# Overparametrized Regression: Ridgeless Interpolation

Advanced Topics in Statistical Learning, Spring 2023

Ryan Tibshirani

## 1 Introduction

Current practice in machine learning suggests that it "works" to design neural networks that are massively overparametrized, and train them without explicit regularization until they interpolate the training data and thus have zero training error. Surprisingly, these models can still have good prediction error.

You might say: so what? We should regularize and the models will stop interpolating and they'll also perform better! However, in some overparametrized settings, it can actually be that tuning over the regularization strength (in an estimator like ridge regression) can suggest that a vanshing amount of regularization is optimal.

There has been some influential and thought-provoking experimental work in support of these phenomena. See Figure 1 for a few examples. How can we understand this through the lens of statistical theory? We can start by understanding what happens in linear models. Even though linear models are "as old as statistics itself", we may learn something new.

### 1.1 What's new here?

Statisticians have been interested in high-dimensional models for a long time. So what's new here? Don't the terms "high-dimensional" (in use for several decades) and "overparametrized" (popularized recently) refer the same thing?

In a sense, the answer is both "yes" and "no". While traditional high-dimensional analyses (like those we studied for the lasso) and newer overparametrized analyses are clearly very related, they are also different in some key ways. In the study of overparametrized models:

- we focus exclusively on out-of-sample prediction error—unlike many traditional high-dimensional regression analyses which focus on in-sample prediction error (and treat out-of-sample prediction error as somewhat of an afterthought);
- we care about how the risk landscape behaves as we vary the regularization strength, and particularly what happens for vanishing explicit regularization—unlike many traditional analyses which study regimes where optimal performance is given by strong explicit regularization.

There are arguably other differences, but those are two of the most salient ones for our discussion.

### 1.2 Ridgeless least squares

Given a response vector $Y \in \mathbb{R}^n$ and predictor matrix $X \in \mathbb{R}^{n \times d}$, as usual, we consider the *minimum $\ell_2$ norm least squares* (or simply min-norm least squares) estimator,

$$\hat{\beta} = (X^\mathsf{T} X)^+ X^\mathsf{T} Y, \tag{1}$$

where $(X^\mathsf{T} X)^+$ denotes the Moore-Penrose pseudoinverse of $X^\mathsf{T} X$. Equivalently, we can write this as

$$\hat{\beta} = \mathrm{argmin} \Big\{ \|\beta\|_2 : \beta \text{ minimizes } \|Y - X\beta\|_2^2 \Big\},$$

which explains its name.

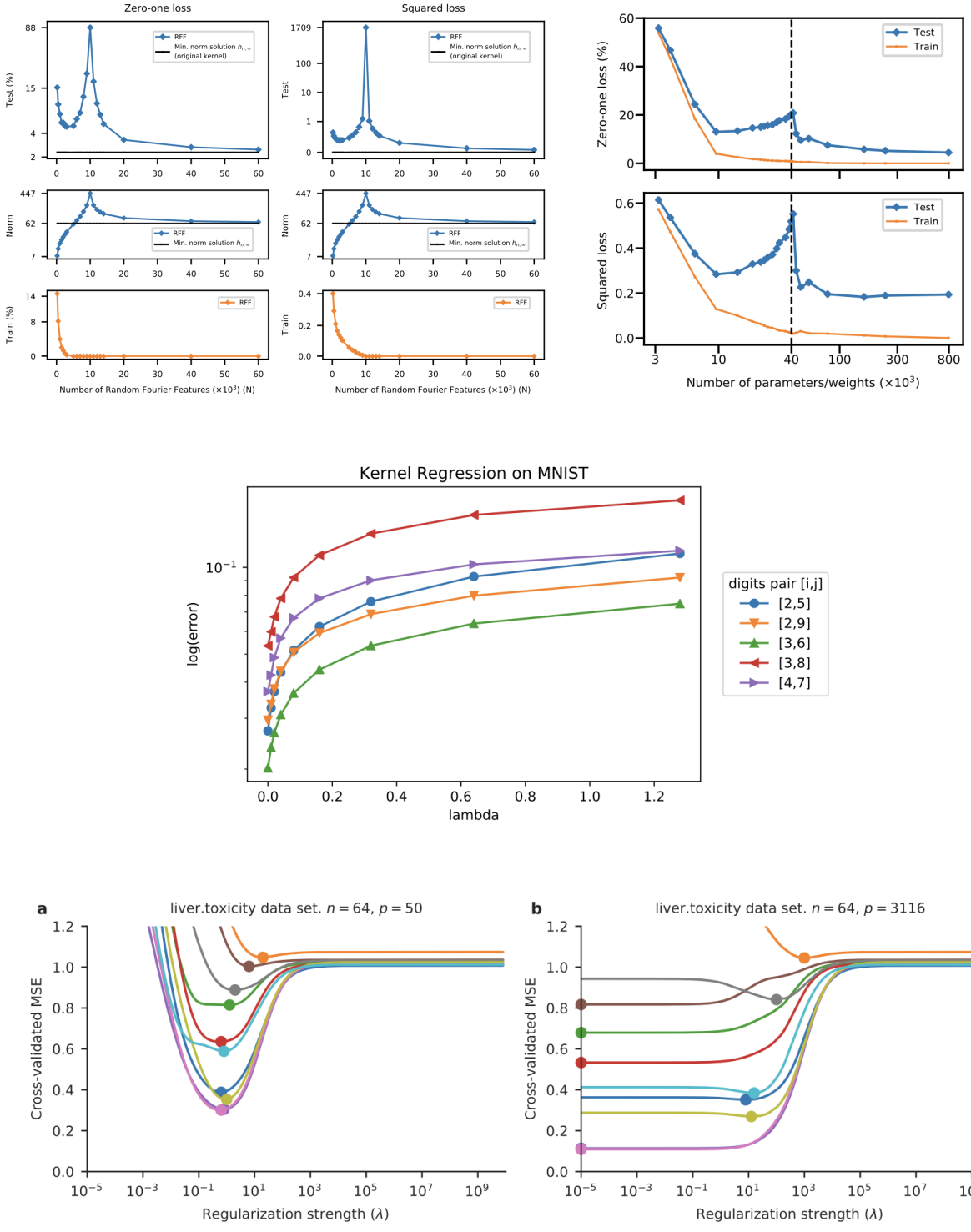


Figure 1: *Experiments demonstrating the motivating points raised in the introduction. Interpolating estimators (with zero training error) can still have good prediction error, as shown in the top row, credit: Belkin et al. (2019). Tuning over regularization levels can suggest that vanishing regularization is optimal even in high dimensions, as shown in the middle row, credit: Liang and Rakhlin (2020), and bottom row, credit: Kobak et al. (2020). The behavior of the prediction risk in the top row is dubbed "double descent".*

