Memorize to Generalize

https://arxiv.org/abs/2202.09889

Benign Overfitting

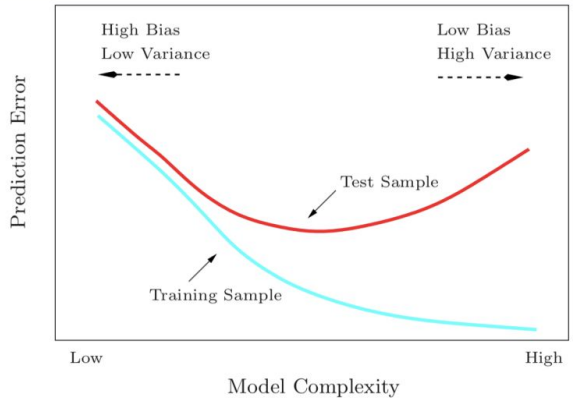
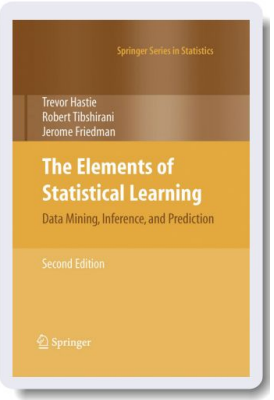


FIGURE 2.11. *Test and training error as a function of model complexity.*

Figure 2.11 shows the typical behavior of the test and training error, as model complexity is varied. The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder. However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error). In

"... interpolating fits... [are] unlikely to predict future data well at all."



Is Benign Overfitting “Benign”

- Privacy concerns
- Benchmarks might not have label noise characteristic of “real-world” data
- Clearly, interpolation **suffices** to learn models with strong generalization, but is it **necessary** in the overparameterized regime?

Is Benign Overfitting “Benign”

- Privacy concerns
- Benchmarks might not have label noise characteristic of “real-world” data
- Clearly, interpolation **suffices** to learn models with strong generalization, but is it **necessary** in the overparameterized regime?

$$\begin{aligned} \underset{\widehat{\theta}\in\mathcal{H}}{\text{minimize}} \quad & \mathsf{Pred}\left(\widehat{\theta}\right) := \mathbb{E}\left[(x^\top\widehat{\theta} - y)^2 \mid X\right] \\ \text{subject to} \quad & \mathsf{Train}\left(\widehat{\theta}\right) := \frac{1}{n}\mathbb{E}\left[\left\|X\widehat{\theta} - Y\right\|_2^2 \mid X\right] \geq \epsilon^2 \,, \end{aligned}$$

Overall Summary

- Framework: random design linear regression with independent noise
- Setting 1: estimator linear in *y* (least norm, ridge, etc.)
- Setting 2: square integrable estimators.

Findings

- Asymptotic characterization of optimal solution to problem
- **Memorization of label noise is necessary for generalization**
  - The threshold for optimal prediction risk goes to zero asymptotically faster than the variance of label noise
  - Consequence: fit linear regression models to accuracy much better than the noise floor

Problem Formulation

$$\mathsf{Train}_{X,\theta}(\widehat{\theta}) = \frac{1}{n}\mathbb{E}_w[\|X\widehat{\theta} - y\|_2^2 \mid X, \theta],$$
$$\mathsf{Pred}_{X,\theta}(\widehat{\theta}) = \mathbb{E}_{x,w}[(x^\top\theta - x^\top\widehat{\theta})^2 \mid X, \theta],$$

$$\mathsf{Train}_X(\widehat{\theta}) = \mathbb{E}_\theta[\mathsf{Train}_{X,\theta}(\widehat{\theta})]$$
$$\mathsf{Pred}_X(\widehat{\theta}) = \mathbb{E}_\theta[\mathsf{Pred}_{X,\theta}(\widehat{\theta})].$$

Problem Formulation

$$\begin{aligned} &\underset{\widehat{\theta}\in\mathcal{H}}{\text{minimize}}\ \mathsf{Pred}_X\left(\widehat{\theta}\right) \\ &\text{subject to}\ \mathsf{Train}_X\left(\widehat{\theta}\right) \geq \epsilon^2 \end{aligned}$$

$$\mathsf{Cost}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}(0)} \mathsf{Pred}_X\left(\widehat{\theta}\right),$$

$$\overline{\mathsf{Cost}}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \mathsf{Pred}_X\left(\widehat{\theta}_{\mathsf{ols}}\right).$$

