Benign Overfitting in Linear Regression

Senem Işık

Recap

Why does deep learning work so well in practice?

(even though classical theory often doesn't explain it well)

Deep learning often generalizes well even when:

- models are **huge** (large norms, many parameters, a lot of layers)
- training error goes to (near) **zero**

Classical "more parameters → overfitting → low test error" intuition often doesn't match practice.

Why does deep learning work so well in practice?

(even though classical theory often doesn't explain it well)

Deep learning often generalizes well even when:

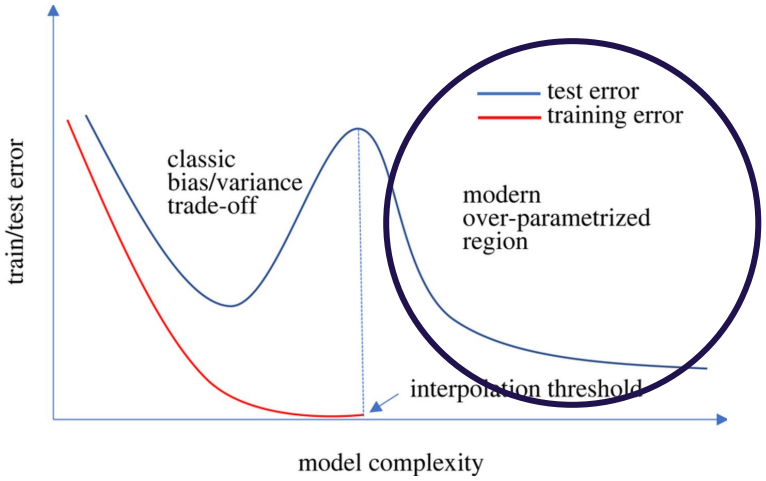
- models are **huge** (large norms, many parameters, a lot of layers)
- training error goes to (near) **zero**

Classical "more parameters → overfitting → low test error" intuition often doesn't match practice.

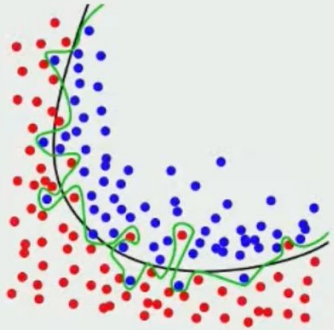
Double descent

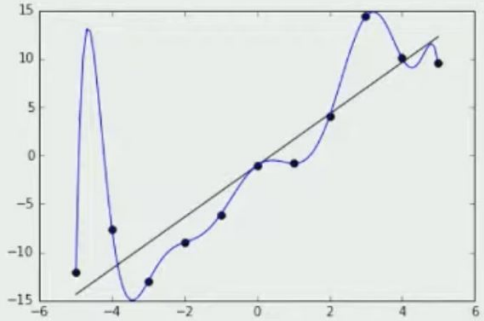
As model size increases:

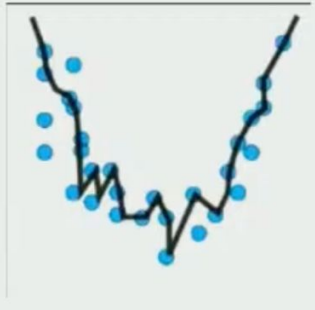
- error can drop,
- spike near the interpolation threshold,
- then drop again in the over-parameterized regime

