Excess risk of OLS, proof

Proof: Using our previous computations:

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star &= \mathbb{E}\left[\left\|\hat{\theta} - \theta^\star\right\|_{\hat{\Sigma}}^2\right] \\
&= \mathbb{E}\left[\operatorname{trace}\left((\hat{\theta} - \theta^\star)^\top \hat{\Sigma}(\hat{\theta} - \theta^\star)\right)\right] && \text{(definition of } \|\cdot\|_{\hat{\Sigma}}\text{)} \\
&= \mathbb{E}\left[\operatorname{trace}\left((\hat{\theta} - \theta^\star)(\hat{\theta} - \theta^\star)^\top \hat{\Sigma}\right)\right] && \text{(cyclic property of the trace)} \\
&= \operatorname{trace}\left(\operatorname{Var}(\hat{\theta})\hat{\Sigma}\right) && \text{(linearity)} \\
&= \operatorname{trace}\left(\frac{\sigma^2}{n}\hat{\Sigma}^{-1}\hat{\Sigma}\right) && \text{(variance computation)} \\
&= \frac{\sigma^2}{n}\operatorname{trace}\left(\mathsf{I}_d\right)
\end{aligned}
$$

□

3.4. Ridge regression

Introduction

- **Reminder:** when $n \approx d$, OLS does not fare too good
- even more complicated when $d > n$
- yet, this is a common occurrence
- **Possible solution:** L^2 regularization

Definition: let $\lambda > 0$. With our notation, the ridge least-squares estimator $\hat{\theta}_\lambda$ is defined as the minimizer of

$$\frac{1}{n} \left\| Y - \Phi\theta \right\|^2 + \lambda \left\| \theta \right\|^2 .$$

- one can easily show the following:

Proposition: we have $\hat{\theta}_\lambda = \frac{1}{n}(\hat{\Sigma} + \lambda \, \mathrm{I}_d)^{-1} \Phi^\top Y$.

A note on invertibility

- in the previous proposition we inverted the matrix $M := \hat{\Sigma} + \lambda \mathrm{I}_d$
- **Why can we do that?**
- $\hat{\Sigma}$ is positive semi-definite, $\lambda \mathrm{I}_d$ "pushes" the spectrum in $\mathbb{R}_+^\star$
- more rigorously, if M was not invertible, one would have

$$\det\left(\frac{1}{n}\Phi^\top\Phi + \lambda \mathrm{I}_d\right) = 0\,.$$

- meaning that $-\lambda$ would be an eigenvalue of $\Phi^\top\Phi$: this is not possible
- **Note:** this was the main motivation when first introduced[5]

[5] Hoerl, Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, 1970

Fixed design analysis

- as with OLS, we can compute the expected excess risk
- only a bit more complicated because of the regularization...
- bias-variance decomposition still holds:

Proposition (ridge bias-variance decomposition): Let $\hat{\theta}_\lambda$ as before. Under assumption I and II,

$$\mathbb{E}[\mathcal{R}(\hat{\theta}_\lambda)] - \mathcal{R}^\star = \left\| \mathbb{E}[\hat{\theta}_\lambda] - \theta^\star \right\|_{\hat{\Sigma}}^2 + \mathbb{E}\left[\left\| \hat{\theta}_\lambda - \mathbb{E}[\hat{\theta}_\lambda] \right\|_{\hat{\Sigma}}^2 \right]$$

- *Proof:* did not depend on $\hat{\theta}$'s exact expression □

Rewriting $\mathbb{E}[\hat{\theta}_\lambda]$

- we will then use the following:

Lemma: Let $\hat{\theta}_\lambda$ be the ridge regressor. Assume that I and II hold. Then

$$\mathbb{E}[\hat{\theta}_\lambda] = \theta^\star - \lambda(\hat{\Sigma} + \lambda \, \mathrm{I}_d)^{-1}\theta^\star \,.$$

- *Proof:*

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_\lambda] &= \mathbb{E}\left[\frac{1}{n}(\hat{\Sigma} + \lambda \, \mathrm{I}_d)^{-1}\Phi^\top Y\right] && (\text{def. of } \hat{\theta}_\lambda) \\
&= \mathbb{E}\left[\frac{1}{n}(\hat{\Sigma} + \lambda \, \mathrm{I}_d)^{-1}\Phi^\top (\Phi\theta^\star + \varepsilon)\right] && (\text{assumption I}) \\
&= \frac{1}{n}(\hat{\Sigma} + \lambda \, \mathrm{I}_d)^{-1}\Phi^\top \Phi\theta^\star && (\text{linearity} + \varepsilon \text{ centered})
\end{aligned}$$