An alternative name for (1) is the "ridgeless" least squares estimator. This is because it can be written as the limit of the ridge estimator for vanishing regularization:

$$\hat{\beta} = \lim_{\lambda \to 0} (X^T X + \lambda I)^{-1} X^T Y = \lim_{\lambda \to 0} \operatorname*{argmin}_{\beta} \ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2.$$

When $X$ has rank $d$, the min-norm least squares estimator reduces to $\hat{\beta} = (X^T X)^{-1} X^T Y$, the usual least squares estimator. Importantly, when $X$ has rank $n$, this estimator interpolates the training data: $y_i = x_i^T \hat{\beta}$, for $i = 1, \ldots, n$.

### 1.3 Connection to gradient descent

There is an interesting connection between gradient descent and the min-norm least squares estimator. Initialize $\beta^{(0)} = 0$, and consider running gradient descent on the least squares loss, yielding iterates

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

where we take the step size $t > 0$ to be small enough. Then gradient descent converges to the min-norm least squares solution in (1):

$$\lim_{k \to \infty} \beta^{(k)} = \hat{\beta}.$$

Why? It's quite simple: the updates $\beta^{(k)}$, $k = 1, 2, 3, \ldots$ always lie in the row space of $X$; hence their limit (guaranteed to exist for small enough $t > 0$) must also lie in the row space of $X$; and the min-norm least squares solution is the unique least squares solution with this property.

The same result (and proof) carries over to stochastic gradient descent, and any other variants of gradient descent whose updates remain in the row space of $X$. Since these algorithms comprise the defacto standard for training neural networks, one can argue that that minimum $\ell_2$ norm solutions arise naturally as interesting objects of study based on practical conventions (and successes) in machine learning.

## 2 Problem setup

We will assume the usual linear model

$$Y = X\beta_0 + \epsilon, \tag{2}$$

where $\epsilon \in \mathbb{R}^n$ has i.i.d. entries with mean zero and variance $\sigma^2$, and $\epsilon \perp\!\!\!\perp X$. The conditions we will assume on the features $X \in \mathbb{R}^{n \times d}$ are as follows:

(A1) $X = Z\Sigma^{1/2}$, where the entries of $Z \in \mathbb{R}^{n \times d}$ are i.i.d. with zero mean, unit variance, and finite $8 + \eta$ moment, for some $\eta > 0$;

(A2) the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ has eigenvalues bounded away from 0 and $\infty$, and satisfies $F_\Sigma \xrightarrow{d} H$, as $n, d \to \infty$;

(A3) $d/n \to \gamma \in (1, \infty)$ as $n, d \to \infty$.

To emphasize, we are considering here $\gamma > 1$, called the *overparametrized* regime. As in the ridge lecture, we will be interested in the out-of-sample prediction risk, conditional on $X$,

$$\mathrm{Risk}_X(\hat{\beta}; \beta_0) = \mathbb{E}\big[(x_0^T \hat{\beta} - x_0^T \beta_0)^2 \,\big|\, X\big], \tag{3}$$

where $x_0 = \Sigma^{1/2} z_0$ is i.i.d. to the rows of $X$. We will study the risk of min-norm least squares in (1).

### 2.1 Summary

Here is a summary of what will happen in linear models, based roughly on the development of results in Hastie et al. (2022). At the outset, we should note that there are a lot of other interesting results from other recent papers that we will not be covering. This includes some work that goes beyond linear feature models and budges closer what actually happens in neural networks. For nice surveys on the recent explosion of work on overparametrization theory, see Bartlett et al. (2021); Belkin (2021); Dar et al. (2021).

0. In the underparametrized regime ($\gamma < 1$), the risk is purely variance (there is no bias), and does not depend on $\beta_0$ or $\Sigma$. Indeed, recall that we already showed in the last lecture that

$$\text{Risk}_X(\hat{\beta}; \beta_0) \overset{\text{as}}{\to} \sigma^2 \frac{\gamma}{1-\gamma}. \tag{4}$$

   This blows up as $\gamma \to 1$ from below.

1. In the overparametrized regime ($\gamma > 1$), the risk is composed of both bias and variance, and generally depends on $\Sigma$ or $\beta_0$. The asymptotic risk descends from its asymptote at $\gamma = 1$, but there is no longer a simple explicit formula that describes its behavior with $\gamma$ in full generality.

2. In the isotropic case $\Sigma = I$, we can derive a simple formula for the limiting risk for fixed $\beta_0$. This case already exhibits some interesting and informative properties. For example, in a misspecified model (where the mean $\mathbb{E}[y_i|x_i]$ is no longer linear in $x_i$ in (2)), the risk can attain its global minimum at $\gamma \in (1, \infty)$.

   However, this case fails to shed light on other important aspects that we seek to understand; in particular, optimal regularization strength in ridge regression in this isotropic case is always be positive.

3. In the case of general $\Sigma$, we can derive an explicit formula for the asymptotic variance. As in the ridge analysis, this is no longer closed-form, but still we can learn various things. For example, in specific covariance structures, we can study the variance term as a function of correlation strength.

   The behavior of the bias is much more complex. With a prior on $\beta_0$, we can get back an explicit asymptotic formula (as in the ridge lecture), but this again fails to shed light on important aspects we would like to study, as optimal regularization strength in ridge regression is again always positive.

4. Thus to expose when and how ridgeless regularization ($\lambda \to 0$) can be optimal, we must study the bias for general $\Sigma$, and fixed $\beta_0$. This is a lot more challenging, but it can be done. The optimality of ridgeless regularization can rigorously confirmed in a latent space model (which is a kind of misspecified model).

5. Finally, though still much more challenging, concrete progress can be made outside of the linear feature model. In certain nonlinear settings, some aspects of the lessons from linear feature models are preserved. You'll explore this (numerically) on the homework.

