# Lecture 04: Introduction to High Dimensional Asymptotics and Random Matrix Theory

Geoff Pleiss

In the previous class, we made an intuitive justification for why the double descent phenomenon occurs in high-dimensional ridgeless regression. We now proceed with a more formal analysis of risk as a function of the number of parameters.

Over the next two lectures, we will derive an asymptotic result for ridge regression, a result which will have implications for neural networks as we will see later in the course. This lecture will focus on a light introduction to the mathematical tools needed to derive an asymptotic result. The following lecture will apply these tools to the ridge risk.

---

## 1) Ridge Regression Problem Setup

To begin, let $\hat{\boldsymbol{\theta}}_\lambda$ be the ridge regression estimator with regularization parameter $\lambda$ trained on the dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, which we vectorize as $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$. We assume that

$$\boldsymbol{x} \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}), \qquad y = \boldsymbol{\theta}^{*\top}\boldsymbol{x} + \sigma\epsilon, \qquad \epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \tag{1}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, and $\sigma^2 > 0$ are fixed.[1] To simplify analysis, we will assume that the amount of regularization scales with $n$, i.e.:

$$\hat{\boldsymbol{\theta}}_\lambda = (\boldsymbol{X}^\top\boldsymbol{X} + n\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

We can write $\boldsymbol{X}^\top\boldsymbol{X}$ as a scaled **empirical covariance matrix**:

$$\tfrac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} := \hat{\boldsymbol{\Sigma}}, \qquad \mathbb{E}[\hat{\boldsymbol{\Sigma}}] = \boldsymbol{\Sigma}.$$

The risk of $\hat{f}_\lambda(\boldsymbol{x}) = \boldsymbol{x}^\top\hat{\theta}_\lambda$ can be factorized into bias and variance terms (as on the problem set):

$$\mathcal{R}(\hat{\boldsymbol{\theta}}_\lambda) = \underbrace{\mathbb{E}\left[\left(\boldsymbol{\theta}^* - \mathbb{E}[\hat{\boldsymbol{\theta}}_\lambda]\right)^\top\boldsymbol{\Sigma}\left(\boldsymbol{\theta}^* - \mathbb{E}[\hat{\boldsymbol{\theta}}_\lambda]\right)\right]}_{\mathcal{B}(\hat{\boldsymbol{\theta}}_\lambda)=\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}_\lambda - \mathbb{E}[\hat{\boldsymbol{\theta}}_\lambda]\right)^\top\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\theta}}_\lambda - \mathbb{E}[\hat{\boldsymbol{\theta}}_\lambda]\right)\right]}_{\mathcal{V}(\hat{\boldsymbol{\theta}}_\lambda)=\text{Var}}.$$

Plugging in $\hat{\boldsymbol{\theta}}_\lambda = (\boldsymbol{X}^\top\boldsymbol{X} + n\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\theta}^* + \sigma\boldsymbol{\epsilon}) = (\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I})^{-1}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\theta}^* + \frac{\sigma}{n}\boldsymbol{X}^\top\boldsymbol{\epsilon})$, where $\sigma\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}^*$, and simplifying, we have

$$\begin{aligned}
\mathcal{B}(\hat{\boldsymbol{\theta}}_\lambda) &= \boldsymbol{\theta}^{*\top}\mathbb{E}\left[\left(\boldsymbol{I} - (\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I})^{-1}\hat{\boldsymbol{\Sigma}}\right)\boldsymbol{\Sigma}\left(\boldsymbol{I} - (\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I})^{-1}\hat{\boldsymbol{\Sigma}}\right)\right]\boldsymbol{\theta}^* \\
&= \lambda^2\boldsymbol{\theta}^{*\top}\mathbb{E}\left[\left(\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I}\right)^{-1}\right]\boldsymbol{\theta}^*, && \text{(Woodbury on } (\boldsymbol{I} - (\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I})^{-1}\hat{\boldsymbol{\Sigma}})\text{)} \\
\mathcal{V}(\hat{\boldsymbol{\theta}}_\lambda) &= \frac{\sigma^2}{n^2}\mathbb{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{X}\left(\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon}\right] \\
&= \frac{\sigma^2}{n}\operatorname{Tr}\mathbb{E}\left[\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I}\right)^{-1}\right]. && \text{(independence of } \boldsymbol{X}, \boldsymbol{\epsilon}\text{, cyclic prop. of Tr)}
\end{aligned}$$

---

[1] Technically, we only assume that $\boldsymbol{x}$ and $\epsilon$ are sampled from **sub-Gaussian** distributions with the moments given in Eq. (1).

