High-Dimensional Regression: Ridge

Advanced Topics in Statistical Learning, Spring 2024

Ryan Tibshirani

Note: we're following the context, problem setup, notation, etc. from the last lecture on high-dimensional regression.

1 Ridge basics

We'll jump right into some basic properties of *ridge regression*, which recall, for a predictor matrix $X \in \mathbb{R}^{n \times d}$ and response vector $Y \in \mathbb{R}^n$, is defined by

$$\underset{\beta}{\text{minimize}} \;\; \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2, \tag{1}$$

for a tuning parameter $\lambda \geq 0$. Unlike the lasso and best subset selection (which replace the $\|\beta\|_2^2$ penalty above with $\|\beta\|_1$ and $\|\beta\|_0$, respectively), recall, the solution in the ridge problem (1) is generically dense: it has all nonzero components, for any $\lambda > 0$.

Yes, above we said *the* solution in (1), and that was intentional. For any $\lambda > 0$, the criterion in the problem (1) is strictly convex, due to the $\lambda \|\beta\|_2^2$ term. It therefore always has a unique solution (regardless of X). There is not much else to say, except to emphasize the contrast here to the lasso problem, which, recall, does not always admit a lasso solution, though we were able to do some basic analysis to show that it has a unique solution whenever X has columns in general position.

More than just being unique, the solution in (1) has an explicit form,

$$\hat{\beta} = (X^\mathsf{T} X + \lambda I)^{-1} X^\mathsf{T} Y, \tag{2}$$

for $\lambda > 0$. As another way to see the uniqueness claim, note that $X^\mathsf{T} X + \lambda I$ is always invertible whenever $\lambda > 0$, since its smallest eigenvalue is at least λ.

Now, there are many "facets" of the ridge regression estimator, by which we mean, many perspectives from which to view it. We'll walk through a number of these now (there are many others we don't cover!), before moving to the tools needed to analyze the risk of ridge regression.

1.1 Principal components view

A standard way to view the ridge regression estimator is through the lens of how it acts in principal components space. Let $X = U \Sigma V^\mathsf{T}$ be a singular value decomposition of X. Then we can write the vector of ridge fitted values as:

$$\begin{aligned} X\hat{\beta} &= X(X^\mathsf{T} X + \lambda I)^{-1} X^\mathsf{T} Y \\ &= U \Sigma^2 (\Sigma^2 + \lambda I)^{-1} U^\mathsf{T} Y \\ &= \sum_{j=1}^n \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j u_j^\mathsf{T} Y, \end{aligned}$$

where u_j denotes the j^{th} column of U, and σ_j the j^{th} diagonal element of Σ. Note that the least squares projection of Y on X (always well-defined, even if $d > n$) is

$$UU^\mathsf{T} Y = \sum_{j=1}^n u_j u_j^\mathsf{T} Y.$$

Comparing the second-to-last display to the last one, we can see that ridge (compared to least squares) performs shrinkage, but not uniformly so in all directions—it shrinks more along the principal component directions u_j that correspond to low variance (small σ_j), and less along the directions u_j that correspond to high variance (large σ_j).

1.2 Bayesian view

Another standard way to view ridge regression is as a Bayes estimator in a normal-normal model for the likelihood and prior. That is, for fixed X, consider the model

$$\begin{aligned} \beta &\sim N(0, \frac{1}{\lambda} I), \\ Y|\beta &\sim N(X\beta, I). \end{aligned} \tag{3}$$

Then, following standard calculations in Bayesian inference, where we often use $p(\cdot)$ to denote the density of its argument, it holds that

$$p(\beta|Y) \propto p(Y|\beta)\, p(\beta),$$

where $\propto$ means "proportional to" and ignores terms not depending on β. Again by standard calculations, the right-hand side above can be identified as proportional to a Gaussian density. Therefore (because it is Gaussian), its mean is equal to its mode, and the Bayes estimator is the maximum a posteriori (MAP) estimator:

$$\begin{aligned} \mathbb{E}[\beta|Y] &= \operatorname*{argmax}_\beta \; p(Y|\beta)\, p(\beta) \\ &= \operatorname*{argmin}_\beta \; -\log p(Y|\beta) - \log p(\beta) && (4) \\ &= \operatorname*{argmin}_\beta \; \frac{1}{2}\|Y - X\beta\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2, && (5) \end{aligned}$$

that is, the Bayes estimator in this model is simply the ridge estimator.

We note that the lasso can also be written as a MAP estimator in a Bayesian model, where the prior in (3) is now Laplace rather than Gaussian. However, the MAP estimator is *not* the Bayes estimator (posterior expectation) in this particular model: the posterior is no longer Gaussian and no longer has the property of symmetry around its mode that would imply its mode and its mean are the same.

1.3 Kernel view

Another interesting view stems from what is called the push-through matrix identity:

$$(aI + UV)^{-1} U = U(aI + VU)^{-1}. \tag{6}$$

for a, U, V such that the products are well-defined and the inverses exist. We can obtain this from

$$U(aI + VU) = (aI + UV)U,$$

followed by multiplication by $(aI + UV)^{-1}$ on the left and the right. Applying the identity (6) to (2) with $a = \lambda$, $U = X^\mathsf{T}$, and $V = X$, we obtain an alternative form for the ridge solution:

$$\hat{\beta} = X^\mathsf{T}(XX^\mathsf{T} + \lambda I)^{-1} Y. \tag{7}$$

This is often referred to as the kernel form of the ridge estimator. From (7), we can see that the ridge fit can be expressed as

$$X\hat{\beta} = XX^\mathsf{T}(XX^\mathsf{T} + \lambda I)^{-1} Y.$$

What does this remind you of? This is precisely $K(K + \lambda I)^{-1} Y$ where $K = XX^\mathsf{T}$, which, recall, is the fit from RKHS regression with a linear kernel $k(x, z) = x^\mathsf{T} z$. Therefore we can think of RKHS regression as generalizing ridge regression, by replacing the standard linear inner product with a general kernel. (Indeed, RKHS regression is often called kernel ridge regression.)