## 2.2 Plan-of-attack?

In general, there are two ways to proceed to derive the risk of the ridgeless regression estimator (1). The first is to take asymptotic results for ridge regression (like those we derived in the previous lecture) and then let $\lambda \to 0$, which requires being careful about why we can exchange limits (as $\lambda \to 0$ and $n, d \to \infty$).

The second is to start with the bias and variance expressions for min-norm least squares, and then carry out asymptotic calculations directly. Often, these calculations require us to "ridge-ify" some of the matrix functionals that involve pseudoinverses, and then let the auxiliary ridge parameter tend to zero. So they will typically bear strong similarites to the ridge calculations anyway. Curiously, the "ridge-ified" functionals in the second approach need not be *exactly* the same as those in the asymptotic calculations for ridge regression (though of course we must end up with the same answers with either approach).

We will follow the second approach to analyze the ridgeless estimator in the isotropic case, $\Sigma = I$, if only to give you more exposure to the elegant and powerful nature of random matrix theory. For general $\Sigma$ we do not give as many proof details, but will tend to follow the first approach, deriving ridgeless limits from ridge limits. In preparation for calculations to come, it is helpful to rewrite the ridgeless estimator (1) as $\hat{\beta} = (X^\mathsf{T} X/n)^+ X^\mathsf{T} Y/n$, and helpful to record the bias and variance components of its risk:

$$B_X(\hat{\beta}; \beta_0) = \beta_0^\mathsf{T} (I - \hat{\Sigma}^+ \hat{\Sigma}) \Sigma (I - \hat{\Sigma}^+ \hat{\Sigma}) \beta_0 \tag{5}$$

$$V_X(\hat{\beta}) = \frac{\sigma^2}{n} \text{tr}(\hat{\Sigma}^+ \Sigma), \tag{6}$$

where $\hat{\Sigma} = X^\mathsf{T} X/n$, which follows from calculations similar to those in the underparametrized case.

# 3 Isotropic $\Sigma$, fixed $\beta_0$

We consider the isotropic case, $\Sigma = I$. We will assume that $\|\beta_0\|_2 = r$ (which is a constant that does not vary with $n, d$), for the true signal vector in (2). Below we analyze the bias and variance separately.

**Bias analysis.** When $\Sigma = I$, the bias (5) becomes

$$\begin{aligned} B_X(\hat{\beta}; \beta_0) &= \beta_0^\mathsf{T}(I - \hat{\Sigma}^+\hat{\Sigma})\beta_0 \\ &= \lim_{\rho \to 0} \beta_0^\mathsf{T}\big[I - (\hat{\Sigma} + \rho I)^{-1}\hat{\Sigma}\big]\beta_0 \\ &= \lim_{\rho \to 0} \rho\beta_0^\mathsf{T}(\hat{\Sigma} + \rho I)^{-1}\beta_0, \end{aligned} \tag{7}$$

where in the second line we used the general fact $(A^\mathsf{T}A)^+A^\mathsf{T} = \lim_{\rho \to 0}(A^\mathsf{T}A + \rho I)^{-1}A^\mathsf{T}$, to $A = X/\sqrt{n}$, and in the third line we added and subtracted $\rho I$ to trailing $\hat{\Sigma}$ in the matrix product. (All limits $\rho \to 0$ here and henceforth are to be interpreted as from above.)

By the generalized MP theorem from Rubio and Mestre (2011) (transcribed in Theorem 2 in the ridge lecture notes using the language of deterministic equivalents), we know that

$$\beta_0^\mathsf{T}(\hat{\Sigma} + \rho I)^{-1}\beta_0 \quad \text{and} \quad \beta_0^\mathsf{T}(a_n I + \rho I)^{-1}\beta_0 = \frac{r^2}{a_n + \rho} \quad \text{have the same asymptotic limit,}$$

where $a_n$ satisfies a certain fixed-point equation. To compute the limit of $a_n$, we note that also

$$\frac{1}{d}\mathrm{tr}\big[(\hat{\Sigma} + \rho I)^{-1}\big] \quad \text{and} \quad \frac{1}{d}\mathrm{tr}\big[(a_n I + \rho I)^{-1}\big] = \frac{1}{a_n + \rho} \quad \text{have the same asymptotic limit,}$$

and by the standard MP asymptotics, the left-hand side converges almost surely to $m_F(-\rho)$, the Stieltjes transform of the MP law, evaluated at $-\rho$. Thus we have $1/(a_n + \rho) \to m_F(-\rho)$, and

$$\beta_0^\mathsf{T}(\hat{\Sigma} + \rho I)^{-1}\beta_0 \overset{\text{as}}{\to} r^2 m_F(-\rho).$$

Returning to (7), some calculations involving the Moore-Osgood theorem (whose details we omit) allow us to switch the order of the limits as $n, d \to \infty$ and $\rho \to 0$, yielding

$$\begin{aligned} B_X(\hat{\beta}; \beta_0) &\overset{\text{as}}{\to} r^2 \lim_{\rho \to 0} \rho m_F(-\rho) \\ &= r^2 \lim_{\rho \to 0} \frac{-(1 - \gamma + \rho) + \sqrt{(1 - \gamma + \rho)^2 + 4\gamma\rho}}{2\gamma} \\ &= r^2 \frac{-(1 - \gamma) + (\gamma - 1)}{2\gamma} \\ &= r^2\Big(1 - \frac{1}{\gamma}\Big), \end{aligned} \tag{8}$$

where to compute the limit we have used the explicit form of the Stieltjes transform of the MP law (as covered in the ridge lecture notes).

