Lecture 5: Basics on Generalization Error

Last lecture:
- we constructed linear approximation spaces in high, infinite-dimensional inputs with kernels.
- efficient learning algorithms (regularized least-squares)

Today:
- Basic tools to control generalization error.

Reminder: Next week guest lecture by Alberto Bietti

→ Focus on ERM (Empirical Risk Minimisation). Let $\nu = \nu(x,y)$ be the joint data distribution over $X \times Y$.

Goal: Learn $f: X \to Y$ that minimises a risk

$$R(f) = \mathbb{E}_\nu [\ell(y, f(x))] \qquad \ell: \text{squared-loss, logistic loss, etc.}$$

→ ERM: Consider a normed function class $\mathcal{F} \subset \{f: X \to Y\}$ and consider an estimator (ERM) of the form

$$\hat{f} \in \underset{\|f\|_{\mathcal{F}} \leq \delta}{\operatorname{argmin}} \ \hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)).$$

→ Recall the fundamental decomposition of the risk:

$$R(\hat{f}) - \inf_g R(g) = \underbrace{R(\hat{f}) - \inf_{f \in \mathcal{F}_\delta = \mathcal{F}} R(f)}_{\text{estimation error}} + \underbrace{\inf_{f \in \mathcal{F}} R(f) - \inf_g R(g)}_{\text{approximation error}}$$

from now on.

→ Estimation error decomposed in terms of ERM:

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) = R(\hat{f}) - R(g) = R(\hat{f}) - \hat{R}(\hat{f}) + \hat{R}(\hat{f}) - \hat{R}(g) + \hat{R}(g) - R(g)$$

$$\leq \sup_{f \in \mathcal{F}} \left(R(f) - \hat{R}(f) \right) + \sup_{f \in \mathcal{F}} \left(\hat{R}(f) - R(f) \right)$$

where $g \in \operatorname{argmin}_{f \in \mathcal{F}} R(f)$

→ When $\hat{f}$ is not the exact minimizer, then we need to add optim. error.

→ Q: What is the expected / intuitive behavior of this estimation error term?

→ Recall that at fixed f, we have

$\hat{R}(f)$ is unbiased estimator of $R(f)$

$|\hat{R}(f) - R(f)|$ are $O(1/\sqrt{n})$ from LLN (Law of Large numbers)

CLT: approximate Gaussianity.

Q: How to go from the (LLN / CLT) view at fixed f to

(i) a uniform deviation bound?

(ii) How to give non-asymptotic guarantees?

Finite Function Classes. Assume that the loss is bounded a.s. ($\hat{R}(f) \leq A$ with proba 1 - / $R(f) \leq A$)

· The simplest tool is the union bound:

$$\mathbb{P}\left(R(\hat{f}) - R(f) \geq t\right) \leq \mathbb{P}\left(2 \sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| \geq t\right)$$

$$\left(\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(|\hat{R}(f) - R(f)| \geq t/2\right)\right)$$

↳ union bound.

→ For any fixed $f \in \mathcal{F}$, $(y, x) \sim \nu$

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

$\ell(y, f(x))$ is a bounded RV.

$\hat{R}(f)$ is an average of bounded, iid random variables.

Prop (Hoeffding ineq): For $z_1 \dots z_n$ iid random variables in $[0, c]$, then $\forall t \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^{n} z_i - \mathbb{E}z\right| \geq t\right) \leq 2 \exp\left(-2nt^2/c^2\right)$$

so using Hoeffding

$$\mathbb{P}\left(|\hat{R}(f) - R(f)| \geq t/2\right) \leq 2 \exp\left(\frac{-nt^2}{2A^2}\right)$$

$\left(\ell(y, f(x)) \leq A \text{ a.s. } \forall f \in \mathcal{F}.\right)$ for each $f \in \mathcal{F}$.

