MIT 9.520/6.7910
Statistical Learning Theory and Applications

Class 9: Learning bounds for linear least squares

Lorenzo Rosasco

ML algorithms design so far

- ERM+bias/regularization+optimization
- Implicit regularization
- Linear models
- Nonlinear models: neural nets+kernels

Theory so far

$$\min_{f \in \mathcal{T}} L(f) \quad \mapsto \quad \min_{f \in \mathcal{H}} \widehat{L}(f)$$

- Representation: $\mathcal{H}, \mathcal{B}$ etc.
- Optimization:

$$\min_{f \in \mathcal{H}} \widehat{L}(f).$$

- **Generalization**: $L(\widehat{f})$.

Outline

Recap statistical learning theory

Linear least squares

Error decomposition

Approximation Error

Estimation error
- Estimation Error I: probabilistic estimates
- Estimation Error II: deterministic estimates

Final bound

Ghost track: improved approximation results

The supervised learning problem

- $\mathcal{X} \times \mathbb{R}$ probability space, with distribution P.
- $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ *loss function*.

Problem: Solve

$$L(f_P) = \min_{f \in \mathcal{T}} L(f), \qquad L(f) = \mathbb{E}_{(x,y)\sim P}[\ell(y, f(x))],$$

given

$$S_n = (x_1, y_1), \ldots, (x_n, y_n) \sim P^n.$$

ERM

$$S_n \longmapsto \hat{f} = \hat{f}_S \in \mathcal{H}$$

ERM

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \widehat{L}(f), \qquad \widehat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)).$$

Note: an estimator is a random variable.

How good is a solution?

$$S_n \to \widehat{f}$$

- Practice: **test error**[1]

$$\widehat{L}_{\mathsf{Test}}(\widehat{f}) = \frac{2}{n} \sum_{i=n/2+1}^{n} \ell(y_i, \widehat{f}(x_i)).$$

- Theory: **excess risk**

$$L(\widehat{f}) - L(f_P).$$

Note: these are also random variables.

[1] Use $n/2$ points for $\widehat{f}$.

Outline

Recap statistical learning theory

Linear least squares

Error decomposition

Approximation Error

Estimation error
- Estimation Error I: probabilistic estimates
- Estimation Error II: deterministic estimates

Final bound

Ghost track: improved approximation results

Linear least squares

- Linear models

$$f_w(x) = w^\top x,$$

(analysis applies to $f_w(x) = \langle w, \Phi(x) \rangle$).

- Linear least squares

$$(y - w^\top x)^2.$$

Lots of nice formulas!

- Bounded data

$$|y| \leq M, \qquad \|x\| \leq \kappa \quad \text{a.s.}$$

Notation for the risks

$$f_w(x) = w^\top x$$

Expected risk

$$L(w) \;=\; \mathbb{E}[(y - w^\top x)^2].$$

Empirical risk

$$\widehat{L}(w) \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2.$$

Rewriting ridge regression

$$\widehat{w}_\lambda = \arg\min_{w \in \mathbb{R}^d} \widehat{L}(w) + \lambda \|w\|^2$$

Then

$$\widehat{w}_\lambda = (\widehat{\Sigma} + \lambda I)^{-1} \widehat{h}$$

where

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top, \qquad \widehat{h} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i.$$

Best target in the model

$$L(w^\dagger) = \min_{w \in \mathbb{R}^d} L(w)$$

- Always true with finite features $f(x) = \langle w, \Phi(x) \rangle$, with $\Phi(x) \in \mathbb{R}^p$.

- It is an assumption for infinite-dimensional models[2].

[2] Think of learning a function in a Sobolev space with the Gaussian kernel.

Normal equation

$$L(w^\dagger) = \min_{w \in \mathbb{R}^d} L(w)$$

$$\Sigma w^\dagger = h.$$

where

$$\Sigma = \mathbb{E}[xx^\top], \qquad h = \mathbb{E}[xy].$$

We define $w^\dagger$ as the minimum-norm solution.

Normal equation: proof

$$\Sigma w^\dagger = h$$

Indeed,

$$\nabla L(w^\dagger) = 0$$

and

$$\nabla L(w) = 2\ \mathbb{E}\big[x(x^\top w - y)\big] = 2(\Sigma w - h).$$

The minimal-norm solution is the unique one in $\text{Null}(\Sigma)^\perp$.

