# Asymptotics of Ridge (less) Regression under General Source Condition

**Dominic Richards**
University of Oxford

**Jaouad Mourtada**
CREST, ENSAE

**Lorenzo Rosasco**
MaLGa, Università di Genova, IIT
CBMM-MIT

## Abstract

We analyze the prediction error of ridge regression in an asymptotic regime where the sample size and dimension go to infinity at a proportional rate. In particular, we consider the role played by the structure of the true regression parameter. We observe that the case of a general deterministic parameter can be reduced to the case of a random parameter from a structured prior. The latter assumption is a natural adaptation of classic smoothness assumptions in nonparametric regression, which are known as source conditions in the the context of regularization theory for inverse problems. Roughly speaking, we assume the large coefficients of the parameter are in correspondence to the principal components. In this setting a precise characterisation of the test error is obtained, depending on the inputs covariance and regression parameter structure. We illustrate this characterisation in a simplified setting to investigate the influence of the true parameter on optimal regularisation for overparameterized models. We show that interpolation (no regularisation) can be optimal even with bounded signal-to-noise ratio (SNR), provided that the parameter coefficients are larger on high-variance directions of the data, corresponding to a more regular function than posited by the regularization term. This contrasts with previous work considering ridge regression with isotropic prior, in which case interpolation is only optimal in the limit of infinite SNR.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. 

## 1 Introduction

Understanding the generalisation properties of overparameterized model is a key question in machine learning, recently popularized by the study of neural networks with millions and even billions of parameters. These models perform well in practice despite perfectly fitting (interpolating) the data, a property that seems at odds with classical statistical theory (Zhang et al., 2016). This observation has lead to the investigation of the generalisation performance of methods that achieve zero training error (interpolators) (Liang et al., 2020; Belkin et al., 2018a, 2019b, 2018b, 2020) and, in the context of linear least squares, the unique least norm solution to which gradient descent converges (Hastie et al., 2019; Bartlett et al., 2020; Mitra, 2019; Belkin et al., 2020; Ghorbani et al., 2019; Muthukumar et al., 2020; Gerbelot et al., 2020; Nakkiran et al., 2020). Overparameterized linear models, where the number of variables exceed the number of points, are arguably the simplest and most natural setting where interpolation can be studied. Moreover, in some specific regimes, neural networks can be approximated by suitable linear models (Jacot et al., 2018; Du et al., 2019a,b; Allen-Zhu et al., 2019; Chizat et al., 2019).

The learning curve (test error versus model capacity) for interpolators has been shown to possibly exhibit a characteristic "Double Descent" (Advani et al., 2020; Belkin et al., 2019a) shape, where the test error decreases after peaking at an "interpolating" threshold, that is, the model capacity required to interpolate the data. The regime beyond this threshold naturally captures the settings of neural networks (Zhang et al., 2016), and thus, has motivated its investigation (Mei and Montanari, 2019; Spigler et al., 2019; Nakkiran et al., 2020). Indeed, for least squares regression, sharp characterisations double descent have been obtained for the least norm interpolating solution in the case of isotropic or auto-regressive covariates (Hastie et al., 2019; Belkin et al., 2020) and random features (Mei and Montanari, 2019).

For least squares regression the structure of the features and data can naturally influence performance. Within kernel regression (or inverse problems), for instance, it is often assumed that the parameter of interest is regular with respect to a given basis so as to to ensure a well-posed problem (Engl et al., 1996; Mathé and Pereverzev, 2003; Bauer et al., 2007). Meanwhile for neural networks, inductive biases can be encoded in the network architecture e.g. convolution layers for image classification (LeCun et al., 1989, 1998). In each case, the problem is made easier by leveraging (through model design) that data encountered in practice exhibits lower dimensional structure owing to, for example, a set of simple physical laws governing the data generation. In contrast, the least squares models investigated beyond the interpolation threshold have focused on cases where the true regression parameter is isotropic (Dobriban et al., 2018; Hastie et al., 2019), which is a single instance in the range possible of alignements between the parameter and population covariance. This has left open the natural questions of whether additional structure within the data generating distribution can be responsible for determining when interpolating is optimal.

In this work we investigate the performance of ridge regression, and its ridgeless limit, in a high dimensional asymptotic regime with a non-isotropic parameter. We show that one can naturally reduce to a parameter sampled from a prior that, in short, encodes how the signal strength is distributed across the principal components of the covariates. This structure has long been recognized as relevant in the statistics literature (Jolliffe, 1982), and is analogous to standard smoothness condition used within kernel regression and inverse problems, see e.g. (Engl et al., 1996; Mathé and Pereverzev, 2003; Bauer et al., 2007).

Specifically, a prior function encodes the parameter's norm when it is projected onto eigenspaces of the covariates population covariance. Thus, it represents how aligned the ground truth is to the principle components in the data. When considering the expected test error of ridge regression, this assumption can then encode *any* deterministic parameter (Proposition 1). Following the classic name in inverse problems, we call these assumptions *source conditions*.

Given this assumption, we then study the test error of ridge regression in a high-dimensional asymptotic regime when the number of samples and ambient dimension go to infinity in proportion to one another. The limits of resulting quantities are then characterised by utilising tools from asymptotic Random Matrix Theory (Bai and Silverstein, 2010; Ledoit and Péché, 2011; Dobriban et al., 2018; Hastie et al., 2019), with results specifically developed to characterise the influence of the prior function. This provides a natural and intuitive framework for studying the limiting test error of ridge regression, characterised by the signal to noise ratio, regularisation, overparmeterisation, and now, the structure of the regression parameter as encoded by the source condition.

We then illustrate our general framework and results in a simplified setting that highlights the role of model misspecification and its effect on prediction error and regularisation. Specifically, we consider a population covariance with two types of eigenvectors: *strong features*, associated with a common large eigenvalue (hence favored by the ridge estimator), as well as *weak features*, with a common smaller eigenvalue. This model is an idealization of a realistic structure for distributions, with some parts of the signal (associated for instance to high smoothness, or low-frequency components) easier to estimate than other, higher-frequency components. The use of source conditions allows to study situations where the true coefficients are either more or less aligned with the principal components, than implicitly postulated by the ridge estimator, a form of model misspecification which affects predictive performance. This encodes the difficulty of the problem, and allows to distinguish between "easy" and "hard" learning problems. We now summarise this work's primary contributions.