## 2) The Need for High Dimensional Asymptotics and Random Matrix Theory

Unfortunately, these expressions are about as simple as we can make them in closed form. Both $\mathcal{B}(\hat{\boldsymbol{\theta}}_\lambda)$ and $\mathcal{V}(\hat{\boldsymbol{\theta}}_\lambda)$ are expectations of some complex function of the random matrix $\hat{\boldsymbol{\Sigma}}$, and the only functions we can compute in closed form are:

1. $\boldsymbol{E}[\hat{\boldsymbol{\Sigma}}] = \boldsymbol{\Sigma}$ and
2. $\boldsymbol{E}[(\hat{\boldsymbol{\Sigma}})^{-1}] = (\frac{1}{n-d-1})\boldsymbol{\Sigma}^{-1}$, if $n > d+1$ and assuming $\boldsymbol{X}$ is Gaussian.

As in most of statistics, we can assume that some "nice simplifying structure" emerges that simplifies these expressions when the problems "get really large." However, we need to be VERY careful about what we mean by "get really large." If we simply take $n \to \infty$, then $\hat{\boldsymbol{\Sigma}} \to \boldsymbol{\Sigma}$ and our bias and variance both disappear as $\lambda \to 0$. This analysis may be appropriate if we're trying to model linear regression with $n \gg d$, but it's going to be a horrible model for problem where $n \approx d$ or $d > n$.

The correct framework for asymptotically analyzing these problems requires a conceptual leap. We will examine what happens when $n, d \to \infty$ *simultaneously.* In other words, we will assume that $d$ grows linearly with $n$, i.e. $d = \gamma n$ for some fixed $\gamma > 0$, and then we will take $n$ (and $d$) $\to \infty$. This limit is known as the **high-dimensional asymptotic regime** and analyzing it requires tools from **random matrix theory**.

---

## 3) Warm-Up: The Marchenko-Pastur Distribution for Isotropic Data

Imagine for a second that $\boldsymbol{\Sigma} = \boldsymbol{I}$. Then

$$\mathcal{V}(\hat{\boldsymbol{\theta}}_\lambda) = \sigma^2 \operatorname{Tr} \mathbb{E}\left[\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-2}\right] = \sigma^2 \sum_{i=1}^{d} \mathbb{E}\left[\frac{\hat{s}_i}{(\hat{s}_i + \lambda)^2}\right], \tag{2}$$

where $\hat{s}_i$ are the eigenvalues of $\hat{\boldsymbol{\Sigma}}$. Thus, understanding $\mathcal{V}(\hat{\boldsymbol{\theta}}_\lambda)$ as $n, d \to \infty$ requires understanding what happens to the $\hat{s}_i$ as $n, d \to \infty$. We will accomplish this understanding by viewing the summation over eigenvalues through a probabilistic lens.

If we define $\hat{F}(\cdot)$ as the **empirical distribution over eigenvalues**

$$\hat{F}(s) = \frac{1}{d} \sum_{i=1}^{d} \mathbf{1}[s = \hat{s}_i],$$

then we can rewrite Eq. (2) as an expectation:

$$\mathcal{V}(\hat{\boldsymbol{\theta}}_\lambda) = \frac{\sigma^2}{d} \int \frac{s}{(s+\lambda)^2} \, d\hat{F}(s).$$

Just as the central limit theorem tells us that empirical distributions of sums converge to normal distributions, the **Marchenko-Pastur theorem** tells us that the empirical distribution of eigenvalues also converge to a deterministic distribution.

