same empirical distribution, reweighted by the projection of β onto the eigenvectors v_i of the covariance Σ. (Recall the eigenvalue decomposition $\Sigma = \sum_{i=1}^{p} s_i v_i v_i^T$.)

Definition 2 (Predicted bias and variance: ridge regression). For $\gamma \in \mathbb{R}_{>0}$, and $z \in \mathbb{C}_+$ (the set of complex numbers with Im(z) > 0), define $m_n(z) = m\left(z; \widehat{H}_n, \gamma\right)$ as the unique solution of

$$m_n(z) = \int \frac{1}{s[1 - \gamma - \gamma z m_n(z)] - z} d\widehat{H}_n(s) \,. \tag{41}$$

Further define $m_{n,1}(z) = m_{n,1}\left(z; \widehat{H}_n, \gamma\right)$ via

$$m_{n,1}(z) := \frac{\displaystyle\int \frac{s^2[1 - \gamma - \gamma z m_n(z)]}{[s[1 - \gamma - \gamma z m_n(z)] - z]^2} d\widehat{H}_n(s)}{\displaystyle 1 - \gamma \int \frac{zs}{[s[1 - \gamma - \gamma z m_n(z)] - z]^2} d\widehat{H}_n(s)} \,. \tag{42}$$

These definitions are extended analytically to Im(z) = 0 whenever possible. We then define the predicted bias and variance by

$$\mathscr{B}\left(\lambda; \widehat{H}_n, \widehat{G}_n, \gamma\right) := \lambda^2 \left\|\beta\right\|^2 \left(1 + \gamma m_{n,1}(-\lambda)\right) \times \int \frac{s}{\left[\lambda + (1 - \gamma + \gamma\lambda m_n(-\lambda))s\right]^2} d\widehat{G}_n(s), \tag{43}$$

$$\mathscr{V}\left(\lambda; \widehat{H}_n, \gamma\right) := \sigma^2 \gamma \int \frac{s^2\left(1 - \gamma + \gamma\lambda^2 m_n'(-\lambda)\right)}{\left[\lambda + s(1 - \gamma + \gamma\lambda m_n(-\lambda))\right]^2} d\widehat{H}_n(s). \tag{44}$$

We next state our deterministic approximation of the risk.

THEOREM 5. *Let* $M^{-1} \le p/n \le M$, *and Assumption* 1 *hold. Further assume* $\lambda \vee s_{\min}(\Sigma) > 1/M$ *and* $n^{-2/3+1/M} < \lambda < M$. *Let* $\hat{\beta}_\lambda$ *be the ridge estimator of* equation (7).

Then for any constants $D > 0$ (*arbitrarily large*) *and* $\varepsilon > 0$ (*arbitrarily small*), *there exist* $C = C(M, D)$ *such that, with probability at least* $1 - Cn^{-D}$ *the following hold*:

$$R_X\left(\hat{\beta}_\lambda; \beta\right) = B_X\left(\hat{\beta}_\lambda; \beta\right) + V_X\left(\hat{\beta}_\lambda; \beta\right), \tag{45}$$

$$\left|B_X\left(\hat{\beta}_\lambda; \beta\right) - \mathscr{B}\left(\lambda; \widehat{H}_n, \widehat{G}_n, \gamma\right)\right| \le \frac{C\|\beta\|_2^2}{\lambda n^{(1-\varepsilon)/2}}, \tag{46}$$

$$\left|V_X\left(\hat{\beta}_\lambda; \beta\right) - \mathscr{V}\left(\lambda; \widehat{H}_n, \gamma\right)\right| \le \frac{C}{\lambda^2 n^{(1-\varepsilon)/2}}, \tag{47}$$

where $\mathscr{B}$ *and* $\mathscr{V}$ *are given in Definition* 2, *and the first identity is just the general bias-variance decomposition of* equation (5).

The proof of this theorem is deferred to Appendix A.1. As for Theorem 2, similar results were proved in [57, 69], subsequently to a first version of this manuscript that only focused on random β. The same comparison of Remark 2 applies here.

In particular, Theorem 5 establishes nonasymptotic deterministic approximations for the bias $B_X\left(\hat{\beta}_\lambda; \beta\right)$ and variance $V_X(\hat{\beta}; \beta)$. The error terms are uniform over the covariance matrix, and have nearly optimal dependence upon the sample size n. Indeed, a central-limit theorem heuristics suggests fluctuations of order $n^{-1/2}$.

As for the case of min-norm regression, Theorem 5 directly implies a characterization of the asymptotics of bias and variance of ridge regression. This statement is analogous to Theorem 6.

THEOREM 6. *Consider the setting of Theorem* 5. *Further assume* $p/n \to \gamma \in (0, \infty)$, $\widehat{H}_n \Rightarrow H$, $\widehat{G}_n \Rightarrow G$. *Define* $\mathscr{B}_1(\lambda; H, G, \gamma)$ *as in* equation (43), *with* $\|\beta\|_2^2$ *replaced by* 1. *Then, for any* $\lambda > 0$, *almost surely*

$$\frac{1}{\|\beta\|_2^2} B_X\left(\widehat{\beta}_\lambda; \beta\right) \to \mathscr{B}_1(\lambda; H, G, \gamma), \quad V_X\left(\widehat{\beta}_\lambda; \beta\right) \to \mathscr{V}(\lambda; H, \gamma) . \tag{48}$$

The same conclusion holds if instead of Assumption 1(*a*), *the coordinates of z*, $(z_i)_{i \le p}$ *are i.i.d. and satisfy the conditions* $\mathbb{E}z_i = 0$, $\mathbb{E}\left(z_i^2\right) = 1$, $\mathbb{E}\left(|z_i|^{4+\delta}\right) \le C < \infty$.

6.1. Isotropic features.

As a special case, we can consider the simple isotropic model that was already studied in Section 3. Very similar (though not identical) results can be found in Dicker [22], Dobriban and Wager [23].

COROLLARY 6. *Assume the conditions of Theorem* 1 (*well-specified model*, *isotropic features*). *Then for ridge regression estimator in* (7) *as* $n, p \to \infty$, *such that* $p/n \to \gamma \in (0, \infty)$, *it holds almost surely that*

$$R_X\left(\widehat{\beta}_\lambda; \beta\right) \to r^2 \lambda^2 m'(-\lambda) + \sigma^2 \gamma(m(-\lambda) - \lambda m'(-\lambda)) . \tag{49}$$

Here, $m(z)$ *is given by* equation (41), *which in this case has the explicit solution* $m(z) = \left[1 - \gamma - z - \sqrt{(1 - \gamma - z)^2 - 4\gamma z}\right] / (2\gamma z)$.

Furthermore, the limiting ridge risk is minimized at $\lambda^* = \sigma^2 \gamma / r^2$, *in which case we have the simpler expression* $R_X\left(\widehat{\beta}_\lambda; \beta\right) \to \sigma^2 \gamma m(-\lambda^*)$.

It is easy to recover the formulas in Theorem 1 as a limiting case of equation (49), by using the $z \to 0$ asymptotics $m(z)\ (1 - \gamma)^{-1}\ O(z)$ for $\gamma < 1$ and $m(z)\ (1 - \gamma)^{-1}\ O(z)$ for $\gamma < 1$ and $m(z) = -(\gamma - 1)/(\gamma z) + [(\gamma - 1)\gamma]^{-1} + O(z)$ for $\gamma > 1$.

Figures 7 and 8 compare the risk curves of min-norm least squares to those from optimally-tuned ridge regression, in the well-specified and misspecified settings, respectively. There are two important points to make. The first is that optimally-tuned ridge regression is seen to have strictly better asymptotic risk throughout, regardless of r^2, γ, κ. This should not be a surprise, as by definition optimal tuning should yield better risk than min-norm least squares, which is the special case given by $\lambda \to 0^+$.

The second point is that, in this example, the limiting risk of optimally-tuned ridge regression appears to have a minimum around $\gamma = 1$, and this occurs closer and closer to $\gamma = 1$ as SNR grows. This behavior is interesting, especially because it is antipodal to that of the min-norm least squares risk, and leads us to very different suggestions for practical usage for feature generators: in settings where we apply substantial ℓ_2 regularization (say, using CV tuning to mimic optimal tuning, which the next section shows to be asymptotically equivalent), it seems we want the complexity of the feature space to put us as close to the interpolation boundary ($\gamma = 1$) as possible.

As we will see, the behavior is rather different in the latent space model.

6.2. Latent space model.

As a special application, we consider the latent space model of Section 5.2. It is immediate to specialize equations (43) and (44) to this case. We omit giving giving explicit formulas for brevity, and instead plot the resulting curves for the prediction risk (test error).

In Figure 9, we plot the risk as a function of the overparametrization ration $\gamma = p/n$ for several values of the regularization parameter λ (included the ridgeless limit $\lambda \to 0$). The setting here is analogous to the one of Figure 5. We observe several interesting phenomena:

1. Independently of λ in the probed range, the risk is minimized at large overparametrization $\gamma \gg 1$.
2. As expected, the divergence of the risk at the interpolation threshold $\gamma = 1$ is smoothed out by regularization, and the risk becomes a monotone decreasing function of γ when λ is large enough. Crucially, the optimal amount of regularization (corresponding to the lower envelope of these curves) results in a monotonically decreasing risk.
3. At large overparametrization, the optimal value of the regularization parameter is $\lambda \to 0$.

We confirm the last finding in Figure 10, which plots the risk as a function of λ: the optimal regularization is $\lambda \to 0$. Notice that this is the case despite the fact that the observations are noisy, namely $\sigma_\xi > 0$ strictly.

The optimality of $\lambda \to 0$ has a known intuitive explanation that is worth recalling here. Recall that ridge predictor at point x_0 takes the form

$$\hat{f}_\lambda(x_0) = \left\langle x_0, \hat{\beta}_\lambda \right\rangle = K(x_0, X)(K(X, X) + \lambda I)^{-1} y, \tag{50}$$

where we introduced the kernel matrix $K(X, X) = XX^T/n$, and the vector $K(x_0, X) = x_0^T X/n$. Consider the case in which the covariates X contain noise, as is our case, $X = \overline{X} + \eta U$ where $\overline{X} = ZW^T$ and $u_{ij} \sim N(0, 1)$. (While we are considering $\eta = 1$, it is instructive to regard the noise standard deviation as a parameter.) Then we might expect $K(X, X) \approx K(\overline{X}, \overline{X}) + \eta^2 UU^T \approx K(\overline{X}, \overline{X}) + \eta^2 I_n$. If this approximation holds, the noise acts as an extra ridge term, which can be sufficient to regularize the problem.

To the best of our knowledge, this argument was first presented by Webb [68] and, more explicitly, by Bishop [13]. Recently, Kobak et al. [42] elucidated its role in linear regression, establishing several of its consequences. In a parallel line of work, a closely related idea has recently emerged in the analysis of kernel methods in high dimension [26, 47].

7. Cross-validation.

We analyze the effect of using cross-validation to choose the tuning parameter in ridge regression. In short, we find that choosing the ridge tuning parameter to minimize the leave-one-out cross-validation error leads to the same asymptotic risk as the optimally-tuned ridge estimator. The next subsection gives the details; the following subsection presents a new "shortcut formula" for leave-one-out cross-validation in the overparametrized regime, for min-norm least squares, akin to the well-known formula for underparametrized least squares and ridge regression. We refer to [20, 32] for background on CV and GCV, and to [5] for a more recent review.