- **Asymptotic prediction error under general source condition.** An asymptotic characterisation of the test error under a general source condition on the regression parameter is provided. This required characterizing the limit of certain trace quantities, and provides a natural framework for investigating the performance of ridge regression. (Theorem 1)

- **Interpolating can be optimal even in noisy cases.** In the overparameterised regime, we show that interpolation can lead to smaller risk than any positive choice of the regularisation parameter. This occurs in the favorable situation where the regression parameter is larger in high-variance directions of the data, and the signal-to-noise ratio is large enough (but finite). Previously, for least squares regression with isotropic prior, the optimal regularisation choice was zero only in the limit of infinite signal to noise ratio (Dicker, 2016; Dobriban et al., 2018). (Section 3.1)

Our analysis of the strong and weak features model also provides asymptotic characterisations of a number of phenomena recently observed within the literature. That is, augmenting the data by adding noisy co-ordinates performs implicit regularisation and can

recover the performance of optimally tuned regression restricted to the strong features (Kobak et al., 2020). Also, we show an additional peak occurring in the learning curve beyond the interpolation threshold for the ridgeless bias and variance (Nakkiran et al., 2020). These insights are presented in Sections 3.2 and 3.3, respectively.

The remainder of this work is organized as follows. Section 1.1 covers the related literature. Section 2 describes the setting, and provides the general theorem. Section 3 formally introduces the strong and weak features model, and presents the aforementioned insights. Section 4 gives the conclusion.

### 1.1 Related Literature

Due to the large number of works investigating interpolating methods as well as double descent, we next focus on works that consider the asymptotic regime.

**High-Dimensional Statistics.** Random matrix theory has found numerous applications in high-dimensional statistics (Yao et al., 2015; El Karoui, 2018). In particular, asymptotic random matrix theory has been leveraged to study the predictive performance of ridge regression under a well-specified linear model with an isotropic prior on the parameter, for identity population covariance (Karoui and Kösters, 2011; Karoui, 2013; Dicker, 2016; Tulino et al., 2004) and then general population covariance (Dobriban et al., 2018). More recently, (Mahdaviyeh and Naulet, 2019) considered the limiting test error of the least norm predictor under the spiked covariance model (Johnstone, 2001) where both a subset of eigenvalues and the ratio of dimension to samples diverge to infinity. They show the bias is bounded by the norm of the ground truth projected on the eigenvectors associated to the subset of large eigenvalues. In contrast, our work follows standard assumption in kernel regression or inverse problems literature (Engl et al., 1996; Mathé and Pereverzev, 2003; Bauer et al., 2007), by adding structural assumptions on the parameter through the variation of its coefficients along the covariance basis. Finally, we note the works (Liao and Couillet, 2019a,b) that utilise tools from random matrix theory to characterise the prediction performance of linear estimators in the context of classification.

**Double Descent for Least Squares.** While interpolating predictors (which perfectly fit training data), are classically expected to be sensitive to noise and exhibit poor out-of-sample performance, empirical observations about the behaviour of artificial neural networks (Zhang et al., 2016) challenged this received wisdom. This surprising phenomenon, where interpolators can generalize, has first been shown for some local averaging estimators (Belkin et al., 2019b, 2018a), kernel "ridgeless" regression (Liang et al., 2020), and linear regression, where (Bartlett et al., 2020) characterised the variance of the ridgeless estimator up to universal constants. A "double descent" phenomenon for interpolating predictors, where test error can decrease past the interpolation threshold, has been suggested by (Belkin et al., 2019a).

This double descent curve has motivated a number of works established in the context of asymptotic least squares (Hastie et al., 2019; Mei and Montanari, 2019; Belkin et al., 2020; Xu and Hsu, 2019; Gerbelot et al., 2020; Muthukumar et al., 2020; Nakkiran et al., 2020). The work (Hastie et al., 2019) considers either isotropic or auto-regressive features, while (Louart et al., 2018; Mei and Montanari, 2019) consider Random Features constructed from a non-linear functional applied to the product of isotropic covariates and a random matrix. In (Hastie et al., 2019; Mei and Montanari, 2019) the data is assumed to be generated with an isotropic ground truth with some model mis-specification. The works (Mitra, 2019; Gerbelot et al., 2020; Muthukumar et al., 2020) considers recovery guarantees under sparsity assumptions on the parameter, with (Gerbelot et al., 2020) showing a peak in the test error when the number of samples equals the sparsity of the true predictor. The work (Muthukumar et al., 2020) considers recovery properties of interpolators in the non-asymptotic regime. In contrast to these works, we consider structural assumption on the ground truth in terms of the population covariance that directly follow from standard smoothness conditions in the kernel regression/ inverse problem literature.

The work (Nakkiran et al., 2020) gave empirical evidence showing additional peaks in the test error can occur beyond the interpolation threshold when the covariance and ground truth parameter are misaligned. These empirical observations are verified by the theory in this paper. Along these lines, we also note the concurrent work (Chen et al., 2020) which shows a variety of different learning curves are possible for interpolating least squares regression when the sample size is fixed and dimension of the problem is varied.

**Concurrent Work.** We now review independent work, which appeared in parallel to or since the first version of this paper. The works (Wu and Xu, 2020; Amari et al., 2020) also considers the asymptotic prediction performance of ridge regression with prior assumptions on the parameter. Similar to us, (Wu and Xu, 2020) shows that interpolating is optimal when the parameter is sufficiently "aligned" to the population covariance and the signal to noise ratio is large. Our technical

formulations are formally different but related: they express the alignment between the parameter and the population covariance in terms of the projections of $\beta$ on the eigenvectors of $\Sigma$, whereas we encode it through the source function $\Phi$; the correspondence between the two formulations is obtained through Proposition 1. They also include additional study of the sign of optimal ridge penalty. Meanwhile, (Hastie et al., 2019) has been recently updated to include refined non-asymptotic results that build upon both our work and (Wu and Xu, 2020), also accounting for the structure of the regression parameter along principal directions. They derived a general non-asymptotic bound, controlling the difference between the finite-sample risk and its high-dimensional limit.

## 2 Dense Regression with General Source Condition

In this section we formally introduce the setting as well as the main theorem. Section 2.1 introduces the linear regression setting. Section 2.2 shows the prior assumption we consider can encapsulate a general ground truth predictor. Section 2.3 introduces the functionals that arise from asymptotic random matrix theory. Section 2.4 presents the main theorem.

### 2.1 Problem Setting

We start by introducing the linear regression setting and the general source condition.