1.4 Noise features view

The last view we cover is likely the least well-known, but it is still very interesting. This is due to Kobak et al. (2020). Suppose that we append to the columns of our feature matrix $X \in \mathbb{R}^{n \times d}$ a large number of "noise" features $Z \in \mathbb{R}^{n \times D}$ where D is very large and each Z_{ij} is stochastic with mean zero and variance τ^2, independent of Y. Denote such an augmented feature matrix by

$$\tilde{X} = [X; Z] \in \mathbb{R}^{n \times (d+D)},$$

and consider performing least squares regression—with no explicit ridge regularization—of Y on $\tilde{X}$. As we are considering the large D limit, we will inevitably have $d + D > n$ at some point, so we'll need to amend the least squares estimator because it is nonunique and has the pathologies we already discussed. To do so, we'll take the *minimum ℓ_2 norm least squares solution*,

$$\tilde{\beta} = (\tilde{X}^\mathsf{T} \tilde{X})^+ \tilde{X}^\mathsf{T} Y. \tag{8}$$

We'll also simply call this the min-norm solution (especially when it is unambiguous from the context that the norm we're talking about is ℓ_2). We'll spend a lot more time discussing min-norm least squares when we talk about overparametrization theory, in the next lecture.

Now we rewrite (8) in kernel form. By even simpler arguments than those in the previous subsection—just using the fact tha the generalized inverse and transpose operations commute in general, $(A^+)^\mathsf{T} = (A^\mathsf{T})^+$—we can rewrite the min-norm solution (assuming that $d + D > n$ and $\tilde{X}$ has full row rank) as

$$\begin{aligned}
\tilde{\beta} &= \tilde{X}^\mathsf{T} (\tilde{X} \tilde{X}^\mathsf{T})^{-1} Y \\
&= [X; Z]^\mathsf{T} (X X^\mathsf{T} + Z Z^\mathsf{T})^{-1} Y \\
&= [X; \tau \tilde{Z}]^\mathsf{T} (X X^\mathsf{T} + \tau^2 \tilde{Z} \tilde{Z}^\mathsf{T})^{-1} Y,
\end{aligned}$$

where in the last line we have rescaled the noise features so that each $\tilde{Z}_{ij} = \tau^{-1} Z_{ij}$ has zero mean and unit variance. If we take $\tau = \sqrt{\lambda / D}$, then by the law of large numbers,

$$\frac{\lambda}{D} \tilde{Z} \tilde{Z}^\mathsf{T} \overset{\text{as}}{\to} \lambda I, \quad \text{as } D \to \infty.$$

Therefore if we let $\tilde{\beta}_{[d]}$ denote the first d components of the min-norm least squares solution in (8) (which has total dimension $d + D$), then

$$\tilde{\beta}_{[d]} \overset{\text{as}}{\to} X^\mathsf{T} (X X^\mathsf{T} + \lambda I)^{-1} Y, \quad \text{as } D \to \infty,$$

the limit here being the ridge solution from regressing Y on X with tuning parameter λ. Similarly, for any fixed $\tilde{x} = (x, z) \in \mathbb{R}^{d+D}$,

$$\tilde{x}^\mathsf{T} \tilde{\beta} \overset{\text{as}}{\to} x^\mathsf{T} X^\mathsf{T} (X X^\mathsf{T} + \lambda I)^{-1} Y, \quad \text{as } D \to \infty,$$

the limit here being the ridge prediction from regressing Y on X with tuning parameter λ. In other words, min-norm least squares on an augmented design where we augment the given features with noise features—whose variance vanishes appropriately as the number of them grows—reproduces ridge regression!

2 Random matrix theory

In this section, we will introduce some basic results in random matrix theory (RMT), which will serve as the backbone that we will use to analyze the risk of ridge regression in the rest of the lecture. As an editorial remark, you'll often hear "random matrix theory" to describe two generally related but different flavors of results: asymptotic and non-asymptotic. Below we will be talking about asymptotic RMT. And as per our usual comment, there is a lot to learn about random matrix theory and we're only really covering the tip of the iceberg; to learn more, well beyond what we cover here, see, e.g., Tulino and Verdu (2004); Bai and Silverstein (2010), which are definitive references for the type of asymptotic results that we study. It is also worth noting that RMT is a very active field of research, and new and important—arguably, even foundational—results seem to be still emerging right now.

2.1 Proportional asymptotics

We'll be working in what is known as a *proportional asymptotics* model, where the dimension d and number of samples n diverge proportionally. That is,

$$\frac{d}{n} \to \gamma \in (0, \infty), \quad \text{as } n, d \to \infty.$$

The quantity γ is often called the aspect ratio.

We assume that the rows of our predictor matrix $X \in \mathbb{R}^{n \times d}$, denoted $x_i \in \mathbb{R}^d$, $i = 1, \dots, n$, are i.i.d., and importantly, each one is a "rotation" of a vector of i.i.d. random variables, so that we can write:

$$x_i = \Sigma^{1/2} z_i, \quad \text{where } z_i \in \mathbb{R}^d \text{ has i.i.d. entries with zero mean, unit variance.} \tag{9}$$

We'll allow the feature covariance matrix Σ to be more or less arbitrary (subject to some limiting conditions, described shortly), and the distribution of each entry of z_{ij} to be more or less arbitrary (subject to some moment conditions, described shortly). Therefore, the condition (9) looks like a very weak assumption on the features.

While this is certainly true in some sense, it is also worth noting that (9) is a not completely innocuous assumption. In particular, left multiplication by the square root matrix $\Sigma^{1/2}$ performs a kind of averaging operation. Consequently, the entries x_{ij} can either have long-tailed distributions (for Σ close to the identity, and z_{ij} having heavy tails), or have complex dependence structures (for Σ far from the identity), but not both, since then the averaging will mitigate any long tail of the distribution of z_{ij}.

Note that we can express (9) succinctly as $X = Z\Sigma^{1/2}$, where $Z \in \mathbb{R}^{n \times d}$ has i.i.d. entries z_{ij}.

Why exact asymptotics? As a foreshadowing of what is to come, we will derive *exact* asymptotic expressions for the risk of ridge regression, under a proportional asymptotics model. This stands in contrast to the risk theory we developed for the lasso in last lecture, as well as theory we developed in nonparametric regression in previous lectures. All of our theory here was expressed in terms of risk bounds, which we inspected primarily for their dependence on n, d. In the proportional asymptotics regime, this will simply not do. This is because, under proportional asymptotics, *essentially all estimators will have constant risk*, both interesting ones—like ridge, and trivial ones—like the null estimator $\hat{\beta} = 0$. Therefore, deriving *exact* formulae for asymptotic risk is of critical importance, as we can no longer distinguish estimators in terms of rates.

