Empirical Risk Minimization I - Rademacher complexity

Clément Lalanne

Where we stopped

The excess risk of the empirical risk minimizer is controlled by the uniform deviation, and the uniform deviation concentrates around its expectation for any class. What is left is the expectation itself.

The object. $\mathbb{E}_S[\sup_{f \in \mathcal{F}}(\widehat{R}_S(f) - R(f))]$ and its mirror image.

The plan. Compare the sample to an independent copy of itself, which turns the deviation into a symmetric quantity. Attach random signs, which turns it into a measure of how well the class fits pure noise. That measure is the Rademacher complexity, and for constrained linear models it is a norm computation.

The outcome. A bound on the excess risk with no finiteness assumption, no covering argument, and an explicit constant.

Notation for this lecture

Write $z = (x, y)$ and $z_i = (x_i, y_i)$, and attach to each predictor the function it induces on pairs,

$$h_f(z) = \ell(y, f(x)), \quad \mathcal{H} = \ell \circ \mathcal{F} = \{h_f : f \in \mathcal{F}\},$$

so that $\widehat{R}_S(f) = \frac{1}{n}\sum_{i \le n} h_f(z_i)$ and $R(f) = \mathbb{E}_{z \sim \mathbb{P}}[h_f(z)]$.

Everything below is stated for an arbitrary class $\mathcal{H}$ of functions of z, because nothing uses the fact that h came from a loss and a predictor. That structure returns only in Part III, where the contraction principle strips the loss away again.

Warm up: a finite class

Hypothesis 1. $\ell(y, f(x)) \in [0, \ell_\infty]$ for every $f \in \mathcal{F}$ and every (x, y) in the support of $\mathbb{P}$. In force for the whole lecture.

Hypothesis 2. $|\mathcal{F}| < +\infty$. In force for this slide and the next two, and nowhere else.

$$
\begin{aligned}
\mathbb{P}_S(\sup_{f \in \mathcal{F}} |\widehat{R}_S(f) - R(f)| \geq t) &\leq \sum_{f \in \mathcal{F}} \mathbb{P}_S(|\widehat{R}_S(f) - R(f)| \geq t) && \text{union bound, } \mathcal{F} \text{ finite} \\
&\leq \sum_{f \in \mathcal{F}} 2 \exp\left(-\frac{2nt^2}{\ell_\infty^2}\right) && \text{Hoeffding, } f \text{ fixed, Hypothesis 1} \\
&= 2|\mathcal{F}| \exp\left(-\frac{2nt^2}{\ell_\infty^2}\right) =: \delta
\end{aligned}
$$

Each $\widehat{R}_S(f)$ is an average of n i.i.d. variables in $[0, \ell_\infty]$, and f is fixed inside the sum, which is what makes the second line legitimate.

The bound for a finite class

Theorem. Under Hypotheses 1 and 2, for every $\delta \in (0,1)$, with probability at least $1-\delta$ over $S \sim \mathbb{P}^{\otimes n}$,

$$\sup_{f \in \mathcal{F}} |\widehat{R}_S(f) - R(f)| \le \frac{\ell_\infty}{\sqrt{2n}} \sqrt{\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}.$$

Corollary. Feeding this into the decomposition of the introductory lecture, for an exact empirical risk minimizer $\hat{f}_S$, with probability at least $1-\delta$,

$$R(\hat{f}_S) \le \inf_{f \in \mathcal{F}} R(f) + \frac{2\ell_\infty}{\sqrt{2n}} \sqrt{\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}.$$

Why this is not enough

Rate. The deviation is of order $\sqrt{\log|\mathcal{F}|/n}$. Doubling the sample halves the squared deviation, whereas squaring the size of the class only doubles it.

Complexity appears logarithmically. A richer $\mathcal{F}$ lowers $\inf_{f\in\mathcal{F}} R(f)$ and raises $\log|\mathcal{F}|$, which is the first quantitative form of the estimation and approximation tension.

