Lecture 06: Benign Overfitting

Geoff Pleiss

In the last lecture we saw that risk of ridgeless linear regression when $\gamma = d/n \geq 1$ is given by

$$\mathcal{R}(\hat{\boldsymbol{\theta}}_0) \;\approx\; \overbrace{\underbrace{\mathcal{E}_0 \left(\textstyle\sum_i (1 - \mathcal{L}_i)^2 \, v_i^2 \right)}_{\text{signal residual}}}^{\text{bias}^2} \;+\; \overbrace{(\mathcal{E}_0 - 1) \underbrace{\sigma^2}_{\text{noise fit}}}^{\text{variance}}, \qquad \mathcal{E}_0 := \frac{n}{\underbrace{n - \sum_i \mathcal{L}_i^2}_{\text{overfit. coeff.}}}. \tag{1}$$

Given the eigendecomposition $\boldsymbol{Q}\boldsymbol{S}\boldsymbol{Q}^\top = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$, the quantities v_i are given by the equation $\boldsymbol{\theta}^* = \sum_{i=1}^d v_i s_i^{-1/2} \boldsymbol{q}_i$, and the $\mathcal{L}_i$ are the eigenvalues s_i weighted by the *implicit reguarlization* parameter κ.

$$\mathcal{L}_i = \frac{s_i}{s_i + \lambda}, \qquad \kappa : \left[\sum_{i=1}^d \mathcal{L}_i = \sum_{i=1}^d \frac{s_i}{s_i + \kappa} = n \right] \tag{2}$$

We went through the high level steps of a high-dimensional asymptotic proof, where $\approx$ signified that the difference between the two quantities goes to zero almost surely as $n, d \to \infty$, (assuming that there is some sequence of Σ_d that converge in Steiltjes transform). However, Eq. (1) holds for various other notions of $\approx$, including high probability bounds for finite n and d [Bartlett et al., 2020].

Though the implicit regularization parameter is implicitly defined by the self-consistency equation (2), we saw that $\kappa > 0$ when $\gamma = d/n > 1$ and $\kappa = 0$ when $\gamma = 1$. Moreover,

- $\mathcal{E}_0 \geq 1$ when $\gamma \geq 1$, increasing as γ increases.
- $\mathcal{E}_0 \to \infty$ as $\kappa \to 0$, suggesting *infinite risk* at the interpolation threshold.

---

1) Can Interpolation Be Consistent?

Eq. (1) helps to explain the double descent curve (we gain "implicit regularization" with increased overparameterization), but it has troubling implications for the consistency of interpolating models.

Consider an RKHS $\mathcal{H}$ with a kernel $k(\boldsymbol{x}, \boldsymbol{x}')$. As we discussed in a previous lecture, we can think of kernel regression as linear regression with infinitly many features, where the features come from (a linear combination of) the eigenepansion $k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^\infty s_i \boldsymbol{q}_i(\boldsymbol{x}) \boldsymbol{q}_i(\boldsymbol{x}')$. If the true data generating function $f^*(\boldsymbol{x})$ lives in $\mathcal{H}$ (i.e. if $y = f^*(\boldsymbol{x}) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$) then we would hope that any learning algorithm would produce a **consistent estimator** $\hat{f}$ where $\hat{f} \to f^*(\boldsymbol{x})$ (almost surely) as $n \to \infty$.

If $\hat{f}$ were the ridgeless kernel regressor (read: the ridgeless linear regressor with infinitly many features), its risk would be approximated by Eq. (1) with $d = \infty$.[1] then we would hope that $\mathcal{R}(\hat{f}) \to \mathcal{R}(f^*)$ as $n \to \infty$. However, it is not immediately obvious how to achieve consistency with Eq. (1).