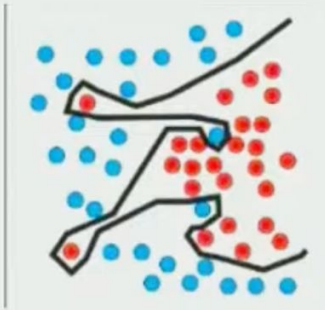


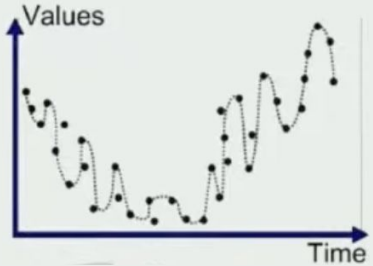
Classical ML Story for Overfitting

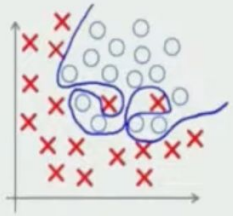


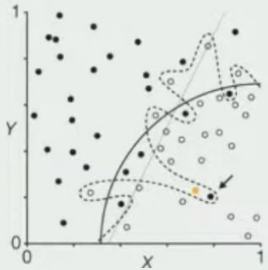


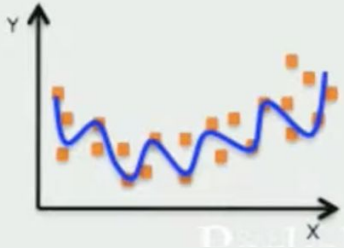




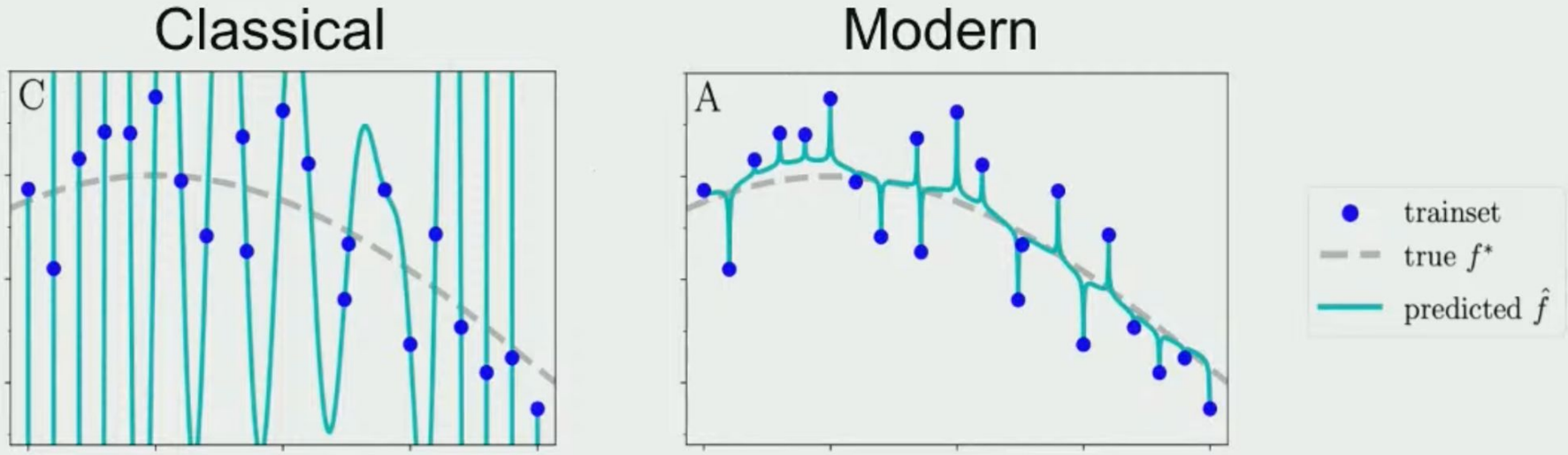








Overfitting can be **benign**



As long as the noisy data points do not change the general trend, since, at test time, we will sample the exact noisy point with very low probability, we are fine!

Benign Overfitting

Benign Overfitting in Linear Regression

Benign Overfitting in Linear Regression

Peter L. Bartlett[a,b], Philip M. Long[c], Gábor Lugosi[d], and Alexander Tsigler[a]

[a]Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley CA 94720-3860
[b]Computer Science Division, UC Berkeley, 387 Soda Hall, Berkeley CA 94720-1776
[c]Google
[d]Economics and Business, Pompeu Fabra University; ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain; Barcelona Graduate School of Economics

One sentence takeaway

In **high-dimensional linear regression**,
perfectly fitting noisy data can still yield vanishing test error
(benign overfitting)
but only under specific covariance/spectral conditions