Rewriting $\mathbb{E}[\hat{\theta}_\lambda]$

- now, by definition of $\hat{\Sigma}$,

$$\mathbb{E}[\hat{\theta}_\lambda] = (\hat{\Sigma} + \lambda \, \mathrm{I}_d)^{-1} \hat{\Sigma} \theta^\star \, .$$

- finally, since for any matrix A

$$(A + \lambda \, \mathrm{I})^{-1} A = \mathrm{I} - \lambda (A + \lambda \, \mathrm{I})^{-1} \, ,$$

we deduce the result. □

Excess risk

Proposition (ridge excess risk): assume I and II, let $\hat{\theta}_\lambda$ as before. Then

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta}_\lambda)\right] - \mathcal{R}^\star = \lambda^2 \left(\theta^\star\right)^\top (\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2}\hat{\Sigma}\theta^\star + \frac{\sigma^2}{n}\mathrm{trace}\left(\hat{\Sigma}^2(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2}\right) .$$

- **Remark (i):** when $\lambda \to 0$, we recover the OLS result
- **Remark (ii):** we have an exact description of the bias / variance evolution w.r.t. λ (!)
- **Remark (iii):** bias increases with λ, variance decreases, $\lambda = 0$ not optimal (in general)
- **Remark (iv):** the quantity $\mathrm{trace}\left(\hat{\Sigma}^2(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-2}\right)$ is called "degrees of freedom" $\approx$ implicit number of parameters

Excess risk, proof

- *Proof:* we plug the alternative expression of $\mathbb{E}[\hat{\theta}_\lambda]$ into the bias / variance decomposition
- the bias term is clear, variance yields

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\theta}_\lambda - \mathbb{E}[\hat{\theta}_\lambda]\right\|_{\hat{\Sigma}}^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n}(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-1}\Phi^\top \varepsilon\right\|_{\hat{\Sigma}}^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\mathrm{trace}\left(\varepsilon^\top \Phi(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-1}\Phi^\top \varepsilon\right)\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\mathrm{trace}\left(\Phi^\top \varepsilon\varepsilon^\top \Phi(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-1}\right)\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(trace cyclic property)} \\
&= \frac{\sigma^2}{n}\mathrm{trace}\left(\hat{\Sigma}(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda \, \mathsf{I}_d)^{-1}\right) . \qquad \left(\mathbb{E}\left[\varepsilon\varepsilon^\top\right] = \sigma^2 \, \mathsf{I}_d\right)
\end{aligned}$$

Excess risk, proof

- finally, since

$$(\hat{\Sigma} + \lambda \mathrm{I}_d)(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1} = (\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}(\hat{\Sigma} + \lambda \mathrm{I}_d) = \mathrm{I}_d \,,$$

  we deduce that

$$\hat{\Sigma}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1} = (\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\hat{\Sigma} \left(= \mathrm{I}_d - \lambda(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1} \right) .$$

- together with the trace cyclic property, this allows us to write

$$\mathrm{trace}\left(\hat{\Sigma}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-1} \right) = \mathrm{trace}\left(\hat{\Sigma}^2(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-2} \right)$$

  and to conclude. □

Choice of regularization

Proposition (choice of regularization parameter): Assume that I and II hold. Set

$$\lambda^\star := \frac{\sigma \operatorname{trace}(\hat{\Sigma})^{1/2}}{\|\theta^\star\| \sqrt{n}}$$

as regularization parameter. Then

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta}_{\lambda^\star})\right] - \mathcal{R}^\star \leq \frac{\sigma \operatorname{trace}(\hat{\Sigma})^{1/2} \|\theta^\star\|}{\sqrt{n}} .$$

- **Remark (i):** of course, in practice, we know neither σ, nor $\theta^\star$...
- **Remark (ii):** $\lambda^\star$ maybe not optimal for the true risk
- **Remark (iii):** slower rate of convergence, but σ instead of σ^2