Assumptions

Assumption A1 (Proportional asymptotics and spherical prior)**.** The dimension $d := d(n)$ satisfies $d/n \to \gamma \in (1, \infty)$. The data matrix $X = [x_1\ x_2\ \cdots\ x_n]^\top \in \mathbb{R}^{n \times d}$, where $X := X(n) = (x_{ij}(n))_{i \in [n], j \in [d]}$ forms a triangular array of random variables with independent rows. There is a deterministic sequence of symmetric positive definite matrices $\Sigma := \Sigma(n) \in \mathbb{R}^{d \times d}$ such that $X = Z\Sigma^{\frac{1}{2}}$, where $Z = (z_{ij})_{i \in [n], j \in [d]}$ and z_{ij} are i.i.d. random variables with distribution independent of n such that $\mathbb{E}[z_{ij}] = 0$, $\mathsf{Var}(z_{ij}) = 1$, and $\mathbb{E}[z_{ij}^4] \leq M$ for a universal constant M. In addition, we assume θ has prior independent of the data X, y, with zero mean and variance $\mathsf{Var}(\theta) = I_d/d$.

Assumption A2 (Linear estimators)**.** The hypothesis class consists of all linear estimators, i.e.,

$$\mathcal{H} = \left\{ \widehat{\theta}(X, y) = Ay\,, A := A(X) \in \mathbb{R}^{d \times n} \right\},$$

where A may depend on the features X but not the labels y.

Isotropic Setting

$$\underset{A \in \mathbb{R}^{d \times n}}{\text{minimize}} \quad \mathcal{P}(A) = \frac{1}{d} \|AX - I\|_F^2 + \sigma^2 \|A\|_F^2$$

$$\text{subject to} \quad \mathcal{T}(A) = \frac{1}{nd} \|XAX - X\|_F^2 + \frac{\sigma^2}{n} \|XA - I\|_F^2 \geq \epsilon^2 .$$

- Notice that this problem has quadratic objective and a quadratic constraint – this lets us leverage strong duality

Some Linear Algebra

$$
\begin{aligned}
\mathsf{Train}_X(\widehat{\theta}) &= \mathbb{E}_\theta\left[\mathsf{Train}_{X,\theta}(\widehat{\theta})\right] = \frac{1}{n}\mathbb{E}_{\theta,w}\left[\|(XA-I)\,(X\theta+w)\|_2^2 \mid X\right] \\
&= \frac{1}{n}\mathsf{Tr}\left(\mathbb{E}_{\theta,w}\left[(X\theta+w)^\top (XA-I)^\top (XA-I)\,(X\theta+w) \mid X\right]\right) \\
&= \frac{1}{n}\mathsf{Tr}\left((XA-I)X\theta\theta^\top X^\top (XA-I)^\top\right) + \frac{\sigma^2}{n}\mathsf{Tr}\left((XA-I)(XA-I)^\top\right) \\
&= \frac{1}{nd}\|XAX-X\|_F^2 + \frac{\sigma^2}{n}\|XA-I\|_F^2\ .
\end{aligned}
$$

Some Linear Algebra

$$
\begin{aligned}
\mathsf{Pred}_X(\widehat{\theta}) &= \mathbb{E}_\theta\left[\mathsf{Pred}_{X,\theta}(\widehat{\theta})\right] = \mathbb{E}_{\theta,w}\left[\|(AX-I)\theta + Aw\|_\Sigma^2 \mid X\right] \\
&= \mathsf{Tr}\left(\mathbb{E}_{\theta,w}\left[((AX-I)\theta + Aw)^\top \Sigma\left((AX-I)\theta + Aw\right) \mid X\right]\right) \\
&= \mathsf{Tr}\left(\mathbb{E}_\theta\left[\Sigma(AX-I)\theta\theta^\top(AX-I)^\top \mid X\right]\right) + \sigma^2\mathsf{Tr}\left(A^\top\Sigma A\right) \\
&= \frac{1}{d}\left\|\Sigma^{\frac{1}{2}}(AX-I)\right\|_F^2 + \sigma^2\left\|\Sigma^{\frac{1}{2}}A\right\|_F^2,
\end{aligned}
$$