**Variance analysis.** When $\Sigma = I$, the variance analysis is actually quite easy (relatively speaking). From (6), we have

$$V_X(\hat{\beta}; \beta_0) = \sigma^2 \sum_{i=1}^n \frac{1}{s_i(X^\mathsf{T}X)},$$

where $s_i(X^\mathsf{T}X/n)$, $i = 1, \ldots, n$ denote the nonzero eigenvalues of $X^\mathsf{T}X/n$. Then because $X^\mathsf{T}X$ and $XX^\mathsf{T}$ always have the same nonzero eigenvalues, we may write this as

$$V_X(\hat{\beta}; \beta_0) = \sigma^2 \sum_{i=1}^n \frac{1}{s_i(X^\mathsf{T}X)} = \frac{\sigma^2 n}{d}\frac{1}{n}\mathrm{tr}\big[(XX^\mathsf{T}/d)^{-1}\big].$$

The rightmost term above is precisely the form of the variance in the underparametrized case, but with $X^\mathsf{T}$ in place of $X$. Thus, denoting $\tau = n/d < 1$, we conclude from our previous analysis that

$$V_X(\hat{\beta}; \beta) \xrightarrow{\text{as}} \sigma^2 \frac{\tau}{1-\tau} = \sigma^2 \frac{1}{\gamma - 1}, \tag{9}$$

**Putting it together.** Adding the bias (8) and variance results (9) together, we get

$$\text{Risk}_X(\hat{\beta}; \beta_0) \xrightarrow{\text{as}} r^2 \Big(1 - \frac{1}{\gamma}\Big) + \sigma^2 \frac{1}{\gamma - 1}. \tag{10}$$

This has quite a striking profile:

*The out-of-sample risk of least squares descends from its asymptote at the interpolation threshold $\gamma = 1$. The bias grows with $\gamma$, but the variance decreases with $\gamma$.*

The calculation above explains why the variance decreases with $\gamma$: we found that the variance of min-norm least squares on $n$ samples and $d$ features is precisely the variance of *ordinary* least squares on $d$ samples and $n$ features. See Figure 2 for visualization of the risk profile over $\gamma$.

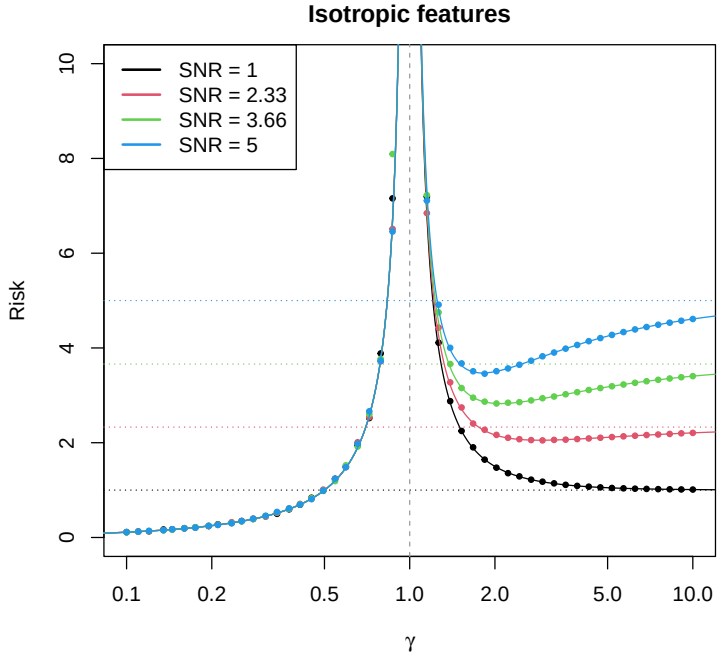


Figure 2: *The asymptotic risk of least squares* (4) *for* $\gamma < 1$ *and min-norm least squares* (10) *for* $\gamma > 1$*, where the latter assumes* $\Sigma = I$*, the isotropic case. Different colors represent different values of* $\text{SNR} = r^2/\sigma^2$*. The points denote finite-sample risks, computed from i.i.d. standard Gaussian features with* $n = 200$ *and* $d = [\gamma n]$ *for varying* $\gamma$*. Credit: Hastie et al. (2022).*

# 4 Misspecified setting

The latter analysis portrays an interesting feature of ridgeless regression in the overparametrized regime: its variance is well-controlled past the interpolation threshold $\gamma = 1$. However, it fails to exhibit any real benefit to doing ridgeless regression in the overparametrized regime, because the global minimum of the risk is always at $\gamma = 0$. We also do not see double descent in Figure 2, as the first descent over $\gamma \in (0, 1)$ never happens. This is due to the fact that we are assuming, for any $n, d$, a true linear model in $d$ features, which makes it hard to reason about what happens across different dimensions $d$.

We therefore pursue a small but important change to the model in (2): we instead assume that

$$Y = X\beta_0 + W\theta_0 + \epsilon, \tag{11}$$

Everything is the same as before, with respect to the distributions of $X$ and $\epsilon$. But now we have an additional part of the mean function that is driven by $W$, which we treat as a matrix of *unobserved* features (that we do not have in hand). Thus we still just carry out regression of $Y$ on $X$, as in (1). In this context, we call (11) a *misspecified* model, and analogous to (3), we are now interested in risk defined as:

$$\mathrm{Risk}_X(\hat{\beta}; \beta_0, \theta_0) = \mathbb{E}\big[\big(x_0^\mathsf{T}\hat{\beta} - \mathbb{E}[y_0|x_0, w_0]\big)^2 \,\big|\, X\big], \tag{12}$$

where $(x_0, w_0, y_0)$ is an i.i.d. draw from the same joint distribution as in (11).

## 4.1 Risk decomposition

We can always write

$$\mathrm{Risk}_X(\hat{\beta}; \beta_0, \theta_0) = \underbrace{\mathbb{E}\big[\big(x_0^\mathsf{T}\hat{\beta} - \mathbb{E}[y_0|x_0]\big)^2 \,|\, X\big]}_{L_X(\hat{\beta}; \beta_0, \theta_0)} + \underbrace{\mathbb{E}\big[\big(\mathbb{E}[y_0|x_0] - \mathbb{E}[y_0|x_0, w_0]\big)^2\big]}_{M(\beta_0, \theta_0)}, \tag{13}$$

which is verified by simply verified by first conditioning on $x_0$, then adding an subtracting $\mathbb{E}[y_0|x_0]$ inside the square in the definition of $R_X(\hat{\beta}; \beta_0, \theta_0)$ in (12), and expanding, and noting that the cross term is zero:

$$\mathbb{E}\big[\big(x_0^\mathsf{T}\hat{\beta} - \mathbb{E}[y_0|x_0]\big) \,|\, X, x_0\big]\mathbb{E}\big[\big(\mathbb{E}[y_0|x_0] - \mathbb{E}[y_0|x_0, w_0]\big) \,|\, x_0\big] = 0.$$

In other words, from (13), we learn that the risk in the misspecified setting decomposes into two terms: $L_X(\hat{\beta}; \beta_0, \theta_0)$, measuring how well we can predict the conditional mean of $\mathbb{E}[y_0|x_0]$, and $M_X(\hat{\beta}; \beta_0, \theta_0)$, measuring how far apart $\mathbb{E}[y_0|x_0]$ and $\mathbb{E}[y_0|x_0, w_0]$ are. We call the latter term the *misspecification bias* (note that it does not depend at all on $X$ or $\hat{\beta}$).