Choice of regularization, proof

- we take for granted that all eigenvalues of $\lambda(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-2}\hat{\Sigma}$ are smaller than $1/2$
- as a consequence:

$$\begin{aligned} B &= \lambda^2 (\theta^\star)^\top (\hat{\Sigma} + \lambda \mathrm{I}_d)^{-2} \hat{\Sigma} \theta^\star \\ &= \lambda (\theta^\star)^\top \left[(\hat{\Sigma} + \lambda \mathrm{I}_d)^{-2} \hat{\Sigma} \right] \theta^\star \\ &\leq \frac{\lambda}{2} \left\| \theta^\star \right\|^2 . \end{aligned}$$

- in the same fashion:

$$\begin{aligned} V &= \frac{\sigma^2}{n} \operatorname{trace}\left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda \mathrm{I}_d)^{-2} \right) \\ &= \frac{\sigma^2}{\lambda n} \operatorname{trace}\left(\hat{\Sigma} \left(\lambda (\hat{\Sigma} + \lambda \mathrm{I}_d)^{-2} \hat{\Sigma} \right) \right) \leq \frac{\sigma^2}{\lambda n} \operatorname{trace}\left(\hat{\Sigma} \right) . \end{aligned}$$

Proof, ctd.

- putting both bounds together, we get

$$\mathbb{E}\left[\hat{\mathcal{R}}(\hat{\theta}_\lambda)\right] - \mathcal{R}^\star \leq \frac{\lambda}{2}\left\|\theta^\star\right\|^2 + \frac{\sigma^2}{2\lambda n}\mathrm{trace}\left(\hat{\Sigma}\right) .$$

- minimizing in λ yields

$$\lambda^\star = \frac{\sigma \mathrm{trace}\left(\hat{\Sigma}\right)^{1/2}}{\left\|\theta^\star\right\|\sqrt{n}} ,$$

as expected. □

Dimension free bound?

- recall that our upper bound reads

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta}_{\lambda^\star})\right] - \mathcal{R}^\star \leq \frac{\sigma \operatorname{trace}(\hat{\Sigma})^{1/2} \left\|\theta^\star\right\|}{\sqrt{n}} .$$

- no explicit dependency in d
- under some assumptions (e.g., sparsity), $\left\|\theta^\star\right\| \ll d$
- moreover, if $\left\|\varphi(x)\right\| \leq R$,

$$\begin{aligned} \operatorname{trace}\left(\hat{\Sigma}\right) &= \sum_{j=1}^{d} \hat{\Sigma}_{j,j} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \varphi(x_i)_j^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\|\varphi(x_i)\right\|^2 \leq R^2 . \end{aligned}$$

3.5. Random design analysis

Random design analysis

- back to random design: (X_i, Y_i) i.i.d. from some distribution p on $\mathcal{X} \times \mathcal{Y}$
- **Goal:** prove the same excess risk bound (i.e., $\approx \frac{\sigma^2 d}{n}$)
- **Important:** we make the same assumptions, transposed to the random design setting:
  - **Assumption I:** $\exists \theta^\star \in \mathbb{R}^d$ such that

$$\forall i \in [n], \qquad Y_i = \varphi(X_i)^\top \theta^\star + \varepsilon_i \,,$$

  - **Assumption II:** the noise distribution of ε_i is independent from that of X_i, $\mathbb{E}[\varepsilon_i] = 0$, and $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$.
- notable consequence of our assumptions:

$$\mathbb{E}[Y_i \mid X_i] = \varphi(X_i)^\top \theta^\star \,.$$

Excess risk

- the excess risk has a similar decomposition:

Proposition (excess risk for random design least-squares regression): Assume that I and II hold. Then $\mathcal{R}^\star = \sigma^2$, and

$$\forall \theta \in \mathbb{R}^d, \qquad \mathcal{R}(\theta) - \mathcal{R}^\star = \|\theta - \theta^\star\|_\Sigma^2 \ ,$$

where $\Sigma := \mathbb{E}\left[\varphi(X)\varphi(X)^\top\right]$.