Covariance matrix

$$\Sigma = \mathbb{E}[xx^\top].$$

This is the covariance (or second-order) matrix. It is:

- Symmetric.
- Positive semidefinite.

Excess risk and norm

$$L(w) - L(w^\dagger) = \|\Sigma^{1/2}(w - w^\dagger)\|^2$$

Proof: adding and subtracting $x^\top w^\dagger$:

$$\begin{aligned} L(w) &= \mathbb{E}\big[(y - x^\top w^\dagger + x^\top(w^\dagger - w))^2\big] \\ &= \mathbb{E}[(y - x^\top w^\dagger)^2] + \mathbb{E}\big[(x^\top(w - w^\dagger))^2\big] + 2\,\mathbb{E}\big[(y - x^\top w^\dagger)\, x^\top(w - w^\dagger)\big]\,. \end{aligned}$$

First term is $L(w^\dagger)$ and the last term is zero since

$$\mathbb{E}\big[(y - x^\top w^\dagger)\, x\big] = h - \Sigma w^\dagger = 0.$$

Finally,

$$L(w) - L(w^\dagger) = (w - w^\dagger)^\top \Sigma (w - w^\dagger) = \|\Sigma^{1/2}(w - w^\dagger)\|^2.$$

Excess risk of least squares

$$L(\widehat{w}_\lambda) - L(w^\dagger) = \|\Sigma^{1/2}(\widehat{w}_\lambda - w^\dagger)\|^2.$$

where

$$\widehat{w}_\lambda = (\widehat{\Sigma} + \lambda I)^{-1}\,\widehat{h}, \qquad\qquad \Sigma w^\dagger = h.$$

Matrix norm and spectral theorem

Let A be a symmetric positive semidefinite matrix with eigenvalues $\lambda_i(A)$.

Then

$$\|A\| = \max_i \lambda_i(A),$$

and for any continuous $f : [0, \infty) \to \mathbb{R}$

$$\|f(A)\| = \max_i |f(\lambda_i(A))|.$$

Moreover

$$\|A\| \ \leq \ \|A\|_{\mathsf{F}} \ \leq \ \mathrm{Tr}(A),$$

where the Frobenius norm is $\|A\|_{\mathsf{F}}^2 = \langle A, A \rangle_{\mathsf{F}} = \mathrm{Tr}(A^\top A)$.

Outline

Recap statistical learning theory

Linear least squares

Error decomposition

Approximation Error

Estimation error
- Estimation Error I: probabilistic estimates
- Estimation Error II: deterministic estimates

Final bound

Ghost track: improved approximation results

Decomposition

$$\|\Sigma^{1/2}(\widehat{w}_\lambda - w^\dagger)\|^2$$

$$\widehat{w}_\lambda - w^\dagger = \widehat{w}_\lambda - w_\lambda + w_\lambda - w^\dagger$$

where

$$w_\lambda = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \; L(w) + \lambda \|w\|^2$$

and

$$\widehat{w}_\lambda = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \; \widehat{L}(w) + \lambda \|w\|^2, \qquad w^\dagger = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \; L(w).$$

Decomposition (cont.)

$$\|\Sigma^{1/2}(\widehat{w}_\lambda - w^\dagger)\|^2$$

$$\widehat{w}_\lambda - w^\dagger = \widehat{w}_\lambda - w_\lambda + w_\lambda - w^\dagger$$

where

$$w_\lambda = (\Sigma + \lambda I)^{-1} h$$

and

$$\widehat{w}_\lambda = (\widehat{\Sigma} + \lambda I)^{-1} \widehat{h} \qquad \Sigma w^\dagger = h.$$

Approximation and estimation errors[3]

$$\|\Sigma^{1/2}(\widehat{w}_\lambda - w^\dagger)\|^2, \qquad \widehat{w}_\lambda - w^\dagger = \widehat{w}_\lambda - w_\lambda + w_\lambda - w^\dagger$$

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w^\dagger)\right\|^2 \leq 2\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\|^2 + 2\left\|\Sigma^{1/2}(w_\lambda - w^\dagger)\right\|^2$$

- Estimation/Variance (stochastic)

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\|^2.$$

- Approximation/Bias (deterministic)

$$\left\|\Sigma^{1/2}(w_\lambda - w^\dagger)\right\|^2.$$

[3] Aka Bias-Variance decomposition

Outline

Recap statistical learning theory

Linear least squares

Error decomposition

Approximation Error

Estimation error
- Estimation Error I: probabilistic estimates
- Estimation Error II: deterministic estimates