7.1. Limiting behavior of CV tuning.

Given the ridge regression solution $\hat{\beta}_\lambda$ in (7), trained on (x_i, y_i), $i = 1, \ldots, n$, denote by $\hat{f}_\lambda$ the corresponding ridge predictor, defined as $\hat{f}_\lambda(x) = x^T\hat{\beta}_\lambda$ for $x \in \mathbb{R}^p$. Additionally, for each $i = 1, \ldots, n$, denote by $\hat{f}_\lambda^{-i}$ the ridge predictor trained on all but ith data point (x_i, y_i).[5] Recall that the *leave-one-out cross-validation* (leave-one-out CV, or simply CV) error of the ridge solution at a tuning parameter value λ is

$$\mathrm{CV}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_\lambda^{-i}(x_i)\right)^2. \tag{51}$$

We typically view this as an estimate of the out-of-sample prediction error $\mathbb{E}\left(y_0 - x_0^T\hat{\beta}_\lambda\right)^2$, where the expectation is taken over everything that is random: the training data (x_i, y_i), $i = 1, \ldots, n$ used to fit $\hat{\beta}_\lambda$, as well as the independent test point (x_0, y_0). Note also that, when we observe training data from the model (1), (2), and when (x_0, y_0) is drawn independently according to the same process, we have the relationship

$$\mathbb{E}\left(y_0 - x_0^T\hat{\beta}_\lambda\right)^2 = \sigma^2 + \mathbb{E}\left(x_0^T\beta - x_0^T\hat{\beta}_\lambda\right)^2 = \sigma^2 + \mathbb{E}\left[R_X\left(\hat{\beta}_\lambda;\beta\right)\right],$$

where $R_X\left(\hat{\beta}_\lambda;\beta\right)\big] = \mathbb{E}\left[\left(x_0^T\beta - x_0^T\hat{\beta}_\lambda\right)^2 \mid X\right]$ is the conditional prediction risk, which has been our focus throughout.

Recomputing the leave-one-out predictors $\hat{f}_\lambda^{-i}$, $i = 1, \ldots, n$ can be burdensome, especially for large n. Importantly, there is a well-known "shortcut formula" that allows us to express the leave-one-out CV error (51) as a weighted average of the training errors,

[5] To be precise, this is $\hat{f}^{-i}(x) = x^T\left(X_{-i}^T X_{-i} + n\lambda I\right)^{-1} X_{-i}^T y_{-i}$, where X_{-i} denotes X with the ith row removed, and y_{-i} denotes y with the ith removed. Arguably, it may seem more natural to replace the factor of n here by a factor of $n - 1$; we leave the factor of n as is because it simplifies the presentation in what follows, but we remark that the same asymptotic results would hold with $n - 1$ in place of n.

$$\mathrm{CV}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}_\lambda(x_i)}{1 - (S_\lambda)_{ii}}\right)^2, \tag{52}$$

where $S_\lambda = X(X^TX + n\lambda)^{-1}X^T$ is the ridge smoother matrix. This identity is an immediate consequence of the Sherman–Morrison–Woodbury formula. In the next subsection, we give an extension to the case $\lambda = 0$ and rank $(X) = n$, that is, to min-norm least squares.

The next result shows that, for isotropic features, the CV error of a ridge estimator converges almost surely to its prediction error. The focus on isotropic features is only for simplicity: a more general analysis is possible but is not pursued here. The proof, given in Appendix A.4.6, relies on the shortcut formula (52). In the proof, we actually first analyze generalized cross-validation (GCV), which turns out to be somewhat of an easier calculation (see the proof for details on the precise form of GCV), and then relate leave-one-out CV to GCV.

Theorem 7. *Assume the a isotropic prior, namely* $\mathbb{E}(\beta) = 0$, $\mathrm{Cov}(\beta)\ r^2 I_p/p$, *and the data model* (1), (2). *Assume that* $x \sim P_x$ *has i.i.d. entries with zero mean, unit variance, and a finite moment of order* $4 + \eta$, *for some* $\eta > 0$. *Then for the CV error* (51) *of the ridge estimator in* (7) *with tuning parameter* $\lambda > 0$, *as* $n, p \to \infty$, *with* $p/n \to \gamma \in (0, \infty)$, *it holds almost surely that*

$$\mathrm{CV}_n(\lambda) - \sigma^2 \to \sigma^2\gamma(m(-\lambda) - \lambda(1 - \alpha\lambda)m'(-\lambda)),$$

where $m(z)$ *denotes the Stieltjes transform of the Marchenko–Pastur law* F_γ (*as in Corollary* 6), *and* $\alpha = r^2/(\sigma^2\gamma)$. *Observe that the right-hand side is the asymptotic risk of ridge regression from Theorem* 5. *Moreover, the above convergence is uniform over compacts intervals excluding zero. Thus if* λ_1, λ_2 *are constants with* $0 < \lambda_1 \le \lambda^* \le \lambda_2 < \infty$, *where* $\lambda^* = 1/\alpha$ *is the asymptotically optimal ridge tuning parameter value, and we define* $\lambda_n = \arg\min_{\lambda \in [\lambda_1, \lambda_2]} \mathrm{CV}_n(\lambda)$, *then the expected risk of the CV-tuned ridge estimator* $R_X(\hat{\beta}) := \mathbb{E}_\beta R_X(\hat{\beta}; \beta)\ \hat{\beta}_{\lambda_n}$ *satisfies, almost surely*

$$R_X\left(\hat{\beta}_{\lambda_n}\right) \to \sigma^2\gamma m(-1/\alpha),$$

with the right-hand side above being the asymptotic risk of optimally-tuned ridge regression. Further, the exact same set of results holds for GCV.

Similar results were obtained for various linear smoothers in Li [45, 46], for the lasso in the high-dimensional (proportional) asymptotics in Miolane and Montanari [51], and for general smooth penalized estimators in Xu et al. [70]. The latter paper covers ridge regression as a special case, and gives more precise results (convergence rates), but assumes more restrictive conditions. After submission of this paper, consistency of CV and GVC was proved in a significantly more general setting in [54], in particular dispensing with the assumption of random β.

The key implication of Theorem 7, in the context of the current paper and its central focus, is that the CV-tuned or GCV-tuned ridge estimator has the same asymptotic performance as the optimally-tuned ridge estimator. In other words, the ridge curves in Figures 1, 7 and 8 can be alternatively viewed as the asymptotic risk of ridge under CV tuning.

7.2. Shortcut formula for ridgeless CV.

We extend the leave-one-out CV shortcut formula (52) to work when $p > n$ and $\lambda = 0+$, that is, for min-norm least squares. In this case, both the numerator and denominator are zero in each summand of (52). To circumvent this, we can use the so-called "kernel trick" to rewrite the ridge regression solution (7) with $\lambda > 0$ as

$$\hat{\beta}_\lambda = X^T\left(XX^T + n\lambda I\right)^{-1}y. \tag{53}$$

Using this expression, the shortcut formula for leave-one-out CV in (52) can be rewritten as

$$\mathrm{CV}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left[\left(XX^T + n\lambda I\right)^{-1}y\right]_i^2}{\left[\left(XX^T + n\lambda I\right)^{-1}\right]_{ii}^2}.$$

Taking $\lambda \to 0^+$ yields the shortcut formula for leave-one-out CV in min-norm least squares (assuming without a loss of generality that rank(X) = n),

$$\mathrm{CV}_n(0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left[\left(XX^T\right)^{-1}y\right]_i^2}{\left[\left(XX^T\right)^{-1}\right]_{ii}^2}. \tag{54}$$

In fact, the exact same arguments given here still apply when we replace XX^T by a positive definite kernel matrix K (i.e., $K_{ij} = k(x_i, x_j)$ for each $i, j = 1, \ldots, n$, where k is a positive definite kernel function), in which case (54) gives a shortcut formula for leave-one-out CV in kernel ridgeless regression (the limit in kernel ridge regression as $\lambda \to 0^+$). We also remark that, when we include an unpenalized intercept in the model, in either the linear or kernelized setting, the shortcut formula (54) still applies with XX^T or K replaced by their doubly-centered (row- and column-centered) versions, and the matrix inverses replaced by pseudoinverses.

8. Nonlinear model.

Our analysis in the previous sections assumed $x_i = \Sigma^{1/2}z_i$, with z_i a vector with independent entries.

In this section, we test universality on one simple example. We observe data as in (1), (2), but now $x_i = \varphi(Wz_i) \in \mathbb{R}^p$, where $z_i \in \mathbb{R}^d$ has i.i.d. entries from $N(0, 1)$, for $i = 1, \ldots, n$. Also, $W \in \mathbb{R}^{p \times d}$ has i.i.d. entries from $N(0, 1/d)$. Finally, $\varphi : \mathbb{R} \to \mathbb{R}$ is an activation function acting entrywise on vectors.

We first consider the case of purely nonlinear activations, namely activation functions that are uncorrelated with linear functions: $\mathbb{E}\{\varphi(G)\} = \mathbb{E}\{G\varphi(G)\} = 0$ for $G \sim N(0, 1)$. In this case, the second-order statistics of the features x_i match the ones of the isotropic model and the same happen for the asymptotic of the risk. We then consider more general activations, and show that the asymptotic variance depends on the activation through the value of $c_1 = \mathbb{E}\{G\varphi(G)\}^2$.

8.1. Limiting risk for purely nonlinear activations.

Notice that, conditionally on W, the vectors $x_i = \varphi(Wz_i)$, $i \le n$ are independent. However, they do not have independent coordinates. For instance, if $\varphi(t) = at^2 + b$, we can reconstruct z_i from the first $2d$ coordinates of x_i and, therefore, the remaining $p - 2d$ coordinates of x_i are a function of the first $2d$.

Nevertheless, the next theorem shows that if φ is purely nonlinear (in the sense that $\mathbb{E}\{\varphi(G)\} = \mathbb{E}\{G\varphi(G)\} = 0$), then the feature matrix X behaves "as if" it had i.i.d. entries, in that the asymptotic bias and variance are exactly as in the linear isotropic case; recall equation (10). In other words, this theorem provides a rigorous confirmation of the universality hypothesis stated in the Introduction.

THEOREM 8. *Assume the model* (1), (2), *where each* $x_i = \varphi(W z_i) \in \mathbb{R}^p$, *for* $z_i \in \mathbb{R}^d$ *having i.i.d. entries from* $N(0, 1)$, $W \in \mathbb{R}^{p \times d}$ *having i.i.d. entries from* $N(0, 1/d)$ (*with* W *independent of* z_i), *and is* φ *an activation function that acts componentwise. Assume that* $|\varphi(x)| \le c_0(1 + |x|)^{c_0}$ *for a constant* $c_0 > 0$. *Also, for* $G \sim N(0, 1)$, *assume that the following standardization conditions hold*: $\mathbb{E}[\varphi(G)] = 0$ *and* $\mathbb{E}\left[\varphi(G)^2\right] = 1$, $\mathbb{E}[G\varphi(G)] = 0$. *Consider the limit* $n, p, d \to \infty$, *with* $p/n \to \gamma$ *and* $d/p \to \psi \in (0, \infty)$.

Then for $\gamma > 1$, *the variance satisfies, almost surely*

$$\lim_{\lambda \to 0^+} \lim_{n, p, d \to \infty} V_X\left(\hat{\beta}_\lambda; \beta\right) = \frac{\sigma^2}{\gamma - 1},$$

which is precisely as in the case of linear isotropic features; recall Theorem 1. *Also, under a isotropic prior, namely* $\mathbb{E}(\beta) = 0$, $\mathrm{Cov}(\beta) = r^2 I_p/p)$, *the Bayes bias* $B_X\left(\hat{\beta}_\lambda\right) := \mathbb{E}_\beta B_X\left(\hat{\beta}_\lambda; \beta\right)$ *satisfies, almost surely*

$$\lim_{\lambda \to 0^+} \lim_{n, p, d \to \infty} B_X\left(\hat{\beta}_\lambda\right) = \begin{cases} 0 & for\ \gamma < 1, \\ r^2(1 - 1/\gamma) & for\ \gamma > 1, \end{cases}$$

which is again as in the case of linear isotropic features; recall Theorem 1.

The proof of Theorem 8 is lengthy and will be sketched shortly. We notice that the definition of $V_X\left(\hat{\beta}_\lambda; \beta\right)$ and $B_X\left(\hat{\beta}_\lambda\right)$ is conditional on X. In fact, we will prove that the stated limits hold asymptotically almost surely, conditionally both on W and on the covariates x_i.