In order for $\mathcal{R}(\hat{f}) \to \mathcal{R}(f^*) = 0$ as $n \to \infty$, we would need both the bias and variance terms to go to zero. It's easy to see that the bias term vanishes, even without the asymptotic form of Eq. (1). Imagine for

[1] We have to be careful with what we mean by "approximated" here. Recall that we derived Eq. (1) by taking $n, d \to \infty$ *simultaneously*, so it doesn't make sense to set one of them to infinity while keeping the other finite. Nevertheless, alternative analyses of $\mathcal{R}(\hat{\boldsymbol{\theta}}_0)$ that set $d \to \infty$ and keeps n finite arrive at the same equation but where the $\approx$ has different meaning.

starters that $d > n$ is finite. Using the notation from previous lectures, the bias term of overparameterized ridgeless regression is equal to

$$\mathcal{B}(\hat{\boldsymbol{\theta}}_0) = \mathbb{E}\Big[\Big(\boldsymbol{x}^\top \overbrace{(\boldsymbol{\theta}^* - \mathbb{E}[\hat{\boldsymbol{\theta}}_0])}^{\boldsymbol{a}}\Big)^2\Big] = \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a},$$

and recalling that

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_0] = \mathbb{E}[\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1} \overbrace{\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}]}^{\boldsymbol{X}\boldsymbol{\theta}}],$$

we have that $\boldsymbol{a} = \mathbb{E}[\boldsymbol{I} - \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}]\boldsymbol{\theta}^*$. The matrix inside the expectation is an orthogonal projection onto the nullspace of $\boldsymbol{X}$. As $n \to d$, the nullspace vanishes and so $\boldsymbol{a} \to 0$. This same logic holds if $d = \infty$ (i.e. if we are working with kernels); the bias term will vanish as $n \to \infty$.

The variance term is more complicated. In order for the variance to vanish, we would need $\mathcal{E}_0 \to 1$ as $n \to \infty$. However, $\kappa = 0$ when $n = d$ in our high-dimensional asymptotic analysis, and so $\mathcal{E}_0 = \infty$ when $\kappa = 0$! Indeed, this equation depicts why statisticians historically thought that interpolating estimators should be avoided at all cost. In practice, it is not challenging to construct a kernel interpolator with this diverging risk. Fig. 1 shows a scenarios (right) where $\mathcal{E}_0 \to \infty$ as $n \to \infty$; i.e. adding more data actually makes the generalization worse!