Final bound

Ghost track: improved approximation results

Approximation and estimation errors[4]

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w^\dagger)\right\|^2 \leq 2\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\|^2 + 2\left\|\Sigma^{1/2}(w_\lambda - w^\dagger)\right\|^2$$

- Estimation/Variance (stochastic)

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\|^2$$

- **Approximation/Bias** (deterministic)

$$\left\|\Sigma^{1/2}(w_\lambda - w^\dagger)\right\|^2$$

[4] Aka Bias-Variance decomposition

Approximation Error

For $\lambda \leq \kappa$

$$\|\Sigma^{1/2}(w_\lambda - w^\dagger)\|^2 \leq \lambda \|w^\dagger\|^2.$$

Proof: $w_\lambda = (\Sigma + \lambda I)^{-1} h$, $\Sigma w^\dagger = h$, so $w_\lambda = (\Sigma + \lambda I)^{-1}\Sigma w^\dagger$. Then[5]

$$w_\lambda - w^\dagger = \big((\Sigma + \lambda I)^{-1}\Sigma - I\big)w^\dagger = (\Sigma + \lambda I)^{-1}(\Sigma - (\Sigma + \lambda I))w^\dagger = -\lambda(\Sigma + \lambda I)^{-1}w^\dagger.$$

Hence

$$\|\Sigma^{1/2}(w_\lambda - w^\dagger)\| \leq \lambda \|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}\| \, \|w^\dagger\|$$

Since

$$\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}\| = \max_i \frac{\sqrt{\lambda_i}}{\lambda_i + \lambda} = \frac{1}{2\sqrt{\lambda}} \quad (\lambda \leq \kappa),$$

then

$$\|\Sigma^{1/2}(w_\lambda - w^\dagger)\| \leq \sqrt{\lambda}\, \|w^\dagger\|.$$

[5] $I = (\Sigma + \lambda I)^{-1}(\Sigma + \lambda I)$.

Remarks

$$\|\Sigma^{1/2}(w_\lambda - w^\dagger)\|^2 \leq \lambda \|w^\dagger\|^2.$$

- The result is dimension free and sharp.

- It can be improved if we allow a dependence on $\lambda_{\min}(\Sigma)$ or add assumptions.

Improved approximation error

$$\|\Sigma^{1/2}(w_\lambda - w^\dagger)\|^2 \leq \lambda \|w^\dagger\|^2.$$

- It can be improved if we allow a dependence on[6] $\lambda_{\min}(\Sigma)$. If $\lambda \leq \lambda_{\min}(\Sigma)$, then

$$\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}\| \leq \frac{\sqrt{\lambda_{\min}(\Sigma)}}{\lambda_{\min}(\Sigma) + \lambda} \quad \Rightarrow \quad \|\Sigma^{1/2}(w_\lambda - w^\dagger)\| \leq \frac{\lambda}{\sqrt{\lambda_{\min}(\Sigma)}} \|w^\dagger\|.$$

- It can also be improved if $w^\dagger$ is *aligned* with Σ,

$$w^\dagger = \Sigma^\alpha v^\dagger \quad \Rightarrow \quad \|\Sigma^{1/2}(w_\lambda - w^\dagger)\| \leq \lambda^{\alpha+1/2} \|v^\dagger\|.$$

  This is called the source condition.

[6] $\lambda_{\min}(\Sigma)$ is always zero for infinite dimensional features!

Outline

Recap statistical learning theory

Linear least squares

Error decomposition

Approximation Error

Estimation error
- Estimation Error I: probabilistic estimates
- Estimation Error II: deterministic estimates

Final bound

Ghost track: improved approximation results

Approximation and estimation errors

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w^\dagger)\right\|^2 \;\leq\; 2\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\|^2 + 2\left\|\Sigma^{1/2}(w_\lambda - w^\dagger)\right\|^2$$

- **Estimation/Variance (stochastic)**

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\|^2$$

- Approximation/Bias (deterministic) ✓

$$\left\|\Sigma^{1/2}(w_\lambda - w^\dagger)\right\|^2 \leq \lambda \left\|w^\dagger\right\|^2.$$

Estimation error

For $\lambda_n \leq \lambda \leq \kappa$, w.h.p.[7]

$$\left\| \Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda) \right\|^2 \leq \frac{C}{\lambda n}$$

Key idea:

$$\widehat{w}_\lambda = (\widehat{\Sigma} + \lambda I)^{-1}\widehat{h}, \qquad w_\lambda = (\Sigma + \lambda I)^{-1}h$$

should be close if

$$\widehat{h} \approx h \quad \text{and} \quad \widehat{\Sigma} \approx \Sigma.$$

[7] With high probability.