Setup

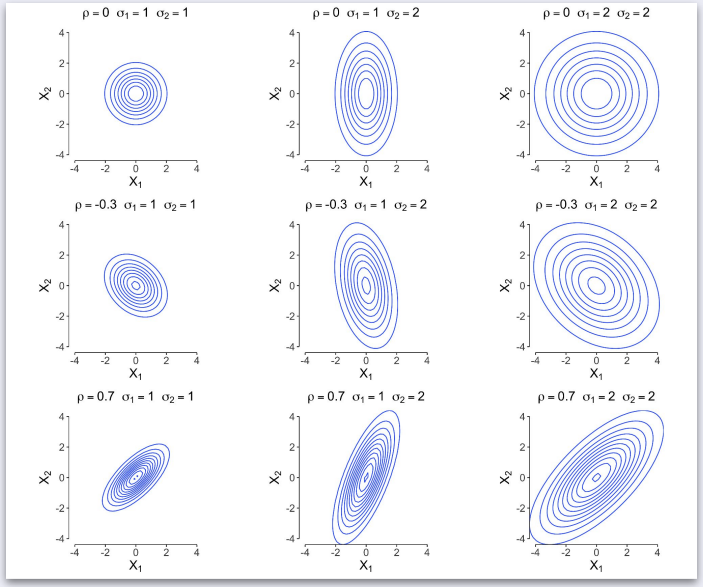
Simple Prediction Setting: Linear Regression

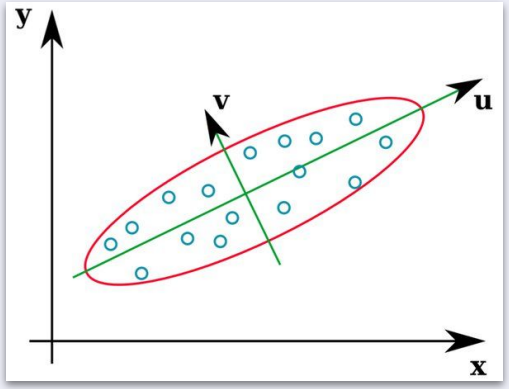
- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- (x, y) subgaussian, mean zero, well-specified: $\mathbb{E}[y|x] = x^\top \theta^*$.
- Define:

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \qquad \text{(assume } \lambda_1 \geq \lambda_2 \geq \cdots\text{)}$$

Intuition for eigenvalues + eigenvectors

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \qquad (\text{assume } \lambda_1 \geq \lambda_2 \geq \cdots)$$





Setup

Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- (x, y) subgaussian, mean zero, well-specified: $\mathbb{E}[y|x] = x^\top \theta^*$.
- Define:

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \qquad \text{(assume } \lambda_1 \geq \lambda_2 \geq \cdots\text{)}$$

$$\theta^* := \arg\min_\theta \mathbb{E}\left(y - x^\top \theta\right)^2,$$

$$\sigma^2 := \mathbb{E}(y - x^\top \theta^*)^2.$$

A bit more general than linear regression

We can capture **both linear** and **polynomial regression**

- A simple way to do that is to do a **feature expansion**
- + have a linear function over the “expanded” features

Eg.

- Feature expansion: $x \rightarrow (1, x, x^2, x^3, \ldots) = f(x)$
- Polynomial regression := linear regression on f(x)

Our estimator

Overparameterized regime: $n \ll d = p$

Regularized linear regression

$$\min \quad \lambda\|\theta\|^2 + \frac{1}{n}\|X\theta - y\|^2,$$

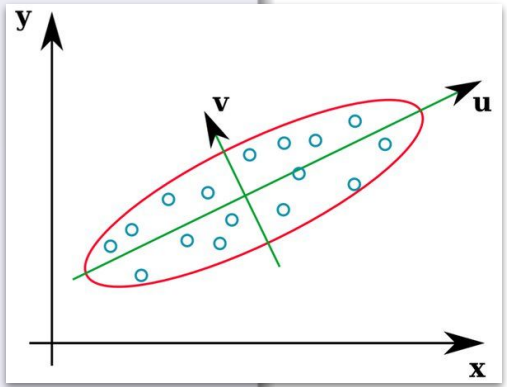
$$\begin{aligned} \min \quad & \|X\theta - y\|^2 \\ \text{s.t.} \quad & \|\theta\| \le b, \end{aligned}$$

$$\begin{aligned} \min \quad & \|\theta\| \\ \text{s.t.} \quad & \frac{1}{n}\|X\theta - y\|^2 \le c. \end{aligned}$$

Quantity of interest

Excess prediction error

$$R(\hat{\theta}) := \mathbb{E}_{(x,y)}\left(y - x^\top \hat{\theta}\right)^2 - \underbrace{\min_\theta \mathbb{E}\left(y - x^\top \theta\right)^2}_{\text{optimal prediction error}}$$



All in one

Overfitting regime: n ≪ d = p

Overfitting regime

- We consider situations where $\min_\beta \|X\beta - y\|^2 = 0$.
- Estimator $\hat{\theta} = (X^\top X)^\dagger X^\top y$ solves

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \text{s.t.} \quad & \|X\theta - y\|^2 = \min_\beta \|X\beta - y\|^2 = 0. \end{aligned}$$

- Hence, $y_1 = x_1^\top \hat{\theta}, \ldots, y_n = x_n^\top \hat{\theta}$.