The origin of the conditions $\mathbb{E}\{\varphi(G)\} = \mathbb{E}\{G\varphi(G)\} = 0$ can be easily explained (throughout $G \sim N(0, 1)$). In summary, these conditions ensure that the first- and second-order statistics of $x_i = \varphi(Wz_i)$ approximately match those of the isotropic model. To illustrate this point, let $i \neq j$, and assume that the corresponding rows of W (denoted by w_i^T and w_j^T) have unit norm (this will only be approximately true, but simplifies our explanation). We then have $\mathbb{E}_z\left\{\varphi\left(w_i^T z_1\right)\right\} = \mathbb{E}_z\left\{\varphi\left(w_j^T z_1\right)\right\} = \mathbb{E}\{\varphi(G)\} = 0$. Further,

$$\mathbb{E}_z\left\{x_{1,i}x_{1,j} \mid W\right\} = \mathbb{E}_z\left\{\varphi\left(w_i^T z_1\right)\varphi\left(w_j^T z_1\right) \mid W\right\} = \mathbb{E}\{\varphi(G_1)\varphi(G_2)\}, \tag{55}$$

where G_1, G_2 are jointly Gaussian with unit variance and covariance $w_i^T w_j$. Denoting by $\varphi(x) = \sum_{k \geq 0} \lambda_k(\varphi) h_k(x)$ the decomposition of φ into orthonormal Hermite polynomials, we thus obtain

$$\mathbb{E}\left\{x_{1,i}x_{1,j} \mid W\right\} = \sum_{k=0}^{\infty} \lambda_k^2(\varphi)\left(w_i^T w_j\right)^k, \tag{56}$$

Since $\lambda_0(\varphi) = \mathbb{E}\{\varphi(G)\} = 0$, $\lambda_1(\varphi) = \mathbb{E}\{G\varphi(G)\} = 0$, we have $\mathbb{E}\left\{x_{1,i}x_{1,j} \mid W\right\} = O\left(\left(w_i^T w_j\right)^2\right) = O(1/d)$. In other words, the population covariance $\mathbb{E}\left\{x_1 x_1^T \mid W\right\}$ has small entries out-of diagonal, and in fact $\left\|\mathbb{E}\left\{x_1 x_1^T \mid W\right\} - I_p\right\|_{\mathrm{op}} = o_P(1)$ [26].

Even if the nonlinear model $x_i = \varphi(Wz_i)$ matches the second-order *population* statistics of the isotropic model, it is far from obvious that the asymptotics of the risk is the same. Indeed the coordinates of vector x_i are highly dependent. Theorem 8 confirms that—despite dependence—the risk is asymptotically the same, thus providing a concrete example of the general universality phenomenon.

Figure 11 compares the asymptotic risk curve from Theorem 8 to that computed by simulation, using an activation function $\varphi_{\mathrm{abs}}(t) = a(|t| - b)$, where $a = \sqrt{\pi/(\pi-2)}$ and $b = \sqrt{2/\pi}$ are chosen to meet the standardization conditions. This activation function is purely nonlinear, that is, it satisfies $\mathbb{E}[G\varphi_{\mathrm{abs}}(G)] = 0$ for $G \sim N(0, 1)$, by symmetry. Again, the agreement between finite-sample and asymptotic risks is excellent. Notice in particular that, as predicted by Theorem 8, the risk depends only on p/n and not on d/n.

8.2. Limiting variance for general activations.

Consider now the case of a general activation with vanishing mean $\mathbb{E}\{\varphi(G)\} = 0$, but nonvanishing linear component $\mathbb{E}\{G\varphi(G)\}^2 = c_1$, and normalized so that $\mathbb{E}\left\{\varphi(G)^2\right\} = 1$. Following the same argument as in the last section, we obtain the following approximation for the conditional covariance of x_i given W:

$$\left\|\mathbb{E}\left\{x_1 x_1^T \mid W\right\} - \widetilde{\Sigma}_{X \mid W}\right\|_{\mathrm{op}} = o_P(1), \quad \widetilde{\Sigma}_{X \mid W} = (1 - c_1)I_p + c_1 W W^T. \tag{57}$$

In other words, the conditional covariance is well approximated by the covariance of the the latent space model of Sections 5.2 and 6.2. Let us emphasize that the conditional covariance is what matters (not the unconditional one) because the x_i's are independent only conditional on W.

In Appendix A.5, we derive the asymptotic variance for this model. The next result confirms once more the universality scenario outlined above. In words, the asymptotic variance of the nonlinear model $x_i = \varphi(Wz_i)$ coincides with the asymptotic variance of the corresponding linear model $x_i \sim N(0, \Sigma_W)$.

Theorem 9. *In the setting of Theorem* 8, *assume* $\mathbb{E}[\varphi(G)] = 0$ *and* $\mathbb{E}\left[\varphi(G)^2\right] = 1$, $\mathbb{E}[G\varphi(G)]^2 = c_1$, *and let* $V_X\left(\hat{\beta}_\lambda; \beta\right)$ *denote the corresponding variance of ridge regression.*

Further, consider a different problem with $\tilde{x}_i \sim N\left(0, \tilde{\Sigma}_{X \mid W}\right)$, $\tilde{\Sigma}_{X \mid W} := (1 - c_1)I_p + c_1 W W^T$, *and denote by* $V_{\tilde{X}}\left(\hat{\beta}_\lambda; \beta\right)$ *denote the corresponding variance of ridge regression. Assume p, n, d,* $\to \infty$ *with* $p/n \to \gamma \in (1, \infty$, *and* $d/p \to \psi \in (0, \infty)$. *Then we have, almost surely*

$$\lim_{\lambda \to 0^+} \lim_{n, p, d \to \infty} V_X\left(\hat{\beta}_\lambda; \beta\right) = \lim_{\lambda \to 0^+} \lim_{n, p, d \to \infty} V_{\tilde{X}}\left(\hat{\beta}_\lambda; \beta\right).$$

Remark 5. After the last result was proved and a preprint posted online, the techniques developed here were significantly sharpened and generalized in [49], which proved in particular that the same universality result holds for the bias term as well. We notice that both Theorem 8 and Theorems 9 as well as its generalization in [49] assume a particularly simple model for the latent covariates z_i. While this simple model simplifies the proof, the result is likely to generalize to other models, for example, z_i with independent sub-Gaussian entries.

Remark 6. Notice that the assumption of isotropic W in Theorem 8 and Theorem 9 is realistic as this is the standard choice of random features models. On the other hand, it is an interesting research question to which extent these results generalize to other distributions of z_i, for example, $z_i \sim N(0, \Sigma_Z)$. We believe that the results obtained here extend to that case as well as long as $p/n \to \gamma \in (1, \infty)$, and $d/p \to \psi \in (0, \infty)$ and Σ_Z has a positive fraction of eigenvalues of the same order as its largest eigenvalue (as requested for Σ in Assumption 1). In that case of course, one has to replace the above formula for $\tilde{\Sigma}_{X \mid W}$ by $\tilde{\Sigma}_{X \mid W} = \mathbb{E}_z\left\{\varphi(Wz)\varphi(Wz)^T\right\}$.

8.3. Proof outline for Theorem 8.

We define $\gamma_n = p/n$ and $\psi_n = d/p$. Recall that as $n, p, d \to \infty$, we have $\gamma_n \to \gamma$ and $\psi_n \to \psi$. To reduce notational overhead, we will generally drop the subscripts from γ_n, ψ_n, writing these simply as γ, ψ, since their meanings should be clear from the context. Let $N = p + n$ and define the symmetric matrix $A(s) \in \mathbb{R}^{N \times N}$, for $s \geq 0$, with the block structure:

$$A(s) = \begin{bmatrix} sI_p & \frac{1}{\sqrt{n}}X^T \\ \frac{1}{\sqrt{n}}X & 0_n \end{bmatrix}, \tag{58}$$

where $I_p \in \mathbb{R}^{p \times p}$ and $0_n \in \mathbb{R}^{n \times n}$ are the identity and zero matrix, respectively. As we will see, this matrix allows to construct the traces of interest by taking suitable derivatives of its resolvent.

We introduce the following resolvents (as usual, these are defined for $\mathrm{Im}(\xi) > 0$ and by analytic continuation, whenever possible, for $\mathrm{Im}(\xi) = 0$):

$$m_{1,n}(\xi, s) = \mathbb{E}\left\{(A(s) - \xi I_N)^{-1}_{1,1}\right\} = \mathbb{E}M_{1,n}(\xi, s),$$

$$M_{1,n}(\xi, s) = \frac{1}{p}\mathrm{Tr}_{[1,p]}\left\{(A(s) - \xi I_N)^{-1}\right\},$$

$$m_{2,n}(\xi, s) = \mathbb{E}\left\{(A(s) - \xi I_N)^{-1}_{p+1,p+1}\right\} = \mathbb{E}M_{2,n}(\xi, s),$$

$$M_{2,n}(\xi, s) = \frac{1}{n}\mathrm{Tr}_{[p+1,p+n]}\left\{(A(s) - \xi I_N)^{-1}\right\}.$$

Here and henceforth, we write $[i, j] = \{i + 1, \ldots, i + j\}$ for integers i, j. We also write $M^{-1}_{ij} = \left(M^{-1}\right)_{ij}$ for a matrix M, and $\mathrm{Tr}_S(M) = \sum_{i \in S} M_{ii}$ for a subset S. The equalities in the first and third lines above follow by invariance of the distribution of $A(s)$ under permutations of $[1, p]$ and $[p + 1, p\ n]$. Whenever clear from the context, we will omit the arguments from block matrix and resolvents, and write $A = A(s)$, $m_{1,n} = m_{1,n}(\xi, s)$ and $m_{2,n} = m_{2,n}(\xi, s)$.

The next lemma characterizes the asymptotics of $m_{1,n}$, $m_{2,n}$.

Lemma 3. *Assume the conditions of Theorem* 8. *Consider* $\mathrm{Im}(\xi) > 0$ *or* $\mathrm{Im}(\xi)\ 0$, $\mathrm{Re}(\xi) < 0$, *with* $s \geq t \geq 0$. *Let* m_1 *and* m_2 *be the unique solutions of the following quadratic equations*:

$$m_2 = (-\xi - \gamma m_1)^{-1}, \quad m_1 = (-\xi - s - m_2)^{-1}, \tag{59}$$

subject to the condition of being analytic functions for $\mathrm{Im}(z) > 0$, *and satisfying* $|m_1(z, s)|$, $|m_2(z, s)| \leq 1/\mathrm{Im}(z)$ *for* $\mathrm{Im}(z) > C$ (*with* C *a sufficiently large constant*). *Then, as* $n, p, d \to \infty$, *such that* $p/n \to \gamma$ *and* $d/p \to \psi$, *we have almost surely* (*and in* L^1),

$$\lim_{n,p,d \to \infty} M_{1,n}(\xi, s) = m_1(\xi, s), \tag{60}$$

$$\lim_{n,\,p,\,d \to \infty} M_{2,\,n}(\xi, s) = m_2(\xi, s)\,. \tag{61}$$

The proof of this lemma is given in Appendix A.5.2. As a corollary of the above, we obtain that the asymptotical empirical spectral distribution of the empirical covariance $\hat{\Sigma} = X^T X / n$ matches the one for the independent entries model, and is hence given by the Marchenko–Pastur law (a result already obtained in Pennington and Worah [55]). We state this formally using the Stieltjes transform

$$R_n(z) = \frac{1}{p}\mathrm{Tr}\left(\left(\hat{\Sigma} - zI_p\right)^{-1}\right). \tag{62}$$

Corollary 7. *Assume the conditions of Theorem* 8. *Consider* $\mathrm{Im}(z) > 0$. *As* $n, p, d \to \infty$, *with* $p/n \to \gamma$ *and* $d/p \to \psi$, *we have* (*almost surely and in* L^1) $R_n(\xi) \to r(\xi)$ *where r is nonrandom and coincides with the Stieltjes transform of the Marchenko–Pastur law, namely*

$$r(z) = \frac{1 - \gamma - z - \sqrt{(1 - \gamma - z)^2 - 4\gamma z}}{2\gamma z}\,. \tag{63}$$

We refer to Appendix A.5.4 for a proof of this corollary. The next lemma connects the above resolvents computed in Lemma 3 to the variance of min-norm least squares, hence completing our proof outline.