And the bound is usually vacuous. Every class used in practice is infinite: linear predictors, neural networks, a single real hyperparameter. The bound reads $+\infty$. The union bound is the step to blame, and replacing it is what the rest of this lecture does.

Part I. Symmetrisation

The ghost sample

Let $S' = (z'_1, \dots, z'_n)$ be i.i.d. with the same law as S and independent of it: a second sample that is never observed.

$$\begin{aligned} \tfrac{1}{n}\sum_{i\le n} h(z_i) - \mathbb{E}_z[h(z)] &= \tfrac{1}{n}\sum_{i\le n} h(z_i) - \mathbb{E}_{S'}[\tfrac{1}{n}\sum_{i\le n} h(z'_i)] && S' \text{ has the same law as } S \\ &= \mathbb{E}_{S'}[\tfrac{1}{n}\sum_{i\le n}(h(z_i) - h(z'_i))] && \text{the first term does not depend on } S' \end{aligned}$$

The unknown expectation $\mathbb{E}_z[h(z)]$ has been replaced by an average over data we do not have. Nothing has been gained yet for a single h, but the right hand side is an expectation, and a supremum of expectations is at most the expectation of the supremum.

Symmetrisation, step one

$$\begin{aligned}\sup_{h\in\mathcal{H}}\left(\tfrac{1}{n}\textstyle\sum_i h(z_i) - \mathbb{E}_z[h(z)]\right) &= \sup_{h\in\mathcal{H}}\ \mathbb{E}_{S'}\left[\tfrac{1}{n}\textstyle\sum_i (h(z_i) - h(z_i'))\right] \quad \text{previous slide}\\ &\le \mathbb{E}_{S'}\left[\sup_{h\in\mathcal{H}} \tfrac{1}{n}\textstyle\sum_i (h(z_i) - h(z_i'))\right] \quad \sup\mathbb{E} \le \mathbb{E}\sup\end{aligned}$$

Taking $\mathbb{E}_S$ on both sides,

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left(\tfrac{1}{n}\textstyle\sum_i h(z_i) - \mathbb{E}_z[h(z)]\right)\right] \le \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \tfrac{1}{n}\textstyle\sum_i (h(z_i) - h(z_i'))\right].$$

Symmetrisation, step two

Let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher, $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = \tfrac{1}{2}$, independent of S and S'.

The key observation. Fix h and put $d_i = h(z_i) - h(z_i')$. Since z_i and z_i' are i.i.d., d_i and $-d_i$ have the same law, and the pairs are independent across i. Hence the whole vector $(d_1, \dots, d_n)$ has the same law as $(\epsilon_1 d_1, \dots, \epsilon_n d_n)$, and this holds jointly for all $h \in \mathcal{H}$ at once.

$$
\begin{aligned}
&\mathbb{E}_{S,S'}\big[\sup_h \tfrac{1}{n}\textstyle\sum_i (h(z_i) - h(z_i'))\big] \\
&= \mathbb{E}_{S,S',\epsilon}\big[\sup_h \tfrac{1}{n}\textstyle\sum_i \epsilon_i (h(z_i) - h(z_i'))\big] && \text{equality in law} \\
&\le \mathbb{E}_{S,\epsilon}\big[\sup_h \tfrac{1}{n}\textstyle\sum_i \epsilon_i h(z_i)\big] + \mathbb{E}_{S',\epsilon}\big[\sup_h \tfrac{1}{n}\textstyle\sum_i (-\epsilon_i) h(z_i')\big] && \sup(a+b) \le \sup a + \sup b \\
&= 2\mathbb{E}_{S,\epsilon}\big[\sup_h \tfrac{1}{n}\textstyle\sum_i \epsilon_i h(z_i)\big] && -\epsilon \overset{d}{=} \epsilon,\ S' \overset{d}{=} S
\end{aligned}
$$