## 4.2 Simplest analysis

In the simplest case, we can take the distribution of the unobserved features $W$ to be independent of $X$ (and $\epsilon$), with the covariances of $X, W$ each being the identity—in keeping with the isotropic setting just studied. This allows to rewrite (12) as

$$Y = X\beta_0 + \delta,$$

where the entries of $\delta = W\theta_0 + \epsilon$ are still i.i.d. with mean zero, and their variance is $\|\theta_0\|_2^2 + \sigma^2$. Denote

$$r^2 = \|\beta_0\|_2^2 + \|\theta_0\|_2^2 \quad \text{and} \quad \kappa = \|\beta_0\|_2^2/r^2,$$

which represents the total signal energy and the the fraction of the signal energy captured by the observed features, respectively. Then $L_X(\hat{\beta}; \beta_0, \theta_0)$ behaves exactly as we computed previously, in (4) for $\gamma < 1$ and (10) for $\gamma > 1$, after we make the substitutions:

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

Furthermore, we can easily calculate the misspecification bias as:

$$M(\beta_0, \theta_0) = \mathbb{E}[(w_0^\mathsf{T}\theta_0)^2] = r^2(1-\kappa).$$

Putting these results together leads to the following conclusion:

$$\mathrm{Risk}_X(\hat{\beta}; \beta_0; \theta_0) \overset{\text{as}}{\to} \begin{cases} r^2(1-\kappa) + \big(r^2(1-\kappa) + \sigma^2\big)\frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2(1-\kappa) + r^2\kappa\big(1 - \frac{1}{\gamma}\big) + \big(r^2(1-\kappa) + \sigma^2\big)\frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases} \tag{14}$$

### 4.3 Interpretation

To interpret the risk profiles (14) in the misspecified setting, we will need to specify a relationship between $\kappa$ and $\gamma$. Since adding features should generally improve our approximation capacity, it is reasonable to model $\kappa = \kappa(\gamma)$ as an increasing function of $\gamma$. Figure 3 gives an example with a polynomial decay, $1 - \kappa(\gamma) = (1+\gamma)^{-a}$. Notably, we can see a clear double descent in the risk curve, and for certain values of $a$ (such as that plotted in green), we find that the global min of the risk occurs at $\gamma > 1$.

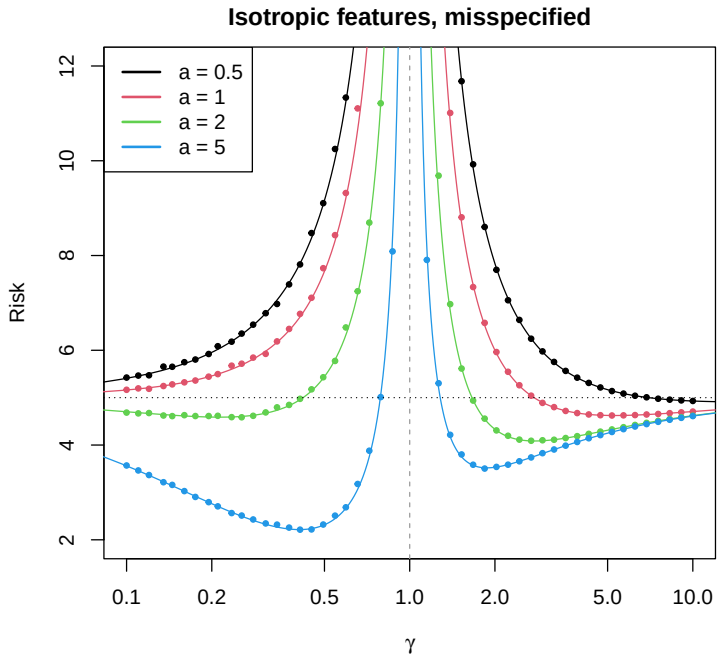


Figure 3: *The asymptotic risk of min-norm least squares in the misspecified setting* (14)*, with isotropic observed and unobserved feature covariances, and* $1 - \kappa(\gamma) = (1+\gamma)^{-a}$*. We set* $\mathrm{SNR} = r^2/\sigma^2 = 5$*. Different colors represent different values of* $a$*. As in Figure 2, the points denote finite-sample risks, computed from* $n = 200$ *and* $d = [\gamma n]$ *for varying* $\gamma$*. Credit: Hastie et al. (2022).*

As a concluding remark in this misspecified setting, we note that the dimension of the unobserved features (the number of columns of $W$) enters nowhere in these calculations, so we can effectively think of it as infinite. This provides us with a nice interpretation: we have an infinitely wide matrix $[X; W]$ that governs the behavior of $Y$ in the model (2). As $d$ grows, we observe more and more columns of this matrix, which improves our approximation capacity. This is an analogy to what a feature generator can do for us.

## 5 General $\Sigma$, random $\beta_0$

In the general $\Sigma$ case, just as in the ridge analysis, the bias in (5) is especially difficult to calculate. We can make progress by placing a spherical prior on $\beta_0$, such that

$$\mathbb{E}[\beta_0 \beta_0^\mathsf{T}] = \frac{r^2}{d} I. \tag{15}$$

**Bias analysis.** By taking a limit as $\lambda \to 0$ in the Bayes bias result from the ridge lecture (or by starting from (5), integrating over $\beta_0$, and following similar arguments to the isotropic case above), one can show

$$B_X(\hat{\beta}) \overset{\text{as}}{\to} \frac{r^2}{\gamma} \frac{1}{v_F(0)}, \tag{16}$$

where $v_F$ is the companion Stieltjes transform of the limiting spectral distribution $F = F(H, \gamma)$ from the Marchenko-Pastur theorem (and $v_F(0) = \lim_{\rho \to 0} v_F(-\rho)$ exists under our assumptions).

**Variance analysis.** Again by taking a limit as $\lambda \to 0$ in the variance calculation from the ridge lecture (or by starting from (6), and following similar arguments to the isotropic case above), one can show

$$V_X(\hat{\beta}) \overset{\text{as}}{\to} \sigma^2 \bigg( \frac{v_F'(0)}{v_F(0)^2} - 1 \bigg), \tag{17}$$

where again $v_F$ is the companion Stieltjes transform of the limiting spectral distribution $F = F(H, \gamma)$ from the MP theorem (and $v_F'(0)/v_F(0)^2 = \lim_{\rho \to 0} v_F'(-\rho)/v_F(-\rho)^2$ exists under our assumptions). We emphasize that the variance calculation in (17) does not depend on the prior (15) and is fully general.