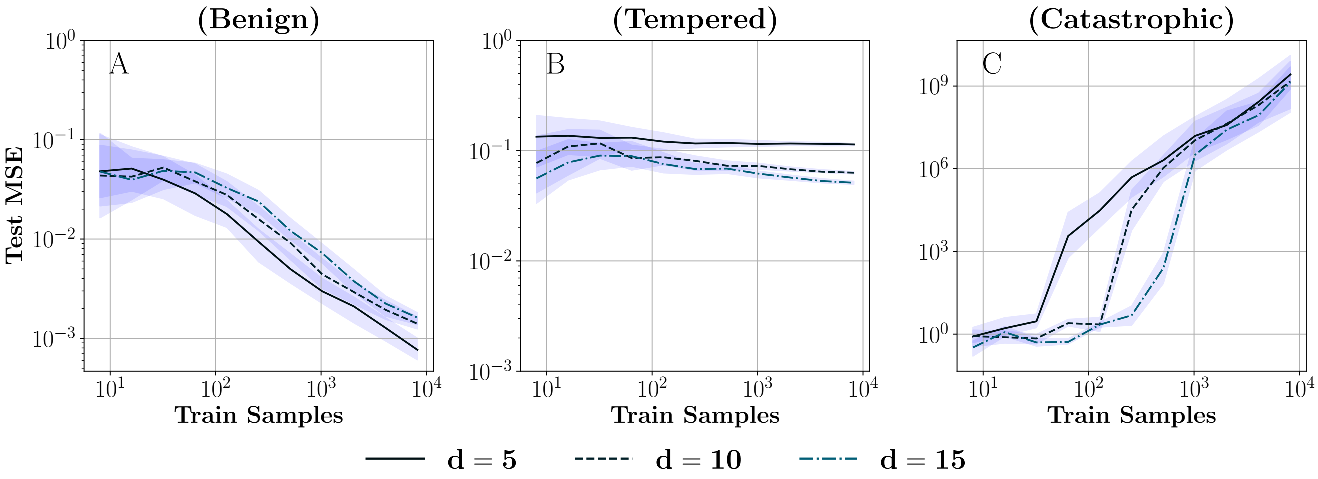

Figure 1: A depiction of the three limiting scenarios for the risk of kernel interpolators. **Left: benign overfitting**: $\mathcal{R} \to 0$ as $n \to \infty$. **Middle: tempered overfitting**: $\mathcal{R} \to c$ for some $0c < \infty$. **Right: catastrophic overfitting**: $\mathcal{R} \to \infty$. All three scenarios are possible depending on the spectrum of the kernel function. (Figure reproduced from Mallinar et al. [2022].)

Our only hope for a consistent (or, at the very least, a non-catastrophic) estimator is for κ to decay at a slower rate than n grows. More specifically, we need the gap between $\sum \mathcal{L}_i^2$ and $\sum \mathcal{L}_i = n$, the two terms in the denominator of $\mathcal{E}_0$ to increase (or at least stay constant) as $n \to \infty$. Surprisingly, we will manage this rate for for most kernels! In certain scenarios, $\mathcal{E}_0$ not only remains finite as $n \to 0$, but also converges to 1 and thus $\lim_{n\to\infty} \mathcal{R}(\hat{\boldsymbol{\theta}}_0) = 0$. The discovery of interpolators that **benignly overfit** despite memorizing the noise present in the training data is one of the more surprising statistical findings of the last 5 years [Bartlett et al., 2020].

2) Strategy and Mathematical Tools

Our derivation will largely follow that of Mallinar et al. [2022], though it is worth noting that their analysis is non-rigorous (see Footnote 1 for an explanation). Bartlett et al. [2020] provides a more complicated analysis that maintains rigor by avoiding asymptotics. Both analyses largely use the same mathematical idea which we outline below.

The convergence (or divergence) of Eq. (1) will entirely depend on the limiting behaviour of $\mathcal{L}_i$. Our key technique will be to divide the $\mathcal{L}_i$ into a "head" and a "tail," bounding the "head" behaviour while controlling the "tail." For example, consider the self-consistency equation for kernel ridgeless regression:

$$\sum_{i=1}^{\infty} \mathcal{L}_i = \sum_{i=1}^{\infty} \frac{s_i}{s_i + \kappa} = n.$$

If we choose some constant $\zeta \in \mathbb{N}$, then we can divide this sum into

$$\underbrace{\sum_{i=1}^{\zeta} \frac{s_i}{s_i + \kappa}}_{\text{head}} + \underbrace{\sum_{i>\zeta} \frac{s_i}{s_i + \kappa}}_{\text{tail}} = n. \tag{3}$$

It's not immediately obvious why this separation is useful. However, assuming that $s_1 > s_2 > \ldots$, and noting that $s_i/(s_i + \kappa) \leq 1$ and $s_i/(s_i + \kappa) < s_i/\kappa$, we can obtain a simple upper bound on κ that depends on η: and defining $\zeta = n(1 - \eta)$ for some[2] $\eta \in (0, 1)$

$$\underbrace{\sum_{i=1}^{n(1-\eta)} \frac{s_i}{s_i + \kappa}}_{\leq n(1-\eta)} + \underbrace{\sum_{i>n(1-\eta)} \frac{s_i}{s_i + \kappa}}_{<\sum s_i/\kappa} = n. \quad \Longrightarrow \quad \kappa < \frac{1}{n\eta} \underbrace{\sum_{i>n(1-\eta)} s_i}_{:=c_\eta}. \tag{4}$$