Rademacher complexity

Definition. Let $\mathcal{H}$ be a class of real valued functions and $S = (z_1, \dots, z_n)$. The **empirical Rademacher complexity** of $\mathcal{H}$ given S, and the **Rademacher complexity** of $\mathcal{H}$ under $\mathbb{P}$, are

$$\mathfrak{R}_S(\mathcal{H}) = \mathbb{E}_\epsilon[\sup_{h \in \mathcal{H}} \tfrac{1}{n} \textstyle\sum_{i \le n} \epsilon_i h(z_i)], \quad \mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_{S \sim \mathbb{P}^{\otimes n}}[\mathfrak{R}_S(\mathcal{H})].$$

What it measures. The signs ϵ are pure noise, independent of the labels. The supremum asks how well some member of $\mathcal{H}$ can correlate with that noise on the observed points. A class that can fit any sign pattern has complexity of order $\sup_h \sup_i |h(z_i)|$; a class that cannot has complexity going to zero with n.

The symmetrisation theorem

Theorem. Let $\mathcal{H}$ be a class of functions and $z_1, \dots, z_n$ i.i.d. Then

$$\mathbb{E}_S\big[\sup_{h \in \mathcal{H}} \big(\tfrac{1}{n} \textstyle\sum_i h(z_i) - \mathbb{E}_z[h(z)] \big)\big] \le 2\mathfrak{R}_n(\mathcal{H}),$$

$$\mathbb{E}_S\big[\sup_{h \in \mathcal{H}} \big(\mathbb{E}_z[h(z)] - \tfrac{1}{n} \textstyle\sum_i h(z_i) \big)\big] \le 2\mathfrak{R}_n(\mathcal{H}).$$

Proof. The first inequality is the chain of the last two slides. The second is the same chain applied to $-\mathcal{H} = \{-h : h \in \mathcal{H}\}$, together with $\mathfrak{R}_n(-\mathcal{H}) = \mathfrak{R}_n(\mathcal{H})$, which holds because $-\epsilon$ has the law of ϵ. ∎

Remark. Applied to $\mathcal{H} = \ell \circ \mathcal{F}$, this is exactly the pair of expectations left unevaluated at the end of the previous lecture.

Three properties

Scaling. $\mathfrak{R}_S(c\mathcal{H}) = |c|\mathfrak{R}_S(\mathcal{H})$ for $c \in \mathbb{R}$.

Proof. $\frac{1}{n}\sum_i \epsilon_i (ch)(z_i) = c \cdot \frac{1}{n}\sum_i \epsilon_i h(z_i)$. For $c \geq 0$ this concludes the proof, and for $c < 0$ relabelling $\epsilon \mapsto -\epsilon$, which doesn't change the law allows to conclude.

Translation. $\mathfrak{R}_S(\mathcal{H} + h_0) = \mathfrak{R}_S(\mathcal{H})$ for a single fixed h_0. Constants are free: only the *variation* of the class is measured.

Proof. $\sup_h \frac{1}{n}\sum_i \epsilon_i (h + h_0)(z_i) = \frac{1}{n}\sum_i \epsilon_i h_0(z_i) + \sup_h \frac{1}{n}\sum_i \epsilon_i h(z_i)$, and $\mathbb{E}_\epsilon$ kills the first term.

Convex hull. $\mathfrak{R}_S(\mathrm{conv}\mathcal{H}) = \mathfrak{R}_S(\mathcal{H})$. Averaging predictors does not increase complexity, which is the reason ensembles do not pay for their size.

Proof. At fixed ϵ, $h \mapsto \frac{1}{n}\sum_i \epsilon_i h(z_i)$ is linear, and a linear functional on a convex set is maximized at an extreme point, so the supremum over $\mathrm{conv}\mathcal{H}$ is already attained on $\mathcal{H}$.

The contraction principle

Proposition. Let $A \subset \mathbb{R}^n$ and let $g_1, \dots, g_n : \mathbb{R} \to \mathbb{R}$ each be G-Lipschitz. Then

$$\mathbb{E}_\epsilon\big[\sup_{a \in A} \tfrac{1}{n} \textstyle\sum_{i \le n} \epsilon_i g_i(a_i)\big] \le G\,\mathbb{E}_\epsilon\big[\sup_{a \in A} \tfrac{1}{n} \textstyle\sum_{i \le n} \epsilon_i a_i\big].$$

The proof is on the next slide. See also Ledoux and Talagrand (1991, Theorem 4.12) or Bach (2024).