**Linear Regression.** We consider prediction in a random-design linear regression setting with Gaussian covariates. Let $\beta^\star \in \mathbb{R}^d$ denote the true regression parameter, $\Sigma \in \mathbb{R}^{d\times d}$ the population covariance, and $\sigma^2 > 0$ the noise variance. We consider an i.i.d. dataset $\{(x_i, y_i)\}_{1\le i\le n}$ such that for $i = 1, \dots, n$,

$$y_i = \langle \beta^\star, x_i\rangle + \sigma\epsilon_i, \qquad x_i \sim \mathcal{N}(0, \Sigma), \tag{1}$$

and the noise satisfies $\mathbf{E}[\epsilon_i|x_i] = 0$, $\mathbf{E}[\epsilon_i^2|x_i] = 1$. In what follows, let $Y = (y_1, \dots, y_n), \epsilon = (\epsilon_1, \dots, \epsilon_n) \in \mathbb{R}^n$, and the design matrix $X \in \mathbb{R}^{n\times d}$. Given the $n$ samples the objective is to derive an estimator $\beta \in \mathbb{R}^d$ that minimises the error of predicting a new response. For a fixed parameter $\beta^\star$, the test risk is then $R(\beta) = \mathbf{E}[(\langle x, \beta\rangle - y)^2] = \|\Sigma^{1/2}(\beta - \beta^\star)\|_2^2 + \sigma^2$, where the expectation is with respect to a new response sampled according to (1). We consider ridge regression (Hoerl and Kennard, 1970; Tikhonov, 1963), defined for $\lambda > 0$ by

$$\widehat{\beta}_\lambda := \Big(\frac{X^\top X}{n} + \lambda I\Big)^{-1}\frac{X^\top Y}{n}. \tag{2}$$

**Source Condition.** We consider an average-case analysis where the parameter $\beta^\star$ is random, sampled with covariance encoded by a *source function* $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$, which describes how coefficients of $\beta^\star$ vary along eigenvectors of $\Sigma$. Specifically, denote by $\{(\tau_j, v_j)\}_{1\le j\le d}$ the eigenvalue-eigenvector pairs of $\Sigma$, ordered so that $\tau_1 \ge \tau_2 \ge \cdots \ge \tau_d \ge 0$, and let $\Phi(\Sigma) = \sum_{i=1}^d \Phi(\tau_i) v_i v_i^\top$. For $r > 0$ the parameter $\beta^\star$ is such that

$$\mathbf{E}[\beta^\star] = 0, \qquad \mathbf{E}[\beta^\star(\beta^\star)^\top] = \frac{r^2}{d}\Phi(\Sigma). \tag{3}$$

For estimators linear in $Y$ (such as ridge regression), the expected risk only depends on the first two moments of the prior on $\beta^\star$, hence one can assume a Gaussian prior $\beta^\star \sim \mathcal{N}(0, r^2\Phi(\Sigma)/d)$. Under prior (3), $\Phi(\Sigma)^{-1/2}\beta^\star$ has isotropic covariance $I/d$, so that $\mathbf{E}\|\Phi(\Sigma)^{-1/2}\beta^\star\|^2 = 1$. This means that the coordinate $\beta_j := \langle\beta^\star, v_j\rangle$ of $\beta^\star$ in the $j$-th direction has standard deviation $\sqrt{\Phi(\tau_j)/d}$. We note that, as $d \to \infty$, $\beta^\star$ has a "dense" high-dimensional structure, where the number of its components grows with $d$, while their magnitude decreases proportionally. This prior is an average-case, high-dimensional analogue of the standard *source condition* considered in inverse problems and nonparametric regression (Mathé and Pereverzev, 2003; Bauer et al., 2007), which describes the behaviour of coefficients of $\beta^\star$ along the eigenvector basis of $\Sigma$. In the special case $\Phi(x) = x^\alpha$, $\alpha \ge 0$, one has $\mathbf{E}\|\Sigma^{-\alpha/2}\beta^\star\|^2 = r^2$. For a Gaussian prior, $\Sigma^{-\alpha/2}\beta^\star \sim \mathcal{N}(0, r^2 I/d)$, which is rotation invariant with squared norm distributed as $r^2\chi^2_d/d$ (converging to $r^2$ as $d \to \infty$), hence "close" to the uniform distribution on the sphere of radius $r$. In Section 2.2 we show, when considering the expected test error, that this source assumption can then encode any deterministic ground truth parameter.

**Easy and Hard Problems.** The case of a constant function $\Phi(x) \equiv 1$ corresponds to an isotropic prior under the Euclidean norm used for regularisation, and has been studied by (Dicker, 2016; Dobriban et al., 2018; Hastie et al., 2019). In this case (see Remark 1 below), properly-tuned ridge regression (in terms of $r^2$) is optimal in terms of average risk. The influence of $\Phi$ can be understood in terms of the average signal strength in eigen-directions of $\Sigma$. Specifically, let $v_j$ be an eigenvector of $\Sigma$, with associated eigenvalue $\tau_j$. Then, given $\beta^\star$, the signal strength in direction $v_j$ (namely, the contribution of this direction to the signal) is $\mathbf{E}_x\langle\langle\beta^\star, v_j\rangle v_j, x\rangle^2 = \tau_j\langle\beta^\star, v_j\rangle^2$, and its expectation over $\beta^\star$ is $\tau_j\Phi(\tau_j)$. When $\Phi$ is increasing, strength along direction $v_j$ decays faster as $\tau_j$ decreases, than postulated by the ridge regression penalty. In this sense, the problem is lower-dimensional, and hence "easier" than for constant $\Phi$; likewise, a decreasing $\Phi$ is associated to

a slower decay of coefficients, and therefore a "harder", higher-dimensional problem. While our results do not require this restriction, it is natural to consider functions $\Phi$ such that $\tau \mapsto \tau\Phi(\tau)$ is non-decreasing, so that principal components (with larger eigenvalue) carry more signal on average; otherwise, the norm used by the ridge estimator favours the wrong directions. In this respect, the hardest prior is obtained for $\Phi(\tau) = \tau^{-1}$, corresponding to the isotropic prior in the prediction norm induced by $\Sigma$: for this un-informative prior, all directions have same signal strength. Finally, note that in the standard nonparametric setting of reproducing kernel Hilbert spaces, source conditions are related to smoothness of the regression function (Steinwart et al., 2009).