Thus κ decays at a rate of $O(1/n)$. We can also provide an lower bound:

$$n = \sum_{i=1}^{n(1+\eta)} \underbrace{\frac{s_i}{s_i + \kappa}}_{\geq \frac{s_{n(1+\eta)}}{s_{n(1+\eta)}+\kappa}} + \underbrace{\sum_{i>n(1+\eta)} \frac{s_i}{s_i + \kappa}}_{>0} > n(1+\eta)\frac{s_{n(1+\eta)}}{s_{n(1+\eta)} + \kappa} \quad \Longrightarrow \quad \kappa > \eta s_{n(1+\eta)}. \tag{5}$$

In other words, the split in Eq. (3) allows us to compute *rates of convergence* for various terms in Eq. (1). We will get even more precise rates once we start considering how fast the s_i eigenvalues decay and consider specific values of ζ or η.

---

3) The Curious Case of Benign Overfitting

Let's assume that the rate of decay of the eigenvalues s_i is given by

$$s_i = i^{-1} \log^{-\alpha} i \tag{6}$$

for some $\alpha > 0$. Note that this is just about the slowest rate of eigenvalue decay that we can have while still having $\sum_{i=1}^{\infty} s_i < \infty$.[3] If we consider $\zeta = n/\log\log n$ in Eq. (3), then we have

$$\sum_{i=1}^{n} \mathcal{L}_i^2 = \sum_{i=1}^{\frac{n}{\log\log n}} \underbrace{\frac{s_i^2}{(s_i + \kappa)^2}}_{\leq 1} + \sum_{i>\frac{n}{\log\log n}} \underbrace{\frac{s_i^2}{(s_i + \kappa)^2}}_{<\sum s_i^2/\kappa^2} = \frac{n}{\log\log n} + \underbrace{\left(\kappa^{-2} \sum_{i>\frac{n}{\log\log n}} \frac{1}{i^2 \log^{2\alpha} i}\right)}_{\text{tail}}.$$

[2] Mallinar et al. [2022] use γ rather than η for this constant; here we use η to avoid confusion with $\gamma = d/n$.
[3] Note that $\lim_{\alpha \to 0} \sum_{i=1}^{\infty} i^{-1} \log^{-\alpha} i = \sum_{i=1}^{\infty} 1/i = \infty$.

Using the lower bound in Eq. (5) with $\eta = 1$ (so that $\kappa \geq \lambda_{2n} = (2n)^{-1} \log^{-\alpha}(2n)$) in conjunction with some additional clever bounding, Mallinar et al. [2022, Appx. A] show that the tail in the sum above goes to zero as $n \to \infty$. Thus,

$$\mathcal{E}_0 = \frac{n}{n - O\left(\frac{n}{\log \log n}\right)} \to 1.$$

Thus the variance term in Eq. (1) goes to zero as $n \to \infty$, and so $\mathcal{R}(\hat{\boldsymbol{\theta}}_0) \to 0$. Magically, we have achieved the balance of $\mathcal{E}_0 \to 1$ despite the former blowing up when $\kappa = 0$.

3.1 Intuition

How does it happen that $\kappa = 0$ when $n = d$ but $\mathcal{E}_0 \to 1$ when $d = \infty$ and $n \to \infty$? You should be somewhat skeptical of this result, since we are applying a limiting analysis to quantities that are asymptotic in nature. Fortunately Bartlett et al. [2020] provide a more rigorous analysis that avoids asymptotics, deriving instead results that hold for finite n with high probability. There analysis finds that

$$\mathcal{V}(\hat{\boldsymbol{\theta}}_0) \asymp \frac{k_n^*}{n} - \frac{n}{R_n^{(k^*)}(\Sigma)}, \tag{7}$$

where k_n^* is a notion of **effective dimensionality** of the problem[4] and $R_n^{(k^*)}(\Sigma)$ is a measure of the **effective rank** of the tail of the covariance matrix Σ. Intuitively, we can think of k_n^* as the number of "relevant" (high eigenvalue) features there are while $R_n^{(k^*)}(\Sigma)$ represents the relative strength of the "irrelevant" (low eigenvalue) features. Roughly, we can think that the largest k_n^* eigenvalues of Σ are mostly fitting signal and the noise is distributed amongst the remaining smaller eigenvalues. Crucially, both numbers depend on n—the amount of data we have limits the amount of signal we can learn. With more data we fit more relevant features, leaving fewer irrelevant features to distribute the noise amongst.