**Theorem 1** (Marchenko and Pastur [1967])**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ have i.i.d. sub-Gaussian entries with mean 0 and variance 1. Assuming the ratio $\gamma = d/n \in (0, 1]$ is fixed, the empirical distribution of eigenvalues $\hat{F}(s)$ of $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}$ converges (in distribution) to the* **Marchenko-Pastur distribution** *$F(s)$ as $n, d \to \infty$:*

$$\lim_{n,d \to \infty} \hat{F}(s) = F(s), \qquad \frac{dF(s)}{ds} = \begin{cases} \frac{1}{2\pi\gamma s}\sqrt{(s_+ - s)(s - s_-^2)} & s \in [s_-, s_+], \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

*where $s_\pm = (1 \pm \sqrt{\gamma})^2$.*

Note that we can arrive at a density for eigenvalues for the $\gamma > 1$ (i.e. $d < n$) by recognizing that the non-zero eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ are the same as the non-zero eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$.

It may be strange to think about a continuous distribution over eigenvalues of a matrix! To gain an intuitive understanding, consider a histogram of eigenvalues of $\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}$. For large $n, d$, this histogram can be smoothed out with the Marchenko-Pastur density. See Fig. 1 for an illustration.

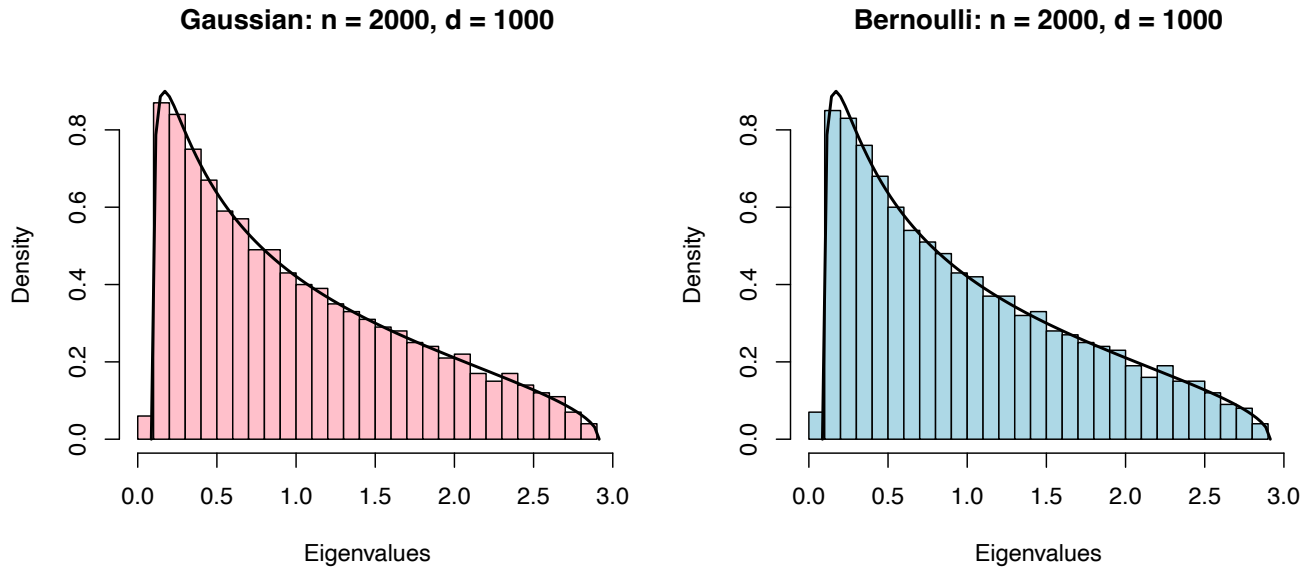


Figure 1: Histogram of eigenvalues of $\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}$ for $n = 2000$, $d = 1000$, when the entries of $\boldsymbol{X}$ are drawn i.i.d. from a Gaussian distribution (left) versus a Bernoulli distribution (right) with zero mean and unit variance. The values of the histogram roughly correspond to the Marchenko-Pastur density. (Figure reproduced from Tibishriani [2023].)

Therefore, the variance for isotropic ridge regression can be approximated by its high-dimensional asymptotic limit:

$$\mathcal{V}(\hat{\boldsymbol{\theta}}_\lambda) = \frac{\sigma^2}{d} \int \frac{s}{(s+\lambda)^2}\, d\hat{F}(s) \approx \frac{\sigma^2}{d} \int \frac{s}{(s+\lambda)^2} \frac{dF(s)}{ds}\, ds,$$

which is now an analytic one-dimensional integral that we can numerically solve!