so, if $g \in \operatorname{argmin}_{f \in \mathcal{F}} R(f)$,

$$\mathbb{P}\left(R(\hat{f}) - R(g) \geq t\right) \leq \sum_{f \in \mathcal{F}} 2 \exp\left(\frac{-nt^2}{2A^2}\right) =$$

$$= 2|\mathcal{F}| \cdot \exp\left(\frac{-nt^2}{2A^2}\right)$$

→ Now, we set $\delta = 2|\mathcal{F}| \exp\left(\frac{-nt^2}{2A^2}\right)$ and solve for t;

→ We conclude that with probability at least $1-\delta$,

$$R(\hat{f}) - R(g) \leq \frac{2A}{\sqrt{n}} \left(\sqrt{\log \frac{2|\mathcal{F}|}{\delta}} \right)$$

$$\sqrt{\log |\mathcal{F}|} + \sqrt{\log \delta^{-1}}$$

corresponds to looking at "extremal" statistics

↳ term that comes from concentration of measure.

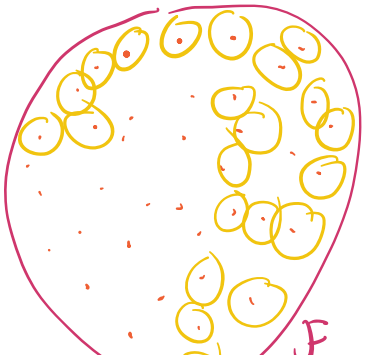
- Remark: Going from expected excess risk $\mathbb{E}[R(\hat{f}) - R(g)]$ to high-probability bounds uses "classic" tools; eg MacDiarmid ineq.

- Remark: The term $\sqrt{\log |\mathcal{F}|}$ appears when computing the expectation of the max of $|\mathcal{F}|$ zero-mean, subgaussian indep random variables.

$$\mathbb{E} \max_i |Z_i| \leq \sigma \sqrt{2 \log n}$$

Q: How to extend to infinite function classes?

idea: let's grid the domain!



→ Assume both $R, \hat{R}$ are β-Lipschitz with respect to a distance metric d over

$\forall f, \tilde{f}: \quad |R(f) - R(\tilde{f})| \leq \beta \cdot \text{dist}(f, \tilde{f})$.

$f_\theta \quad f_{\tilde{\theta}} \quad (\theta, \tilde{\theta})$

→ Fix $\varepsilon > 0$. Assume there exists m_ε elements $f_1 \dots f_{m_\varepsilon}$ such that $\sup_{f \in F} \inf_i d(f, f_i) \leq \varepsilon$. The smallest m_ε satisfying this property is the covering number of F.

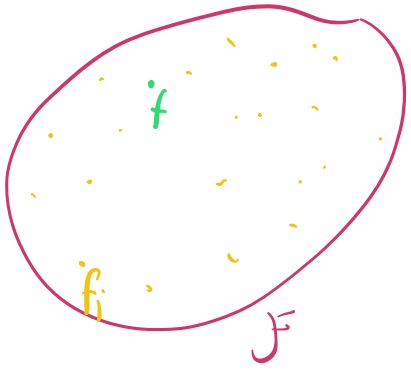
→ Q: m_ε blows up as $\varepsilon \to 0$. For parametrised function classes $F = \{ f_\theta ; \theta \in \Theta \subseteq \mathbb{R}^d \}$ we expect

$m_\varepsilon \sim \varepsilon^{-\tilde{d}}$ ($\|\theta\| \leq D$)

$\tilde{d}$ is the "intrinsic" dimension of F.

→ Let's proceed trying to reuse our previous finite analysis:

$\sup_{f \in F} |\hat{R}(f) - R(f)| \leq$



$\leq \sup_{f \in F} \{ |\hat{R}(f) - \hat{R}(f_i)| + |\hat{R}(f_i) - R(f_i)| + |R(f_i) - R(f)| \}$

(≤) $2\beta \cdot \text{dist}(f, f_i) + \sup_i |\hat{R}(f_i) - R(f_i)|$

↳ R and $\hat{R}$ are β-Lipschitz.