- **Intuition:** $\hat{\Sigma}$ is replace by its expectation, which is Σ
- (recall that $\hat{\Sigma} = \frac{1}{n}\Phi^\top\Phi$)

Excess risk, proof

- **Proof:** let (X_0, Y_0) be a "new" observation, with noise ε_0

$$\begin{aligned}\mathcal{R}(\theta) &= \mathbb{E}\left[(Y_0 - \theta^\top \varphi(X_0))^2\right] \\ &= \mathbb{E}\left[(\varphi(X_0)^\top \theta^\star + \varepsilon_0 - \theta^\top \varphi(X_0))^2\right] \qquad \text{(AI)} \\ &= \mathbb{E}\left[(\varphi(X_0)^\top \theta^\star - \theta^\top \varphi(X_0))^2\right] + 2\mathbb{E}\left[\varepsilon_0 (\theta^\star - \theta)^\top \varphi(X_0)\right] + \mathbb{E}\left[\varepsilon_0^2\right]\end{aligned}$$

- by independence, and since the noise is centered,

$$\mathbb{E}\left[\varepsilon_0 (\theta^\star - \theta)^\top \varphi(X_0)\right] = \mathbb{E}\left[\varepsilon_0\right] \mathbb{E}\left[(\theta^\star - \theta)^\top \varphi(X_0)\right] = 0\,.$$

- now we can conclude:

$$\begin{aligned}\mathcal{R}(\theta) &= \mathbb{E}\left[((\theta^\star - \theta)^\top \varphi(X_0))^2\right] + \mathbb{E}\left[\varepsilon_0^2\right] && \text{(AII)} \\ &= (\theta - \theta^\star)^\top \mathbb{E}\left[\varphi(X_0)\varphi(X_0)^\top\right](\theta - \theta^\star) + \sigma^2 && \text{(linearity)} \\ &= (\theta - \theta^\star)^\top \Sigma (\theta - \theta^\star) + \sigma^2\,. \quad \square && \text{(definition of } \Sigma\text{)}\end{aligned}$$

Excess risk of OLS

- we now use the previous result to investigate $\hat{\theta}$:

Proposition: Assume that I and II hold. Assume further that $\hat{\Sigma}$ is almost surely invertible. Then the expected excess risk of the OLS estimator is equal to

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star = \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{trace}\left(\Sigma\hat{\Sigma}^{-1}\right)\right] .$$

- **Remark (i):** $\hat{\Sigma}$ has the same definition, but is now a *random* quantity
- **Remark (ii):** under reasonable assumptions (e.g., density), $\hat{\Sigma}$ is almost surely invertible
- **Intuition:** $\det(\hat{\Sigma}) = 0$ is a “zero-measure” condition

Excess risk of OLS, proof

- from the definition of $\hat{\theta}$,

$$\hat{\theta} = \frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top Y = \frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top(\Phi\theta^\star + \varepsilon) = \theta^\star + \frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\,.$$

- using the previous result:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star &= \mathbb{E}\left[\left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\right)^\top \Sigma \left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\right)\right] \\
&= \mathbb{E}\left[\operatorname{trace}\left(\Sigma\left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\right)\left(\frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\varepsilon\right)^\top\right)\right] \qquad \text{(cyclic property)} \\
&= \frac{1}{n^2}\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi\hat{\Sigma}^{-1}\right)\right]
\end{aligned}$$

Excess risk of OLS, proof ctd.

- now we use properties of the conditional expectation:

$$\begin{aligned}
\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi\hat{\Sigma}^{-1}\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi\hat{\Sigma}^{-1}\right) \mid X_1,\dots,X_n\right]\right] \\
&\qquad \text{(tower property)} \\
&= \mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\mathbb{E}\left[\varepsilon\varepsilon^\top \mid X_1,\dots,X_n\right]\Phi\hat{\Sigma}^{-1}\right)\right] \\
&\qquad (\Phi, \hat{\Sigma} \text{ are } X_1,\dots,X_n\text{-measurable}) \\
&= \mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\mathbb{E}\left[\varepsilon\varepsilon^\top\right]\Phi\hat{\Sigma}^{-1}\right)\right] \quad \text{(independence)} \\
&= \sigma^2\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\Phi^\top\Phi\hat{\Sigma}^{-1}\right)\right] \qquad \left(\mathbb{E}\left[\varepsilon\varepsilon^\top\right] = \sigma^2 \, \mathrm{I}_d\right) \\
&= \sigma^2\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\right)\right].
\end{aligned}$$

□

Gaussian design

- to be more precise, we need to specify a distribution for the $\varphi(X_i)$s

Proposition: Assume that I and II hold. Assume further that $\varphi(X) \sim \mathcal{N}(0, \Sigma)$. Then the expected risk of OLS is given by

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^\star = \frac{\sigma^2 d}{n - d - 1} .$$

- **Remark:** we (nearly) recover the $\sigma^2 d/n$ bound from fixed design!

