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

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

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

Concentration inequalities

- informally speaking: random variable is "close" to its expectation with high probability
- **Example:** Markov, Chebyshev
- more involved:

Proposition (Hoeffding's inequality): let $Z_1, \ldots, Z_n$ be independent random variables such that $Z_i \in [0,1]$ almost surely, then, for any $t \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}\left[Z_i\right])\right| \geq t\right) \leq 2\exp\left(-2nt^2\right) .$$

Single function

- assume that $\mathcal{H} = \{f_0\}$ and ℓ a bounded loss function
- then we can control

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

- indeed, since the observations are i.i.d., we can use Hoeffding on the $Z_i := \ell(Y_i, f_0(X_i))$
- common expectation $= \mathcal{R}(f_0)$
- for any $\delta \in (0, 1/2)$,

$$\mathbb{P}\left(\left| \hat{\mathcal{R}}(f_0) - \mathcal{R}(f_0) \right| \geq \frac{1}{\sqrt{2n}} \sqrt{\log \frac{1}{\delta}} \right) \leq 2\exp\left(-2n \frac{1}{2n} \log 1/\delta \right) = 2\delta .$$

Single function

- scaling by ℓ_∞, we obtain:

Proposition: Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. observations of p and f_0 be a fixed predictor. Then, for any $\delta \in (0, 1/2)$, with probability greater than $1 - 2\delta$,

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

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, $\mathbb{P}\left(\varepsilon_i = \pm 1\right) = 1/2$).

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. $\square$

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

Proof of the proposition

- recall the decomposition of the estimation error:

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

- by symmetrization:

$$\mathbb{E}\left[\mathcal{R}(\hat{f})\right] - \inf_{f \in \mathcal{H}} \mathcal{R}(f) \leq 4 R_n(\mathcal{H}) .$$

- set $\mathcal{F} := \{f_\theta, \|\theta\| \leq D\}$. Since the loss is L-Lipschitz, by contraction (*see exercise*),

$$R_n(\mathcal{H}) \leq L R_n(\mathcal{F}) .$$

- by previous computation,

$$R_n(\mathcal{F}) \leq \frac{DR}{\sqrt{n}} .$$

□