**Inspecting the asymptotics.** We note that the some of the key dependence of (16), (17) on $\gamma$ is hidden in the Stieltjes transform terms $v_F(0)$, $v_F'(0)$, which themselves depends on $\gamma$ (since $F$ does). While these asymptotic limit cannot be computed in closed-form for general covariance models (general $H$), they can be computed numerically by solving the Silverstein equation. This is done in Hastie et al. (2022), in order to probe the asymptotic limits, and better understand how they behave. This reveals some interesting phenomena, such as the fact that stronger correlations can increase the variance, but decrease the Bayes bias. See Figure 4.

**Where this fall short.** The Bayes model studied here falls short in one key way. It does not explain why ridgeless regression is statistically interesting above and beyond ridge regression. In the Bayes model studied here, recall (from the last lecture), we know that the asymptotically optimal ridge tuning parameter is $\lambda^* = \sigma^2\gamma/r^2$, regardless of the feature covariance model. This is always positive, and depends only on the SNR. Thus in order to really understand the phenomena in Figure 1, i.e., to understand when and how the ridge risk landscape can be minimized at $\lambda = 0$, we must tackle the beast: calculate the ridge bias for general $\Sigma$ and fixed $\beta_0$.

# 6 General $\Sigma$, fixed $\beta_0$

For general $\Sigma$, and a fixed $\beta_0$, the ridge bias where (recall) $\hat{\beta}_\lambda = (X^\mathsf{T}X + n\lambda I)^{-1}X^\mathsf{T}Y$ for $\lambda > 0$, and

$$B_X(\hat{\beta}_\lambda; \beta_0) = \lambda^2 \beta_0^\mathsf{T} (\hat{\Sigma} + \lambda I)^{-1} \Sigma (\hat{\Sigma} + \lambda I)^{-1} \beta_0, \tag{18}$$

and especially the ridgeless bias (5), are formidable calculations.

The variance calculations in the ridge or ridgeless risk expansions that we computed in the previous general $\Sigma$ analyses (in the ridge lecture, and in the last section) were done under the guise of a prior assumption on $\beta_0$, but this assumption did not actually matter for the variance terms, so they were already carried out in full generality anyway.

For the ridge bias, the main challenge is that the functional $(\hat{\Sigma} + \lambda I)^{-1}\Sigma(\hat{\Sigma} + \lambda I)^{-1}$ does not has a transparent deterministic equivalent; and similarly for the ridgeless bias. Only recently have results started to appear on the asymptotics of the ridge and ridgeless bias terms for general $\Sigma$ and $\beta_0$, see, e.g., Wu and Xu (2020); Richards et al. (2021); Hastie et al. (2022). We mainly follow the calculations in the latter paper (as they are, in a sense, the most general anyway), but in keeping with our style in these lectures thus far, we translate them to the language of deterministic equivalents. We only give some part of the details; for the remainder, a supplementary note can be found on the course website.

**Ridge bias analysis.** We start with the ridge bias calculation. We seek to use the generalized Marchenko-Pastur theorem from Rubio and Mestre (2011) (Theorem 2 in the ridge lecture notes), but it is not as of yet clearly applicable. There is one key trick: we rewrite (18) as

$$\lambda^2 (\hat{\Sigma} + \lambda I)^{-1} \Sigma (\hat{\Sigma} + \lambda I)^{-1} = -\frac{d}{d\rho} \Big\{ \lambda (\hat{\Sigma} + \lambda I + \rho\lambda\Sigma)^{-1} \Big\} \Big|_{\rho = 0}. \tag{19}$$