Consequence. If $\ell(y, \cdot)$ is G-Lipschitz for every y in the support, then applying this conditionally on S with $g_i = \ell(y_i, \cdot)$ and $A = \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\}$,

$$\mathfrak{R}_S(\ell \circ \mathcal{F}) \le G\,\mathfrak{R}_S(\mathcal{F}), \quad \text{hence} \quad \mathfrak{R}_n(\ell \circ \mathcal{F}) \le G\,\mathfrak{R}_n(\mathcal{F}).$$

Warning on the constant. The functions g_i must be allowed to differ with i, since $\ell(y_i, \cdot)$ depends on the label. The variant with absolute values inside both suprema is also true but costs $2G$ instead of G, so we avoid it throughout.

The contraction principle, proof

By scaling, $G = 1$; the $\frac{1}{n}$ divides both sides and is dropped. Fix the last coordinate, condition on $\epsilon_1, \dots, \epsilon_{n-1}$, and set $\phi(a) = \sum_{i<n} \epsilon_i g_i(a_i)$.

$$
\begin{aligned}
&\mathbb{E}_{\epsilon_n} \sup_a (\phi(a) + \epsilon_n g_n(a_n)) \\
&= \tfrac{1}{2} \sup_a (\phi(a) + g_n(a_n)) + \tfrac{1}{2} \sup_{a'} (\phi(a') - g_n(a'_n)) && \epsilon_n = \pm 1 \\
&= \tfrac{1}{2} \sup_{a,a'} (\phi(a) + \phi(a') + g_n(a_n) - g_n(a'_n)) && \text{independent suprema} \\
&\le \tfrac{1}{2} \sup_{a,a'} (\phi(a) + \phi(a') + |a_n - a'_n|) && g_n \text{ is 1-Lipschitz} \\
&= \tfrac{1}{2} \sup_{a,a'} \big((\phi(a) + a_n) + (\phi(a') - a'_n)\big) && |a_n - a'_n| = \pm(a_n - a'_n), \text{ relabel} \\
&= \mathbb{E}_{\epsilon_n} \sup_a (\phi(a) + \epsilon_n a_n) && \text{line 1 backwards}
\end{aligned}
$$

Iterating on the other coordinates finishes the proof. ■

The line $\le$ is the only assumption that is used: A, the g_i, and n are otherwise free.

Finite classes, revisited

Lemma (Massart). Let $A \subset \mathbb{R}^n$ be finite and nonempty, and $\rho = \max_{a \in A} \|a\|_2$. Then

$$\mathbb{E}_\epsilon\big[\max_{a \in A} \tfrac{1}{n} \textstyle\sum_{i \le n} \epsilon_i a_i\big] \le \dfrac{\rho\sqrt{2 \log |A|}}{n}.$$

Each direction is sub-gaussian. Fix $a \in A$ and write $W_a = \sum_{i \le n} \epsilon_i a_i$. Each $\epsilon_i a_i$ is centered with values in $[-|a_i|, |a_i|]$, an interval of length $2|a_i|$, so Hoeffding's lemma gives $\epsilon_i a_i \in \mathcal{G}(a_i^2)$. The ϵ_i are independent, so the proxies add: $W_a \in \mathcal{G}(\|a\|_2^2) \subseteq \mathcal{G}(\rho^2)$.