Lemma 4. *Assume the conditions of Theorem* 8. *Let* m_1, m_2 *be the asymptotic resolvents given in Lemma* 3. *Define*

$$m(\xi, s) = \gamma m_1(\xi, s) + m_2(\xi, s)\,.$$

Then for $\gamma \neq 1$, $\partial_x m(\xi, x)|_{x=0}$ *as a simple pole at* $\xi = 0$, *and hence admits a Taylor–Laurent expansion around* $\xi = 0$, *whose coefficients will be denoted by* D_{-1}, D_0,

$$-\partial_x m(\xi, x)\big|_{x = 0} = \frac{D_{-1}}{\xi^2} + D_0 + O\left(\xi^2\right). \tag{64}$$

Here, each coefficient is a function of γ, ψ: $D_{-1} = D_{-1}(\gamma, \psi)$, $D_0 = D_0(\gamma, \psi)$. *Furthermore, for the ridge regression estimator* $\hat{\beta}_\lambda$ *in* (7), *as* $n, p, d \to \infty$, *such that* $p/n \to \gamma \in (0, \infty)$, $d/p \to \psi \in (0, 1)$, *the following ridgeless limit holds almost surely*:

$$\lim_{\lambda \to 0^+} \lim_{n,\,p,\,d \to \infty} V_X\left(\hat{\beta}_\lambda; \beta\right) = D_0\,.$$

The proof of this lemma can be found in Appendix A.5.3. Theorem 8 follows by evaluating the formula in Lemma 4, by using the result of Lemma 3. We refer to the Appendix in the Supplementary Material for details [36].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

The authors are grateful to Brad Efron, Rob Tibshirani and Larry Wasserman for inspiring us to work on this in the first place. RT sincerely thanks Edgar Dobriban for many helpful conversations about the random matrix theory literature, in particular, the literature trail leading up to Proposition 2, and the reference to Rubio and Mestre [60]. We are also grateful to Dmitry Kobak for bringing to our attention [42] and clarifying the implications of this work on our setting.

Funding.

TH was partially supported by NSF Grants DMS-1407548, NSF IIS-1837931 and NIH 5R01-EB001988-21.

AM was partially supported by NSF Grants DMS-1613091, NSF CCF-1714305, NSF IIS-1741162 and ONR N00014-18-1-2729.

RT was partially supported by NSF Grant DMS-1554123.

REFERENCES

[1]. Adlam B and Pennington J (2020). The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. Available at http://proceedings.mlr.press/v119/adlam20a/adlam20a.pdf.

[2]. Advani MS and Saxe AM (2017). High-dimensional dynamics of generalization error in neural networks. Available at arXiv:1710.03667.

[3]. Ali A, Kolter JZ and Tibshirani RJ (2019). A continuous-time view of early stopping for least squares. Int. Conf. Artif. Intell. Stat 22.

[4]. Allen-Zhu Z, Li Y and Song Z (2019). A convergence theory for deep learning via overparameterization. International Conference on Machine Learning. PMLR. Availablae at http://proceedings.mlr.press/v97/allen-zhu19a/allen-zhu19a.pdf.

[5]. Arlot S and Celisse A (2010). A survey of cross-validation procedures for model selection. Stat. Surv 4 40–79. 10.1214/09-SS054

[6]. Bartlett PL, Long PM, Lugosi G and Tsigler A (2020). Benign overfitting in linear regression. Proc. Natl. Acad. Sci. USA 117 30063–30070. 10.1073/pnas.1907378117 [PubMed: 32332161]

[7]. Bartlett PL, Montanari A and Rakhlin A (2021). Deep learning: A statistical viewpoint. Acta Numer 30 87–201. 10.1017/S0962492921000027

[8]. Belkin M, Hsu D, Ma S and Mandal S (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proc. Natl. Acad. Sci. USA 116 15849–15854. 10.1073/pnas.1903070116 [PubMed: 31341078]

[9]. Belkin M, Hsu D and Xu J (2020). Two models of double descent for weak features. SIAM J. Math. Data Sci 2 1167–1180. 10.1137/20M1336072

[10]. Belkin M, Ma S and Mandal S (2018). To understand deep learning we need to understand kernel learning. 22nd International Conference on Artificial Intelligence and Statistics. PMLR. Available at http://proceedings.mlr.press/v80/belkin18a/belkin18a.pdf.

[11]. Belkin M, Rakhlin A and Tsybakov AB (2018). Does data interpolation contradict statistical optimality? Available at arXiv:1806.09471.

[12]. Bing X, Bunea F, Strimas-Mackey S and Wegkamp M (2021). Prediction under latent factor regression: Adaptive PCR, interpolating predictors and beyond. J. Mach. Learn. Res 22 177. 10.22405/2226-8383-2021-22-1-177-187

[13]. Bishop CM (1995). Training with noise is equivalent to Tikhonov regularization. Neural Comput 7 108–116.

[14]. Bunea F, Strimas-Mackey S and Wegkamp M (2020). Interpolating predictors in high-dimensional factor regression. Available at arXiv:2002.02525.

[15]. Chatterji NS and Long PM (2021). Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. J. Mach. Learn. Res 22 129.

[16]. Chen SS, Donoho DL and Saunders MA (1998). Atomic decomposition by basis pursuit. SIAM J. Sci. Comput 20 33–61. 10.1137/S1064827596304010

[17]. Cheng X and Singer A (2013). The spectrum of random inner-product kernel matrices. Random Matrices Theory Appl. 2 1350010. 10.1142/S201032631350010X

[18]. Chizat L and Bach F (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. Adv. Neural Inf. Process. Syst 31.

[19]. Chizat L and Bach F (2019). A note on lazy training in supervised differentiable programming. Advances in Neural Information Processing Systems 32 2933–2943.

[20]. Craven P and Wahba G (1978/79). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math 31 377–403. 10.1007/BF01404567

[21]. Daubechies I (1988). Time-frequency localization operators: A geometric phase space approach. IEEE Trans. Inf. Theory 34 605–612. 10.1109/18.9761

[22]. Dicker LH (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. Bernoulli 22 1–37. 10.3150/14-BEJ609

[23]. Dobriban E and Wager S (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. Ann. Statist 46 247–279. 10.1214/17-AOS1549

[24]. Du SS, Lee JD, Li H, Wang L and Zhai X (2018). Gradient descent finds global minima of deep neural networks. International conference on machine learning. PMLR.

[25]. Du SS, Zhai X, Poczos B and Singh A (2018). Gradient descent provably optimizes over-parameterized neural networks. Available at arXiv:1810.02054.

[26]. El Karoui N (2010). The spectrum of kernel random matrices. Ann. Statist 38 1–50. 10.1214/08-AOS648

[27]. Fan Z and Montanari A (2019). The spectral norm of random inner-product kernel matrices. Probab. Theory Related Fields 173 27–85. 10.1007/s00440-018-0830-4

[28]. Geiger M, Jacot A, Spigler S, Gabriel F, Sagun L, D'ascoli S, Biroli G, Hongler C and Wyart M (2020). Scaling description of generalization with number of parameters in deep learning. J. Stat. Mech. Theory Exp 2 023401. 10.1088/1742-5468/ab633c

[29]. Gerace F, Loureiro B, Krzakala F, Mézard M and Zdeborová L (2020). Generalisation error in learning with random features and the hidden manifold model. International Conference on Machine Learning. PMLR.

[30]. Ghorbani B, Mei S, Misiakiewicz T and Montanari A (2021). Linearized two-layers neural networks in high dimension. Ann. Statist 49 1029–1054. 10.1214/20-aos1990

[31]. Goldt S, Reeves G, Mezard M, Krzakala F and Zdeborová L (2020). The gaussian equivalence of generative models for learning with two-layer neural networks. Available at arXiv:2006.14709.

[32]. Golub GH, Heath M and Wahba G (1979). Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21 215–223. 10.2307/1268518

[33]. Goodfellow I, Bengio Y and Courville A (2016). Deep Learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

[34]. Gunasekar S, Lee J, Soudry D and Srebro N (2018). Characterizing implicit bias in terms of optimization geometry. In 35th International Conference on Machine Learning, ICML 2018 2932–2955. International Machine Learning Society (IMLS).

[35]. Gunasekar S, Lee JD, Soudry D and Srebro N (2018). Implicit bias of gradient descent on linear convolutional networks. In Advances in Neural Information Processing Systems 9461–9471.

[36]. Hastie T, Montanari A, Rosset S and Tibshirani RJ (2022). Supplement to "Surprises in high-dimensional ridgeless least squares interpolation." 10.1214/21-AOS2133SUPP

[37]. Hastie T, Rosset S, Tibshirani R and Zhu J (2003/04). The entire regularization path for the support vector machine. J. Mach. Learn. Res 5 1391–1415.

[38]. Hu H and Lu YM (2020). Universality laws for high-dimensional learning with random features. Available at arXiv:2009.07669.

[39]. Jacot A, Gabriel F and Hongler C (2018). Neural tangent kernel: Convergence and generalization in neural networks. Adv. Neural Inf. Process. Syst 31.

[40]. Ke W and Thrampoulidis C (2020). Benign overfitting in binary classification of gaussian mixtures. arXiv preprint. Available at arXiv:2011.09148.

[41]. Knowles A and Yin J (2017). Anisotropic local laws for random matrices. Probab. Theory Related Fields 169 257–352. 10.1007/s00440-016-0730-4

[42]. Kobak D, Lomond J and Sanchez B (2020). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. J. Mach. Learn. Res 21 169.

[43]. Ledoit O and Péché S (2011). Eigenvectors of some large sample covariance matrix ensembles. Probab. Theory Related Fields 151 233–264. 10.1007/s00440-010-0298-3

[44]. Lee J, Xiao L, Schoenholz SS, Bahri Y, Novak R, Sohl-Dickstein J and Pennington J (2020). Wide neural networks of any depth evolve as linear models under gradient descent. J. Stat. Mech. Theory Exp 12 124002. 10.1088/1742-5468/abc62b

[45]. Li K-C (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. Ann. Statist 14 1101–1112. 10.1214/aos/1176350052

[46]. Li K-C (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. Ann. Statist 15 958–975. 10.1214/aos/1176350486

[47]. Liang T and Rakhlin A (2020). Just interpolate: Kernel “ridgeless” regression can generalize. Ann. Statist 48 1329–1347. 10.1214/19-AOS1849

[48]. Liang T, Rakhlin A and Zhai X (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In Conference on Learning Theory 2683–2711.

[49]. Mei S and Montanari A (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. Comm. Pure Appl. Math To appear.

[50]. Mei S, Montanari A and Nguyen P-M (2018). A mean field view of the landscape of two-layer neural networks. Proc. Natl. Acad. Sci. USA 115 E7665–E7671. 10.1073/pnas.1806579115 [PubMed: 30054315]

[51]. Miolane L and Montanari A (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. Ann. Statist 49 2313–2335. 10.1214/20-aos2038

[52]. Montanari A, Ruan F, Sohn Y and Yan J (2019). The generalization error of maxmargin linear classifiers: High-dimensional asymptotics in the overparametrized regime. Available at arXiv:1911.01544.

[53]. Muthukumar V, Narang A, Subramanian V, Belkin M, Hsu D and Sahai A (2021). Classification vs regression in overparameterized regimes: Does the loss function matter? J. Mach. Learn. Res 22 222.

[54]. Patil P, Wei Y, Rinaldo A and Tibshirani R (2021). Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In International Conference on Artificial Intelligence and Statistics 3178–3186. PMLR.