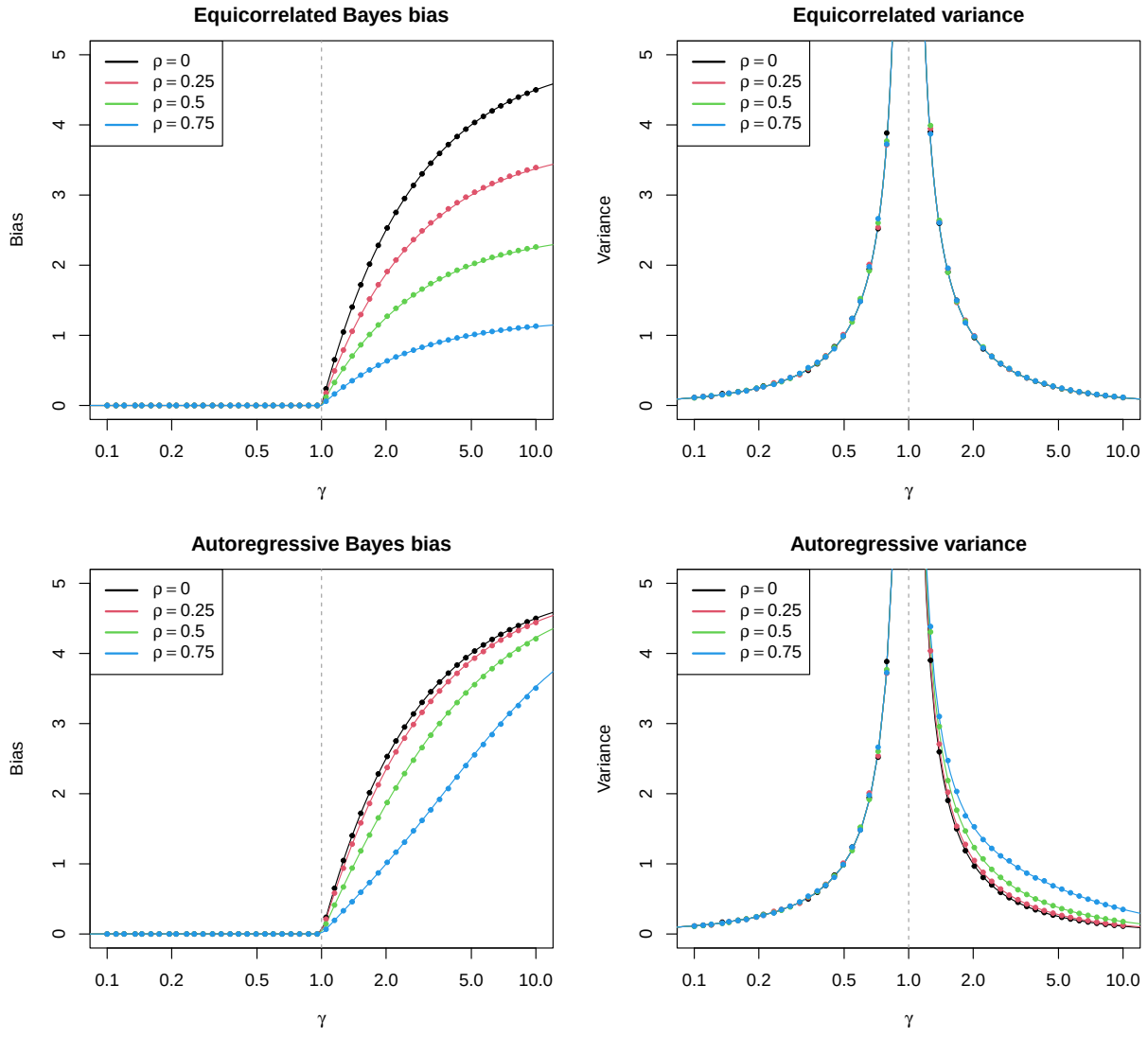


Figure 4: *Asymptotic Bayes bias and variance of min-norm least squares for different feature covariance models. The top row shows an equicorrelated feature model, where* $\Sigma_{ij} = 1$ *if* $i = j$ *and* $\rho$ *otherwise. In this model the asymptotics can be done in closed-form: we find the bias decreases with* $\rho$*, whereas the variance does not actually depend on* $\rho$*. The bottom row shows an autocorrelated feature model, where* $\Sigma_{ij} = \rho^{|i-j|}$*. Here the asymptotics cannot be done in closed-form but they can be efficiently computed numerically: we find that the bias decreases with* $\rho$*, and the variance increases with* $\rho$*. Credit: Hastie et al. (2022).*

Note that this is similar in spirit to the way we have "ridge-ified" various functionals above that involve pseudoinverses, but it is technically different: the auxiliary parameter $\rho$ now multiplies $\lambda\Sigma$ (instead of simply multiplying the identity matrix, as above). Checking (19) can be done using the following general fact from matrix calculus:

$$\frac{dA^{-1}}{d\rho} = -A^{-1}\frac{dA}{d\rho}A^{-1}, \tag{20}$$

for an invertible matrix $A$ (where $dA/d\rho$ is to be understood elementwise). We can thus seek to compute a deterministic equivalent for the functional inside the derivative on the right-hand side in (19), then differentiate with respect to $\rho$ and setting $\rho = 0$.

Towards this end, we rewrite it once more

$$\begin{aligned}\lambda(\hat{\Sigma} + \lambda I + \rho\lambda\Sigma)^{-1} &= \lambda\bigg(\Sigma^{1/2}\frac{Z^T Z}{n}\Sigma^{1/2} + \lambda(I+\rho\Sigma)\bigg)^{-1} \\ &= (I+\rho\Sigma)^{-1/2}\lambda(\hat{\Sigma}_\rho + \lambda I)^{-1}(I+\rho\Sigma)^{-1/2},\end{aligned} \tag{21}$$

where we define

$$\hat{\Sigma}_\rho = \Sigma_\rho^{1/2}\frac{Z^T Z}{n}\Sigma_\rho^{1/2} \quad \text{and} \quad \Sigma_\rho = (I+\rho\Sigma)^{-1/2}\Sigma(I+\rho\Sigma)^{-1/2}.$$

Now the middle term in (21) has a deterministic equivalent by the generalized MP theorem,

$$\lambda(\hat{\Sigma}_\rho + \lambda I)^{-1} \asymp (c_n\Sigma_\rho + I)^{-1}, \tag{22}$$

where $c_n$ solves a particular fixed-point equation. Some calculations (detailed in the supplementary notes) show that, after differentiating (22) with respect to $\rho$ and taking $\rho = 0$, we get the following conclusion: assuming $\|\beta_0\|_2$ remains bounded, the ridge bias in (18) satisfies

$$\big|B_X(\hat{\beta}_\lambda;\beta_0) - (1+c_n)\beta_0^T\Sigma(b_n\Sigma+I)^{-2}\beta_0\big| \overset{\text{as}}{\to} 0, \tag{23}$$

where $b_n, c_n$ solve (recalling $\gamma_n = d/n$):

$$\frac{1}{b_n} = \lambda + \frac{\gamma_n}{d}\mathrm{tr}\big[\Sigma(b_n\Sigma+I)^{-1}\big] \tag{24}$$

$$c_n = \frac{\gamma_n \mathrm{tr}[\Sigma^2(b_n\Sigma+I)^{-2}]/d}{b_n^{-2} - \gamma_n\mathrm{tr}[\Sigma^2(b_n\Sigma+I)^{-2}]/d}. \tag{25}$$

There are different parametrizations available for these fixed-point equations (more later) but we choose to use the one above because it allows to send $\lambda \to 0$ in a graceful way, which we do next.

**Ridgeless bias analysis.** To derive ridgeless asymptotics, we let $\lambda \to 0$ in (23), (24), (25). The supplement gives details on why this is valid. This yields for the ridgeless estimator $\hat{\beta} = (X^TX)^+X^TY$,

$$\big|B_X(\hat{\beta};\beta_0) - (1+\tilde{c}_n)\beta_0^T\Sigma(\tilde{b}_n\Sigma+I)^{-2}\beta_0\big| \overset{\text{as}}{\to} 0, \tag{26}$$

where $\tilde{b}_n, \tilde{c}_n$ solve

$$\frac{1}{\tilde{b}_n} = \frac{\gamma_n}{d}\mathrm{tr}\big[\Sigma(\tilde{b}_n\Sigma+I)^{-1}\big] \tag{27}$$

$$\tilde{c}_n = \frac{\gamma_n \mathrm{tr}[\Sigma^2(\tilde{b}_n\Sigma+I)^{-2}]/d}{\tilde{b}_n^{-2} - \gamma_n\mathrm{tr}[\Sigma^2(\tilde{b}_n\Sigma+I)^{-2}]/d}. \tag{28}$$

**"Semi-asymptotic" formulations.** The fixed-point equations above can understood from an asymptotic point of view, as follows. Recall the Silverstein equation, which the uniquely defines the companion Stieltjes transform $v_F$ of the limiting spectral measure $F$ in the Marchenko-Pastur theorem,

$$\frac{1}{v_F(-\lambda)} = \lambda + \gamma\int\frac{s}{sv_F(-\lambda)+1}\,dH(s). \tag{29}$$

Writing $b$ for the limit of $b_n$, we can see that as $n, d \to \infty$, the the fixed-point equation (24) converges to the Silverstein equation, with the relationship $b = v_F(-\lambda)$. (Note that we used a similar argument in ridge calculations in the last lecture, where we encountered a reparametrization of the fixed-point equation (24) with $a_n = \lambda b_n$.) In other words, to be clear, we have learned that $b_n \to v_F(-\lambda)$.