→ Since Bounded random variables in particular are subgaussian

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)|\right] \le 2\beta\varepsilon + \mathbb{E}\left\{\sup_{i=1,\dots,m_\varepsilon} |\underbrace{\hat{R}(f_i) - R(f_i)}_{\text{std is } \frac{A}{\sqrt{n}}}|\right\}$$

$$\le 2\beta\varepsilon + A\sqrt{\frac{2\log(2m_\varepsilon)}{n}}$$

→ So, if $m_\varepsilon \sim \varepsilon^{-d}$, then optimising this upper bound wrt ε gives $\varepsilon \sim 1/\sqrt{n}$ and a bound

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)|\right] = O\left(\sqrt{\left(\frac{d}{n}\right)\log n}\right)$$

* → Dependency on d; is this real or artifact of the proof?

→ $\ell(y, f(x))$ bounded for eg. squared loss.
+ Lipschitz.
it is Lipschitz because $\|f\| \le \widetilde{A}$
and (y,x) is bounded a.s.

A: We have been a bit sloppy.

→ V-C dimension is one tool that gives a generalised notion of cardinality, via the "shattering" notion.
↳ Not very popular these days to analyse large overparametrised models such as NNs.

→ Instead, we will use <u>Rademacher Complexity</u>

→ Let $z = (x, y) \in Z$ and consider a class $\mathcal{H}$ of functions from Z to $\mathbb{R}$

functions from $\mathcal{Z}$ to $\mathbb{R}$.

$$H = \{ (x,y) \mapsto \ell(y, f(x)) \text{ for } f \in \mathcal{F} \}.$$

→ we want to control $\sup_{f \in \mathcal{F}} R(f) - \hat{R}(f) =$

$$= \sup_{h \in H} \{ \mathbb{E} h(z) - \frac{1}{n} \sum_{i=1}^{n} h(z_i) \}$$

Def: (Rademacher complexity) $R_n(H)$ given data $D = \{ z_1 \dots z_n \}$ is

$$R_n(H) := \mathbb{E}_{D,\varepsilon} \left(\sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \cdot h(z_i) \right)$$

where $\varepsilon_1 \dots \varepsilon_n$ are Rademacher independent r.v.
$\varepsilon_i = \begin{cases} +1 & \text{with prob } 1/2 \\ -1 & 1/2 \end{cases}$

↳ Measuring how well can functions in H correlate with random labels.

Proposition (Symmetrisation) we have

$$\mathbb{E} \left[\sup_{h \in H} \left(\frac{1}{n} \sum_i h(z_i) - \mathbb{E} h(z) \right) \right] \leq 2 \cdot R_n(H)$$

and $$\mathbb{E} \left[\sup_{h \in H} \left(\mathbb{E} h(z) - \frac{1}{n} \sum_i h(z_i) \right) \right] \leq 2 \cdot R_n(H)$$

Proof: Let $D' = \{ z_1' \dots z_n' \}$ independent copy of D.
$\varepsilon_1 \dots \varepsilon_n$ iid Rademacher rvs.

$$\left(\mathbb{E}(\cdot \mid D) = \mathbb{E}(\cdot) \right)$$

$$\mathbb{E}\left[\sup_{h\in H}\left(\mathbb{E}\,h(z) - \frac{1}{n}\sum_{i=1}^n h(z_i)\right)\right] = \left(\mathbb{E}\,h(z_i|D) = h(z_i)\right)$$

$$= \mathbb{E}\left[\sup_{h\in H}\left(\frac{1}{n}\sum_{i=1}^n\left(\mathbb{E}(h(z_i'|D) - h(z_i)\right)\right)\right]$$

$$\left(\begin{matrix}\mathbb{E}(h(z_i)|D)\\ \| \\ h(z_i)\end{matrix}\right) \leftarrow = \mathbb{E}\left[\sup_{h\in H}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(h(z_i') - h(z_i)\,|D\right)\right)\right]$$