The maximum over A. For every $\lambda > 0$,

$$e^{\lambda \mathbb{E}_\epsilon[\max_a W_a]} \le \mathbb{E}_\epsilon\big[e^{\lambda \max_a W_a}\big] \le \sum_{a \in A} \mathbb{E}_\epsilon\big[e^{\lambda W_a}\big] \le |A| e^{\lambda^2 \rho^2 / 2} \qquad \text{Jensen, then max} \le \text{sum}$$

$$\mathbb{E}_\epsilon[\max_a W_a] \le \frac{\log |A|}{\lambda} + \frac{\lambda \rho^2}{2} = \rho\sqrt{2 \log |A|} \qquad \text{at } \lambda = \sqrt{2 \log |A|}/\rho$$

Divide by n. ■

Finite classes, the rate recovered

Corollary. If $\mathcal{H}$ is finite and $|h(z)| \le \ell_\infty$ for every h and every z in the support, then

$$\mathfrak{R}_S(\mathcal{H}) \le \ell_\infty \sqrt{\frac{2\log|\mathcal{H}|}{n}}, \quad \text{hence} \quad \mathbb{E}_S[\sup_{h\in\mathcal{H}}(\tfrac{1}{n}\textstyle\sum_i h(z_i) - \mathbb{E}_z[h(z)])] \le 2\ell_\infty\sqrt{\tfrac{2\log|\mathcal{H}|}{n}}.$$

Proof. Apply Massart's lemma to $A = \{(h(z_1),\dots,h(z_n)) : h \in \mathcal{H}\}$, which has at most $|\mathcal{H}|$ elements and $\rho \le \ell_\infty\sqrt{n}$. Then symmetrisation. ■

Reading it. The same $\sqrt{\log|\mathcal{F}|/n}$ rate as the union bound of the warm up, with a worse constant and in expectation rather than with high probability. The gain is that nothing in the definition of $\mathfrak{R}_n$ needs $\mathcal{H}$ to be finite: the union bound is a special case, not a competitor.

Finite classes, and their convex hulls

The convex hull property and the bound just proved were established separately. Put together, they cover a large family of infinite classes.

Corollary. If $\mathcal{H} = \operatorname{conv}\{h_1, \dots, h_N\}$ and $|h_k(z)| \le \ell_\infty$ for every $k \le N$ and every z in the support, then

$$\mathfrak{R}_S(\mathcal{H}) \le \ell_\infty \sqrt{\frac{2 \log N}{n}}.$$

Proof. $\mathfrak{R}_S(\operatorname{conv}\{h_1, \dots, h_N\}) = \mathfrak{R}_S(\{h_1, \dots, h_N\})$ by the convex hull property, and the previous slide applies to a class of N elements. ■

What it buys. The class is infinite, and usually infinite dimensional, yet it is charged only for the number of its extreme points, and only through a logarithm. Boundedness is needed on the h_k alone, since a convex combination of functions bounded by ℓ_∞ is again bounded by ℓ_∞. A polytope costs the logarithm of its number of vertices, whatever its dimension.

Part II. Constrained linear models

The class and its complexity

Definition. With a feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$ and a norm Ω on $\mathbb{R}^d$,

$$\mathcal{F} = \{f_\theta : x \mapsto \theta^\top \varphi(x) \ , \ \Omega(\theta) \le D\}.$$

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{F}) &= \mathbb{E}_{S,\epsilon}\Big[\sup_{\Omega(\theta)\le D} \tfrac{1}{n}\textstyle\sum_{i\le n} \epsilon_i \theta^\top \varphi(x_i)\Big] \\
&= \mathbb{E}_{S,\epsilon}\Big[\sup_{\Omega(\theta)\le D} \tfrac{1}{n}\epsilon^\top \Phi\theta\Big] && \Phi \in \mathbb{R}^{n\times d} \text{ with rows } \varphi(x_i)^\top \\
&= \frac{D}{n}\mathbb{E}_{S,\epsilon}[\Omega^*(\Phi^\top \epsilon)] && \sup_{\Omega(\theta)\le D} \langle u, \theta\rangle = D\Omega^*(u)
\end{aligned}$$

The last line is the definition of the dual norm Ω^*. The geometry of the constraint set enters only here, and the rest is a computation with $\Phi^\top \epsilon = \sum_{i\le n} \epsilon_i \varphi(x_i)$, a random vector with independent symmetric summands.

The Euclidean case

Take $\Omega = \|\cdot\|_2$, self dual, so $\Omega^* = \|\cdot\|_2$.