Gaussian design, proof

- define $Z := \Sigma^{-1/2}\varphi(X)$
- properties of Gaussian vectors: $Z \sim \mathcal{N}(0, \mathrm{I}_d)$
- we see that

$$\mathbb{E}\left[\operatorname{trace}\left(\Sigma\hat{\Sigma}^{-1}\right)\right] = \operatorname{trace}\left(\mathbb{E}\left[\Sigma(\Sigma^{1/2}Z\Sigma^{1/2}Z^\top)^{-1}\right]\right)$$
$$= \operatorname{trace}\left(\mathbb{E}\left[(ZZ^\top)^{-1}\right]\right) .$$

- $(ZZ^\top)^{-1}$ has the *inverse Wishart distribution*
- we read in the tables:

$$\mathbb{E}\left[(ZZ^\top)^{-1}\right] = \frac{1}{n-d-1}\mathrm{I}_d$$

and conclude. □

4. Generalization bounds

Reminder: risk decomposition

- **Reminder:**

$$\mathcal{R}(f) - \mathcal{R}^\star = \left[\mathcal{R}(f) - \inf_{h \in \mathcal{H}} \mathcal{R}(h)\right] + \left[\inf_{h \in \mathcal{H}} \mathcal{R}(h) - \mathcal{R}^\star\right]$$

$$\text{excess risk} = \text{estimation error} + \text{approximation error}$$

- **Estimation error:**
  - always non-negative
  - random if there is randomness in the creation of f
  - characterizes how much we loose by picking the wrong predictor in a given class
- **Approximation error:**
  - deterministic, does not depend on f, **only on the class of functions** $\mathcal{H}$
  - characterizes how much we loose by restricting ourselves to a given class

Decomposition of the estimation error

- **Notation (i):** $f_{\mathcal{H}} \in \arg\min_{f \in \mathcal{H}} \mathcal{R}(f)$, best predictor in our function class
- **Notation (ii):** $\hat{f}$ empirical risk minimizer
- **Useful decomposition:**

$$
\begin{aligned}
\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) &= \mathcal{R}(\hat{f}) - \mathcal{R}(f_{\mathcal{H}}) && \text{(def. of } f_{\mathcal{H}}) \\
&= \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) + \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f_{\mathcal{H}}) + \hat{\mathcal{R}}(f_{\mathcal{H}}) - \mathcal{R}(f_{\mathcal{H}}) \\
&\leq \sup_{f \in \mathcal{H}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} + \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f_{\mathcal{H}}) + \sup_{f \in \mathcal{H}} \left\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right\}
\end{aligned}
$$

- middle term is ≤ 0 by definition, and we get

$$
\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) \leq 2 \sup_{f \in \mathcal{H}} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right| .
$$

Decomposition of the estimation error, ctd.

- **Remark (i):** no more dependency in $\hat{f}$, we only need to control functions (but we do need uniform control)
- **Remark (ii):** if $\hat{f}$ not global minimizer, say

$$\hat{\mathcal{R}}(\hat{f}) \leq \inf_{f\in\mathcal{H}} \hat{\mathcal{R}}(f) + \varepsilon\,,$$

  we need to add ε to our bound
- **Remark (iii):** bound usually grows with size of $\mathcal{H}$ and decreases with n

4.1. Uniform bounds via concentration

Single function

- when there is a single function f_0 in $\mathcal{H}$, we have already seen how to control

$$\sup_{f\in\mathcal{H}} \left|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\right| = \hat{\mathcal{R}}(f_0) - \mathcal{R}(f_0) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) - \mathbb{E}\left[\ell(Y, f(X))\right] .$$

- indeed, since the observations are i.i.d., we can use Hoeffding's inequality (Exercise sheet 1):

Proposition: for any $\delta \in (0, 1/2)$, with probability greater than $1 - \delta$,

$$\mathcal{R}(f_0) - \hat{\mathcal{R}}(f_0) < \frac{\ell_\infty\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}} ,$$

where ℓ_∞ is an upper bound on $\ell(Y_i, f(X_i))$.