$\sup\mathbb{E} \le \mathbb{E}\sup$ $\leftarrow (\le)\ \mathbb{E}\left[\mathbb{E}\left(\sup_{h\in H}\left(\frac{1}{n}\sum_{i=1}^n h(z_i') - h(z_i)\right)\right)\Big| D\right]$

$$= \mathbb{E}\sup_{h\in H}\frac{1}{n}\sum_{i=1}^n\left(h(z_i') - h(z_i)\right)$$

symmetry of z, z' and ε_i. $\leftarrow (=)\ \mathbb{E}\sup_{h\in H}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(h(z_i') - h(z_i)\right)$

$$\le 2\,\mathbb{E}\sup_{h\in H}\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(z_i) = 2R_n(H) \quad \square$$

$\left(2\,\mathrm{var}(X) = \mathbb{E}\,\|X - X'\|^2 \text{ with } X' \text{ an independent copy of } X\right)$

→ An important property of Rademacher complexity is that it can be easily composed with Lipschitz maps:

Prop (Contraction Principle of Rademacher complexity)
Given any functions $b, a_i : \Theta \to \mathbb{R}$ and $\varphi_i : \mathbb{R} \to \mathbb{R}$ β-Lipschitz, we have

$$\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta} b(\theta) + \sum_{i=1}^n \varepsilon_i\,\varphi(a_i(\theta))\right] \le$$

$$\le \beta\,\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta} b(\theta) + \sum_{i=1}^n \varepsilon_i\,a_i(\theta)\right]$$

→ For us, this means that if $\tilde{y} \mapsto \ell(y, \tilde{y})$ is β-Lipschitz a.s., then

$$\mathbb{E}_\varepsilon \left[\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \, \ell(y_i, f(x_i)) \,\middle|\, D \right] \leq$$

$$\leq \beta \cdot \mathbb{E}_\varepsilon \left[\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \, f(x_i) \right]$$

In other words, $R_n(H) \leq \beta \cdot R_n(F)$.

Example: Linear models with norm constraint. ← L_p norm

$$F = \left\{ f_\theta(x) = \langle \theta, \varphi(x) \rangle \text{ , with } \|\theta\|_p \leq D \right\}$$

with $\varphi(x) \in \mathbb{R}^d$ with $d \gg 1$

Let's write $\Phi = (\varphi(x_1) \dots \varphi(x_n)) \in \mathbb{R}^{n \times d}$

Let's now compute $R_n(F)$:

$$R_n(F) = \mathbb{E}_{\varepsilon, D} \left[\sup_{\|\theta\|_p \leq D} \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \, \theta^T \varphi(x_i) \right) \right] =$$

$\varepsilon = (\varepsilon_1 \dots \varepsilon_n)$

$$= \mathbb{E} \sup_{\|\theta\|_p \leq D} \frac{1}{n} \theta^T \Phi \varepsilon$$

$$= \frac{D}{n} \mathbb{E} \sup_{\|\theta\|_p \leq 1} \langle \theta, \Phi^T \varepsilon \rangle = \frac{D}{n} \mathbb{E} \| \Phi^T \varepsilon \|_p^* =$$

↳ Dual Norm

$$\left(\|x\|^* = \sup_{\|\theta\| \leq 1} \langle \theta, x \rangle \right)$$

$$= \frac{D}{n} \mathbb{E} \| \Phi^T \varepsilon \|_q \quad \text{with } \frac{1}{p} + \frac{1}{q} = 1$$

→ when $p=2$, then $q=2$,

$$\frac{D}{n}\mathbb{E}\|\phi^T\varepsilon\|_2 \overset{\text{Jensen}}{\leq} \frac{D}{n}\sqrt{\mathbb{E}\|\phi^T\varepsilon\|_2^2}$$

$$= \frac{D}{n}\sqrt{\mathbb{E}\langle \phi\phi^T, \varepsilon\varepsilon^T\rangle}$$

$\mathrm{Cov}(\varepsilon) = Id.$

$$= \frac{D}{n}\sqrt{\mathbb{E}\|\Phi\|_F^2} = \frac{D}{n}\sqrt{\mathbb{E}\sum_{i=1}^n \|\varphi(x_i)\|^2}$$

$$= \frac{D}{\sqrt{n}}\sqrt{\mathbb{E}_x\|\varphi(x)\|^2}$$

↳ There is no longer a dependency on dimension!