Staring at this equation we see that benign overfitting requires a very delicate balance. We need $k_n^* \leq o(n)$—i.e. we need the effective dimensionality to be less than the true dimensionality. However, we also need $R_n^{(k^*)}(\Sigma) \geq \omega(n)$—i.e. we need the relative strength of the "irrelevant" features to become larger. Intuitively, we want to "spread out the noise" as much as possible over this tail so that it does not concentrate in any single feature. If the eigenvalues in this tail decay too quickly, then the noise will concentrate in a few relatively important dimensions increasing the influence of this noise.

Peter Bartlett, the lead author of Bartlett et al. [2020], outlines this intuition very well in a lecture from NeurIPS 2021 Bartlett [2021].

3.2 Other Mechanisms for Achieving Benign Overfitting

The decay rate of $s^{-1} \log^{-\alpha} s$ that yields benign overfitting is a very slow rate of decay; it is the slowest rate of decay that still allows for $\sum_{i=1}^{\infty} s_i < \infty$. Unfortunately as we will soon see it is the only rate of decay for fixed $d = \infty$ that allows for interpolating benign overfitting.

However, with an infinitessimal amount of ridge regularization, we can also achieve benign overfitting. If

[4] There is a relation between k_n^* and κ; see [Misiakiewicz and Montanari, 2023, Ch 2.4] for details.

the amount of ridge λ shrinks with n the bias term will still vanish. Moreover:

$$\begin{aligned}
\sum_i \mathcal{L}_i^2 &= \sum_i \frac{s_i^2}{(s_i+\kappa(\lambda))^2} \\
&< \sum_i \frac{s_i^2}{(s_i+\lambda)^2} && (\kappa(\lambda) > \lambda, \text{ from the previous lecture}) \\
&= \sum_i^{n^{1/2}} \underbrace{\frac{s_i^2}{(s_i+\lambda)^2}}_{<1} + \sum_{i>n^{1/2}} \underbrace{\frac{s_i^2}{(s_i+\lambda)^2}}_{<\frac{s_i^2}{\lambda^2}} \\
&= n^{1/2} + \sum_{i>n^{1/2}} \underbrace{\frac{s_i^2}{\lambda^2}}_{<s_n^{1/2}\frac{s_i}{\lambda^2}} = n^{1/2} + \frac{s_{n^{1/2}}}{\lambda^2} \underbrace{\sum_{i>n^{1/2}}}_{c} s_i,
\end{aligned}$$

where $c < \infty$ because the eigenvalues are summable. Setting $\lambda = \sqrt{s_{n^{1/2}}}$ thus allows λ to decay with n in a way that $\sum_i \mathcal{L}_i^2 < o(n)$ and thus $\mathcal{E}_0$ and the variance go to zero.

---

4) Tempered and Catastrophic Overfitting

Most kernel interpolators however will not be consistent estimators. Using similar techniques as what we used above, Mallinar et al. [2022] show that

- If the kernel eigenvalues decay at a **power law** rate of $s_i = i^{-\alpha}$ for some $1 < \alpha < \infty$, then $\mathcal{E}_0 \to \alpha$. The bias will disappear, but the variance will converge to $(\alpha - 1)\sigma^2$. We refer to this non-zero but non-infinite limiting risk as **tempered overfitting**. Note that most kernels have spectra that fall in this category. See Fig. 1 (middle) for a visual depiction of this scenario.
- If the kernel eigenvalues decay exponentially, such as $s_i = i^{-\log i}$, then $\mathcal{E}_0 \to \infty$. See Fig. 1 (right) for a visual depiction of this scenario. This scenario, which is the classic interpolation boogeyman, requires extremely strong assumptions of smoothness.