Main question

Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top \theta^*$ and $y - x^\top \theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of x.

     Not a problem, even in high dimensions ($p > n$).
  2. $\hat{\theta}$ is corrupted by the noise in $y_1, \ldots, y_n$.

     Problematic.
- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?

(1) Inevitable error that comes from **small training size**

(2) The contribution of **overfitting to noise** (!)

Main result

Theorem

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if

Important object

Definition (Effective Ranks)

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of Σ.
For $k \geq 0$, if $\lambda_{k+1} > 0$, define the effective ranks

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

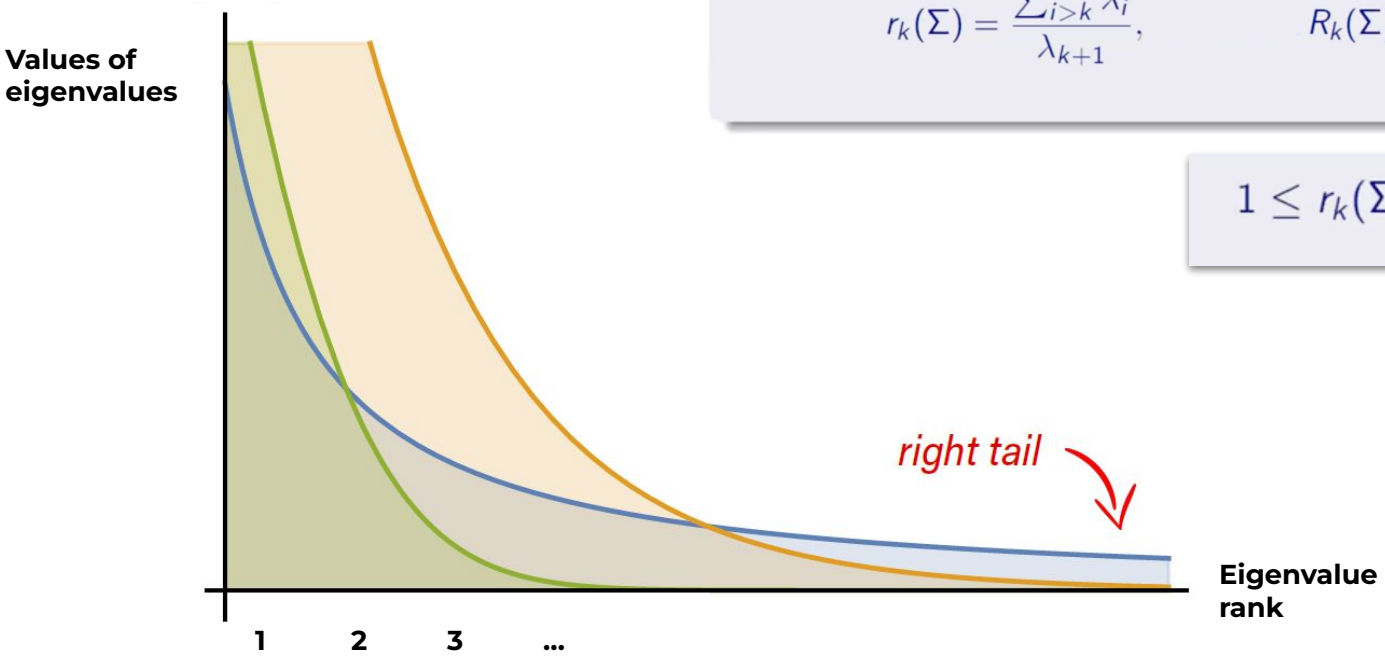
Example in picture

Definition (Effective Ranks)

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of Σ.
For $k \geq 0$, if $\lambda_{k+1} > 0$, define the effective ranks