What of the limit of $c_n$? First write

$$1 + c_n = \frac{b_n^{-2}}{b_n^{-2} - \gamma_n \mathrm{tr}[\Sigma^2(b_n\Sigma + I)^{-2}]/d}. \tag{30}$$

Now go back and differentiate (29) with respect to $\lambda$; this gives, using the matrix calculus fact (20) once again for the right-hand side,

$$\frac{v_F'(-\lambda)}{v_F(-\lambda)^2} = 1 + \gamma v_F'(-\lambda) \int \frac{s^2}{(s v_F(-\lambda) + 1)^2} \, dH(s).$$

Simply solving for $v_F'(-\lambda)$, we get

$$v_F'(-\lambda) = \frac{1}{\frac{1}{v_F(-\lambda)^2} - \gamma \int \frac{s^2}{(s v_F(-\lambda)+1)^2} \, dH(s)}.$$

Therefore we can see from (30) (and the fact that $b_n \to v_F(-\lambda)$) that $1 + c_n \to v_F'(-\lambda)/v_F(-\lambda)^2$.

Putting this together, instead of (23), we can write the ridge bias in "semi-asymptotic" form (where we reduce $b_n, c_n$ to their asymptotic limits) as:

$$\left| B_X(\hat{\beta}_\lambda; \beta_0) - \frac{v_F'(-\lambda)}{v_F(-\lambda)^2} \beta_0^\mathsf{T} \Sigma \big(v_F(-\lambda)\Sigma + I\big)^{-2} \beta_0 \right| \overset{\text{as}}{\to} 0. \tag{31}$$

By analogous reasoning, instead of (26), we can write the ridgeless bias in "semi-asymptotic" form as:

$$\left| B_X(\hat{\beta}_\lambda; \beta_0) - \frac{v_F'(0)}{v_F(0)^2} \beta_0^\mathsf{T} \Sigma \big(v_F(0)\Sigma + I\big)^{-2} \beta_0 \right| \overset{\text{as}}{\to} 0. \tag{32}$$

**Inspecting the asymptotics.** The formulae in (31) (or (23), (24), (25) for the finite-sample versions) and (32) (or (26), (27), (28)) are the most general ones that we have seen thus far for the ridge and ridgeless bias terms, respectively. Combined with our previous general variance calculations, this completes the picture for the asymptotic risk of ridge and ridgeless regression. For concreteness, the results (writing the bias terms in "semi-asymptotic" form) are as follows:

$$\left| \mathrm{Risk}_X(\hat{\beta}_\lambda; \beta_0) - \left[ \frac{v_F'(-\lambda)}{v_F(-\lambda)^2} \beta_0^\mathsf{T} \Sigma \big(v_F(-\lambda)\Sigma + I\big)^{-2} \beta_0 + \sigma^2 \left( \frac{v_F'(-\lambda)}{v_F(-\lambda)^2} - 1 \right) \right] \right| \overset{\text{as}}{\to} 0, \tag{33}$$

$$\left| \mathrm{Risk}_X(\hat{\beta}; \beta_0) - \left[ \frac{v_F'(0)}{v_F(0)^2} \beta_0^\mathsf{T} \Sigma \big(v_F(0)\Sigma + I\big)^{-2} \beta_0 + \sigma^2 \left( \frac{v_F'(0)}{v_F(0)^2} - 1 \right) \right] \right| \overset{\text{as}}{\to} 0. \tag{34}$$

To be clear, the result in (33) holds for any $\gamma$ and that in (34) holds for $\gamma > 1$.

Unfortunately, these are not really closed-form, since they rely not only on the companion Stieltjes transform $v_F$ of the limiting spectral distribution $F$ from the MP theorem (which is not closed-form for general $H$), but moreover, the bias terms depend on the alignment of true signal vector $\beta_0$ with the population covariance matrix $\Sigma$, in a complex way.

Nonetheless, we can inspect what happens for particular feature models. For example, an interesting finding, as studied empirically in Kobak et al. (2020), occurs in an equicorrelated feature model, where $\Sigma_{ij} = 1$ if $i = j$ and $\rho$ otherwise. When $\beta_0$ is aligned with the top eigenvector of $\Sigma$, then one will find that for a large enough aspect ratio $\gamma_n = d/n$ (and large enough SNR), the optimal ridge tuning parameter will be zero. This phenomenon should be mathematically verifiable from (33), (34), and I will add the details to these notes at a later point. Todo.

This relates to, but is simpler than, an asymptotic calculation from Hastie et al. (2022) for a latent feature model. Todo: details here or in its own "paragraph"?

**Degrees of freedom perspective.** Very recently, Bach (2023) provided a nice re-interpretation of (31), (32) and the corresponding variance results from the perspective of (effective) degrees of freedom. I will also add the details to these notes at a later point. Todo.

# References


Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. arXiv: 2303.01372, 2023.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: A statistical viewpoint. *Acta Numerica*, pages 87–201, 2021.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *Proceeding of the National Academy of Sciences*, 116(32):15849–15954, 2019.

Misha Belkin. Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, pages 203–248, 2021.

Yehuda Dar, Vidya Muthukumar, and Richard Baraniuk. A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. arXiv: 2109.02355, 2021.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional "ridgeless" least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

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*, 21:1–16, 2020.

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

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Probability Letters*, 81(5):592–602, 2011.

Denny Wu and Ji Xu. On the optimal weighted $\ell_2$ regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, 2020.