[55]. Pennington J and Worah P (2017). Nonlinear random matrix theory for deep learning. Adv. Neural Inf. Process. Syst 30.

[56]. Rakhlin A and Zhai X (2019). Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In Conference on Learning Theory 2595–2623. PMLR.

[57]. Richards D, Mourtada J and Rosasco L (2020). Asymptotics of ridge (less) regression under general source condition. Available at arXiv:2006.06386.

[58]. Rosset S, Zhu J and Hastie T (2003/04). Boosting as a regularized path to a maximum margin classifier. J. Mach. Learn. Res 5 941–973.

[59]. Rotskoff GM and Vanden-Eijnden E (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. Available at arXiv:1805.00915.

[60]. Rubio F and Mestre X (2011). Spectral convergence for a general class of random matrices. Statist. Probab. Lett 81 592–602. 10.1016/j.spl.2011.01.004

[61]. Serdobolskii VI (2008). Multiparametric Statistics. Elsevier, Amsterdam.

[62]. Sirignano J and Spiliopoulos K (2020). Mean field analysis of neural networks: A law of large numbers. SIAM J. Appl. Math 80 725–752. 10.1137/18M1192184

[63]. Spigler S, Geiger M, D'ascoli S, Sagun L, Biroli G and Wyart M (2019). A jamming transition from under- to over-parametrization affects generalization in deep learning. J. Phys. A 52 474001. 10.1088/1751-8121/ab4c8b

[64]. Tao T (2012). Topics in Random Matrix Theory. Graduate Studies in Mathematics 132. Amer. Math. Soc., Providence, RI. 10.1090/gsm/132

[65]. Tibshirani RJ (2015). A general framework for fast stagewise algorithms. J. Mach. Learn. Res 16 2543–2588.

[66]. Tsigler A and Bartlett PL (2020). Benign overfitting in ridge regression. Available at arXiv:2009.14286.

[67]. Tulino AM and Verdu S (2004). Random matrix theory and wireless communications. Foundations and Trends in Communications and Information Theory 1 1–182.

[68]. Webb AR (1994). Functional approximation by feed-forward networks: A least-squares approach to generalization. IEEE Trans. Neural Netw 5 363–371. [PubMed: 18267804]

[69]. Wu D and Xu J (2020). On the optimal weighted ℓ_2 regularization in overparameterized linear regression. Adv. Neural Inf. Process. Syst 33 10112–10123.

[70]. Xu J, Maleki A, Rad KR and Hsu D (2021). Consistent risk estimation in moderately high-dimensional linear regression. IEEE Trans. Inf. Theory 67 5997–6030. 10.1109/TIT.2021.3095375

[71]. Zhang C, Bengio S, Hardt M, Recht B and Vinyals O (2016). Understanding deep learning requires rethinking generalization. Available at arXiv:1611.03530.

[72]. Zhang C, Bengio S and Singer Y (2019). Are all layers created equal? Available at arXiv:1902.01996.

[73]. Zou D, Cao Y, Zhou D and Gu Q (2018). Stochastic gradient descent optimizes over-parameterized deep ReLU networks. Available at arXiv:1811.08888.

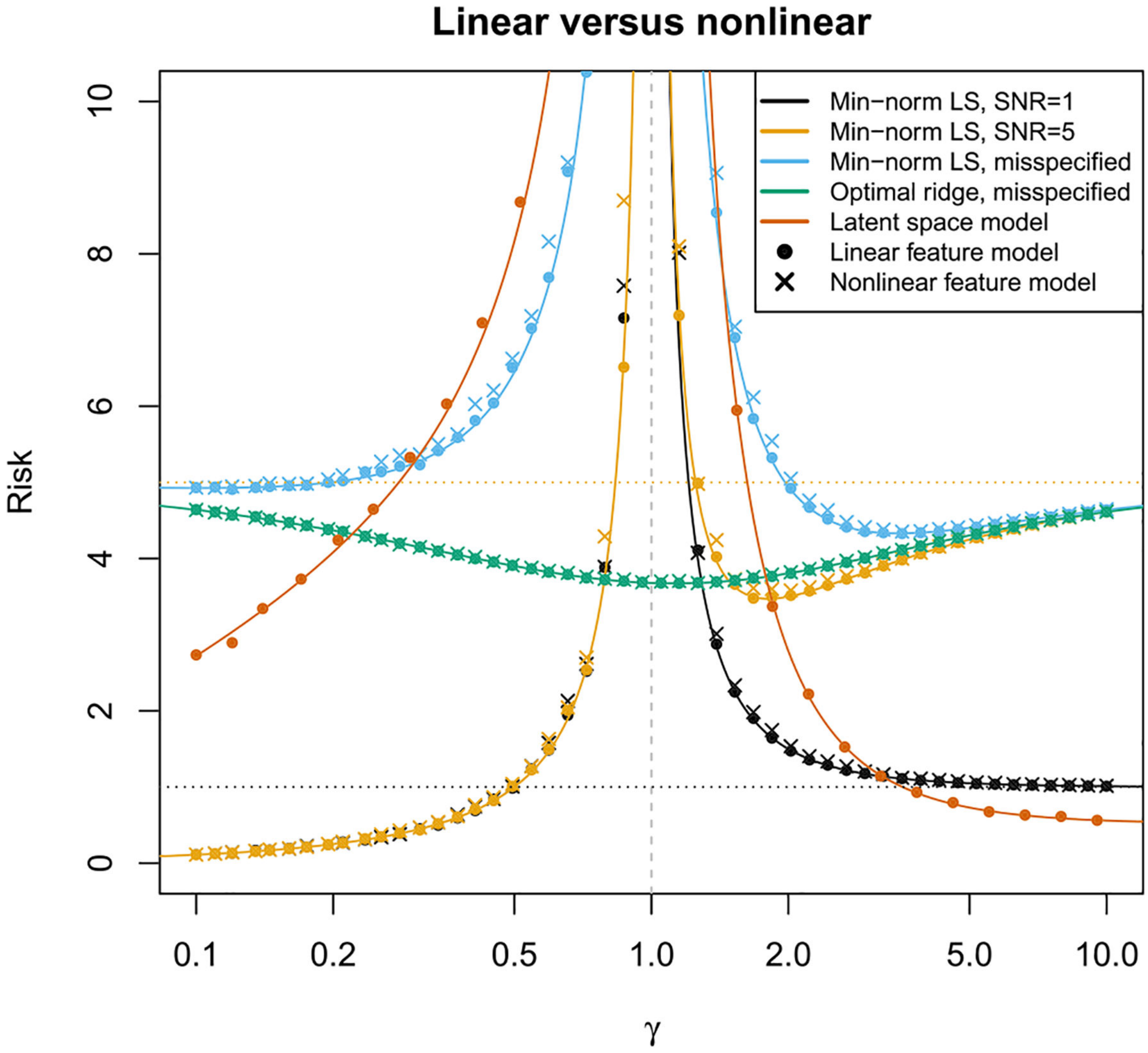

Fig. 1.

Asymptotic risk curves for the linear feature model, as a function of the limiting aspect ratio γ. Black and yellow: risks for min-norm least squares in the isotropic well-specified model, for SNR = 1 and SNR = 5, respectively. These two match for $\gamma < 1$ but differ for $\gamma > 1$. The null risks for SNR = 1 and SNR = 5 are marked by the dotted black and yellow lines, respectively. Light blue: risk for a misspecified model with significant approximation bias ($a = 1.5$ in (27)), when SNR = 5. Green: optimally-tuned (equivalently, CV-tuned) ridge regression, in the same misspecified setup as for the light blue. Red: latent space model of Section 5.2, with $r = 7$, $\sigma = 0$. The points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ. Meanwhile, the "×" points mark finite-sample risks for a nonlinear feature model, with $n = 200$, $p = [\gamma n]$, $d = 100$ *and* $X = \varphi(ZW^T)$, where Z has i.i.d. $N(0, 1)$ entries, W has i.i.d. $N(0, 1/d)$ entries and $\varphi(t) = a(|t| - b)$ is a "purely nonlinear" activation function, for constants a, b. Theorem 8 predicts that this nonlinear risk should converge to the linear risk with p features (regardless of d).

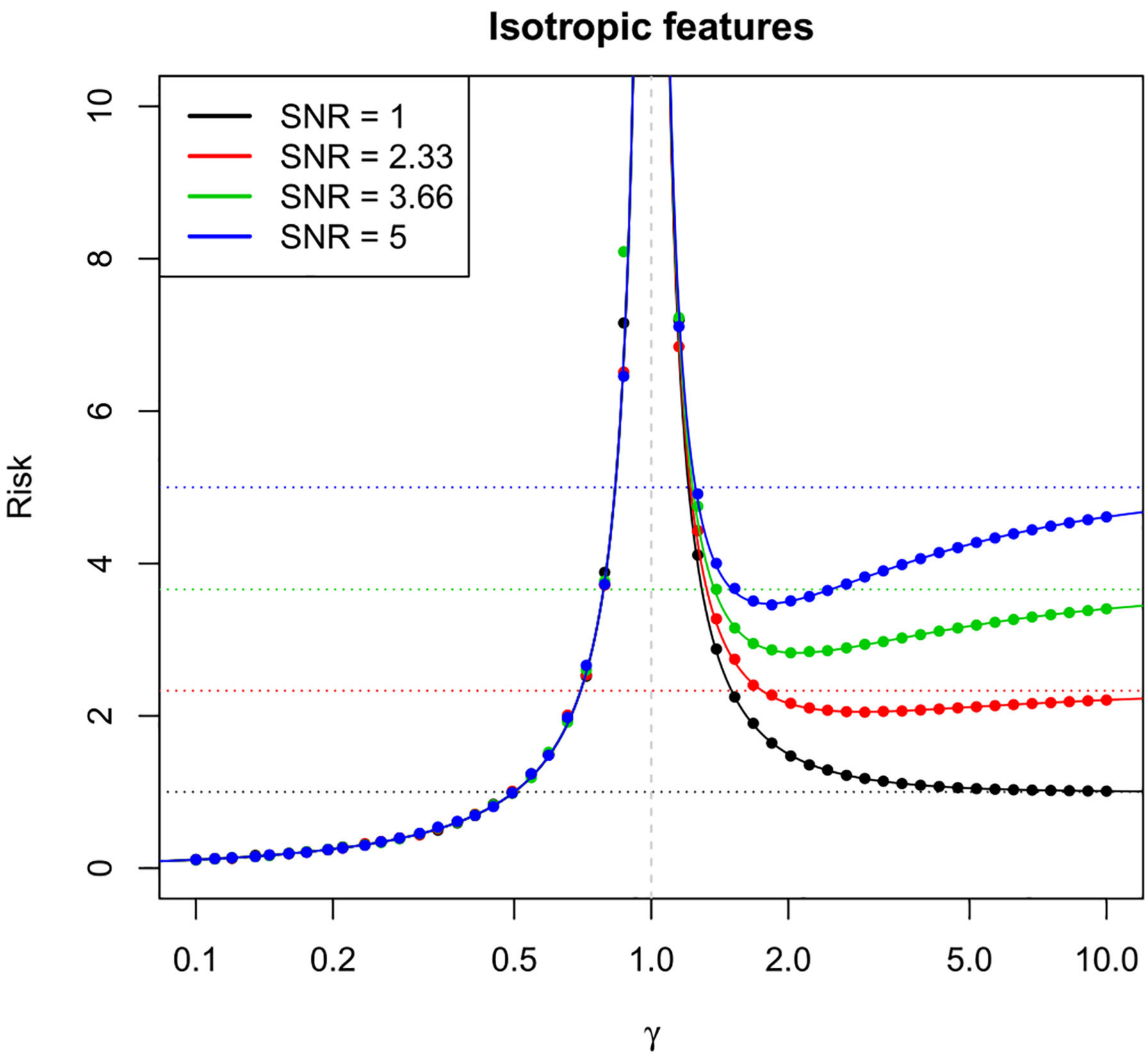

Fig. 2.