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

$$1 \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma).$$



Example in picture

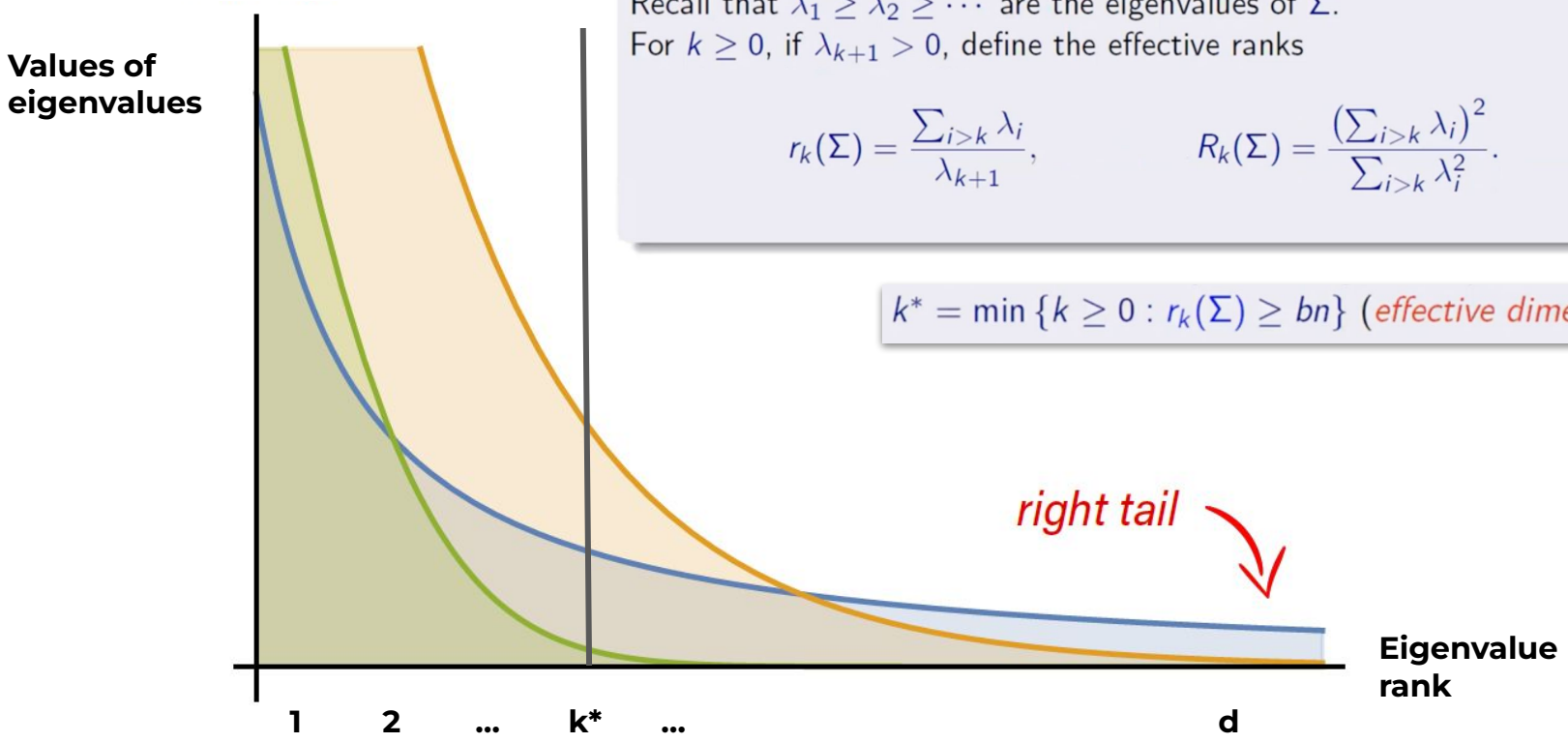
Overfitting regime: n ≪ d = p

Definition (Effective Ranks)

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of Σ.
For $k \geq 0$, if $\lambda_{k+1} > 0$, define the effective ranks

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

$$k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\} \text{ (effective dimension)}$$



Theorem

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*),

1. With high probability,

$$R(\hat{\theta}) \leq c\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right),$$

2. With some independence properties,

$$\mathbb{E}R(\hat{\theta}) \geq \frac{1}{c}\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2 \min\left\{\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1\right\}\right).$$

$$\text{bias}(\theta^*, \Sigma, n) = \|\theta^*_{k+1:\infty}\|^2_{\Sigma_{k+1:\infty}} + \|\theta^*_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{\sum_{i>k}\lambda_i}{n}\right)^2.$$

Intuition

We say $\{\Sigma_n\}$ is *asymptotically benign* if

$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

Intuition

We say $\{\Sigma_n\}$ is *asymptotically benign* if

Eigenvalues should decay fast so that their sum is o(n)