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{F}) &= \frac{D}{n}\mathbb{E}[\|\Phi^\top \epsilon\|_2] \\
&\leq \frac{D}{n}\sqrt{\mathbb{E}[\|\Phi^\top \epsilon\|_2^2]} && \text{Jensen, } t \mapsto \sqrt{t} \text{ concave} \\
&= \frac{D}{n}\sqrt{\mathbb{E}[\mathrm{tr}(\Phi\Phi^\top \epsilon\epsilon^\top)]} && \|\Phi^\top \epsilon\|_2^2 = \epsilon^\top \Phi\Phi^\top \epsilon, \text{ trace is cyclic} \\
&= \frac{D}{n}\sqrt{\mathbb{E}_S[\mathrm{tr}(\Phi\Phi^\top)]} && \mathbb{E}_\epsilon[\epsilon\epsilon^\top] = I_n, \ \epsilon \perp S \\
&= \frac{D}{n}\sqrt{n\mathbb{E}_x[\|\varphi(x)\|_2^2]} = \frac{D}{\sqrt{n}}\sqrt{\mathbb{E}_x[\|\varphi(x)\|_2^2]} && \mathrm{tr}(\Phi\Phi^\top) = \textstyle\sum_{i\leq n} \|\varphi(x_i)\|_2^2
\end{aligned}
$$

The dimension d has disappeared. What replaces it is the average squared norm of the features, and the constraint radius D.

The sparse case

Take $\Omega = \|\cdot\|_1$, whose dual is $\|\cdot\|_\infty$, and assume the features are bounded coordinatewise, $|\varphi_j(x)| \le B_\infty$ for every $j \le d$ and every x in the support.

Read off the coordinates. The j-th coordinate of $\Phi^\top \epsilon$ is $\sum_{i \le n} \epsilon_i \varphi_j(x_i)$, so the dual norm formula becomes

$$\mathfrak{R}_n(\mathcal{F}) = \frac{D}{n}\mathbb{E}_{S,\epsilon}[\|\Phi^\top \epsilon\|_\infty] = \frac{D}{n}\mathbb{E}_{S,\epsilon}[\max_{j \le d} |\textstyle\sum_{i \le n} \epsilon_i \varphi_j(x_i)|].$$

Massart bounds a maximum, not a maximum of absolute values. Use $|u| = \max\{u, -u\}$ and double the index set. With the $2d$ vectors

$$A = \{\pm(\varphi_j(x_1), \dots, \varphi_j(x_n)) : j \le d\} \subset \mathbb{R}^n, \quad \max_{j \le d} |\textstyle\sum_i \epsilon_i \varphi_j(x_i)| = \max_{a \in A} \sum_{i \le n} \epsilon_i a_i.$$

The sparse case, applying Massart

Condition on S. Then A is a fixed finite subset of $\mathbb{R}^n$ and only ϵ is random, which is exactly the setting of the lemma.

The two inputs. $|A| \leq 2d$, and each $a \in A$ is a feature column, so

$$\|a\|_2^2 = \sum_{i \leq n} \varphi_j(x_i)^2 \leq n B_\infty^2, \quad \text{hence} \quad \rho = \max_{a \in A} \|a\|_2 \leq B_\infty \sqrt{n}.$$

Massart's lemma then gives, conditionally on S,

$$\frac{1}{n} \mathbb{E}_\epsilon \left[\max_{a \in A} \sum_{i \leq n} \epsilon_i a_i \right] \leq \frac{\rho \sqrt{2 \log |A|}}{n} \leq \frac{B_\infty \sqrt{n} \sqrt{2 \log(2d)}}{n} = B_\infty \sqrt{\frac{2 \log(2d)}{n}}.$$