2.2 Stieltjes transform

In order to state and understand the main result we will use in our analysis, we need to define a few more quantities. Given a symmetric positive semidefinite matrix $A \in \mathbb{R}^{d \times d}$, we define its *spectral distribution* as the empirical distribution of its eigenvalues. This is denoted F_A, and writing $s_i(A)$, $i = 1, \dots, d$ for the eigenvalues of A, we have

$$F_A(t) = \frac{1}{d} \sum_{i=1}^{d} 1\{s_i(A) \le t\}. \tag{10}$$

This will be a useful tool because we can reason about the behavior of large covariance matrices by studying their distribution of their eigenvalues.

Another key tool to introduce is called the *Stieltjes transform*. This takes as input a measure F supported on the set $\mathbb{R}$ of real numbers, and produces a function m_F, defined by

$$m_F(z) = \int \frac{1}{s - z} \, dF(s), \quad \text{for } z \in \mathbb{C} \setminus \text{supp}(dF). \tag{11}$$

The Stieltjes transform has many nice properties. For example, there is a one-to-one correspondence function between Sieltjes functions and probability measures. That is, if F, G are probability measures, then $m_F = m_G \iff F = G$.

There is also is a close connection between the Stieltjes transform of a measure F and its moments, which we denote by

$$\mu_k(F) = \int s^k \, dF(s), \quad k = 0, 1, 2, \ldots.$$

We can see this because, by Taylor expansion,

$$\frac{1}{s-z} = -\frac{1}{z}\frac{1}{1-s/z} = -\frac{1}{z}\sum_{k=0}^{\infty}(s/z)^k,$$

and under appropriate regularity conditions, we can integrate both sides of the above, and exchange the order of the integral and infinite sum, to obtain

$$m_F(z) = -\sum_{k=0}^{\infty} \frac{\mu_k(F)}{z^{k+1}}.$$

Given this connection, it should not be surprising that the Stieltjes transform is also connected to weak convergence. That is, if $m_n = m_{G_n}$ is the Stieltjes transform of G_n, for $n = 1, 2, 3, \ldots$, then

$$m_n \to m_G \text{ as } n \to \infty \implies G_n \xrightarrow{d} G \text{ as} \to \infty.$$

In other words, convergence in Stieltjes transform implies convergence in distribution. The converse is also true: convergence in distribution implies convergence in Stieltjes functions, away from $z = 0$, because the function $s \mapsto 1/(s - z)$ is continuous and bounded when $z \neq 0$.

Lastly, the Stieltjes transform is intricately connected to matrix functionals that appear in ridge regression: for $F_{\hat{\Sigma}}$ denoting the empirical spectral distribution of the sample covariance $\hat{\Sigma} = X^\mathsf{T} X/n$, as in (10), note

$$\begin{aligned} m_{F_{\hat{\Sigma}}}(-\lambda) &= \int \frac{1}{s+\lambda} \, dF_{\hat{\Sigma}}(s) \\ &= \frac{1}{d}\sum_{i=1}^{n} \frac{1}{s_i(\hat{\Sigma}) + \lambda} \\ &= \frac{1}{d} \operatorname{tr}\left[(\hat{\Sigma} + \lambda I)^{-1}\right]. \end{aligned}$$

We can recognize the term inside the trace from the expression from the ridge solution at tuning parameter $n\lambda$: from (2), this is $\hat{\beta} = (\hat{\Sigma} + \lambda I)^{-1} X^\mathsf{T} Y/n$.

2.3 Marchenko-Pastur theorem

We are now ready to state one of the most important results in random matrix theory, which is called the *Marchenko-Pastur theorem*, or MP theorem, due to Marchenko and Pastur (1967). This result was further developed and generalized by many other authors, including Silverstein (1995).

Theorem 1 (Marchenko and Pastur 1967; Silverstein 1995)**.** *Let $X = Z\Sigma^{1/2} \in \mathbb{R}^{n \times d}$, where the entries of Z are i.i.d. from a distribution with zero mean and unit variance. Assume that as $n, d \to \infty$, it holds that $d/n \to \gamma \in (0, \infty)$, and the spectral distribution of Σ converges weakly, $F_\Sigma \xrightarrow{d} H$, where H is supported on $[0, \infty)$. Then, almost surely, the spectral distribution of $\hat{\Sigma} = X^\mathsf{T} X/n$ converges weakly to a deterministic limit, $F_{\hat{\Sigma}} \xrightarrow{d} F$. This limiting distribution $F = F(H, \gamma)$ depends on H and γ only. It can be identified with its Stieltjes transform m_F, which can be described as follows:*

$$m_F(z) + 1/z = \frac{1}{\gamma}(v_F(z) + 1/z), \tag{12}$$

where $v_F(z)$ is the unique solution of the nonlinear equation:

$$-\frac{1}{v_F(z)} = z - \gamma \int \frac{s}{1 + s v_F(z)} \, dH(s). \tag{13}$$

Several remarks are in order. First, the object v_F defined in (12) is called the *companion Stieltjes transform*, and is actually quite a natural object. Though the relationship in (12) may look obscure, you can just think of it through the lens of the following fact: the companion Stieltjes transform of the spectral distribution of $X^\mathsf{T} X/n$ is the name that we give to the Stieltjes transform of the spectral distribution of XX^T/d, that is,

$$v_{F_{X^\mathsf{T} X/n}} = m_{F_{XX^\mathsf{T}/d}}.$$

Second, the equation in (13), which is sometimes called the *Silverstein equation*, is not generally solvable in closed-form (for general H). However, in special cases it is. For example, in the isotropic case $\Sigma = I$, whose spectral distribution $F_\Sigma = \delta_1$ is a point mass at 1, we have of course $H = \delta_1$. In this case, equation (13) is explicitly solveable, and the limiting distribution F in Theorem 1 admits an explicit form as well, which we call the *Marchenko-Pastur law*, or MP law. For $\gamma \le 1$, this law is supported on an interval $[a, b]$, where $a = (1-\sqrt{\gamma})^2$ and $b = (1+\sqrt{\gamma})^2$, and it can be defined by its density

$$\frac{dF(s)}{ds} = \frac{1}{2\pi\gamma s}\sqrt{(b-s)(s-a)}. \tag{14}$$

For $\gamma > 1$, the Marchenko-Pastur law is just as above but has an additional point mass of probability $1-1/\gamma$ at the origin $s = 0$. See Figure 1 for a visualization.