$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

Small number of large eigenvectors

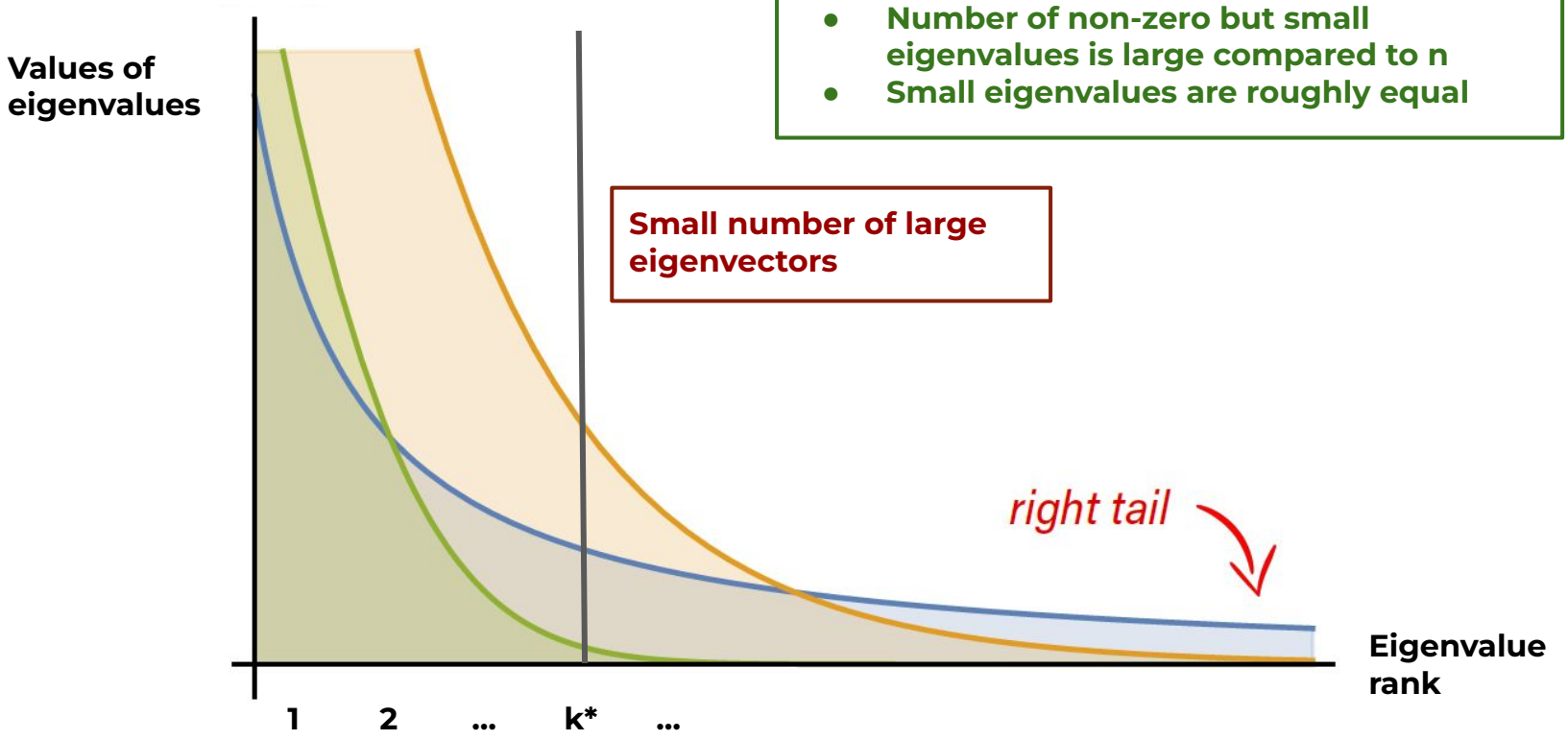
- **Number of non-zero but small eigenvalues is large compared to n**
- **Small eigenvalues are roughly equal**

Sum of eigenvalues must **almost** diverge

Example in picture

Eigenvalues should decay fast so that their sum is o(n)

- Number of non-zero but small eigenvalues is large compared to n
- Small eigenvalues are roughly equal



Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top \theta^*$ and $y - x^\top \theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of x.

