Lecture 05: Risk of Overparameterized Ridge Regression, Effective Regularization

Geoff Pleiss

Recall that our goal is to find a (high-dimensional) asymptotic expression for the risk of ridge regression. The previous lecture introduced tools from random matrix theory which we now use to analyze the risk.

---

1) The Risk of Ridge Regression

Using the same notation as the previous lecture, recall that the risk of ridge regression decomposes into squared bias and variance terms:

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

And recall that the tools of random matrix theory gave us **deterministic equivalents** for expressions of the form $\operatorname{Tr}(\boldsymbol{B}(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1})$ for all $\boldsymbol{B}$ satisfying certain regularity conditions:

$$\lambda \operatorname{Tr}\left(\boldsymbol{B}(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \right) \approx \kappa(\lambda) \operatorname{Tr}\left(\boldsymbol{B}(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I})^{-1} \right), \tag{2}$$

where $\approx$ denotes some notion of convergence as $n, d \to \infty$ (which, for the purposes of this class, we will not rigorously define), and $\kappa(\lambda)$ is the limiting Steiltjes transform of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ which happens to be the solution to the self-consistency/Silverstein equation:

$$\underbrace{\gamma \frac{1}{d} \operatorname{Tr}\left(\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma} + \kappa(\lambda) \boldsymbol{I}\right)^{-1} \right)}_{\frac{1}{n}\sum_{i=1}^d \frac{s_i}{s_i + \kappa(\lambda)}} + \frac{\lambda}{\kappa(\lambda)} = 1, \qquad \kappa(\lambda) = \lim_{n,d\to\infty} \frac{1}{n} \operatorname{Tr}\Big(\left(\tfrac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}\right)^{-1} \Big). \tag{3}$$

---

2) Applying the Deterministic Equivalent

Note that the deterministic equivalent only applies to terms of the form $\operatorname{Tr}(\boldsymbol{\Sigma}(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1})$, and not other terms like $\operatorname{Tr}(\boldsymbol{\Sigma}(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1})$—what we see in the variance expression. So we have to massage the terms in Eq. (1) a bit.

2.1 Variance

A key insight from Hastie et al. [2022] is to use the following identity:

$$\operatorname{Tr}\left[\left(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{\Sigma} \left(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}\right)^{-1} \right] = -\frac{d}{d\lambda} \left\{ \operatorname{Tr}\left[\boldsymbol{\Sigma} \left(\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}\right)^{-1} \right] \right\} \tag{4}$$

where now the inside of the derivative can be replaced with its deterministic equivalent (after CAREFULLY checking that derivative and limit can be interchanged.[1])

$$
\begin{aligned}
&\operatorname{Tr}\left[\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\right] \\
&=\operatorname{Tr}\left[\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}-\lambda\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\right] \\
&=\operatorname{Tr}\left[\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\right]+\lambda\frac{d}{d\lambda}\left\{\operatorname{Tr}\left[\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\right]\right\}
\end{aligned}
$$

Plugging in the deterministic equivalents from Eq. (2) (where we go on faith that we can interchange limits and derivatives), and recalling Eq. (3), we have

$$
\begin{aligned}
\mathcal{V}(\hat{\boldsymbol{\theta}}_\lambda) &= \frac{\sigma^2}{n}\mathbb{E}\operatorname{Tr}\left[\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\right] \\
&= \frac{\sigma^2}{n}\mathbb{E}\left(\operatorname{Tr}\left[\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\right]+\lambda\frac{d}{d\lambda}\left\{\operatorname{Tr}\left[\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\right]\right\}\right) \\
&\approx \frac{\sigma^2}{n}\mathbb{E}\left(\frac{\kappa(\lambda)}{\lambda}\operatorname{Tr}\left[\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}+\kappa(\lambda)\boldsymbol{I}\right)^{-1}\right]+\lambda\frac{d}{d\lambda}\left\{\frac{\kappa(\lambda)}{\lambda}\operatorname{Tr}\left[\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}+\kappa(\lambda)\boldsymbol{I}\right)^{-1}\right]\right\}\right) \\
&= \sigma^2\cancel{\left(\frac{1}{n}\frac{d}{\gamma}\right)}\left[\left(\frac{\kappa(\lambda)}{\lambda}\left(1-\frac{\lambda}{\kappa(\lambda)}\right)\right)+\lambda\frac{d}{d\lambda}\left\{\frac{\kappa(\lambda)}{\lambda}\left(1-\frac{\lambda}{\kappa(\lambda)}\right)\right\}\right] \\
&= \sigma^2\left[\left(\frac{\kappa(\lambda)}{\lambda}-1\right)+\lambda\left(\frac{1}{\lambda}\frac{d\kappa(\lambda)}{d\lambda}-\frac{\kappa(\lambda)}{\lambda^2}\right)\right] \\
&= \sigma^2\left[\frac{d\kappa(\lambda)}{d\lambda}-1\right].
\end{aligned}
$$