↳ Norm-constrained hypothesis spaces are very general.

Summary so far: ∫ generalization bounds that are
(i) non-asymptotic
(ii) in high-probability (not just in expectation).

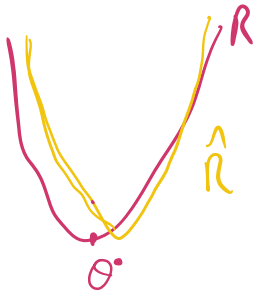
Q: What about asymptotic analysis?

$\theta \in \mathbb{R}^M$

$$R(\theta) = \mathbb{E}\left[\ell(y, f_\theta(x))\right], \quad \hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_\theta(x_i))$$

assume $R, \hat{R}$ are convex wrt θ
(+ R is strictly convex)

→ $\theta^* = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} R(\theta)$, with Hessian $\nabla^2 R(\theta^*) \succ 0$ $(\nabla R(\theta^*) = 0)$.



→ $\hat{\theta}_n \in \underset{\theta}{\operatorname{argmin}} \hat{R}(\theta)$

since $\nabla R(\theta^*) = 0$

$$\nabla \hat{R}_n(\theta^*) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \ell(y_i, f_\theta(x_i))\Big|_{\theta=\theta^*}$$

by the LLN $\nabla \hat{R}(\theta^*) \to 0$

by the CLT $\nabla \hat{R}(\theta^*) \sim N(0, \frac{1}{n} G(\theta^*))$

$$G(\theta^*) = \mathbb{E}\, \nabla_\theta \ell(y, f_\theta(x)) \cdot \nabla_\theta \ell(y, f_\theta(x))\Big|_{\theta=\theta^*}$$

$\hat{\theta}_n \to \theta^*$ thanks to strict convexity.

→ Taylor expansion of $\nabla \hat{R}$ around θ^*:

$$0 = \nabla \hat{R}(\theta_n) \cong \nabla \hat{R}(\theta^*) + \nabla^2 \hat{R}(\theta^-)(\theta_n - \theta^*)$$

↑ because θ_n is a minimum

by LLN, we also have

$$\nabla^2 \hat{R}(\theta^*) \to \nabla^2 R(\theta^*)$$

and thus

$$\theta_n - \theta^* \approx \left[\nabla^2 R(\theta^*)\right]^{-1} \cdot \nabla \hat{R}(\theta^*)$$

→ Since $\nabla \hat{R}(\theta^*)$ is asymp. normal, we also have that $\theta_n - \theta^*$ is approx normal, with mean 0 and

covariance. $\frac{1}{n} \nabla^2 R(\theta^*)^{-1} G(\theta^*) \nabla^2 R(\theta^*)^{-1}$

→ We can use this asymptotic gaussianity to measure excess risk:

$$R(\hat{\theta}_n) = R(\theta^*) + \langle \nabla R(\theta^*), \hat{\theta}_n - \theta^* \rangle + \frac{1}{2} (\theta_n - \theta^*)^T \nabla^2 R(\theta^*) (\theta_n - \theta^*)$$

$$\mathbb{E}\left[R(\hat{\theta}_n) - R(\theta^*)\right] \approx \frac{1}{n} \mathrm{Tr}\left(\nabla^2 R(\theta^*)^{-1} \cdot G(\theta^*)\right]$$

→ NOT an upper bound as before, but the actual expect. value !

→ Fast rate in $O(1/n)$ versus $O(1/\sqrt{n})$!!

↳ In essence, this is a local analysis.