**Remark 1 (Oracle estimator)** *The best linear (in $Y$) estimator in terms of average risk can be described explicitly. It corresponds to the Bayes-optimal estimator under prior $\mathcal{N}(0, r^2\Phi(\Sigma)/d)$ on $\beta^\star$, which writes:*

$$\widetilde{\beta} = \Big(\frac{X^\top X}{n} + \frac{\sigma^2}{r^2}\frac{d}{n}\Phi(\Sigma)^{-1}\Big)^{-1}\frac{X^\top Y}{n}. \tag{4}$$

*This estimator requires knowledge of $\Sigma$ and $r^2\Phi$. In the special case of an isotropic prior with $\Phi \equiv 1$, the oracle estimator is the ridge estimator* (2) *with $\lambda = (\sigma^2 d)/(r^2 n)$.*

### 2.2 Reduction to Source Condition

In this section, we show that the source condition introduced in Section 2.1 is not restrictive, since the general case reduces to it. Specifically, the following proposition shows that the expected error for *any* deterministic $\beta^\star \in \mathbb{R}^d$ is equal to the averaged error according to a prior with covariance of the form $\Phi(\Sigma)$ for some function $\Phi = \Phi_{\beta^\star,\Sigma}$ depending on $\beta^\star$ and $\Sigma$.

**Proposition 1 (Reduction to source condition)** *Consider data generated according to* (1)*. Let $\beta^\star \in \mathbb{R}^d$, $\beta_j = \langle \beta^\star, v_j\rangle$ for $j = 1, \dots, d$ and $\Phi = \Phi_{\beta^\star,\Sigma} : \mathbb{R}^+ \to \mathbb{R}$ be a function such that, for $\tau \in \{\tau_1, \dots, \tau_d\}$,*

$$\Phi(\tau) = \frac{d}{|J(\tau)|}\sum_{j\in J(\tau)} \beta_j^2\,, \tag{5}$$

*where $J(\tau) = \{1 \le j \le d : \tau_j = \tau\}$. Let $\Pi$ be a distribution on $\mathbb{R}^d$ such that $\mathbf{E}_{\beta\sim\Pi}[\beta] = 0$ and $\mathbf{E}_{\beta\sim\Pi}[\beta\beta^\top] = \Phi(\Sigma)/d$. Then, we have*

$$\mathbf{E}_{X,\epsilon}\big[\|\Sigma^{1/2}(\widehat{\beta}_\lambda - \beta^\star)\|_2^2\big] = \mathbf{E}_{\beta\sim\Pi}\mathbf{E}_{X,\epsilon}\big[\|\Sigma^{1/2}(\widehat{\beta}_\lambda - \beta)\|_2^2\big].$$

The equality in Proposition 1 holds for finite samples and deterministic $\beta^\star$ (and $\Sigma$), and provides a reduction to the setting of random $\beta^\star$ used in remaining sections. On a technical side, the equality in Proposition 1 holds for the expected test error, while the remaining results within this work align with prior work (Dobriban et al., 2018) where expectation is taken with respect to the parameter and noise only (conditionally on covariates $X$) i.e. $\mathbf{E}_{\epsilon,\beta^\star}[R(\widehat{\beta}_\lambda) - R(\beta^\star)] = \mathbf{E}_{\epsilon,\beta^\star}[\|\Sigma^{1/2}(\beta - \beta^\star)\|_2^2]$. Note that convergence results on the conditional risk can be integrated under suitable domination assumptions, for instance with positive ridge parameter $\lambda$. In addition, framing our next convergence results in the context of deterministic $\beta^\star$ would lead to consider source functions $\Phi_{\beta^\star,\Sigma} = \Phi_d$ depending on the dimension $d$, and converging to a fixed function $\Phi$ in a suitable sense as $d \to \infty$. For the sake of simplicity, we instead work in the setting of random parameter $\beta^\star$ with a fixed source function $\Phi$.

On another note, the generalised ridge estimator, which penalises with respect to a general covariance $\|P\beta\|_2^2$ for a positive definite matrix $P$, reduces after rescaling to standard ridge regression with an appropriate prior and covariate covariance. Namely, the problem instance with prior, penalisation and covariate covariances ($\Pi$,$P$,$\Sigma$) is equivalent to using $(P^{1/2}\Pi P^{1/2}, I, P^{-1/2}\Sigma P^{-1/2})$ with parameterisation $\widetilde{\beta}^\star = P^{1/2}\beta^\star$, $\widetilde{\beta} = P^{1/2}\beta$ and $\widetilde{X} = XP^{-1/2}$.

### 2.3 Random Matrix Theory

Let us now describe the considered asymptotic regime, as well as quantities and notions from random matrix theory that appear in the analysis.

**High-Dimensional Asymptotics.** We study the performance of the ridge estimator $\widehat{\beta}_\lambda$ under high-dimensional asymptotics (Karoui and Kösters, 2011; Karoui, 2013; Dicker, 2016; Dobriban et al., 2018; Tulino et al., 2004; Bai and Silverstein, 2010), where the number of samples and dimension go to infinity $n, d \to \infty$ proportionally with $d/n \to \gamma \in (0, \infty)$. This setting enables precise characterisation of the risk, beyond the classical regime where $n \to \infty$ with fixed true distribution.

The ratio $\gamma = d/n$ plays a key role. A value of $\gamma > 1$ corresponds to an overparameterised model, with more parameters than samples. Some care is required in interpreting this quantity: indeed, for a fixed sample size $n$, varying $\gamma$ changes $d$ and hence the underlying distribution. Hence, $\gamma$ should not be interpreted as a degree of overparmeterisation. Rather, it quantifies the sample size relatively to the dimension of the problem.

**Random Matrix Theory.** Following standard assumptions (Ledoit and Péché, 2011; Dobriban et al., 2018), assume the spectral distribution of the covari-

ance $\Sigma$ converges almost surely to a probability distribution $H$ supported on $[h_1, h_2]$ for $0 < h_1 \leq h_2 < \infty$. Specifically, denoting the cumulative distribution function of the population covariance eigenvalues as $H_d(\tau) = \frac{1}{d}\sum_{i=1}^{d} \mathbb{1}(\tau)_{[\tau_i,\infty)}$, we have $H_d(\tau) \to H(\tau)$ almost surely as $d \to \infty$.