Asymptotic risk curves in (10) for the min-norm least squares estimator, when r^2 varies from 1 to 5, and $\sigma^2 = 1$. For each value of r^2, the null risk is marked as a dotted line, and the points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from features X having i.i.d. $N(0, 1)$ entries.

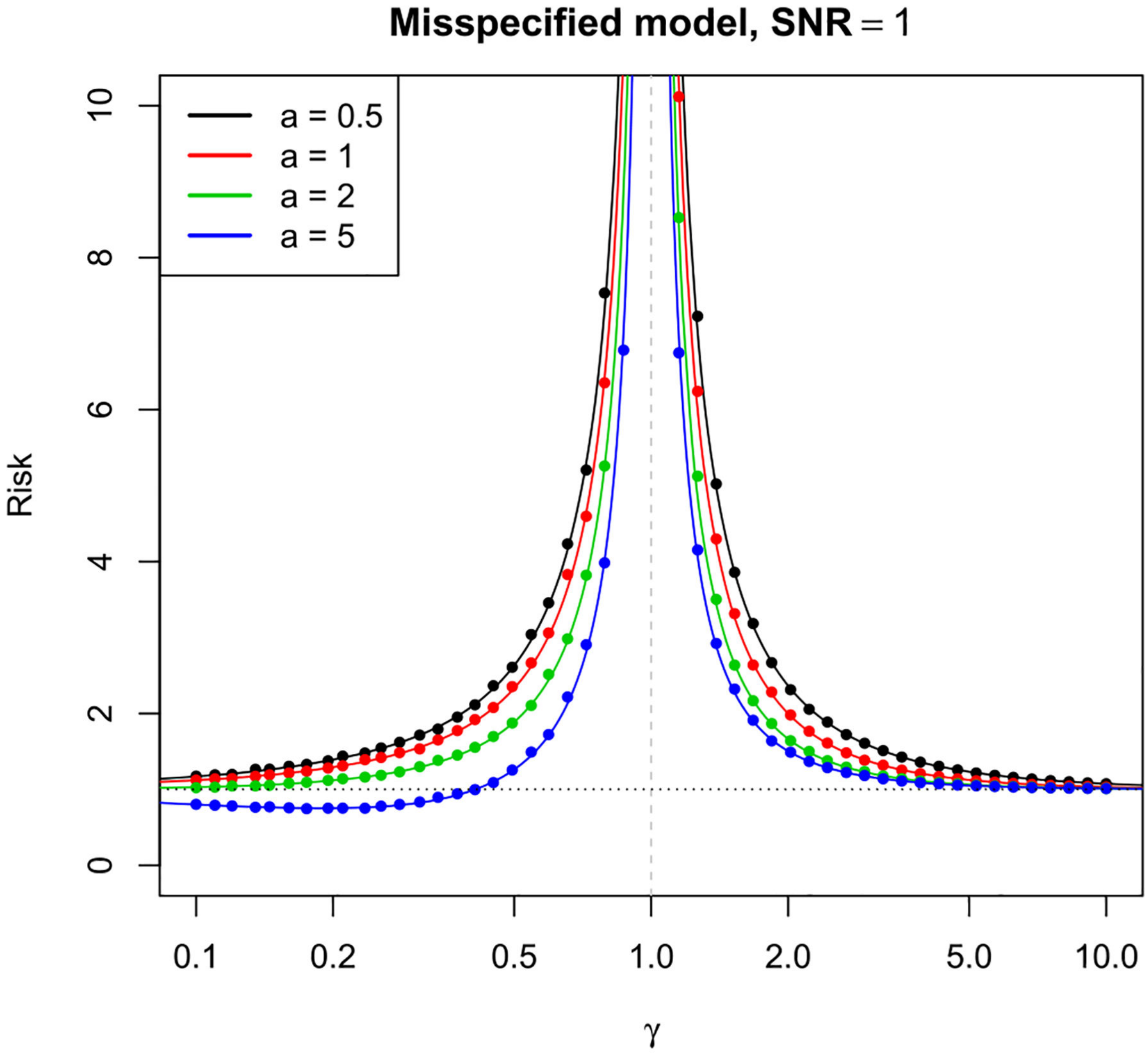

Fig. 3.

Asymptotic risk curves in (28) for the min-norm least squares estimator in the misspecified case, when the approximation bias has polynomial decay as in (27), as a varies from 0.5 to 5. Here, $r^2 = 1$ and $\sigma^2 = 1$, so SNR = 1. The null risk $r^2 = 5$ is marked as a dotted black line. The points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from features X having i.i.d. $N(0, 1)$ entries.

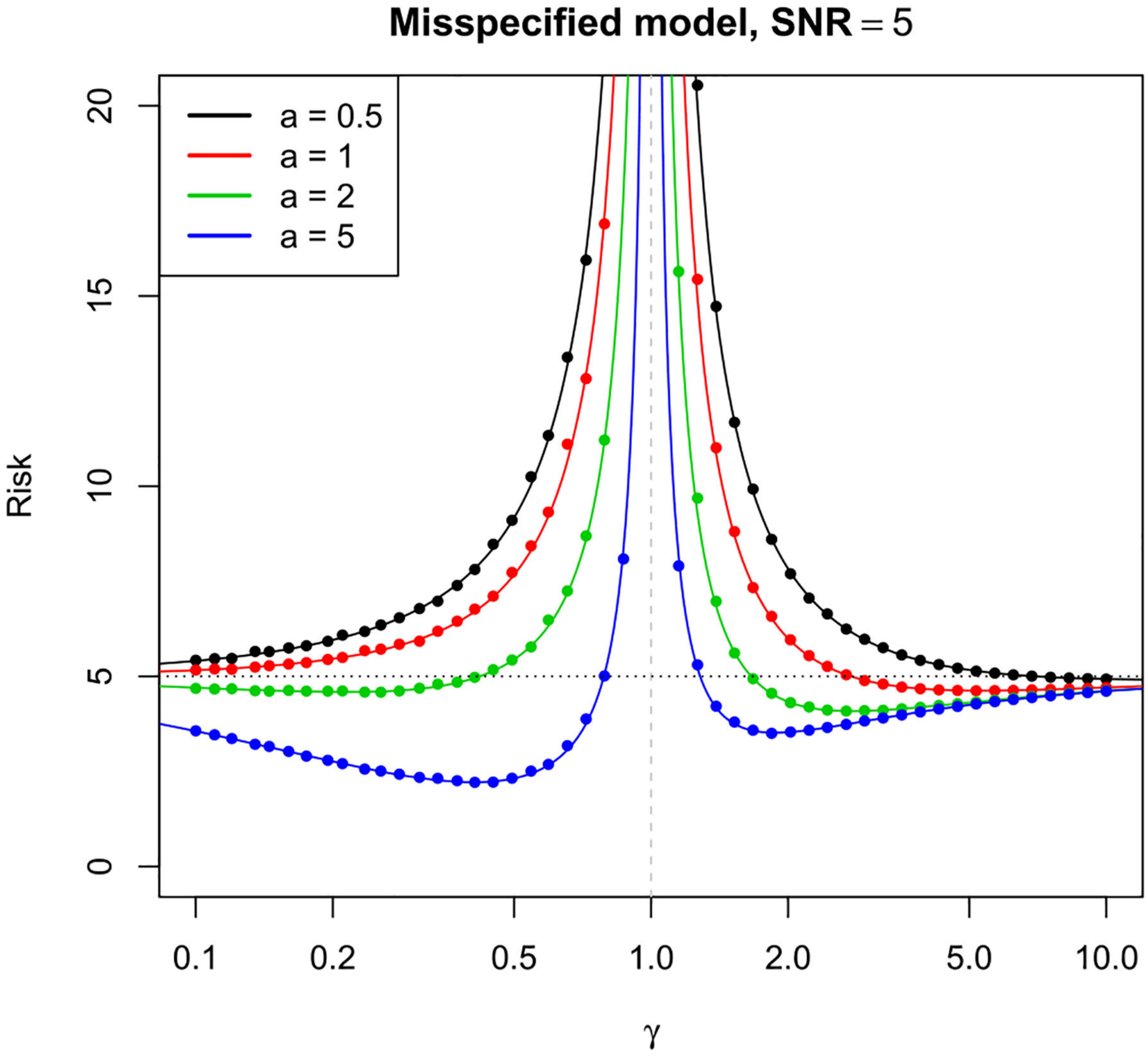

Fig. 4.

Asymptotic risk curves in (28) for the min-norm least squares estimator in the misspecified case, when the approximation bias has polynomial decay as in (27), as a varies from 0.5 to 5. Here, $r^2 = 5$ and $\sigma^2 = 1$, so SNR = 5. The null risk $r^2 = 5$ is marked as a dotted black line. The points are again finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ.

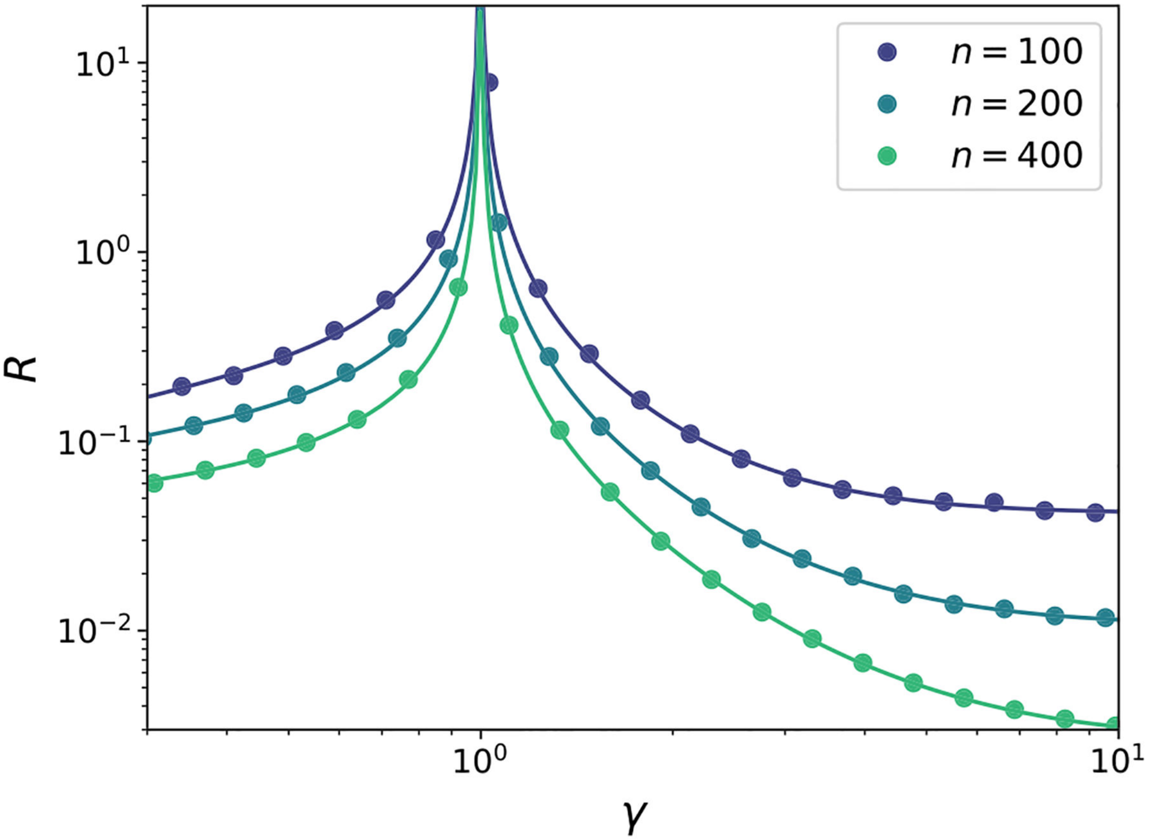

Fig. 5.