Duality Interpretation

Thus we may leverage strong duality [10, Appendix B.1], writing a Lagrangian and solving, to conclude that for some $\rho_n := \rho_n(\epsilon)$ such that $I - \frac{\rho_n}{d} X^\top X \succ 0$, the optimal A for the problem (2) is

$$A(\rho_n) = \left(I - \rho_n \sigma^2 \left(I - \frac{\rho_n}{d} X^\top X \right)^{-1} \right) (X^\top X + d\sigma^2 I)^{-1} X^\top,$$

where ρ_n is the dual optimal value of the Lagrange multiplier associated with the constraint $\mathcal{T}(A) \geq \epsilon^2$. When $\rho_n = 0$, the constraint is inactive, so $A(0)$ is the global minimizer of

Duality Interpretation

Thus we may leverage strong duality [10, Appendix B.1], writing a Lagrangian and solving, to conclude that for some $\rho_n := \rho_n(\epsilon)$ such that $I - \frac{\rho_n}{d} X^\top X \succ 0$, the optimal A for the problem (2) is

$$A(\rho_n) = \left(I - \rho_n \sigma^2 \left(I - \frac{\rho_n}{d} X^\top X \right)^{-1} \right) (X^\top X + d\sigma^2 I)^{-1} X^\top,$$

where ρ_n is the dual optimal value of the Lagrange multiplier associated with the constraint $\mathcal{T}(A) \geq \epsilon^2$. When $\rho_n = 0$, the constraint is inactive, so $A(0)$ is the global minimizer of the unconstrained problem and evidently corresponds to a ridge regression estimate; we have $\mathsf{Cost}_X(\epsilon) = \mathcal{P}(A(\rho_n)) - \mathcal{P}(A(0))$ and $\mathcal{T}(A(\rho_n)) = \epsilon^2$. Substituting $A = A(\rho)$ into $\mathcal{P}(A)$ and $\mathcal{T}(A)$, we obtain

$$\mathcal{P}(A(\rho)) - \mathcal{P}(A(0)) = \frac{\rho^2 \sigma^4}{d} \mathsf{Tr} \left(\left(I - \frac{\rho}{d} X^\top X \right)^{-2} \frac{X^\top X}{d} \left(\frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right),$$

$$\mathcal{T}(A(\rho)) = \frac{\sigma^4}{n} \mathsf{Tr} \left(\left(I - \frac{\rho}{d} X^\top X \right)^{-2} \left(\frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right).$$

Linear Algebra before RMT

$$\mathcal{P}(A(\rho, I); I) - \mathcal{P}(A(0, I); I) = \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\left(I - \frac{\rho}{d}X^\top X\right)^{-2} X^\top X \left(X^\top X + d\sigma^2 I\right)^{-1}\right)$$
$$= \frac{\rho^2\sigma^4}{d/n} \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\left(1 - \rho\lambda_i^2/d\right)^2} \cdot \frac{\lambda_i^2}{d} \cdot \frac{1}{\lambda_i^2/d + \sigma^2}$$

$$\mathcal{T}(A(\rho, I); I) = \frac{d\sigma^4}{n}\mathsf{Tr}\left(\left(I - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X \left(I - \frac{\rho}{d}X^\top X\right)^{-1} \left(X^\top X\right)^\dagger \left(X^\top X + d\sigma^2 I\right)^{-1}\right)$$
$$= \frac{\sigma^4}{n}\mathsf{Tr}\left(\left(I - \frac{\rho}{d}X^\top X\right)^{-1} \frac{X^\top X}{d} \left(I - \frac{\rho}{d}X^\top X\right)^{-1} \left(\frac{X^\top X}{d}\right)^\dagger \left(\frac{X^\top X}{d} + \sigma^2 I\right)^{-1}\right)$$
$$= \sigma^4 \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 - \rho\lambda_i^2/d} \cdot \frac{\lambda_i^2}{d} \cdot \frac{1}{1 - \rho\lambda_i^2/d} \cdot \frac{1}{\lambda_i^2/d} \cdot \frac{1}{\lambda_i^2/d + \sigma^2}$$

Two RMT Lemmas

Lemma A.1 (Marchenko-Pastur law, Bai and Silverstein [4], Thm. 3.4)**.** *Let Z have singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, and let $\frac{1}{d}ZZ^\top$ have spectral distribution with c.d.f.*

$$H_n(s) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\lambda_i^2/d \leq s}.$$