A key quantity utilised within the analysis is the *Stieltjes Transform* of the empirical spectral distribution, defined for $z \in \mathbb{C}\backslash\mathbb{R}_+$ as $\widetilde{m}(z) := d^{-1}\operatorname{Tr}\left(\left(\frac{X^\top X}{n} - zI\right)^{-1}\right)$. Under appropriate assumptions of the covariates $x$ (see for instance (Dobriban et al., 2018)) it is known as $n, d \to \infty$ the Stieltjes Transform of the empirical covariance $\widetilde{m}(z)$ converges almost surely to a Stieltjes transform $m(z)$ that satisfies the following stationary point equation

$$m(z) = \int_0^\infty \frac{1}{\tau(1-\gamma(1+zm(z))) - z} \mathrm{d}H(\tau). \quad (6)$$

For an isotropic covariance $\Sigma = I$ the limiting spectral distribution is a point mass at one, and the above equation can be solved for $m(z)$ where it is the Stieltjes Transform of the Marchenko-Pastur distribution (Marčenko and Pastur, 1967). For general spectral densities, the stationary point equation (6) may not be easily solved algebraically, but still yields insights into the limiting properties of quantities that arise. One tool that we will use to gain insights will be the *companion transform* $v(z)$ which is the Stieltjes transform of the limiting spectral distribution of the Gram matrix $n^{-1}XX^\top$. It is related to $m(z)$ through the following equality $\gamma(m(z) + 1/z) = v(z) + 1/z$ for all $z \in \mathbb{C}\backslash\mathbb{R}_+$. Finally, introduce the $\Phi$-weighted Stieltjes Transform defined for $z \in \mathbb{C}\backslash\mathbb{R}_+$

$$\Theta^\Phi(z) := \int \Phi(\tau) \frac{1}{\tau(1-\gamma(1+zm(z))) - z} \mathrm{d}H(\tau),$$

which is the limit of the trace quantity $d^{-1}\operatorname{Tr}\left(\Phi(\Sigma)(\frac{X^\top X}{n} - zI)^{-1}\right)$ (Ledoit and Péché, 2011).

### 2.4 Main Theorem: Asymptotic Risk under General Source Condition

Let us now state the main theorem of this work, which provides the limit of the ridge regression risk.

**Theorem 1** *Consider the setting described in Section 2.1 and 2.3. Suppose $\Phi$ is a real-valued bounded function defined on $[h_1, h_2]$ with finitely many points of discontinuity and let $v'(z) = \partial v(z)/\partial z$. If $n, d \to \infty$ with $\gamma = d/n \in (0, \infty)$ then almost surely $\mathbf{E}_{\epsilon,\beta^\star}[R(\widehat{\beta}_\lambda) - R(\beta^\star)] \to R_{\mathrm{Asym}}(\lambda)$ where*

$$R_{\mathrm{Asym}}(\lambda) = \underbrace{\sigma^2\Big(\frac{v'(-\lambda)}{v(-\lambda)^2} - 1\Big)}_{Variance} + \underbrace{r^2 \frac{\Theta^\Phi(-\lambda) + \lambda\frac{\partial\Theta^\Phi(-\lambda)}{\partial\lambda}}{v(-\lambda)^2}}_{Bias}.$$

The above theorem characterises the expected test error of the ridge estimator when the sample size and dimension go to infinity $n, d \to \infty$ with $d/n = \gamma \in (0, \infty)$, and $\beta^\star$ is distributed as (3). The asymptotic risk in Theorem 1 is characterised by the relative sample size $\gamma$, the limiting spectral distribution $H$, and the source function $\Phi$ (normalising $\sigma^2 = r^2 = 1$). This provides a general form for studying the asymptotic test error for ridge regression in a dense high-dimensional setting. The source condition affects the limiting bias; to evaluate it we are required to study the limit of the trace quantity $d^{-1}\operatorname{Tr}\left(\Sigma(\frac{X^\top X}{n} - zI)^{-1}\Phi(\Sigma)(\frac{X^\top X}{n} - zI)^{-1}\right)$, which is achieved utilising techniques from both (Chen et al., 2011) and (Ledoit and Péché, 2011) (key steps in proof of Lemma 2 Appendix B). The variance term in Theorem 1 aligns with that seen previously in (Dobriban et al., 2018), as the structure of $\beta^\star$ only influences the bias.

We now give some examples of asymptotic expected risk in Theorem 1 for 3 different structures of $\beta^\star$, namely $\Phi(x) = 1$ (isotropic), $\Phi(x) = x$ (easier case) and $\Phi(x) = x^{-1}$ (harder case).

**Corollary 1** *Consider the setting of Theorem 1. If $n, d \to \infty$ with $\gamma = d/n$, then almost surely*

$$\mathbf{E}_{\epsilon,\beta^\star}[R(\widehat{\beta}_\lambda) - R(\beta^\star)] \to \sigma^2\Big(\frac{v'(-\lambda)}{(v(-\lambda))^2} - 1\Big) + r^2 \begin{cases} \frac{v'(-\lambda)}{\gamma(v(-\lambda))^4} - \frac{1}{\gamma v(-\lambda)^2} & \text{if } \Phi(x) = x \\ \frac{1}{\gamma v(-\lambda)} - \frac{\lambda}{\gamma}\frac{v'(-\lambda)}{(v(-\lambda))^2} & \text{if } \Phi(x) = 1 \\ \frac{2\lambda}{\gamma}\frac{v'(-\lambda)}{v(-\lambda)} + (1-\frac{1}{\gamma})\frac{v'(-\lambda)}{v(-\lambda)^2} - \frac{1}{\gamma} & \text{if } \Phi(x) = 1/x \end{cases}$$

The three choices of source function $\Phi$ in Corollary 1 are cases where the asymptotic bias in Theorem 1 can be expressed in terms of the companion transform and its first derivative. The expression in the case $\Phi(x) = 1$ was previously investigated in (Dobriban et al., 2018), while for $\Phi(x) = x$ the bias aligns with quantities previously studied in (Chen et al., 2011), and thus, can be simply plugged in. For $\Phi(x) = x^{-1}$, algebraic manipulations similar to the $\Phi(x) = x$ case allow $\Theta^\Phi(z)$ to be simplified. Finally, for $\Phi(x) = 1$ it is clear how the bias and variance can be brought together and simplified yielding optimal regularisation choice $\lambda = \sigma^2\gamma/r^2$ (Dobriban et al., 2018), see also Remark 1. As noted in Section 2.1, $\Phi(x) = x^{-1}$ corresponds to a "harder" case, with no favoured direction, while $\Phi(x) = x$ corresponds to an "easier" case with faster coefficient decay.