Latent space model of Section 5.2: test error $R_X(\hat{\beta}; \beta)$ of minimum norm regression as a function of the overparametrization ratio γ. Here, $d = 20$, $r = 1$, $\mu = 1$, $\sigma\xi = 0$ and n varies across various curves. Symbols are averages over 100 realizations; continuous lines report the analytical prediction of Corollary 5.

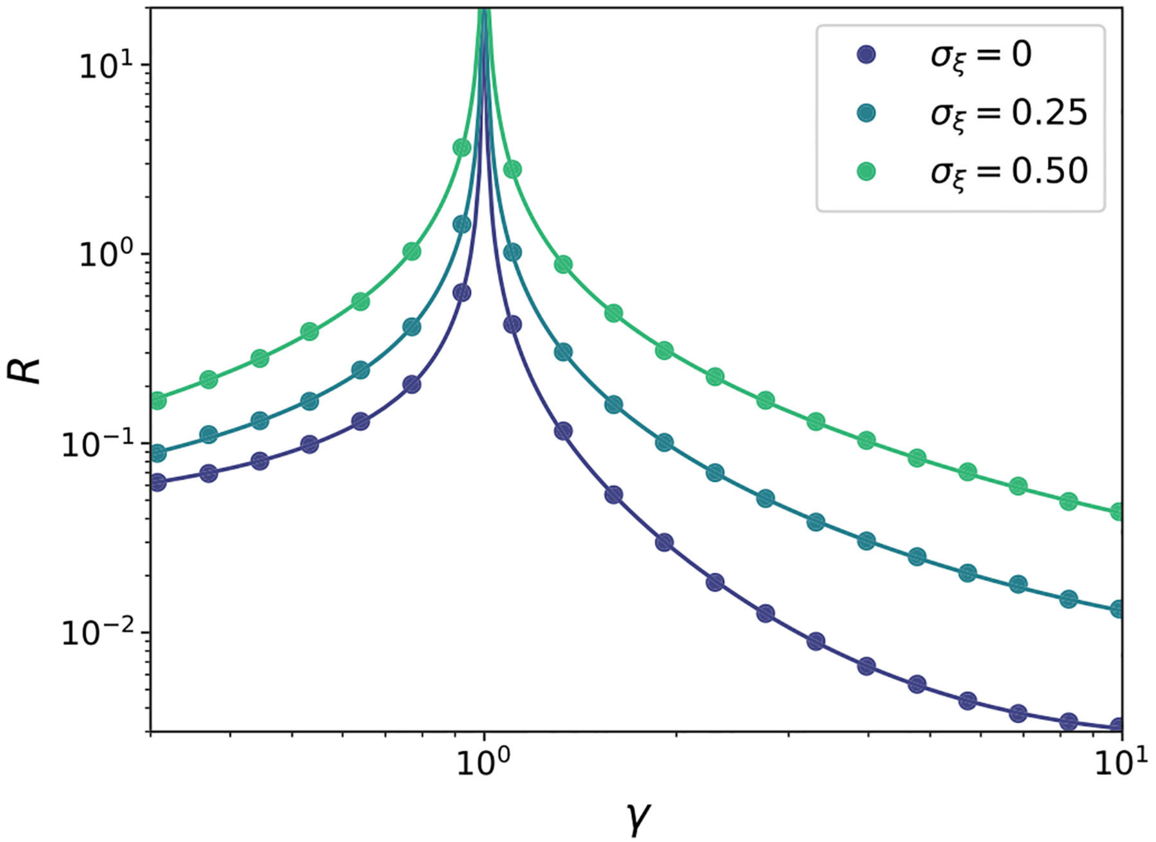

Fig. 6.

Latent space model of Section 5.2: test error $R_X(\hat{\beta}; \beta)$ of minimum norm regression as a function of the overparametrization ratio γ. Here, $d = 20$, $r = 1$, $\mu = 1$, $n = 400$ and the noise variance σ_ξ varies across different curves. Symbols are averages over 100 realizations; continuous lines report the analytical prediction of Corollary 5.

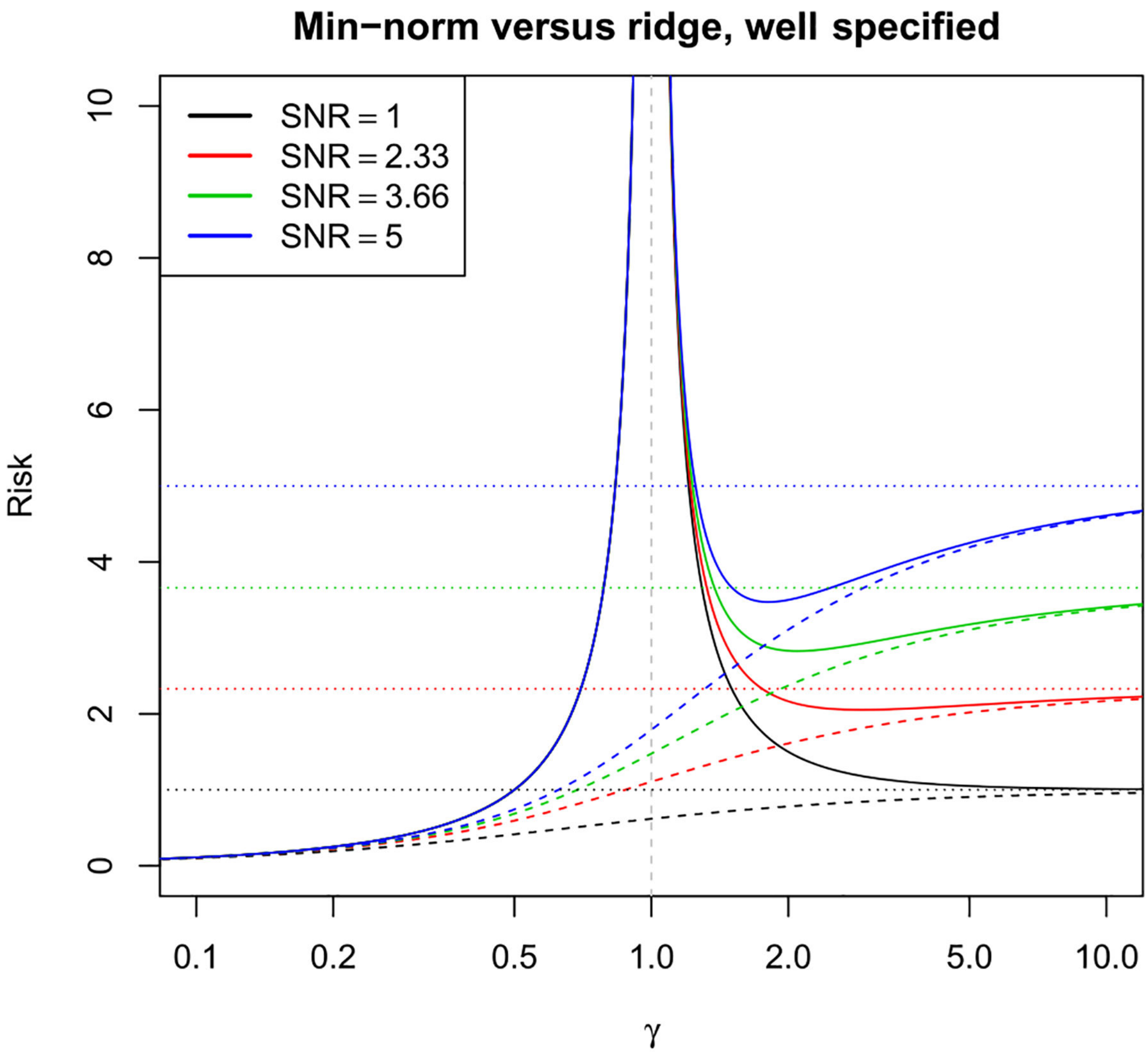

Fig. 7.

Asymptotic risk curves for the min-norm least squares estimator in (10) as solid lines, and optimally–tuned ridge regression (from Theorem 5) as dashed lines. Here, r^2 varies from 1 to 5, and $\sigma^2 = 1$. The null risks are marked by the dotted lines.

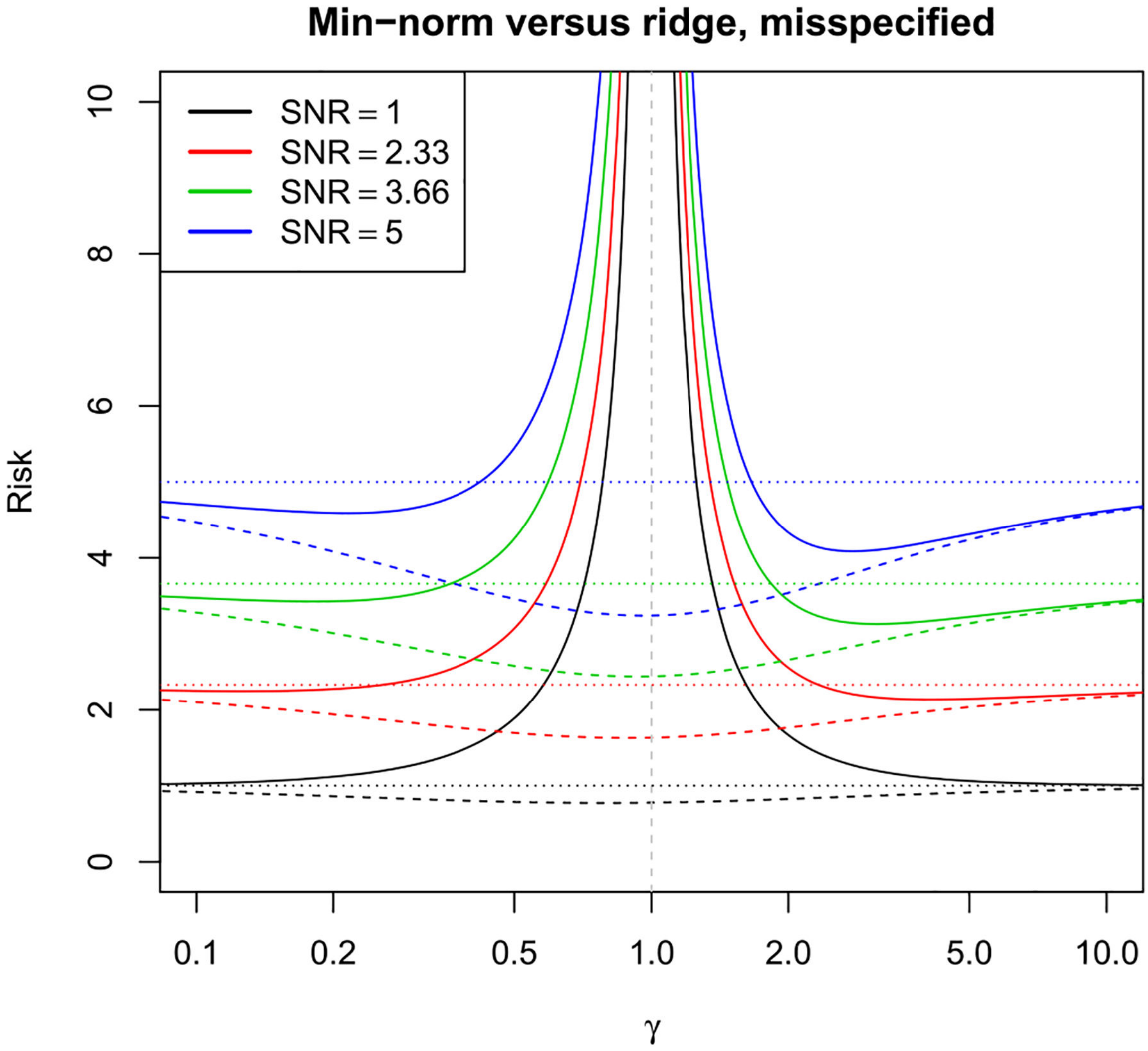

Fig. 8.