Hoeffding inequality: scalar case

Let $Z_1, \ldots, Z_n \in \mathbb{R}$ be independent random variables with

$$|Z_i| \leq C \quad \text{a.s.},$$

and $\mu = \mathbb{E}[Z_i]$. Then, for any $\epsilon > 0$,

$$\Pr\left(\left| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mu \right| \geq \epsilon \right) \leq 2 \exp\left(-\frac{n\epsilon^2}{2C^2} \right).$$

Equivalently, with probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mu \right| \leq C \sqrt{\frac{2 \log(2/\delta)}{n}}.$$

Hoeffding inequality: vector case

Let $Z_1, \dots, Z_n \in \mathcal{H}$ be independent random variables in a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ with

$$\|Z_i\| \leq C \quad \text{a.s.},$$

and $\mu = \mathbb{E}[Z_i]$. Then, for any $\epsilon > 0$,

$$\Pr\left(\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mu \right\| \geq \epsilon \right) \leq 2 \exp\left(-\frac{n\epsilon^2}{2C^2} \right).$$

Equivalently, with probability at least $1 - \delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mu \right\| \leq C \sqrt{\frac{2 \log(2/\delta)}{n}}.$$

Probabilistic estimates

With probability $1-\delta$

$$\left\|\widehat{h}-h\right\| \;\leq\; \kappa M\sqrt{\tfrac{2\log(4/\delta)}{n}}, \qquad \text{and} \qquad \left\|\widehat{\Sigma}-\Sigma\right\|_{\mathsf{F}} \;\leq\; \kappa^2\sqrt{\tfrac{2\log(4/\delta)}{n}}.$$

- $Z_i = x_i y_i$, $\mu = h$, and $\|Z_i\| \leq \kappa M$.
- $Z_i = x_i x_i^\top$, $\mu = \Sigma$, and[8] $\|Z_i\|_{\mathsf{F}} \leq \kappa^2$
- Union bound to control both events simultaneously ($\delta \mapsto \delta/2$).

[8] $\|x_i x_i^\top\|_{\mathsf{F}} \leq \mathrm{Tr}(x_i x_i^\top) = \|x_i\|^2 \leq \kappa^2$.

Key deterministic estimate

$$\widehat{w}_\lambda - w_\lambda = (\widehat{\Sigma} + \lambda I)^{-1}\Big[(\widehat{h} - h) + (\Sigma - \widehat{\Sigma})w_\lambda\Big]$$

By definition

$$\widehat{w}_\lambda - w_\lambda = (\widehat{\Sigma} + \lambda I)^{-1}\widehat{h} \; - \; (\Sigma + \lambda I)^{-1}h$$

Add and subtract $(\widehat{\Sigma} + \lambda I)^{-1}h$

$$\widehat{w}_\lambda - w_\lambda = (\widehat{\Sigma} + \lambda I)^{-1}(\widehat{h} - h) + \Big[(\widehat{\Sigma} + \lambda I)^{-1} - (\Sigma + \lambda I)^{-1}\Big]h$$

Use $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$

$$\widehat{w}_\lambda - w_\lambda = (\widehat{\Sigma} + \lambda I)^{-1}(\widehat{h} - h) + (\widehat{\Sigma} + \lambda I)^{-1}(\Sigma - \widehat{\Sigma})\underbrace{(\Sigma + \lambda I)^{-1}h}_{w_\lambda}.$$

Intermediate bound

$$\widehat{w}_\lambda - w_\lambda = (\widehat{\Sigma} + \lambda I)^{-1}\Big[(\widehat{h} - h) + (\Sigma - \widehat{\Sigma})w_\lambda\Big]$$

Then

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\| \;\leq\; \left\|\Sigma^{1/2}(\widehat{\Sigma} + \lambda I)^{-1}\right\| \left[\left\|\widehat{h} - h\right\| + \left\|\widehat{\Sigma} - \Sigma\right\| \|w_\lambda\|\right]$$

- $\|w_\lambda\| \leq \left\|w^\dagger\right\|$, since

$$\|w_\lambda\| = \left\|(\Sigma + \lambda I)^{-1}h\right\| = \left\|(\Sigma + \lambda I)^{-1}\Sigma w^\dagger\right\| \leq \left\|w^\dagger\right\|.$$