---

5) How Smoothness Affects Overfitting Behaviour

There is a very nice relation between the smoothness implied by the spectral decays and their overfitting behaviours. Intuitively, if we are interpolating a dataset with an infinite amount of noise, then it is very likely that we will get two data points with opposite noise profiles near each other. I.e., for a given input $\boldsymbol{x}$, as $n \to \infty$ we will likely have the responses $\boldsymbol{\theta}^*(\boldsymbol{x} - \boldsymbol{\delta}) + \epsilon$ and $\boldsymbol{\theta}^*(\boldsymbol{x} + \boldsymbol{\delta}) - \epsilon$ in the dataset for some $\boldsymbol{\delta}$ and ϵ. If we have to interpolate both of these points with a smooth function (rapidly decaying eigenvalues), there are lots of constraints on how fast our interpolating function can change. To compensate, our interpolator will have to have strange wiggles to maintain smoothness. Conversely, if we have slow decaying eigenvalues (less smoothness), then we can learn a "rough" bump at $(\boldsymbol{x} - \boldsymbol{\delta})$ and $(\boldsymbol{x} + \boldsymbol{\delta})$ to fit the noise while still learning the overall signal. This intuition is depicted in Fig. 2.

With that, we have finished our study of high-dimensional linear regression. We have learned that it is not as scary as it sounds; overparameterization adds a surprising implicit regularization effect and we can even obtain a consistent estimator while interpolating noise! In the next module, we will relate these results to

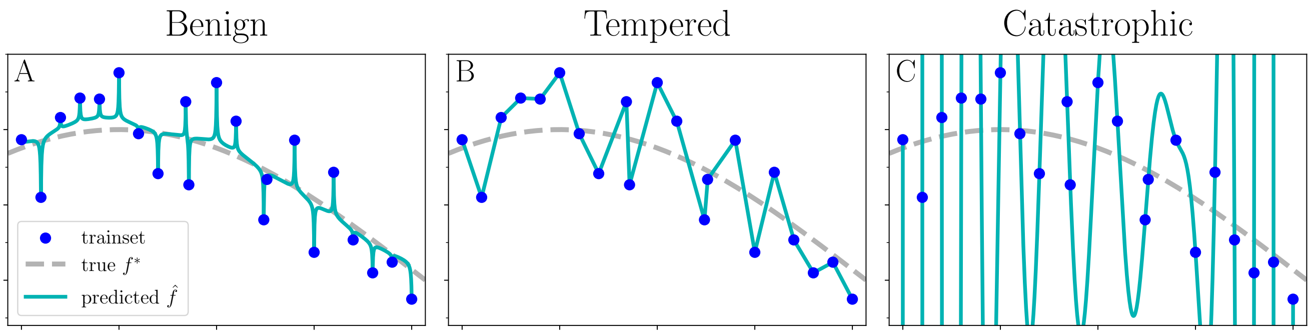

Figure 2: A cartoon depiction of how the smoothness of the kernel function affects the overfitting behaviour. (Figure reproduced from Mallinar et al. [2022].)

neural networks, using these findings as a first-order approximation for why neural networks fit data so well.

References

P. L. Bartlett. Benign overfitting. Posner Lecture at Advances in Neural Information Processing Systems, 2021. URL `https://www.youtube.com/watch?v=zNGm0iXWxYE&ab_channel=ArtificialIntelligence`.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

N. Mallinar, J. Simon, A. Abedsoltan, P. Pandit, M. Belkin, and P. Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. *Advances in Neural Information Processing Systems*, 35:1182–1195, 2022.

T. Misiakiewicz and A. Montanari. Six lectures on linearized neural networks. *arXiv preprint arXiv:2308.13431*, 2023.