Asymptotic risk curves for the min-norm least squares estimator in (28) as solid lines, and optimally–tuned ridge regression (from Theorem 5) as dashed lines, in the misspecified case, when the approximation bias has polynomial decay as in (27), with $a = 2$. Here, r^2 varies from 1 to 5, and $\sigma^2 = 1$. The null risks are marked by the dotted lines.

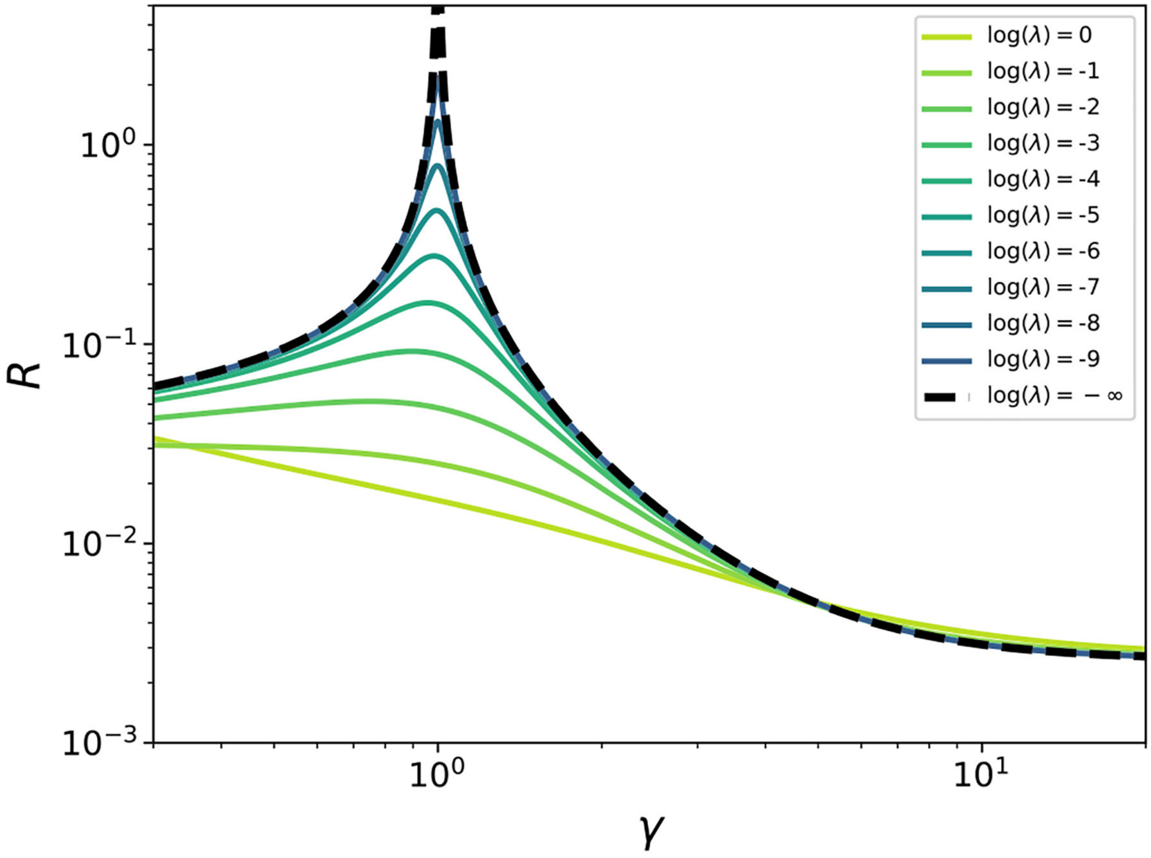

Fig. 9.

Asymptotic risk as a function of the overparametrization ratio $\gamma = p/n$, for ridge regression in the latent space model of Section 6.2. Here, $n = 400$, $d = 20$, $\mu = 1$, $r_\theta = 1$, $\sigma\xi = 0$, and each curve corresponds to a different value of the regularization λ. The dashed curve correspond to the min-norm interpolator (which coincides with the $\lambda \to 0$ limit of ridge regression).

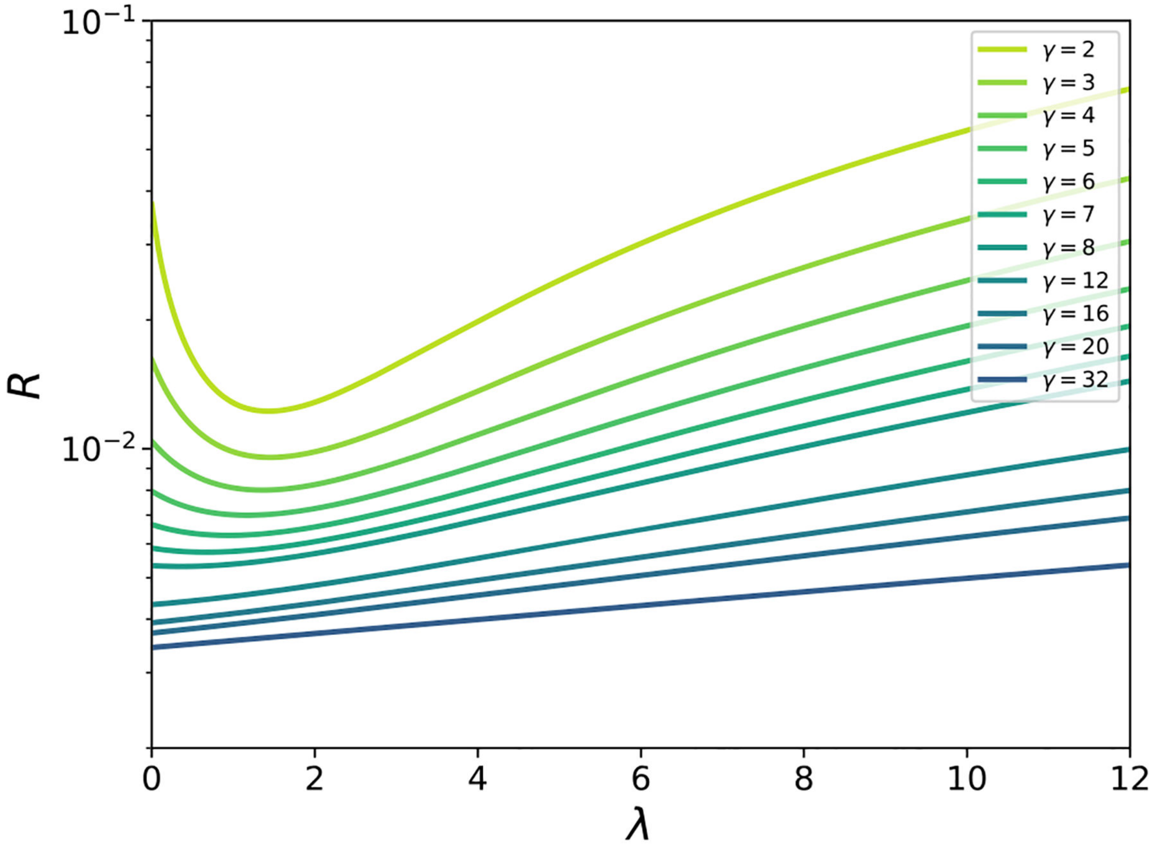

Fig. 10.

Asymptotic risk as a function of the regularization parameter λ, for ridge regression in the latent space model of Section 6.2. Here, $n = 400$, $d = 20$, $\mu = 1$, $r_\theta = 1$, $\sigma_\xi = 0.1$, and each curve corresponds to a different value of the overparametrization ratio $\gamma = p/n$.

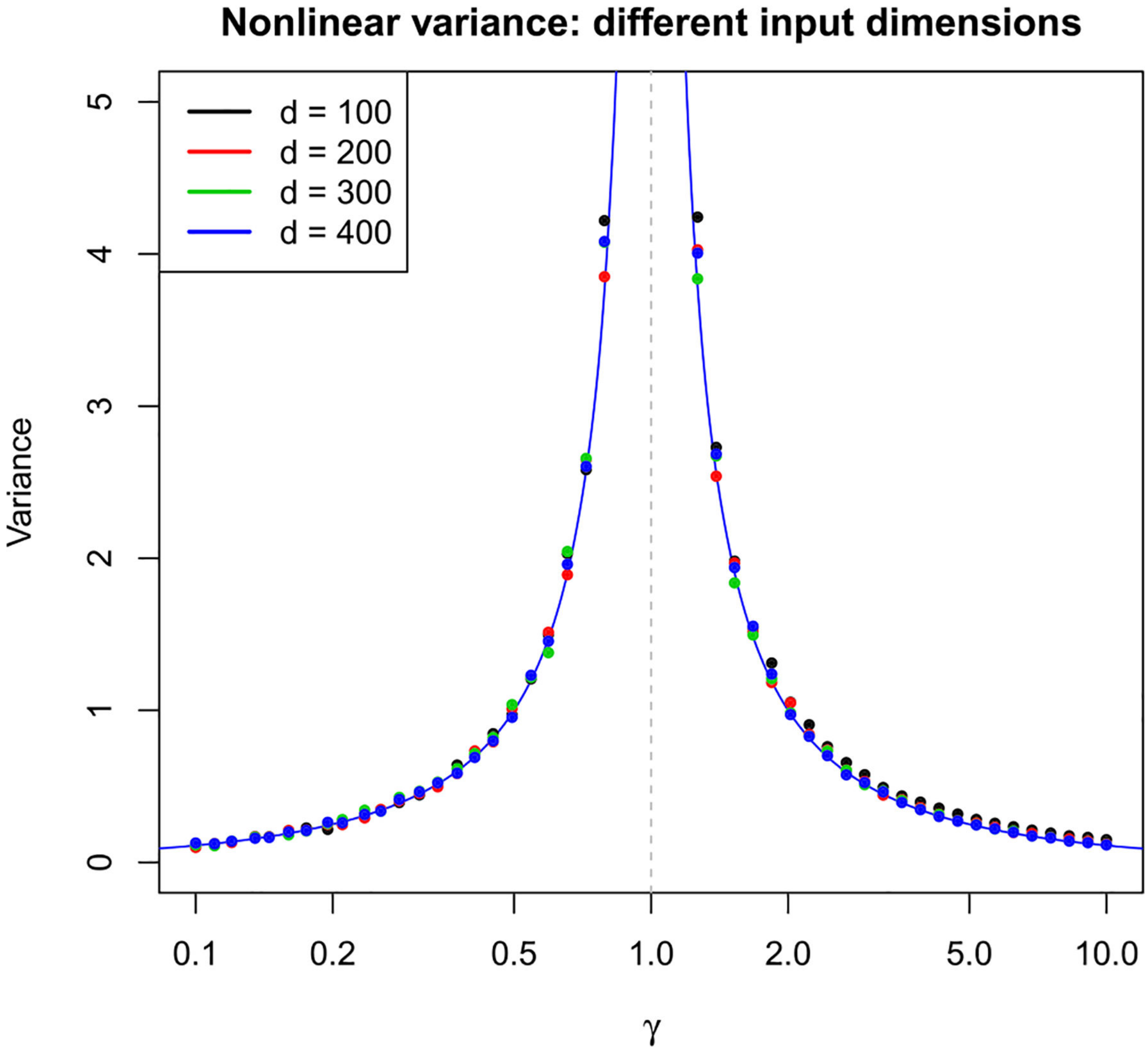

Fig. 11.

Asymptotic variance curves for the min-norm least squares estimator in the nonlinear feature model (from Theorem 8), for the purely nonlinear activation φ_{abs}. Here $\sigma^2 = 1$, and the points are finite-sample risks, with $n = 200$, $p = [\gamma n]$, over various values of γ, and varying input dimensions: $d = 100$ in black, $d = 200$ in red, $d = 300$ in green, and $d = 400$ in black. As before, the features used for finite-sample calculations are $X = \varphi(ZW^T)$, where Z has i.i.d. $N(0, 1)$ entries and W has i.i.d. $N(0, 1/d)$ entries.