## 3 Strong and Weak Features Model

In this section we consider a particular covariance structure, the *strong and weak features model*. Let

$U_1 \in \mathbb{R}^{d_1 \times d}$ and $U_2 \in \mathbb{R}^{d_2 \times d}$ be orthonormal matrices whose rows forms an orthonormal basis of $\mathbb{R}^d$ and $d_1 + d_2 = d$. The covariance is then for $\rho_1 \geq \rho_2 > 0$

$$\Sigma = \rho_1 U_1^\top U_1 + \rho_2 U_2^\top U_2. \tag{7}$$

We call elements of the span of rows of $U_1$ *strong features*, as they are associated to the dominant eigenvalue $\rho_1$. Similarly, $U_2$ is associated to the *weak features*. The size of $U_1, U_2$ go to infinity $d_1, d_2 \to \infty$ with the sample size $n \to \infty$, with $d_i/d \to \psi_i \in (0,1)$ and thus $\psi_1 + \psi_2 = 1$. The limiting population spectral measure is then atomic $dH(\tau) = \psi_1 \delta_{\rho_1} + \psi_2 \delta_{\rho_2}$.

The parameter $\beta^\star$ has covariance $\mathbf{E}[\beta^\star (\beta^\star)^\top] = \frac{r^2}{d}\big(\phi_1 U_1^\top U_1 + \phi_2 U_2^\top U_2\big)$, where $\phi_1, \phi_2$ are the coefficients for each type of feature and the source condition is $\Phi(x) = \phi_1 \mathbb{1}_{x=\rho_1} + \phi_2 \mathbb{1}_{x=\rho_2}$. The coefficients $\phi_1, \phi_2$ encode the composition of the ground truth in terms of strong and weak features, and thus, the difficulty of the estimation problem. The case $\phi_1 = \phi_2$ corresponds to the isotropic prior, while the case $\phi_1 > \phi_2$ corresponds to faster decay and hence an "easier" problem. Specifically, if $\phi_1 > \phi_2$ increases, $\beta^\star$ has faster decay, the problem becomes "easier" since the ground truth is increasingly made of strong features. Therefore, if $\phi_1/\phi_2 \geq 1$ we say the problem is *easy*, while if $\phi_1/\phi_2 < 1$ we say the problem is *hard*.

Under the model just introduced, Theorem 1 provides the following asymptotic characterization for the expected test risk as $n, d \to \infty$

$$R_{\mathrm{Asym}}(\lambda) = \frac{v'(-\lambda)}{v(-\lambda)^2}\Big(\sigma^2 + r^2 \sum_{i=1}^{2} \frac{\phi_i \psi_i \rho_i}{(\rho_i v(-\lambda) + 1)^2}\Big) - \sigma^2.$$

To gain insights into the performance of least squares when data is generated from the strong and weak features model, we now investigate the above limit in the overparameterised setting $\gamma > 1$. The insights are summarised in the following sections. Section 3.1 shows that zero regularisation is optimal for easy problems with high signal to noise ratio. Section 3.2 shows how weak features can be used as a form of regularisation similar to ridge regression. Section 3.3 present findings related to the ridgeless bias and variance.

**Source Condition Reduction for Strong and Weak Features Model.** Following Section 2.2, the case of a general deterministic parameter $\beta^\star \in \mathbb{R}^d$ can be encoded by the strong and weak features setting just described. Namely, let $\beta_1, \beta_2$ be the respective projections of $\beta^\star$ onto rows of $U_1, U_2$. Then $\phi_1 = \|\beta_1\|_2^2 d/d_1 =$ and $\phi_2 = \|\beta_2\|_2^2 d/d_2$, and thus, the coefficients $\phi_1, \phi_2$ align with the norm of the ground truth $\beta^\star$ projected on each bulk.

### 3.1 Interpolating can be optimal in the presence of noise

In this section, we investigate how the true regression function, namely the parameter $\beta^\star$ (through the source condition), affects optimal ridge regularisation. We begin with the following corollary, which, in short, describes when zero regularisation can be optimal. Let us denote the derivative of the asymptotic risk $R_{\mathrm{Asym}}(\lambda)$ with respect to the regularisation $\lambda$ as $R'_{\mathrm{Asym}}(\lambda)$.

**Corollary 2** *Consider the strong and weak features model with $\gamma = 2$, $\psi_1 = \psi_2 = 1/2$ and $\mathbf{E}[\|\beta^\star\|_2^2] = r^2$. If*

$$\underbrace{\frac{r^2}{\sigma^2}\frac{\rho_1\rho_2}{(\sqrt{\rho_1}+\sqrt{\rho_2})^2}}_{\textit{Signal to Noise Ratio}}\Big(\underbrace{\frac{\phi_1\sqrt{\rho_1}+\phi_2\sqrt{\rho_2}}{\sqrt{\rho_1}+\sqrt{\rho_2}} - 1}_{\textit{Alignment}}\Big) \geq 1$$

*then $R'_{\mathrm{Asym}}(0) \geq 0$. Otherwise $R'_{\mathrm{Asym}}(0) < 0$.*

Corollary 2 states that if the ground truth is aligned $\phi_1 > 1$ (the signal concentrates more on strong features) and the signal to noise ratio $r^2/\sigma^2$ is sufficiently large, then the derivative of the asymptotic test error at zero regularisation is positive $R'_{\mathrm{Asym}}(0) \geq 0$. The interpretation being that interpolating is optimal if adding regularisation increases (locally at 0) the test error. The case $\gamma = 2$ and $\psi_1 = \psi_2 = 1/2$ is considered, as the companion transform at zero takes a simple form $v(0) = 1/\sqrt{\rho_1\rho_2}$, allowing the derivatives $v'(0)$ and $v''(0)$, and thus $R'_{\mathrm{Asym}}(0)$, to be tractable.

Looking to Figure 1 plots for the performance of optimally tuned ridge regression (*Left*) and the optimal choice of regularisation parameter (*Right*) against (a monotonic transform) of the eigenvalue ratio $\rho_1/\rho_2$, for a coefficient ratios $\phi_1 \geq \phi_2$ have been given. As shown in the right plot of Figure 1, for a fixed distribution of $X$ (characterised by $\psi_1, \rho_1, \rho_2$) and sample size (characterised by $\gamma$) as the ratio $\phi_1/\phi_2$ increases the optimal regularisation decreases. Following Corollary 2, if the ratio $\phi_1/\phi_2$ is large enough, the optimal ridge regularisation parameter $\lambda$ can be 0, corresponding to ridgeless interpolation. We note that the negative derivative at 0 (Corollary 2) and the right plot of Figure 1, see also (Kobak et al., 2020; Wu and Xu, 2020).