### 3.1 Gaussian Universality

Notably, from Fig. 1 we see that the distribution of eigenvalues doesn't really depend on the distribution of the entries of $\boldsymbol{X}$. The left and right histograms, corresponding to Gaussian and Bernoulli entries (respectively) for $\boldsymbol{Z}$, are nearly identical. This **universality** is a key feature of random matrix theory: in the high-dimensional asymptotic limit, the properties of a random matrix often only depends on its first and second moments and not on the specific distribution of the entries. Therefore, we can often analyze the behaviour of random matrices with Gaussian entries without much loss of generality.

---

## 4) A Hand-Wavy Derivation of the Marchenko-Pastur Theorem with the Stieltjes Transform

How do we arrive at the Marchenko-Pastur distribution for isotropic data? We will need to introduce a concept from probability known as the **Stieltjes transform**. The Stieltjes transform of a symmetric matrix $\boldsymbol{A}$ is a $\mathbb{R} \to \mathbb{R}$ function defined as

$$m_{\boldsymbol{A}}(-\lambda) = \tfrac{1}{d} \operatorname{Tr}\left((\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\right), \qquad \lambda > 0, \quad \boldsymbol{A} \in \mathbb{R}^{d \times d}.$$

More generally, the Stieltjes transform of a probability distribution $F(s)$ is defined as $m_F(-\lambda) = \int \frac{1}{s+\lambda} dF(s)$, and so the Stieltjes transform of a matrix is just the Stieltjes transform of its empirical eigenvalue distribution. Like characteristic functions, Stieltjes transforms are an alternative characterization of a probability measure.[2] Moreover, convergence in Stieltjes transform implies convergence in distribution—i.e. given a sequence of probability distributions $F_n(s)$ and some limiting distribution $F(s)$:

$$m_{F_n}(-\lambda) \to m_F(-\lambda) \quad \iff \quad F_n(s) \to F(s).$$

If we assume that there is limiting matrix $\boldsymbol{A}$ so that, for all $\lambda > 0$,

$$\tfrac{1}{d} \operatorname{Tr}\left(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}\right)^{-1} \to \tfrac{1}{d} \operatorname{Tr}\left(\boldsymbol{A} + \lambda \boldsymbol{I}\right)^{-1}, \tag{4}$$

(i.e. $m_{\hat{\boldsymbol{\Sigma}}}(-\lambda) \to m_{\boldsymbol{A}}(-\lambda)$), then it should also be true that, for all $i \in [1, n]$,

$$\tfrac{1}{d} \operatorname{Tr}\left(\left(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}\right) - \tfrac{1}{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \to \tfrac{1}{d} \operatorname{Tr}\left(\boldsymbol{A} + \lambda \boldsymbol{I}\right)^{-1}, \tag{5}$$

i.e. that $m_{\hat{\boldsymbol{\Sigma}}}(-\lambda) \to m_{\boldsymbol{A}}(-\lambda)$ even if we removed one $\boldsymbol{x}_i$ from the empirical covariance matrix, since each data point contributes infinitesimally to $\hat{\boldsymbol{\Sigma}}$ as $n, d \to \infty$. Following this argument[3], coupled with copious amounts of measure theory, the fact that both Eqs. (4) and (5) hold implies that $\lim_{n,d\to\infty} m_{\hat{\boldsymbol{\Sigma}}}(-\lambda)$ must adhere to the **self-consistency equation**:

$$\tfrac{1}{d} \operatorname{Tr}\Big(\hat{\boldsymbol{\Sigma}}\big(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}\big)^{-1}\Big) \;\longrightarrow\; \tfrac{1}{d} \operatorname{Tr}\Big(\boldsymbol{\Sigma}\big(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}\big)^{-1}\Big), \qquad 1/\kappa(\lambda) := \lim_{n,d\to\infty} \tfrac{1}{n} \operatorname{Tr}\Big(\big(\tfrac{1}{n} \boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}\big)^{-1}\Big), \tag{6}$$

where we recognize $\frac{1}{n} \operatorname{Tr}((\frac{1}{n} \boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I})^{-1})$ as the Steiltjes transform of the empirical distribution of the spectrum of $\frac{1}{n} \boldsymbol{X}\boldsymbol{X}^\top$.