The right side does not depend on S, so it survives $\mathbb{E}_S$. Multiplying by D,

$$\mathfrak{R}_n(\mathcal{F}) \leq D B_\infty \sqrt{\frac{2 \log(2d)}{n}}.$$

The sparse case, discussion

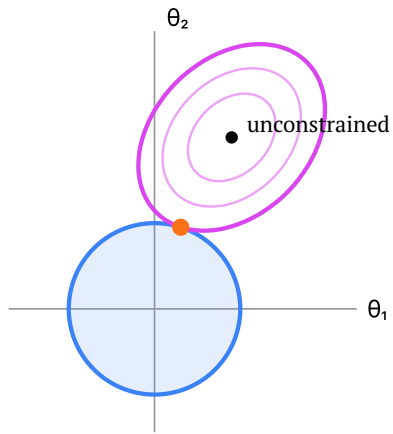
Comparison. In the Euclidean case $\mathbb{E}_x[\|\varphi(x)\|_2^2]$ is typically of order dB_∞^2, so the bound is $DB_\infty\sqrt{d/n}$. Constraining the ℓ_1 norm instead replaces $\sqrt{d}$ by $\sqrt{\log d}$. This is the quantitative argument for the lasso.

The same bound, without the dual norm. The ball $\{\|\theta\|_1 \le D\}$ is the convex hull of the $2d$ points $\pm De_j$, so $\mathcal{F}$ is the convex hull of the $2d$ functions $\pm D\varphi_j$, each bounded by DB_∞. The corollary on convex hulls of finite classes returns $\mathfrak{R}_n(\mathcal{F}) \le DB_\infty\sqrt{2\log(2d)/n}$ immediately, with no dual norm and no Massart applied by hand. Few extreme points leave the noise little to exploit, and those same extreme points are sparse, which is the next slide.

Where each factor came from. D is the constraint radius and scales the whole class. B_∞ is the coordinatewise bound on the features. The $\log$ is Massart's, and it is the only place the dimension appears.

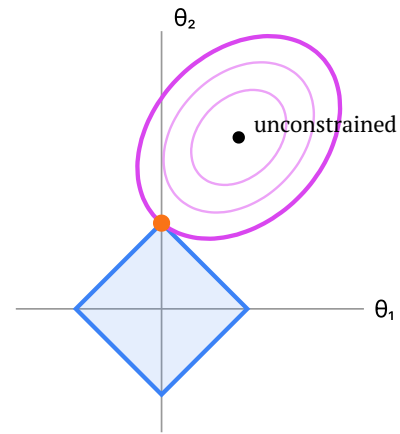
The geometry behind ℓ_1 vs ℓ_2

ℓ_2 constraint



minimizer off the axes, $\theta_1 \neq 0$

ℓ_1 constraint



minimizer on a vertex, $\theta_1 = 0$

Part III. The estimation error of linear models

The theorem

Theorem. Let $\mathcal{F} = \{f_\theta : \|\theta\|_2 \le D\}$ and suppose $\ell(y, \cdot)$ is G-Lipschitz for every y in the support. Let $\hat{f}_S$ be an exact empirical risk minimizer over $\mathcal{F}$. Then

$$\mathbb{E}_S[R(\hat{f}_S)] \le \inf_{\|\theta\|_2 \le D} R(f_\theta) + \frac{2GD}{\sqrt{n}} \sqrt{\mathbb{E}_x[\|\varphi(x)\|_2^2]}.$$

Proof. Write $f^\dagger \in \arg\min_{f \in \mathcal{F}} R(f)$, fixed and independent of S, and decompose as in the previous lecture:

$$\begin{aligned} \mathbb{E}_S[R(\hat{f}_S)] - R(f^\dagger) &= \mathbb{E}_S[R(\hat{f}_S) - \widehat{R}_S(\hat{f}_S)] + \mathbb{E}_S[\widehat{R}_S(\hat{f}_S) - \widehat{R}_S(f^\dagger)] + \mathbb{E}_S[\widehat{R}_S(f^\dagger) - R(f^\dagger)] \\ &\le \mathbb{E}_S[\sup_{f \in \mathcal{F}} (R(f) - \widehat{R}_S(f))] + 0 + 0 \end{aligned}$$