     Not a problem, even in high dimensions ($p > n$).
  2. $\hat{\theta}$ is corrupted by the noise in $y_1, \ldots, y_n$.

     Problematic.
- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?

One sentence answer

In **high-dimensional linear regression**,
perfectly fitting noisy data can still yield vanishing test error
(benign overfitting)
when **many comparable weak features (eigenvalues)** exists

Proof Intuition

1. Start with a bias–variance decomposition for the min-norm interpolator.
2. Variance term involves Tr(Σ) that captures how noise affects prediction accuracy.
3. Then use concentration (this is where *sub-gaussian assumption is useful)*
   + spectral inequalities to bound each term.

Lemma 7. *The excess risk of the minimum norm estimator satisfies*

$$R(\hat{\theta}) \leq 2\theta^{*\top} B\theta^* + c\sigma^2 \log\frac{1}{\delta}\operatorname{tr}(C)$$

with probability at least $1-\delta$ *over* ϵ, *and*

$$\mathbb{E}_{\boldsymbol{\epsilon}} R(\hat{\theta}) \geq \theta^{*\top} B\theta^* + \sigma^2 \operatorname{tr}(C),$$

where

$$B = \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \Sigma \left(I - X^\top \left(XX^\top\right)^{-1} X\right),$$
$$C = \left(XX^\top\right)^{-1} X\Sigma X^\top \left(XX^\top\right)^{-1}.$$

We say $\{\Sigma_n\}$ is *asymptotically benign* if

$$\lim_{n\to\infty} \left(\|\Sigma_n\| \sqrt{\frac{r_0(\Sigma_n)}{n}} + \frac{k_n^*}{n} + \frac{n}{R_{k_n^*}(\Sigma_n)} \right) = 0,$$

where $k_n^* = \min\{k \geq 0 : r_k(\Sigma_n) \geq bn\}$.

Proof Intuition

Inevitable error that comes from small training size

The contribution of **overfitting to noise** in the direction of **important** eigenvectors

The contribution of **overfitting to noise** in the direction of **unimportant** eigenvectors

We say $\{\Sigma_n\}$ is *asymptotically benign* if

$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

where $k_n^*=\min\{k\geq 0: r_k(\Sigma_n)\geq bn\}$.

Example: *Finite dimension*, fast λ_i decay, plus isotropic noise

If

$$\lambda_{k,n} = \begin{cases} e^{-k} + \epsilon_n & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

then Σ_n is benign iff

- $p_n = \omega(n)$,
- $\epsilon_n p_n = o(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$. $(n \ge 40 \implies ne^{-n} < 2^{-52})$

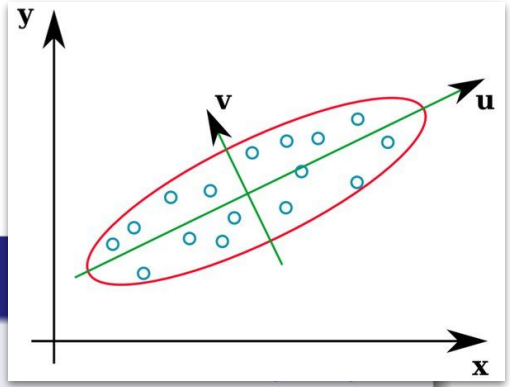
Furthermore, for $p_n = \Omega(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$,

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

Generic phenomenon:

quickly converging λ_i plus noise in all directions, $p_n \gg n$.

Very brittle for **adversarial** noise



Label noise appears in $\hat{\theta}$

We can find a unit norm Δ
such that perturbing an input x by Δ changes the output enormously:
even if $\Delta^\top \theta^* = 0$,

$$\left\|(x+\Delta)^\top \hat{\theta} - x^\top \hat{\theta}\right\|^2 \geq \frac{\sigma}{\sqrt{\lambda_{k^*+1}}} \geq \sqrt{\frac{n}{\mathrm{tr}(\Sigma)}}\sigma.$$

Benign overfitting leads to huge sensitivity.

Conclusion

Far from the regime of a **tradeoff** between fit to training data and complexity.

In linear regression, **a long, flat tail of the covariance eigenvalues** is ***necessary*** and ***sufficient*** for the **minimum norm interpolant** to predict well:

The noise is hidden in many unimportant directions.

Questions

Beyond **minimum euclidean norm interpolant**?

What is the approximately equivalent approach for analyzing **deep neural networks**?

Thank you!