- The critical term is

$$\left\|\Sigma^{1/2}(\widehat{\Sigma} + \lambda I)^{-1}\right\|.$$

Key deterministic estimate II

If $2\kappa^2\sqrt{\frac{2\log(4/\delta)}{n}} \leq \lambda \leq \kappa$, then

$$\left\|\Sigma^{1/2}(\widehat{\Sigma}+\lambda I)^{-1}\right\| \;\leq\; \frac{2}{\sqrt{\lambda}}.$$

Indeed,

$$(\widehat{\Sigma}+\lambda I)^{-1} = \Big[(\Sigma+\lambda I)+(\widehat{\Sigma}-\Sigma)\Big]^{-1} = (\Sigma+\lambda I)^{-1}\Big[I-(\Sigma-\widehat{\Sigma})(\Sigma+\lambda I)^{-1}\Big]^{-1}.$$

Using the Neumann series,

$$\left\|\Big(I-(\Sigma-\widehat{\Sigma})(\Sigma+\lambda I)^{-1}\Big)^{-1}\right\| \;\leq\; \Big(1-\left\|(\Sigma-\widehat{\Sigma})(\Sigma+\lambda I)^{-1}\right\|\Big)^{-1}.$$

Finally

$$\left\|\Sigma^{1/2}(\widehat{\Sigma}+\lambda I)^{-1}\right\| \;\leq\; \underbrace{\left\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}\right\|}_{\leq\frac{1}{\sqrt{\lambda}}} \Big(1-\underbrace{\left\|(\Sigma-\widehat{\Sigma})(\Sigma+\lambda I)^{-1}\right\|}_{\leq\frac{\|\widehat{\Sigma}-\Sigma\|}{\lambda}\leq\frac{1}{2}}\Big)^{-1}.$$

Neumann series argument

Let A be invertible and set $B = A - C$. Assume $\|CA^{-1}\| < 1$. Then

$$I - CA^{-1} \quad \text{is invertible,} \quad (I - CA^{-1})^{-1} = \sum_{k=0}^{\infty} (CA^{-1})^k.$$

Hence B is invertible and

$$B^{-1} = A^{-1}(I - CA^{-1})^{-1}, \quad \left\|B^{-1}\right\| \leq \frac{\|A^{-1}\|}{1 - \|CA^{-1}\|}.$$

In our case:

$$A = \Sigma + \lambda I, \quad B = \widehat{\Sigma} + \lambda I, \quad C = \Sigma - \widehat{\Sigma}.$$

Final estimation error bound

With probability $1-\delta$, if $2\kappa^2\sqrt{\frac{2\log(4/\delta)}{n}} \leq \lambda \leq \kappa$, then

$$\begin{aligned}\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\| &\leq \underbrace{\left\|\Sigma^{1/2}(\widehat{\Sigma}+\lambda I)^{-1}\right\|}_{\leq \frac{2}{\sqrt{\lambda}}} \Bigg[\underbrace{\left\|\widehat{h}-h\right\|}_{\leq \kappa M\sqrt{\frac{2\log(4/\delta)}{n}}} + \underbrace{\left\|\widehat{\Sigma}-\Sigma\right\|}_{\leq \kappa^2\sqrt{\frac{2\log(4/\delta)}{n}}} \left\|w^\dagger\right\| \Bigg] \\ &\leq C\frac{\sqrt{\log(4/\delta)}}{\sqrt{\lambda n}}.\end{aligned}$$

where $C = (2\sqrt{2}(M\kappa + \kappa^2\left\|w^\dagger\right\|))$.

Approximation and estimation errors[9]

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w^\dagger)\right\|^2 \leq 2\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\|^2 + 2\left\|\Sigma^{1/2}(w_\lambda - w^\dagger)\right\|^2$$

- Estimation/Variance (stochastic) ✓

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w_\lambda)\right\|^2 \leq C\frac{\log(4/\delta)}{\lambda n}.$$

- Approximation/Bias (deterministic) ✓

$$\left\|\Sigma^{1/2}(w_\lambda - w^\dagger)\right\|^2 \leq \lambda \left\|w^\dagger\right\|^2.$$

[9] Aka Bias-Variance decomposition

Outline

Recap statistical learning theory

Linear least squares

Error decomposition

Approximation Error

Estimation error
- Estimation Error I: probabilistic estimates
- Estimation Error II: deterministic estimates

Final bound

Ghost track: improved approximation results

Error bound for fixed λ

If $2\kappa^2\sqrt{\frac{2\log(4/\delta)}{n}} \leq \lambda \leq \kappa$, then

$$\left\|\Sigma^{1/2}(\widehat{w}_\lambda - w^\dagger)\right\|^2 \leq C^2\,\frac{\log(4/\delta)}{\lambda n} + 2\lambda\left\|w^\dagger\right\|$$

where $C = 2\sqrt{2}\left(M\kappa + \kappa^2\left\|w^\dagger\right\|\right)$.