**Comparison with the Isotropic Model.** In the case of a parameter $\beta^\star$ drawn from an isotropic prior $\Phi \equiv 1$ (see Section 2.1), the optimal ridge parameter is given by $\lambda = (\sigma^2 d)/(r^2 n)$ (see Remark 1, as well as (Dobriban et al., 2018; Hastie et al., 2019)). This parameter is always positive, and is inversely proportional to the signal-to-noise ratio $r^2/\sigma^2$. Studying the influence of $\beta^\star$ through a general $\phi_1, \phi_2$ shows that (1) optimal regularisation also depends on the coefficient

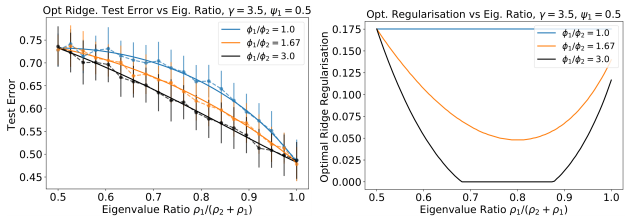


Figure 1: *Left*: Limiting test error for optimally tuned ridge regression as described by $R_{\text{Asym}}(\lambda)$, *Right*: optimal regularisation computed numerically using theory. *Both*: Quantities plotted against Eigenvalue ratio $\rho_1/(\rho_1+\rho_2)$. Problem parameters were $\mathbf{E}[\langle x, \beta^\star\rangle^2] = \rho_1\phi_1\psi_1 + \rho_2\phi_2\psi_2 = 1$, $\mathbf{E}[\|\beta^\star\|_2^2] = r^2(\phi_1\psi_1 + \phi_2\psi_2) = r^2 = 1$, $\sigma^2 = 0.05$, $\gamma = 3.5$ and $\psi_1 = 0.5$. *Left*: Dashed lines indicate simulations with $d = 2^{10}$, 40 replications, noise $\epsilon$ from standard Gaussian and covariance $\Sigma$ diagonal with $\rho_1$ on first $d_1$ co-ordinates and $\rho_2$ on remaining $d_2$.

decay of $\beta^\star$; (2) optimal regularisation can be equal to $\lambda = 0$, which interpolates training data. Finally, let us note that the optimal estimator of Remark 1 (with oracle knowledge of $\Sigma, \Phi$) does *not* interpolate; hence, the optimality of interpolation among the family of ridge estimators arises from a form of "prior misspecification". We believe this phenomenon to extend beyond the specific case of ridge estimators.

### 3.2 The Special Case of Noisy Weak Features

In this section we consider the special case where weak features are pure noise variables, namely $\phi_2 = 0$, while their dimension is large. Such noisy weak features can be artificially introduced to the dataset, to induce an overparameterised problem. We then refer to this technique as *Noisy Feature Regularisation*, and note it corresponds to the design matrix augmentation in (Kobak et al., 2020). Looking to Figure 2, the ridgeless test error is then plotted against the eigenvalue ratio $\rho_2/\rho_1$ (*Left*) and the number of weak features with the tuned eigenvalue ratio (*Right*).

Observe (right plot) as we increase the number of weak features (as encoded by $1/\psi_1$), and tune the eigenvalue $\rho_2$, the performance converges to optimally tuned ridge regression with the strong features only. The left plot then shows the "regularisation path" as a function of the eigenvalue ratio $\rho_2/\rho_1$ for some numbers of weak features $1/\psi_1$.

**Weak Features Can Implicitly Regularise.** The results in Sections 3.1 and 3.2 suggest that weak features can implicitly regularise when the ground truth is associated to a subset of stronger features. Specifically, Section 3.1 demonstrated how this can occur passively in an easy learning problem, with the weak features providing sufficient stability that zero ridge regularisation can be the optimal choice [1]. Meanwhile, in this section we demonstrated an active approach where weak features can purposely be added to a model and tuned similar to ridge. We note the recent work (Jacot et al., 2020) which shows a similar implicit regularisation phenomena for kernel regression.

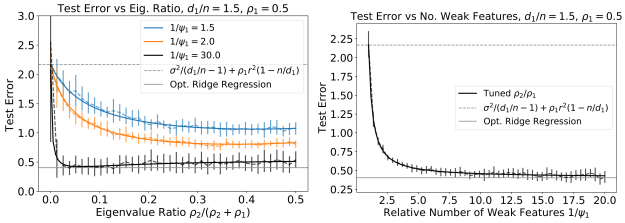


Figure 2: Ridgeless test error for strong and weak features model ($r^2 = \sigma^2 = 1$) against eigenvalue ratio $\rho_2/\rho_1$ (*Left*) and size of noisy bulk $1/\psi_1 = d_1/d$ (*Right*). Solid lines show theory computed using $v(0)$ with $\gamma\psi_1 = d_1/n = 1.5$ and $\rho_1 = 0.5$. Dashed lines are simulations with $d = 2^8$ (*Left*) and $2^{10}$ (*Right*) and 20 replications. *Solid Grey Horizontal Line*: Performance of optimally tuned ridge regression with strong features only.

### 3.3 Ridgeless Bias and Variance

In this section we investigate how the ridgeless bias and variance depend on the ratio of dimension to sample size $\gamma$. Looking to Figure 3 the ridgeless bias and variance is plotted against the ratio of dimension to sample size in the overparameterised regime $\gamma \geq 1$ .

Note an additional peak in the ridgeless bias and variance is observed beyond the interpolation threshold. This has only recently been empirically observed for the test error (Nakkiran et al., 2020), as such, these plots now theoretically verify this phenomenon. The location of the peaks naturally depends on the number of strong and weak features as well as the ambient dimension, as denoted by the vertical lines. Specifically, the peak occurs in the ridgeless bias for the "hard" setting when the number of samples and number of strong features are equal $n = d_1$. Meanwhile, a peak occurs in the ridgeless variance when the number of samples and strong features equal $n = d_1$, and the eigenvalue ratio is large $\rho_1 > \rho_2$. This demonstrates that learning curves beyond the interpolation threshold can have different characteristics due to the interplay between the

[1] Zero regularisation has been shown to be optimal for Random Feature regression with a high signal to noise ratio (Mei and Montanari, 2019) and a misspecified component. The work (Kobak et al., 2020) numerically estimated $R'_{\text{Asym}}(\lambda)$ for a spiked covariance model and found it can be positive.

covariate structure and underlying data. We conjecture this arises due to instabilities of the design matrix Moore-Penrose Pseudo-inverse, akin to the isotropic setting (Belkin et al., 2020). Since variance matches prior work (Dobriban et al., 2018), the additional peak could be previously derived. Meanwhile the peak in the bias here uses of the source condition, and thus, as far as we aware is not encompassed in prior work.