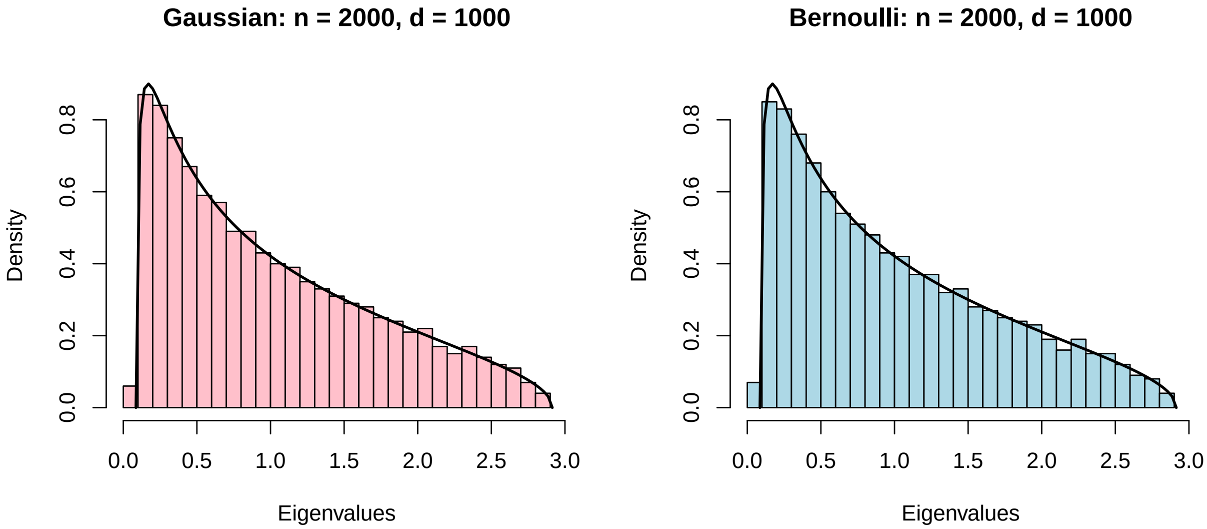

Figure 1: *Empirical verification of the MP theorem when* $n = 2000$, $d = 1000$. *The left panel shows the empirical distribution of eigenvalues of* $X^\mathsf{T} X/n$ *when it has standard Gaussian entries, and the right panel shows the same but when it has standardized Bernoulli entries. The black curve in each panel is the density of the MP law.*

Third, the Marchenko-Pastur theorem displays a remarkable phenomenon called *universality*: no matter the distribution of the elements of Z that give rise to our sample covariance matrix $\hat{\Sigma} = X^\mathsf{T} X/n$ (recall the relationship $X = Z\Sigma^{1/2}$), *we get the same limit* F for the spectral distribution of $\hat{\Sigma}$. This limit only depends on γ and H. So for example, in the isotropic case, we learn that if we populate the entries of X with i.i.d. standardized (zero mean, unit variance) random variables, whether they be Gaussian, Bernoulli, Poisson, t, etc., and plot a histogram of the eigenvalues of $X^\mathsf{T} X/n$ for large n, then it is "very likely" that they will look like they follow (14). See again Figure 1.

Fourth, and last, it is worth emphasizing that the distribution F from Theorem 1 is the *almost sure* limit of eigenvalues of $\hat{\Sigma} = X^\mathsf{T} X/n$. Interpreting this correctly can sometimes be challenging for people learning this material for the first time. Let us be clear about what it does *not* say: the result does not imply that the eigenvalues of $\hat{\Sigma}$ for large n will approximately concentrate around some deterministic number. (This would be the case in classical asymptotics, with $n \to \infty$ and d fixed.) Rather, it means that the eigenvalues will exhibit a *predictable spread* for large n. In other words, when seeking to empirically examine the statement in Theorem 1, as we did in Figure 1, we do not need to average results over repetitions or anything like that, because just a single draw of X should produce eigenvalues that approximately display the predicted spread. And if we simply redrew the Gaussian or Bernoulli entries, then we would (and should) get basically identical-looking plots.

2.4 Deterministic equivalents

It turns out that we can restate the Marchenko-Pastur theorem in modern (or at least, not-so-classical) terms, using what is known as the language of *deterministic equivalents*. Two sequences of (deterministic or random) matrices A_n, B_n, $n = 1, 2, 3, \dots$ of growing dimension are said to be asymptotically equivalent, written as $A_n \asymp B_n$, provided that for all sequences Θ_n, $n = 1, 2, 3, \dots$ that are bounded in trace norm,[1]

$$\operatorname{tr}\big[\Theta_n(A_n - B_n)\big] \to 0, \quad \text{as } n \to \infty.$$

This language gives us a way to cleanly state the MP theorem, as promoted by Dobriban and Sheng (2021) (these authors also develop a "calculus" for deterministic equivalents). The following is a transcription of a result by Rubio and Mestre (2011), that can be viewed as a generalized version of the MP theorem.

Theorem 2 (Rubio and Mestre 2011). *Let $X = Z\Sigma^{1/2} \in \mathbb{R}^{n \times d}$, where the entries of Z are i.i.d. from a distribution with zero mean, unit variance, and finite $8 + \eta$ moment, for some $\eta > 0$. Assume that as $n, d \to \infty$, the aspect ratio $\gamma_n = d/n$ remains bounded away from 0 and ∞, as do the eigenvalues of Σ. Then the resolvent of $\hat{\Sigma} = X^\mathsf{T} X/n$ is asymptotically equivalent to a deterministic matrix, namely:*

$$(\hat{\Sigma} - zI)^{-1} \asymp (a_n \Sigma - zI)^{-1}, \quad \textit{for } z \in \mathbb{C} \setminus \mathbb{R}_+, \tag{15}$$

where a_n is the unique solution of the fixed-point equation:

$$\frac{1}{\gamma_n}\left(\frac{1}{a_n} - 1\right) = \frac{1}{d} \operatorname{tr}\big[\Sigma(a_n\Sigma - zI)^{-1}\big]. \tag{16}$$

We note that this theorem is quite general: it does not require d/n to actually converge to anything, nor require that the spectral distribution of Σ has a limit. Furthermore, asymptotic equivalence implies many interesting things: for example, taking $\Theta_n = I/d$ in the definition of asymptotic equivalence shows that (15) implies convergence in Stieltjes transforms, which effectively recovers the original MP result. However, there is much more we can learn from (15), including convergence of eigenvectors; see Rubio and Mestre (2011) for a discussion.