Then with probability one H_n converges weakly to the c.d.f. H supported on $[\lambda_-, \lambda_+]$, with

$$\lambda_+ := \left(1 + \frac{1}{\sqrt{\gamma}}\right)^2 \quad \text{and} \quad \lambda_- := \left(1 - \frac{1}{\sqrt{\gamma}}\right)^2,$$

and H has density

$$dH(s) = \frac{\gamma}{2\pi}\frac{\sqrt{(\lambda_+ - s)(s - \lambda_-)}}{s}\mathbb{1}_{s\in[\lambda_-,\lambda_+]}ds.$$

Lemma A.2 (Bai-Yin law, Bai and Silverstein [4], Thm. 5.10)**.** *Let the conditions of Lemma A.1 hold, and assume additionally that $\sup_{ij}\mathbb{E}[z_{ij}^4] < \infty$. Then the largest and smallest singular values $\lambda_1 = \lambda_1(Z)$ and $\lambda_n = \lambda_n(Z)$ of Z satisfy*

$$\frac{\lambda_1^2}{d} \overset{a.s.}{\to} \lambda_+ = \left(1 + \frac{1}{\sqrt{\gamma}}\right)^2, \qquad \frac{\lambda_n^2}{d} \overset{a.s.}{\to} \lambda_- = \left(1 - \frac{1}{\sqrt{\gamma}}\right)^2.$$

RMT Time :(

$$\begin{aligned}\mathcal{P}(A(\rho,I);I)-\mathcal{P}(A(0,I);I) &= \frac{\rho^2\sigma^4}{d}\mathsf{Tr}\left(\left(I-\frac{\rho}{d}X^\top X\right)^{-2}X^\top X\left(X^\top X+d\sigma^2 I\right)^{-1}\right)\\ &= \frac{\rho^2\sigma^4}{d/n}\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left(1-\rho\lambda_i^2/d\right)^2}\cdot\frac{\lambda_i^2}{d}\cdot\frac{1}{\lambda_i^2/d+\sigma^2}\\ &= \frac{\rho^2}{d/n}\int\frac{\sigma^4 s}{(1-\rho s)^2(s+\sigma^2)}dH_n(s).\end{aligned}$$

By the assumption that $\rho<\lambda_+^{-1}$, the Bai-Yin law (Lemma A.2) guarantees that $I-\frac{\rho}{d}XX^\top$ is eventually positive definite and with probability one $\lambda_1^2/d\to\lambda_+$. The function $s\mapsto\frac{\sigma^4 s}{(1-\rho s)^2(s+\sigma^2)}$ is thus eventually bounded on the support of H_n. Applying the Marchenko-Pastur law, we deduce

$$\lim_{n\to\infty}\left(\mathcal{P}(A(\rho,I);I)-\mathcal{P}(A(0,I);I)\right)=\frac{\rho^2}{\gamma}\int\frac{\sigma^4 s}{(1-\rho s)^2(s+\sigma^2)}dH(s).$$

RMT Time :(

$$
\begin{aligned}
\mathcal{T}(A(\rho, I); I) &= \frac{d\sigma^4}{n}\mathsf{Tr}\left(\left(I - \frac{\rho}{d}X^\top X\right)^{-1} X^\top X \left(I - \frac{\rho}{d}X^\top X\right)^{-1} \left(X^\top X\right)^\dagger \left(X^\top X + d\sigma^2 I\right)^{-1}\right) \\
&= \frac{\sigma^4}{n}\mathsf{Tr}\left(\left(I - \frac{\rho}{d}X^\top X\right)^{-1} \frac{X^\top X}{d} \left(I - \frac{\rho}{d}X^\top X\right)^{-1} \left(\frac{X^\top X}{d}\right)^\dagger \left(\frac{X^\top X}{d} + \sigma^2 I\right)^{-1}\right) \\
&= \sigma^4 \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 - \rho\lambda_i^2/d} \cdot \frac{\lambda_i^2}{d} \cdot \frac{1}{1 - \rho\lambda_i^2/d} \cdot \frac{1}{\lambda_i^2/d} \cdot \frac{1}{\lambda_i^2/d + \sigma^2} \\
&= \int \frac{\sigma^4}{(1 - \rho s)^2 (s + \sigma^2)} dH_n(s).
\end{aligned}
$$

Applying the Marchenko-Pastur law gives the desired limit.