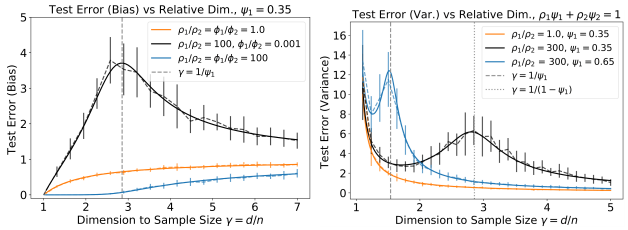


Figure 3: Ridgeless bias and variance for strong and weak feature model plotted against relative dimension $\gamma = d/n$ with various eigenvalue ratios $\rho_1/\rho_2$ and coefficients $\phi_1/\phi_2$. Solid lines are theory computed using $v(0)$ with $\psi_1 = 0.35$, $\mathbf{E}[\langle x, \beta^\star\rangle^2] = \rho_1\phi_1\psi_1 + \rho_2\phi_2\psi_2 = 1$ and $\mathbf{E}[\|\beta^\star\|_2^2] = r^2(\phi_1\psi_1 + \phi_2\psi_2) = 1$. Dashed lines are simulations with $d = 2^8$ and 20 replications.

## 4 Conclusion

In this work, we introduced a framework for studying ridge regression in a high-dimensional regime. We characterised the limiting risk of ridge regression in terms of the dimension to sample size ratio, the spectrum of the population covariance and the coefficients of the true regression parameter along the covariance basis. This extends prior work (Dicker, 2016; Dobriban et al., 2018), that considered an isotropic ground truth parameter. Our extension enables the study of "prior misspecification", where signal strength may decrease faster or slower than postulated by the ridge estimator, and its effect on ideal regularisation.

We instantiated this general framework to a simple structure, with strong and weak features. In this case, we show that "ridgeless" regression with zero regularisation can be optimal among all ridge regression estimators. This occurs when the signal-to-noise ratio is large and when strong features (with large eigenvalue of the covariance matrix) have sufficiently more signal than weak ones. The latter condition corresponds to an "easy" or "lower-dimensional" problem, where ridge tends to over-penalise along strong features. This phenomenon does not occur for isotropic priors, where optimal regularisation is always strictly positive in the presence of noise. Finally, we discussed noisy weak features, which act as a form of regularisation, and concluded by showing additional peaks in ridgeless bias and variance can occur for our model.

Moving forward, it would be natural to consider non-Gaussian covariates. Given universality results in Random Matrix Theory we expect that the results provided here extend to the case of random vectors with independent coordinates (and linear transformations thereof). Other structures for the ground truth and data generating process can be investigated through Theorem 1 by consider different functions $\Phi$ and the population eigenvalue distributions. The tradeoff between prediction and estimation error exhibited by (Dobriban et al., 2018) in the isotropic case can be explored with a general source $\Phi$.

## 5 Acknowledgments

D.R. is supported by the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). Part of this work has been carried out at the Machine Learning Genoa (MaLGa) center, Universita di Genova (IT). L.R. acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR projects FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS - DLV-777826. We would also like to thank the anonymous reviewers for their feedback and suggestions.

### References

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 242–252, 2019. URL `http://proceedings.mlr.press/v97/allen-zhu19a.html`.

Shun-ichi Amari, Jimmy Ba, Roger Grosse, Xuechen Li, Atsushi Nitanda, Taiji Suzuki, Denny Wu, and Ji Xu. When does preconditioning help or hurt generalization? *arXiv preprint arXiv:2006.10732*, 2020.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco.

On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in neural information processing systems*, pages 2300–2311, 2018a.

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

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

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1611–1619, 2019b.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, 2020.

Lin S Chen, Debashis Paul, Ross L Prentice, and Pei Wang. A regularized hotelling's t 2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association*, 106(496):1345–1360, 2011.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016.

Edgar Dobriban, Stefan Wager, et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1675–1685, 2019a. URL `http://proceedings.mlr.press/v97/du19c.html`.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *International Conference on Learning Representations (ICLR)*, 2019b.

Noureddine El Karoui. Random matrices and high-dimensional statistics: beyond covariance matrices. In *Proceedings of the International Congress of Mathematicians*, volume 4, pages 2875–2894, Rio de Janeiro, 2018.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Cédric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for convex penalized linear regression beyond gaussian matrices. *arXiv preprint arXiv:2002.04372*, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640. PMLR, 2020.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

Ian T Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303, 1982.

Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

Noureddine El Karoui and Holger Kösters. Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *arXiv preprint arXiv:1105.1404*, 2011.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020. URL `http://jmlr.org/papers/v21/19-844.html`.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011.

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

Zhenyu Liao and Romain Couillet. On inner-product kernels of high dimensional data. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 579–583. IEEE, 2019a.

Zhenyu Liao and Romain Couillet. A large dimensional analysis of least squares support vector machines. *IEEE Transactions on Signal Processing*, 67(4):1065–1074, 2019b.

Cosme Louart, Zhenyu Liao, Romain Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Yasaman Mahdaviyeh and Zacharie Naulet. Asymptotic risk of least squares minimum norm estimator under the spike covariance model. *arXiv preprint arXiv:1912.13421*, 2019.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4): 457, 1967.

Peter Mathé and Sergei V Pereverzev. Geometry of linear ill-posed problems in variable hilbert scales. *Inverse problems*, 19(3):789, 2003.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Partha P Mitra. Understanding overfitting peaks in generalization error: Analytical risk curves for $l\_2$ and $l\_1$ penalized interpolation. *arXiv preprint arXiv:1906.03667*, 2019.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

Jack W Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

Jack W Silverstein and Patrick L Combettes. Signal detection via spectral theory of large dimensional random matrices. *IEEE Transactions on Signal Processing*, 40(8):2100–2105, 1992.

S Spigler, M Geiger, S d'Ascoli, L Sagun, G Biroli, and M Wyart. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, 2019.

Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, pages 79–93, 2009.

Andrey N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4:1035–1038, 1963.

Antonia M Tulino, Sergio Verdú, et al. Random matrix theory and wireless communications. *Foundations and Trends® in Communications and Information Theory*, 1(1):1–182, 2004.

Denny Wu and Ji Xu. On the optimal weighted l_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*, 2020.

Ji Xu and Daniel J Hsu. On the number of variables to use in principal component regression. In *Advances in Neural Information Processing Systems*, pages 5094–5103, 2019.

Jianfeng Yao, Shurong Zheng, and ZD Bai. *Sample covariance matrices and high-dimensional data analysis*. Cambridge University Press Cambridge, 2015.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.