From sup to expectation

- **Problem:** there is often more than one function in $\mathcal{H}$...
- still possible, using for instance:

Proposition (McDiarmid's inequality): Let $Z_1, \ldots, Z_n$ be independent random variables and F a function such that

$$|F(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n) - F(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c \, .$$

Then

$$\mathbb{P}\left(|F(Z_1, \ldots, Z_n) - \mathbb{E}\left[F(Z_1, \ldots, Z_n)\right]| \geq t\right) \leq 2 \exp\left(-2t^2/(nc^2)\right) \, .$$

Application of McDiarmid

- set $Z_i := (X_i, Y_i)$, and

$$H(Z_1, \ldots, Z_n) := \sup_{f \in \mathcal{H}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} .$$

- Mc Diarmid tells us that, with probability higher than $1 - \delta$,

$$H(Z_1, \ldots, Z_n) - \mathbb{E}\left[H(Z_1, \ldots, Z_n) \right] \leq \frac{\ell_\infty \sqrt{2}}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}} .$$

- getting bound on $\mathbb{E}\left[H(Z_1, \ldots, Z_n) \right]$ automatically yields bound on $\sup_{f \in \mathcal{H}} \left\{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right\}$
- by symmetry, upper bound on $\sup_{f \in \mathcal{H}} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right|$

4.2. Rademacher complexity

Rademacher complexity

- set $Z := (X, Y)$ and $\mathcal{G} := \{(x, y) \mapsto \ell(y, f(x))\}$, with f in some function class $\mathcal{H}$
- **Recall:** we want to bound

$$\sup_{f \in \mathcal{H}} \left\{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \right\} = \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}\left[g(Z)\right] - \frac{1}{n} \sum_{i=1}^{n} g(Z_i) \right\} .$$

- set $\mathcal{D} := \{Z_1, \ldots, Z_n\}$ the data

Definition: We call *Rademacher complexity* of the function class $\mathcal{G}$ the quantity

$$R_n(\mathcal{G}) := \mathbb{E}_{\varepsilon, \mathcal{D}} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(Z_i) \right] ,$$

where the ε_is are independent Rademacher random variables (that is, $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$).

Rademacher complexity, first properties

- **Intuition:** expectation of maximal dot-product with random labels
- measures the *capacity* of the set $\mathcal{G}$

Properties: Rademacher complexity satisfies the following properties:

- if $\mathcal{G} \subset \mathcal{G}'$, then $R_n(\mathcal{G}) \leq R_n(\mathcal{G}')$;
- $R_n(\mathcal{G} + \mathcal{G}') = R_n(\mathcal{G}) + R_n(\mathcal{G}')$;
- $R_n(\alpha\mathcal{G}) = |\alpha| \, R_n(\mathcal{G})$;
- if g_0 is a function, $R_n(\mathcal{G} + \{g_0\}) = R_n(\mathcal{G})$;
- $R_n(\mathcal{G}) = R_n(\mathrm{conv}(\mathcal{G}))$.

Symmetrization

- **Question:** why is it useful?
- Rademacher complexity directly controls expected uniform deviation

Proposition (symmetrization): With the previous notation,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\left[g(Z)\right]\right\}\right] \leq 2R_n(\mathcal{G})\,,$$

and

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right\}\right] \leq 2R_n(\mathcal{G})\,.$$

Symmetrization, proof

- let $\mathcal{D}' := \{Z_1', \ldots, Z_n'\}$ be an independent copy of $\mathcal{D}'$
- in particular, one has $\mathbb{E}\left[g(Z_i') \mid \mathcal{D}\right] = \mathbb{E}\left[g(Z)\right]$
- we write

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right\}\right] = \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(Z_i') \mid \mathcal{D}\right] - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right\}\right]$$
$$= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[g(Z_i') - g(Z_i) \mid \mathcal{D}\right]\right\}\right] .$$

Symmetrization, proof ctd.