Lastly, as a particularly simple and hence notable consequence of Theorem 2, it is shown in Dobriban and Sheng (2021) that we can take $z \to 0$ in (15), which gives

$$\hat{\Sigma}^{-1} \asymp \frac{1}{1 - \gamma_n} \Sigma^{-1}, \tag{17}$$

where we have used the fact (which can easily be verified) that $a_n = 1 - \gamma_n$ solves the fixed-point equation (16) in the case $z = 0$.

3 Least squares analysis

In this section, we will analyze the out-of-sample risk of least squares regression in a proportional asymptotics model, both as a warm-up for ridge regression (the focus of the following sections), but also because it is a certainly important result in its own right—as we alluded to in previous lectures more than once.

[1] The trace norm of $\Theta \in \mathbb{R}^{n \times n}$ is $\|\Theta\|_* = \sum_{i=1}^n |\sigma_i(\Theta)|$, where $\sigma_i(\Theta)$, $i = 1, \dots, n$ are the singular value of Θ. Equivalently, $\|\Theta\|_* = \operatorname{tr}[(\Theta^\mathsf{T}\Theta)^{1/2}]$.

We consider the linear model

$$Y = X\beta_0 + \epsilon, \tag{18}$$

assuming as usual that $\epsilon \in \mathbb{R}^n$ has i.i.d. entries with mean zero and variance σ^2, and $\epsilon \perp\!\!\!\perp X$. To clearly lay out the conditions on the features $X \in \mathbb{R}^{n \times d}$, we assume the following:

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

(A2) the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ has eigenvalues bounded away from 0 and ∞;

(A3) $d/n \to \gamma \in (0, 1)$ as $n, d \to \infty$,

To emphasize, we are placing the restriction here that $\gamma < 1$, called the *underparametrized* regime. We will analyze the asymptotic risk of the ordinary least squares estimator, which we can write as

$$\hat{\beta} = (X^\mathsf{T} X/n)^{-1} X^\mathsf{T} Y/n. \tag{19}$$

One can show that this is almost surely well-defined under the assumptions laid out.[2] Now let $x_0 = \Sigma^{1/2} z_0$ be i.i.d. to the rows of X, and consider the out-of-sample risk of least squares, conditional on X,

$$\mathrm{Risk}_X(\hat{\beta}; \beta_0) = \mathbb{E}\big[(x_0^\mathsf{T} \hat{\beta} - x_0^\mathsf{T} \beta_0)^2 \,\big|\, X\big]. \tag{20}$$

By standard calculations, as covered previously, we can decompose the risk (20) of the least squares estimator (19) into bias and variance terms; the bias term is zero, so the risk is pure variance,

$$\begin{aligned} \mathrm{Risk}_X(\hat{\beta}; \beta_0) &= \frac{\sigma^2}{n} \operatorname{tr}\Big(\mathbb{E}[x_0 x_0^\mathsf{T}]\,(X^\mathsf{T} X/n)^{-1}\Big) \\ &= \frac{\sigma^2}{n} \operatorname{tr}\Big(\Sigma\big[\Sigma^{1/2}(Z^\mathsf{T} Z/n)\Sigma^{1/2}\big]^{-1}\Big) \\ &= \frac{\sigma^2 d}{n} \frac{1}{d} \operatorname{tr}\big[(Z^\mathsf{T} Z/n)^{-1}\big]. \end{aligned} \tag{21}$$

There are now several ways to proceed to compute the limit of above line, all based around the Marchenko-Pastur theorem. Of course, each way must arrive at the same answer, which is

$$\mathrm{Risk}_X(\hat{\beta}; \beta_0) \overset{\text{as}}{\to} \sigma^2 \frac{\gamma}{1-\gamma}, \tag{22}$$

where to be clear, "almost surely" is to be interpreted with respect to the distribution of X. Looking back at (21), clearly $\sigma^2 d/n \to \sigma^2\gamma$, so in order to establish (22) it suffices to prove that

$$\frac{1}{d} \operatorname{tr}\big[(Z^\mathsf{T} Z/n)^{-1}\big] \overset{\text{as}}{\to} \frac{1}{1-\gamma}. \tag{23}$$

Below, we step through three routes for calculating this limit. But first, it is worth emphasizing the behavior of the asymptotic risk profile in (22):

> *The out-of-sample risk of least squares blows up as $\gamma \to 1$ from below, that is, as we grow the aspect ratio until $d = n$ in the limit, least squares regression exhibits catastrophic risk.*

What happens past $\gamma = 1$? The answer may surprise you. We'll return to this in the next lecture.

Marchenko-Pastur theorem, followed by calculus. We can recognize the left-hand side in (23) in terms of the Stieltjes transform of the spectral distribution $F_{Z^\mathsf{T} Z/n}$,

$$\frac{1}{d} \operatorname{tr}\big[(Z^\mathsf{T} Z/n)^{-1}\big] = m_{F_{Z^\mathsf{T} Z/n}}(0).$$

[2] To be more precise, as $n, d \to \infty$ with $d/n \to \gamma \in (0, 1)$, the minimum eigenvalue of $X^\mathsf{T} X/n$ will be almost surely lower bounded away from zero. This follows from what is called the Bai-Yin theorem (Bai and Yin, 1993), along with the fact that Σ has eigenvalues bounded away from zero.

To study the limit of $m_{F_{Z^\mathsf{T} Z/n}}(0)$, we can use the Marchenko-Pastur theorem, transcribed in Theorem 1. This tells us that $F_{Z^\mathsf{T} Z/n}$ converges weakly almost surely to F, the MP law in (14), hence we get convergence of Stieltjes transforms, so

$$m_{F_{Z^\mathsf{T} Z/n}}(0) \overset{\text{as}}{\to} m_F(0).$$

Fortunately, the Stieltjes transform of F in (14) has an explicit form, for real $z > 0$:

$$m(-z) = \frac{-(1-\gamma+z) + \sqrt{(1-\gamma+z)^2 + 4\gamma z}}{2\gamma z}. \tag{24}$$

Since the limit as $z \to 0$ is indeterminate, we can use l'Hôpital's rule to calculate:

$$\begin{aligned} \lim_{z\to 0} m(-z) &= \lim_{z \to 0} \frac{-1 + \frac{1+\gamma+z}{\sqrt{(1-\gamma+z)^2+4\gamma z}}}{2\gamma} \\ &= \frac{-1 + \frac{1+\gamma}{1-\gamma}}{2\gamma} = \frac{1}{1-\gamma}, \end{aligned}$$

which establishes the result in (23).