> **Aside.** It is worth thinking about what $\lim_{d\to\infty} \frac{1}{d} \operatorname{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1})$ even means. $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is a fixed matrix, so what does it to consider a limit over $d$? There are two points worth considering:
>
> - Rigorously defining what $\lim_{d\to\infty} \frac{1}{d} \operatorname{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1})$ means depends on the particular problem setup. Most of the "work" in theory papers using random matrix theory involves setting up a mathematically valid interpretation Eq. (6). For the purposes of this course, we can just assume the hand-wavy interpretation that $\frac{1}{d} \operatorname{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}) \approx \frac{1}{d} \operatorname{Tr}(\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1})$ when $n, d$ are "sufficiently large."
>
> - As $d \to \infty$ and $\boldsymbol{\Sigma}$ becomes infinitely large, we can go back to our notion of kernels. If $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$, then $k(\boldsymbol{x}, \boldsymbol{x}') = \lim_{d\to\infty} \boldsymbol{x}^\top \boldsymbol{x}'$ remains valid and the spectrum of $k(\cdot, \cdot)$ matches the limiting spectrum of $\boldsymbol{\Sigma}$.

[2] To gain intuition for why the Steiltjes transform characterizes a probability measure, note that the Taylor expansion suggests that $\mathbb{E}_s[1/(s - \lambda)] = -\frac{1}{\lambda} \sum_{j=1}^{\infty} \mathbb{E}_s[(s/\lambda)^j]$ and thus the Steiltjes transform determines the distributions' moments.

[3] A hand-wavy version of this argument can be shown by applying the Woodbury formula to $((\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}) - \frac{1}{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top)^{-1}$ and recognizing that $\boldsymbol{x}_i^\top \boldsymbol{B} \boldsymbol{x}_i \approx \operatorname{Tr} \boldsymbol{B}$ for a matrix $\boldsymbol{B}$ that is independent of $\boldsymbol{x}_i$. See [Pedregosa et al., 2021, Part 2] for a simple introduction to the argument.

Using the Woodbury formula, the cyclic property of the trace, and rewriting $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$ we have that

$$\begin{aligned}\tfrac{1}{d}\left[\operatorname{Tr}(\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I})^{-1})\right] = \tfrac{1}{d}\left[\operatorname{Tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top)\right] &= \tfrac{1}{d}\left[\operatorname{Tr}(\boldsymbol{I}_{n\times n} - \lambda\operatorname{Tr}(\tfrac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top+\lambda\boldsymbol{I})\right] \\ &= \tfrac{n}{d}\left[1-\lambda\tfrac{1}{n}\operatorname{Tr}(\tfrac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top+\lambda\boldsymbol{I})\right] \\ &\to \gamma\left[1-\lambda/\kappa(\lambda)\right].\end{aligned}$$

Thus Eq. (6) can be rearranged to be written as

$$\underbrace{\gamma\,\frac{1}{d}\operatorname{Tr}\left(\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}+\kappa(\lambda)\boldsymbol{I}\right)^{-1}\right)}_{\frac{1}{n}\sum_{i=1}^{d}\frac{s_i}{s_i+\kappa(\lambda)}} + \frac{\lambda}{\kappa(\lambda)} = 1, \tag{7}$$

which is known as the **Silverstein equation** [Silverstein, 1995]. In the deep learning literature, this equation is often referred to as the implicit regularization equation [Jacot et al., 2020] for reasons that we will see in next lecture.

We now have a recipe to derive the limiting spectrum of $\hat{\boldsymbol{\Sigma}}$:

1. Given a $\boldsymbol{\Sigma}$ and a $\gamma$, solve for the $\kappa(\lambda)$ function that satisfies Eq. (7).
2. Recognizing from Eq. (6) that $1/\kappa(\lambda)$ is the limiting Steiltjes transform of the spectrum of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$, and that limiting Steiltjes transformations characterize limiting distributions, find the distribution $F(s)$ with a Steiltjes transform that matches $1/\kappa(\lambda)$.
3. $F(s)$ is the limiting distribution of the eigenvalues of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$, which is also the limiting distribution of the non-zero eigenvalues of $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \hat{\boldsymbol{\Sigma}}$.