- We can choose λ balancing these terms.
- We can obtain a bound $O(\frac{1}{\sqrt{n}})$.
- This choice of λ depends on quantities not known in practice, hold-out can be used instead.

Remarks

- The result is optimal in the considered setting.
- Better (fast) rates can be obtained with further assumptions (noise, smoothness eigendecay etc.).
- The bound is valid only for

$$2\kappa^2 \sqrt{\frac{2\log(2/\delta)}{n}} \leq \lambda$$

recent studies extend this analysis beyond this range (interpolation/benign overfitting).

Summary

- Generalization error for linear least squares.
- Dimension-free analysis applying to infinite-dimensional features/kernels.
- Only one of many possible regimes (faster/slower rates possible!).
- Many extensions (implicit regularization, benign overfitting, etc.).

Outline

Recap statistical learning theory

Linear least squares

Error decomposition

Approximation Error

Estimation error
- Estimation Error I: probabilistic estimates
- Estimation Error II: deterministic estimates

Final bound

Ghost track: improved approximation results

Improved approximation error

$$\|\Sigma^{1/2}(w_\lambda - w^\dagger)\|^2 \leq \lambda \|w^\dagger\|^2.$$

- The result is dimension free and sharp.
- It can be improved if we allow a dependence on[10] $\lambda_{\min}(\Sigma)$. If $\lambda \leq \lambda_{\min}(\Sigma)$, then

$$\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}\| \leq \frac{\sqrt{\lambda_{\min}(\Sigma)}}{\lambda_{\min}(\Sigma) + \lambda} \quad \Rightarrow \quad \|\Sigma^{1/2}(w_\lambda - w^\dagger)\| \leq \frac{\lambda}{\sqrt{\lambda_{\min}(\Sigma)}} \|w^\dagger\|.$$

- It can also be improved if $w^\dagger$ is *aligned* with Σ,

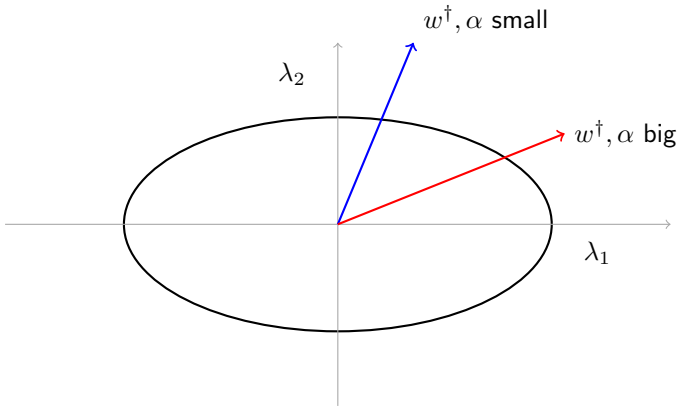
$$w^\dagger = \Sigma^\alpha v^\dagger \quad \Rightarrow \quad \|\Sigma^{1/2}(w_\lambda - w^\dagger)\| \leq \lambda^{\alpha+1/2} \|v^\dagger\|.$$

  This is called the source condition.

[10] $\lambda_{\min}(\Sigma)$ is always zero for infinite dimensional features!

Alignment geometry

$$w^\dagger = \Sigma^\alpha v^\dagger$$



Alignment as smoothness

$$w^\dagger = \Sigma^\alpha v^\dagger$$

- The above condition is equivalent to

$$v^\dagger = \Sigma^{-\alpha} w^\dagger = \sum_i \frac{(v_i^\top w^\dagger)}{\lambda_i^\alpha} v_i \qquad \Leftrightarrow \qquad \sum_i \frac{(v_i^\top w^\dagger)^2}{\lambda_i^{2\alpha}} < \infty$$

- For feature maps of translation/rotation invariant kernels, the source condition corresponds to decay in the Fourier domain.