Gaussian calculation, in expectation. We can actually get away with a simpler calculation. Observe that the Marchenko-Pastur theorem tells us that the limit of the left-hand side in (23) is both *universal* and *almost sure*, and thus it suffices to compute it in the Gaussian case, in expectation. That is, it suffices to study the limit of

$$\frac{1}{d} \operatorname{tr}\Big[\mathbb{E}\big[(Z^\mathsf{T} Z/n)^{-1}\big]\Big], \quad \text{for } Z \text{ having i.i.d. } N(0,1) \text{ entries.}$$

In this case, $Z^\mathsf{T} Z$ is Wishart, and $(Z^\mathsf{T} Z)^{-1}$ is inverse Wishart, so it has a known expectation $\mathbb{E}[(Z^\mathsf{T} Z)^{-1}] = I/(n-d-1)$. This means that

$$\frac{1}{d} \operatorname{tr}\Big[\mathbb{E}\big[(Z^\mathsf{T} Z/n)^{-1}\big]\Big] = \frac{n}{n-d-1} \to \frac{1}{1-\gamma},$$

as desired, which establishes (23).

Deterministic equivalents. The formulation of the Marchenko-Pastur theorem in terms of deterministic equivalents, as transcribed in Theorem 2, leads to the simplest calculation. To be precise, in order to use this result, we must assume a bit more about the distribution of entries of Z: recall that we must assume that it has $8 + \eta$ moments, for some $\eta > 0$. Now recall that an implication of this theorem is that we have the deterministic equivalence (17), in the isotropic case. But we are in this case—there are no appearances of Σ in (22). A direct implication of (17) (just use $\Theta_n = I/d$ in the definition of asymptotic equivalence) is that

$$\frac{1}{d} \operatorname{tr}\big[(Z^\mathsf{T} Z/n)^{-1}\big] \quad \text{and} \quad \frac{1}{1-d/n}\frac{1}{d}\operatorname{tr}(I) = \frac{1}{1-d/n} \quad \text{have the same asymptotic limit,}$$

and we can just read off that right-hand quantity converges to $1/(1-\gamma)$, which again proves (23).

4 Ridge analysis

On to the analysis of ridge regression, which we will break up into two cases: the isotropic case, in which $\Sigma = I$, and the general case, in which Σ is arbitrary (subject to minor restrictions, as usual, like eigenvalues bounded away from 0 and ∞). In the isotropic case, we will be able to analyze the ridge risk with the random matrix theory tools introduced previously. The general Σ case will be more challenging, and there we will simplify the bias calculation by taking underlying signal β_0 to be random, i.e., by imposing a prior on β_0. We will discuss briefly what happens for general Σ and fixed β_0 at the end.

As in the least squares analysis, we will assume the linear model (18), and will use similar assumptions to (A1)–(A3) for the feature model, but considering the full range $\gamma \in (0, \infty)$. To be specfic, we assume:

(B1) $X = Z\Sigma^{1/2}$, 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 $\eta > 0$;

(B2) the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ has eigenvalues bounded away from 0 and ∞, and satisfies $F_\Sigma \overset{d}{\to} H$, as $n, d \to \infty$;

(B3) $d/n \to \gamma \in (0, \infty)$ as $n, d \to \infty$.

We will analyze the asymptotic out-of-sample risk (20) of the ridge estimator. (Recall that this is conditional on X.) For convenience we will reparametrize the ridge estimator as

$$\hat{\beta} = (X^\mathsf{T} X/n + \lambda I)^{-1} X^\mathsf{T} Y/n, \tag{25}$$

which can either be seen as the original ridge estimator in (2) with tuning parameter $n\lambda$, or as the solution in the original ridge problem (1) after rescaling the loss term by $1/n$.

At the outset, we will record the following facts, which you'll verify on the homework. The bias and variance components of the risk (20) of the ridge estimator (25) are:

$$B_X(\hat{\beta}; \beta_0) = \lambda^2 \beta_0^\mathsf{T} (\hat{\Sigma} + \lambda I)^{-1} \Sigma (\hat{\Sigma} + \lambda I)^{-1} \beta_0 \tag{26}$$

$$V_X(\hat{\beta}) = \frac{\sigma^2}{n} \operatorname{tr}\big[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2} \Sigma\big], \tag{27}$$

respectively, where recall $\hat{\Sigma} = X^\mathsf{T} X/n$. It is worth noting that the variance does not depend on β_0.

4.1 Isotropic Σ, fixed β_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 (18). Below we analyze the bias and variance separately.

Bias analysis. When $\Sigma = I$, the bias (26) becomes

$$B_X(\hat{\beta}; \beta_0) = \lambda^2 \beta_0^\mathsf{T} (\hat{\Sigma} + \lambda I)^{-2} \beta_0. \tag{28}$$

The key is to recognize this as the derivative with respect to λ of a certain functional,

$$\beta_0^\mathsf{T} (\hat{\Sigma} + \lambda I)^{-2} \beta_0 = -\frac{d}{d\lambda} \Big\{ \beta_0^\mathsf{T} (\hat{\Sigma} + \lambda I)^{-1} \beta_0 \Big\}. \tag{29}$$

By the deterministic equivalence in (15) from Theorem 2, we know (just take $\Theta_n = \beta_0 \beta_0^\mathsf{T}$ in the definition of deterministic equivalence) that

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

We need to figure out the asymptotic limit of a_n. Instead of trying to solve the fixed-point equation (16), it is easier to "sneak up on the answer", by approaching it this way:

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

and by the standard MP asymptotics, we know that the left-hand side converges almost surely to $m_F(-\lambda)$, the Stieltjes transform (24) of the MP law (14), evaluated at $-\lambda$. Thus we have $1/(a_n + \lambda) \to m_F(-\lambda)$, and