When $\boldsymbol{\Sigma} = \boldsymbol{I}$, $\kappa(\lambda)$ in Eq. (7) admits a closed-form analytic solution, where $1/\kappa(\lambda)$ is exactly equal to the Steiltjes transformation of the Marchenko-Pastur distribution given in Eq. (3).

---

## 5) Strategy for Analyzing High-Dimensional Ridge Risk: Deterministic Equivalents

Unfortuantely, the recipe above will only work in very special cases, since most $\boldsymbol{\Sigma} \neq \boldsymbol{I}$ don't afford closed-form solutions to $\kappa(\lambda)$ in Eq. (7). Luckily, for our purposes, we won't need to actually solve for the limiting spectrum of $\hat{\boldsymbol{\Sigma}}$; we only care about specific *reductions* of the limiting spectrum. Recall that for the variance we need to compute $\operatorname{Tr}(\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}(\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I})^{-1})$. This computation does not require knowing the exact eigenvalues of $\hat{\boldsymbol{\Sigma}}$, but instead requires knowing the summation (trace) of some function of the eigenvalues.

As a starting point, we know an asymptotic limit of the reduction $\operatorname{Tr}(\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I})^{-1})$; it is approximately equal to $\operatorname{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\kappa(\lambda)\boldsymbol{I})^{-1})$. By adding and subtracting $\boldsymbol{I}$ from both sides of Eq. (6), we have that

$$\lambda\operatorname{Tr}\Big(\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I}\Big)^{-1} \;\approx\; \kappa(\lambda)\operatorname{Tr}\Big(\boldsymbol{\Sigma}+\kappa(\lambda)\boldsymbol{I}\Big)^{-1}.$$

More rigorously, we have that

$$\frac{1}{d}\left[\lambda\operatorname{Tr}\Big(\boldsymbol{B}\big(\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I}\big)^{-1}\Big) - \kappa(\lambda)\operatorname{Tr}\Big(\boldsymbol{B}\big(\boldsymbol{\Sigma}+\kappa(\lambda)\boldsymbol{I}\big)^{-1}\Big)\right] \longrightarrow 0, \tag{8}$$

for any $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ under certain regularity conditions [Rubio and Mestre, 2011]. We will denote this asymptotic convergence by the symbol

$$\lambda\Big(\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I}\Big)^{-1} \asymp \kappa(\lambda)\Big(\boldsymbol{\Sigma}+\kappa(\lambda)\boldsymbol{I}\Big)^{-1} \quad\iff\quad \text{Eq. (8) holds for all "regular" } \boldsymbol{B} \in \mathbb{R}^{d\times d}.$$

Eq. (8) gives us a powerful mechanism for analyzing equations involving random matrices. Every time we come across a term that looks like $\lambda \operatorname{Tr}(\boldsymbol{B}(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1})$, we can "swap it out"[4] with its **deterministic equivalent** $\kappa(\lambda) \operatorname{Tr}(\boldsymbol{B}(\boldsymbol{\Sigma} + \kappa(\lambda)\boldsymbol{I})^{-1})$. Once we have arrived at a final expression that involves only deterministic matrices, we can then think about bounding (or numerically solving) for $\kappa(\lambda)$.

We will apply this strategy to the high-dimensional ridge risk in the next lecture.

---

## References


A. Jacot, B. Simsek, F. Spadaro, C. Hongler, and F. Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pages 4631–4640, 2020.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

F. Pedregosa, C. Paquette, T. Trogdon, and J. Pennington. Random matrix theory and machine learning tutorial, 2021. URL `https://random-matrix-learning.github.io/`.

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

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

R. Tibishriani. High-dimensional regression: Ridge, 2023. URL `https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/ridge.pdf`.


[4]For mathematically rigorous results, this "swap" requires appropriate conditions and analysis of limits. However, in this course we will assume that there is some reasonable notion of approximation that allows us to swap these terms.