Main Punchline 1

Theorem 1. *Let Assumption A1 and either Assumption A2 or A2′ hold. Then as* $n \to \infty$,

(i) (***threshold value***) *for* ϵ_σ *defined in Eq. (7),* $\epsilon_\sigma^2 = \frac{\sigma^4}{\sigma^2+1-1/\gamma} + o(\sigma^4)$.

(ii) (***no cost below threshold***) *if* $\epsilon < \epsilon_\sigma$, *then with probability one* $\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) = 0$. *In addition, for the ridge estimator* $\widehat{\theta}_{d\sigma^2} = (X^\top X + d\sigma^2 I)^{-1} X^\top y$, *we have*

$$\lim_{n\to\infty} \left(\min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \mathsf{Pred}_X\left(\widehat{\theta}_{d\sigma^2}\right) \right) = 0\,.$$

(iii) (***cost of not fitting***) *if* $\epsilon \geq \epsilon_\sigma$, *there exists a scalar* $\rho := \rho(\epsilon) \in \left[0, \lambda_+^{-1}\right)$ *that uniquely solves*

$$\int \frac{\sigma^4}{(1-\rho s)^2\,(s+\sigma^2)} dH(s) = \epsilon^2, \tag{8}$$

and with probability one

$$\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) = \frac{\rho^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho s)^2 (s+\sigma^2)} dH(s). \tag{9}$$

For the constants $\mathsf{c} := \frac{2}{\lambda_-^2+\sigma^2}$ *and* $\mathsf{C} := \frac{(1-1/\sqrt{2})^2\lambda_-}{\lambda_+^2\gamma}$, *we have* $\lim_{n\to\infty} \mathsf{Cost}_X(\epsilon) \geq \mathsf{C}\epsilon^2$ *whenever* $\epsilon^2 \geq \mathsf{c}\sigma^4$.

Main Punchline 2

Theorem 2. *Let Assumption A1 and either Assumption A2 or A2′ hold. Then*

(i) ***(interpolation cost)*** *for any* $\epsilon \geq 0$, $\mathsf{Cost}_X(\epsilon) - \overline{\mathsf{Cost}}_X(\epsilon) = \mathsf{Pred}_X(\widehat{\theta}_{\mathsf{ols}}) - \mathsf{Pred}_X(\widehat{\theta}(0))$, *and with probability one*

$$\lim_{n\to\infty} \left(\mathsf{Pred}_X(\widehat{\theta}_{\mathsf{ols}}) - \mathsf{Pred}_X(\widehat{\theta}(0))\right) = \frac{\sigma^4}{\gamma}\int \frac{1}{s(s+\sigma^2)} dH(s) = \frac{\sigma^4}{\gamma\,(1-1/\gamma)^3} + o(\sigma^4).$$

(ii) ***(interpolation threshold)*** *for any* $\sigma > 0$, *there exists a* $\rho = \rho_{\mathrm{ols}} \in (0, \lambda_+^{-1})$ *that uniquely solves*

$$\rho^2 \int \frac{s}{(1-\rho s)^2\,(s+\sigma^2)} dH(s) = \int \frac{1}{s(s+\sigma^2)} dH(s), \tag{10}$$

where for the threshold $\epsilon^2_{\sigma,\mathrm{ols}} := \int \frac{\sigma^4}{(1-\rho_{\mathrm{ols}} s)^2 (s+\sigma^2)} dH(s)$ *we have*

$$\lim_{n\to\infty} \overline{\mathsf{Cost}}_X(\epsilon) \begin{cases} < 0 & \text{if } \epsilon < \epsilon_{\sigma,\mathrm{ols}} \\ = 0 & \text{if } \epsilon = \epsilon_{\sigma,\mathrm{ols}} \\ > 0 & \text{if } \epsilon > \epsilon_{\sigma,\mathrm{ols}}. \end{cases}$$

In comparison to the threshold ϵ_σ *in Eq. (7) and Theorem 1, we have* $\epsilon_\sigma < \epsilon_{\sigma,\mathrm{ols}} \leq \frac{2\lambda_+}{\lambda_-}\epsilon_\sigma$.