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

Returning to (29), some calculations involing Vitali's theorem (whose details we omit) show that we may exchange the order of the derivative and the limit, yielding

$$\beta_0^\mathsf{T} (\hat{\Sigma} + \lambda I)^{-2} \beta_0 \overset{\text{as}}{\to} r^2 m_F'(-\lambda),$$

and finally from (28),

$$B_X(\hat{\beta}; \beta_0) \overset{\text{as}}{\to} \lambda^2 r^2 m_F'(-\lambda). \tag{30}$$

Variance analysis. When $\Sigma = I$, the variance (27) becomes

$$\begin{aligned} V_X(\hat{\beta}) &= \frac{\sigma^2}{n} \operatorname{tr}\big[\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\big] \\ &= \frac{\sigma^2}{n}\Big(\operatorname{tr}\big[(\hat{\Sigma}+\lambda I)^{-1}\big] - \lambda \operatorname{tr}\big[(\hat{\Sigma}+\lambda I)^{-2}\big]\Big), \end{aligned}$$

where in the second line we added and subtracted λI to the leading $\hat{\Sigma}$ inside the trace. Calculation of the limit of the above is now straightforward, given what we have done above for the bias. After multiplying and dividing by d, this is the same as

$$\frac{\sigma^2 d}{n}\bigg(\frac{1}{d}\operatorname{tr}\big[(\hat{\Sigma}+\lambda I)^{-1}\big] - \frac{\lambda}{d}\operatorname{tr}\big[(\hat{\Sigma}+\lambda I)^{-2}\big]\bigg).$$

The first term inside the parentheses has limit $m_F(-\lambda)$ by standard MP asymptotics, and the second term has limit $-\lambda m_F'(-\lambda)$ by the same arguments as above. Thus

$$V_X(\hat{\beta}) \overset{\text{as}}{\to} \sigma^2\gamma\big(m_F(-\lambda) - \lambda m_F'(-\lambda)\big). \tag{31}$$

Putting it together. Adding the bias (30) and variance results (31) together, we get

$$\operatorname{Risk}_X(\hat{\beta};\beta_0) \overset{\text{as}}{\to} \sigma^2\gamma\Big(m_F(-\lambda) - \lambda(1-\alpha\lambda)m_F'(-\lambda)\Big), \tag{32}$$

where we have introduced $\alpha = r^2/(\sigma^2\gamma)$. Note that we can think of this as $\alpha = \mathrm{SNR}/\gamma$, where $\mathrm{SNR} = r^2/\sigma^2$ can be thought of the signal-to-noise ratio for our problem. Recall that m_F is the Stieltjes transform (24) of the MP law (14).

It can be shown that the asymptotically optimal tuning parameter value—the one minimizing the asymptotic risk in (32), is $\lambda^* = 1/\alpha$. This has the general behavior that we would intuitively expect: it shrinks (less regularization) as α grows (higher SNR), or equivalently, it grows (more regularization) as α shrinks (lower SNR). Moreover, the asymptotic risk (32) at the tuning parameter value $\lambda^* = 1/\alpha$ simplifies to

$$\sigma^2\gamma m_F(-1/\alpha) = \sigma^2 \frac{-(1-\gamma+1/\alpha) + \sqrt{(1-\gamma+1/\alpha)^2 + 4\gamma/\alpha}}{2\gamma/\alpha}.$$

It is worth emphasizing that this *does not blow up at* $\gamma = 1$, unlike the asymptotic least squares risk (22). Regularization has saved the day!

4.2 General Σ, random β_0

We consider the general Σ case. We follow the general approach in Dobriban and Wager (2018), but adopt the perspective of deterministic equivalents as suggested by Dobriban and Sheng (2021). We will place a spherical prior on β_0, such that

$$\mathbb{E}[\beta_0\beta_0^\mathsf{T}] = \frac{r^2}{d} I. \tag{33}$$

Note this implies $\mathbb{E}\|\beta_0\|_2^2 = r^2$. Our measure of risk is now a kind of *Bayes* out-of-sample prediction risk,

$$\operatorname{Risk}_X(\hat{\beta}) = \mathbb{E}\big[(x_0^\mathsf{T}\hat{\beta} - x_0^\mathsf{T}\beta_0)^2 \,\big|\, X\big], \tag{34}$$

where to be clear the expectation is over ϵ, x_0, and β_0 (all independent), and conditional on X.

Bias analysis. For the bias, after taking an expectation in (26) over β_0 drawn from (33), we get

$$\begin{aligned} B_X(\hat{\beta}) &= \lambda^2 \mathbb{E}\Big[\beta_0^\mathsf{T}(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\beta_0\Big] \\ &= \frac{\lambda^2 r^2}{d}\operatorname{tr}\big[\Sigma(\hat{\Sigma}+\lambda I)^{-2}\big], \end{aligned} \tag{35}$$

where in the second line we used trace rotation, as in $\beta_0^\mathsf{T} M \beta_0 = \operatorname{tr}(\beta_0 \beta_0^\mathsf{T} M)$ for a matrix M, and invoked the prior (33). To compute the limit of (35), the key, similar to the isotropic case, is to recognize that

$$\frac{1}{d} \operatorname{tr}\left[\Sigma(\hat{\Sigma} + \lambda I)^{-2}\right] = -\frac{d}{d\lambda} \left\{ \frac{1}{d} \operatorname{tr}\left[\Sigma(\hat{\Sigma} + \lambda I)^{-1}\right] \right\}. \tag{36}$$

By the deterministic equivalence in (15) from Theorem 2, we know that

$$\frac{1}{d} \operatorname{tr}\left[\Sigma(\hat{\Sigma} + \lambda I)^{-1}\right] \quad \text{and} \quad \frac{1}{d} \operatorname{tr}\left[\Sigma(a_n \Sigma + \lambda I)^{-1}\right] \quad \text{have the same asymptotic limit.}$$

Again, we need to figure out the asymptotic limit of a_n. Solving the fixed-point equation (16) will not be possible, but nonetheless we can "sneak up on the answer", by rewriting (16) for $z = -\lambda$ as

$$\frac{1}{a_n} = 1 + \frac{\gamma_n}{d} \operatorname{tr}\left[\Sigma(a_n \Sigma + \lambda I)^{-1}\right].$$

What does this remind you of? Recall the Silverstein equation (13); at $z = -\lambda$, this can be rewritten as