2.2 Bias

We can play a similar trick with the bias term, but it's a little more complicated:

1. we want the λ^2 term in the bias to "disappear" (i.e. to be "transformed" into a $\kappa(\lambda)$ as part of the deterministic equivalent) and
2. we have $\boldsymbol{\theta}^{*\top}$ terms in the expression

If we introduce an auxiliary variable ρ so that

$$
\lambda^2\operatorname{Tr}\left[\boldsymbol{\theta}^*\boldsymbol{\theta}^{*\top}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}\right)^{-1}\right] = -\frac{d}{d\rho}\left\{\operatorname{Tr}\left[\lambda\boldsymbol{\theta}^*\boldsymbol{\theta}^{*\top}\left(\hat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I}+\rho\lambda\boldsymbol{\Sigma}\right)^{-1}\right]\right\}\Bigg|_{\rho=0}.
$$

then the term inside the derivative can be massaged into a form that admits a deterministic equivalent,[2] ultimately yielding the following approximation:

$$
\mathcal{B}(\hat{\boldsymbol{\theta}}_\lambda) \approx \left(\frac{d\kappa(\lambda)}{d\lambda}\right)\underbrace{\kappa(\lambda)^2\boldsymbol{\theta}^{*\top}\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{\Sigma}+\kappa(\lambda)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^*}_{c^2}.
$$

[1] See [Hastie et al., 2022] for all the gory technical details.
[2] See [Hastie et al., 2022] for gory details or [Tibshirani, 2023] for a more palatable introduction.

Letting $\boldsymbol{Q}\boldsymbol{S}\boldsymbol{Q}^\top$ be the eigendecomposition of $\boldsymbol{\Sigma}$, let $\boldsymbol{v} = \boldsymbol{S}^{1/2}\boldsymbol{Q}^\top\boldsymbol{\theta}^*$ be the coordinates of $\boldsymbol{\theta}^*$ in the eigenbasis of $\boldsymbol{\Sigma}$ (scaled by the square root of the eigenvalues). Then we can rewrite c^2 as

$$c^2 = \sum_{i=1}^{d} \left(\kappa^2 - \frac{1}{(s_i+\kappa)^2}\right) v_i^2 = \sum_{i=1}^{d} \left(1 - \mathcal{L}_i^2\right) v_i^2, \qquad \mathcal{L}_i := \frac{s_i}{s_i + \kappa(\lambda)}.$$

---

3) Interpretation

Putting these results together, we have that

$$\mathcal{R}(\hat{\boldsymbol{\theta}}_\lambda) = \underbrace{\left(\frac{d\kappa(\lambda)}{d\lambda}\right)}_{\mathcal{E}_0} \left(\overbrace{\sigma^2}^{\text{noise fit}} + \overbrace{\sum_{i=1}^{d} (1-\mathcal{L}_i)^2 v_i^2}^{\text{signal residual}} \right) - \sigma^2. \tag{5}$$

While we have thus far assumed that we are in the ridge scenario ($\lambda > 0$), these results also hold in the ridgeless case (i.e. as we take $\lambda \to 0$).[3] Importantly, $\lambda \to 0$ does not imply that $\kappa(\lambda) \to 0$.

Simon et al. [2023] offers an semantically meaningful interpretations of all these terms.

- v_i is i^{th} *eigenmode coefficient* of $\boldsymbol{\theta}^*$; i.e. the portion of the true signal that aligns with the i^{th} eigenvector of $\boldsymbol{\Sigma}$.
- $\mathcal{L}_i = (s_i)/(s_i + \kappa(\lambda))$ is the *learnability* of the i^{th} eigenmode. It corresponds to the (square root of) the percentage of "signal" in the i^{th} eigenmode that can be learned by the model.
  - Note that, $\sum_{i=1}^d \mathcal{L}_i = \mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \kappa(\lambda))^{-1})$. Thus, by Eq. (3), $\sum_{i=1}^d \mathcal{L}_i \leq n$, with equality in the ridgeless case.
  - In other words, the total learnability of all d eigenmodes is *fixed* at n. We cannot hope to learn all eigenmodes completely when $n < d$.
- The *signal residual* represents the true signal not learned by the ridge parameters.
- The *noise fit* term represents the training response noise that "ends up in" the ridge parameters.
- Finally, $\mathcal{E}_0$ is the *overfitting coefficient*. It is a multiplicative penalty that increases the test error.
  - By differentiating the Silverstein equation in Eq. (3) with respect to λ on both sides:

    $$\frac{d}{d\lambda}\left\{ \kappa(\lambda) \frac{1}{n} \sum_{i=1}^{d} \underbrace{\left(\frac{s_i}{s_i + \kappa(\lambda)}\right)}_{\mathcal{L}_i} + \lambda \right\} = \frac{d}{d\lambda}\left\{\kappa(\lambda)\right\}$$

    and rearranging terms we find that

    $$\mathcal{E}_0 = \frac{d\kappa(\lambda)}{d\lambda} = \frac{n}{n - \sum_{i=1}^d \mathcal{L}_i^2}. \tag{6}$$

  - We have that $\mathcal{E}_0 \to 1$ as $\kappa(\lambda) \to \infty$ (and thus $\mathcal{L}_i \to 0$) and $\mathcal{E}_0 \to \infty$ as $\kappa(\lambda) \to 0$ (and thus $\mathcal{L}_i \to 1$).

[3] Showing that these results hold in the ridgeless case requires a very careful analysis of the limits, which we will ignore for the purposes of this course.

3.1 Implicit Regularization

We will refer to $\kappa(\lambda)$ as the **implicit regularization coefficient**. Much as explicit regularization reduces overfitting/variance at the cost of increased bias, larger $\kappa(\lambda)$ will reduce the overfitting coefficient at the cost of increased signal residual.

$\kappa(\lambda)$ only appears in Eq. (5) through the eigenmode learnabilities $\mathcal{L}_i$ and the only two terms containing $\mathcal{L}_i$ are the overfitting coefficient $\mathcal{E}_0$ and the "signal residual" $\sum_{i=1}^d (1-\mathcal{L}_i)^2 v_i$. To understand how $\kappa(\lambda)$ impacts both of these terms, note that $0 \le \mathcal{L}_i \le 1$ when $\kappa(\lambda) \ge 0$ and thus

$$\textstyle\sum_{i=1}^d \mathcal{L}_i^2 \le \sum_{i=1}^d \mathcal{L}_i \le n.$$

where the first inequality is strict unless $\mathcal{L}_i = 1$. If $\kappa(\lambda) \approx 0$, $\mathcal{L}_i^2$ will go towards 1 and—in the ridgeless case—$\sum \mathcal{L}_i^2 \to n$. The signal residual will be approximately zero but the overfitting coefficient will diverge. Conversely if $\kappa(\lambda)$ is very large, $\mathcal{L}_i^2$ will shrink towards 0 and $\sum \mathcal{L}_i^2 \ll n$. The signal residual will be large but the overfitting coefficient will be nearly 1.

3.2 How $\gamma = d/n$ Affects the Implicit Regularization

In the next lecture we will gain a better sense for how $\kappa(\lambda)$ is affected by the spectrum $\lambda_1, \ldots, \lambda_d$. For now, let's consider some basic rules that depend on γ. Assume that we are in the ridgeless case, so that

$$\textstyle\sum_{i=1}^d s_i/(s_i + \kappa(0)) = n.$$

- In the overparameterized regime ($d > n$, $\gamma > 1$), there are d terms in the summation that must add up to n. Therefore, each term in the summation must be < 1 and so $\kappa(0) > 0$. As $\gamma = d/n$ increases, each term must contribute less to the overall sum, and thus $\kappa(0)$ must increase.
- In the underparameterized regime $d < n$, $\gamma < 1$), we have the opposite scenario. Each of the d terms in the summation must be > 1 to add up to n, so $\kappa(0) < 0$, becoming more negative as $\gamma \to 0$. Curiously we have *negative implicit regularization*!
- When $n = d$, each term in the summation must be exactly 1, implying that $\kappa(0) = 0$. As discussed above, the overfitting coefficient diverges, resulting in *infinite risk!*

These results explain the double descent curve we saw earlier, but they also hint at a potentially troubling scenario. Let's say that we are working in an RKHS—i.e. with $d = \infty$, and we are training an (overparameterized) ridgeless regressor with n data points. As $n \to \infty$, we have $\gamma \to 1$, decreasing our effective regularization. However, as $\gamma \to 1$, we have that $\kappa(0) = 0$ which potentially brings infinite risk. With careful analysis of the limits, we will show that—surprisingly—we often avoid this catastrophic behaviour.

References

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949, 2022.

J. B. Simon, M. Dickens, D. Karkada, and M. R. DeWeese. The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks. *Transactions on Machine Learning Research*, 2023.

R. Tibshirani. Overparameterized regression: Ridgeless interpolation, 2023. URL `https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/ridgeless.pdf`.