- since the sup of expectation is $\leq$ than expectation of the sup,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\mathbb{E}\left[g(Z)\right]-\frac{1}{n}\sum_{i=1}^{n}g(Z_i)\right\}\right]\leq\mathbb{E}\left[\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_i')-g(Z_i))\right\}\mid\mathcal{D}\right]\right]$$
$$=\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_i')-g(Z_i))\right\}\right]$$

  by the tower property.
- we notice that

$$g(Z_i')-g(Z_i)\quad\text{and}\quad\varepsilon_i(g(Z_i')-g(Z_i))\text{ have the same distribution}$$

  (this is what we call symmetrization)

Symmetrization proof, ctd.

- thus

$$\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}(g(Z_i') - g(Z_i))\right\}\right] &= \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(g(Z_i') - g(Z_i))\right\}\right] \\
&\leq \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)\right\}\right] + \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^{n}-\varepsilon_i g(Z_i)\right\}\right] \\
&= 2R_n(\mathcal{G})
\end{aligned}$$

since ε and $-\varepsilon$ have the same distribution. □

Example: linear predictors

- let Ω be a norm on $\mathbb{R}^d$
- assume $\mathcal{H} = \{\theta^\top \varphi(x), \Omega(\theta) \leq D\}$
- then

$$
\begin{aligned}
R_n(\mathcal{H}) &= \mathbb{E}\left[\sup_{\Omega(\theta)\leq D} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \theta^\top \varphi(X_i)\right] \\
&= \mathbb{E}\left[\sup_{\Omega(\theta)\leq D} \frac{1}{n} \varepsilon^\top \Phi\theta\right] \\
&= \frac{D}{n}\mathbb{E}\left[\Omega^\star(\Phi^\top \varepsilon)\right] ,
\end{aligned}
$$

where $\Omega^\star$ is the *dual norm* of Ω:

$$
\Omega^\star(u) := \sup_{\Omega(\theta)\leq 1} u^\top \theta .
$$

Example: linear predictors, ctd.

- when $p \in [1, +\infty)$ and Ω is the p-norm, $\Omega^\star$ is the q-norm with $1/p + 1/q = 1$
- $\Rightarrow$ **Rademacher complexity computations boil down to expected norm computations**
- let us do this for the 2-norm:

$$
\begin{aligned}
R_n(\mathcal{H}) &= \frac{D}{n}\mathbb{E}\left[\left\|\Phi^\top \varepsilon\right\|\right] \\
&\leq \frac{D}{n}\sqrt{\mathbb{E}\left[\left\|\Phi^\top \varepsilon\right\|^2\right]} \qquad \text{(Jensen's inequality)} \\
&= \frac{D}{n}\sqrt{\mathbb{E}\left[\operatorname{trace}\left(\Phi^\top \varepsilon\varepsilon^\top \Phi\right)\right]} \\
&= \frac{D}{n}\sqrt{\mathbb{E}\left[\operatorname{trace}\left(\Phi^\top \Phi\right)\right]} = \frac{D}{n}\sqrt{\sum_{i=1}^{n}\mathbb{E}\left[(\Phi^\top\Phi)_{i,i}\right]} = \frac{D}{n}\sqrt{\sum_{i=1}^{n}\mathbb{E}\left[\left\|\varphi(X_i)\right\|^2\right]} \\
&= \frac{D}{\sqrt{n}}\sqrt{\mathbb{E}\left[\left\|\varphi(x)\right\|^2\right]}
\end{aligned}
$$

$\Rightarrow$ dimension-free bound with the same rate!

Example: linear predictors, ctd.

- we can get a bound on the estimation error:

Proposition: assume that ℓ is L-Lipschitz and continuous. Consider linear predictors with bounded coefficients, that is, $f_\theta(x) = \theta^\top \varphi(x)$ with $\|\theta\| \leq D$. Assume further that $\mathbb{E}\left[\|\varphi(X)\|^2\right] \leq R^2$. Let $\hat{f}$ be the empirical risk minimizer. Then

$$\mathbb{E}\left[\mathcal{R}(\hat{f})\right] \leq \inf_{\|\theta\| \leq D} \mathcal{R}(f_\theta) + \frac{4LRD}{\sqrt{n}} .$$

- **Remark (i):** does not depend on exact expression of the loss
- **Remark (ii):** does not depend on the dimension