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

Writing a for the limit of a_n, note that the second-to-last display converges as $n, d \to \infty$ to the last display, with the relationship $a = \lambda v_F(-\lambda)$. That is, to be clear, we have learned that $a_n \to \lambda v_F(-\lambda)$, where v_F is the companion Stieltjes transform of the limiting spectral distribution F from the MP theorem, and

$$\frac{1}{d} \operatorname{tr}\left[\Sigma(a_n \Sigma + \lambda I)^{-1}\right] = \frac{1}{\gamma_n}\left(\frac{1}{a_n} - 1\right) \to \frac{1}{\gamma}\left(\frac{1}{\lambda v_F(-\lambda)} - 1\right),$$

and therefore

$$\frac{1}{d} \operatorname{tr}\left[\Sigma(\hat{\Sigma} + \lambda I)^{-1}\right] \overset{\text{as}}{\to} \underbrace{\frac{1}{\gamma}\left(\frac{1}{\lambda v_F(-\lambda)} - 1\right)}_{\phi_F(-\lambda)}.$$

Returning to (36), after checking some conditions (whose details we omit), we may exchange the order of the derivative and the limit, yielding

$$\frac{1}{d} \operatorname{tr}\left[\Sigma(\hat{\Sigma} + \lambda I)^{-2}\right] = \phi_F'(-\lambda),$$

and finally from (35),

$$B_X(\hat{\beta}) \overset{\text{as}}{\to} \lambda^2 r^2 \phi_F'(-\lambda). \tag{37}$$

Variance analysis. For the variance (27), by adding and subtracting λI in the leading $\hat{\Sigma}$ in the trace, we get

$$\begin{aligned} V_X(\hat{\beta}) &= \frac{\sigma^2 d}{n}\Big(\operatorname{tr}\left[\Sigma(\hat{\Sigma} + \lambda I)^{-1}\right] - \lambda \operatorname{tr}\left[\Sigma(\hat{\Sigma} + \lambda I)^{-2}\right] \Big) \\ &= \frac{\sigma^2 d}{n}\left(\frac{1}{d} \operatorname{tr}\left[\Sigma(\hat{\Sigma} + \lambda I)^{-1}\right] - \frac{\lambda}{d} \operatorname{tr}\left[\Sigma(\hat{\Sigma} + \lambda I)^{-2}\right] \right). \end{aligned}$$

We can apply the same logic as developed for the bias to each of the two terms above inside the parentheses: the first has limit $\phi_F(-\lambda)$ and the second has limit $-\lambda \phi_F'(-\lambda)$, therefore

$$V_X(\hat{\beta}) \overset{\text{as}}{\to} \sigma^2 \gamma \big(\phi_F(-\lambda) - \lambda \phi_F'(-\lambda)\big). \tag{38}$$

We remark that this variance calculation did not require the prior assumption (33).

Putting it together. Adding the bias (37) and variance results (38) together, we get

$$\mathrm{Risk}_X(\hat{\beta}) \overset{\text{as}}{\to} \sigma^2\gamma\Big(\phi_F(-\lambda) - \lambda(1-\alpha\lambda)\phi_F'(-\lambda)\Big), \tag{39}$$

where as before $\alpha = r^2/(\sigma^2\gamma)$. Recall that

$$\phi_F(z) = -\frac{1}{\gamma}\bigg(\frac{1}{zv_F(z)} + 1\bigg),$$

and v_F is the companion Stieltjes transform of the limit F of the spectral distribution of $\hat{\Sigma}$, as given by Theorem 1. It is worth noting the close similarity between the results in the general Σ case (39) and in the $\Sigma = I$ case (32), where m_F in the latter plays the role of ϕ_F in the former (we have written the asymptotic risk above precisely in this way in order to emphasize this connection).

There is an alternative formulation that we can obtain by simply calculating the derivative of ϕ_F, then reducing it to as simple terms as possible involving the companion Stieltjes transform v_F, which results in:

$$\mathrm{Risk}_X(\hat{\beta}) \overset{\text{as}}{\to} \frac{r^2}{\gamma}\bigg(\frac{1}{v_F(-\lambda)} - \frac{\lambda v_F'(-\lambda)}{v_F(-\lambda)^2}\bigg) + \sigma^2\bigg(\frac{v_F'(-\lambda)}{v_F(-\lambda)^2} - 1\bigg). \tag{40}$$

It is worth being clear that, in the general Σ case, while we were able to obtain an exact expressions for the limiting risk, either (39) or equivalently (40), these are no longer truly closed-form, as the solution v_F to the Silverstein equation (13) does not have an explicit closed-form for general H.

Optimal tuning. Remarkably, despite the lack of a closed-form limiting risk, it is shown in Dobriban and Wager (2018) that the asymptotically optimal tuning parameter value—the one that minimizes the asymptotic risk in (39) or (40), is once again $\lambda^* = 1/\alpha$, regardless of the sequence of covariance matrices Σ (regardless of H). Their argument is too clever to pass by in these notes, and so we outline it here. Specialize to the case where $\epsilon \sim N(0, \sigma^2 I)$ and $\beta_0 \sim N(0, (r^2/d)I)$ in (18) and (33), respectively. As we argued previously (recall (3), (5)), note that the ridge estimator with

$$\lambda_n^* = (\sigma^2 d)/(r^2 n)$$

is the Bayes estimator in this normal-normal model. In fact, it is the unique Bayes estimator, and thus it obtains a smaller Bayes risk than any other estimator. Note that $\lambda_n^* \to \sigma^2\gamma/r^2 = \lambda^*$. Since the limit of the risk in (34) is both universal and almost sure, we can use the optimality of λ_n^* as argued above, along with an equicontinuity argument, to show that $\lambda^* = 1/\alpha$ is optimal in the limit.

4.3 General Σ, fixed β_0?

For a general Σ, the behavior of ridge regression along an arbitrary sequence of fixed signal vectors β_0 can actually be surprisingly exotic. First, it is no longer true that the optimal limiting tuning parameter value λ^* is simply $1/\alpha$, and furthermore, it is no longer true that it is even positive (it may be zero). The reason for this exotic behavior is the bias term (26), specifically the way that it depends on the joint geometry of β_0 and Σ.

The results describing the asymptotic risk here are very recent, and relate to the study of overparametrization, so we will touch on them in the next lecture.

References

Zhidong Bai and Jack Silverstein. *Spectral Analysis of Large Dimensional Random Matrices.* Springer, 2010.

Zhidong Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21(3):1275–1294, 1993.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *Annals of Statistics*, 49(2): 918–943, 2021.

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

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.

Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

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

Jack Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

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