The theorem, end of proof

$$
\begin{aligned}
\mathbb{E}_S[\sup_{f\in\mathcal{F}}(R(f)-\widehat{R}_S(f))] &\le 2\mathfrak{R}_n(\ell\circ\mathcal{F}) && \text{symmetrisation, second direction}\\
&\le 2G\mathfrak{R}_n(\mathcal{F}) && \text{contraction, } \ell(y_i,\cdot) \text{ is } G\text{-Lipschitz}\\
&\le \frac{2GD}{\sqrt{n}}\sqrt{\mathbb{E}_x[\|\varphi(x)\|_2^2]} && \text{Euclidean case of Part II} \quad \blacksquare
\end{aligned}
$$

Remark on the constant. Routing the argument through $2\sup_f|\widehat{R}_S(f)-R(f)|$, as the decomposition of the introductory lecture suggests, gives $4GD$ instead of $2GD$. The factor is recovered by using the one sided supremum, which is legitimate because the two discarded terms are respectively nonpositive and exactly zero. The two sided route also forces the version of the contraction principle with absolute values, whose constant is $2G$.

If the minimization is not exact, add $\epsilon_{\text{opt}} = \mathbb{E}_S[\widehat{R}_S(\hat{f}_S) - \inf_{f\in\mathcal{F}}\widehat{R}_S(f)]$ to the right hand side.

The high probability version

Theorem. Assume in addition $\ell(y, f(x)) \in [0, \ell_\infty]$ on the support, for every $f \in \mathcal{F}$. Then for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$R(\hat{f}_S) \le \inf_{\|\theta\|_2 \le D} R(f_\theta) + \frac{2GD}{\sqrt{n}}\sqrt{\mathbb{E}_x[\|\varphi(x)\|_2^2]} + 2\ell_\infty\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof. Same decomposition, kept pathwise. The first term is at most $\sup_f (R(f) - \widehat{R}_S(f))$, which by the concentration lecture exceeds its expectation by at most $\ell_\infty\sqrt{\log(2/\delta)/(2n)}$ with probability $1-\delta/2$, and that expectation is bounded as above. The third term concerns the single fixed $f^\dagger$, so Hoeffding gives the same quantity with probability $1-\delta/2$. A union bound over the two events. $\blacksquare$

Reading the result

No dimension, no cardinality. The bound involves the constraint radius D, the Lipschitz constant G, and the second moment of the features. A class with infinitely many members, even in infinite dimension, has a finite complexity provided it is constrained.

The rate is $n^{-1/2}$, the same as the central limit theorem for a single f, and the same as the finite class bound. What changes between the three is the constant, not the rate.

The trade-off is explicit. Enlarging D decreases $\inf_{\|\theta\|_2 \le D} R(f_\theta)$, the approximation error, and increases $2GD\sqrt{\cdot}/\sqrt{n}$, the estimation error. The underfitting and overfitting curves of the introductory lecture are now two named terms of one inequality.

In practice. The hinge and logistic losses are 1-Lipschitz in their second argument, so $G = 1$ for the convexified classification problem of the convexification lecture.

Regression. The square loss $\ell(y, u) = (y - u)^2$ is not Lipschitz on $\mathbb{R}$, but contraction only ever compares values taken by members of $\mathcal{F}$. If $\|\varphi(x)\|_2 \le B_2$ then $|f_\theta(x)| \le DB_2$ by Cauchy and Schwarz, and if $|y| \le M$ then $\ell(y, \cdot)$ is G-Lipschitz on that range with $G = 2(M + DB_2)$. The theorem applies and gives an excess risk at most $4DB_2(M + DB_2)/\sqrt{n}$: the same $n^{-1/2}$, with a constant now quadratic in D, because the Lipschitz constant grows with the class it has to cover.

What to remember

Symmetrisation. Compare the sample to a ghost sample, attach random signs, and the deviation becomes a correlation with noise: $\mathbb{E}_S[\sup_h(\frac{1}{n}\sum_i h(z_i) - \mathbb{E}_z[h(z)])] \leq 2\mathfrak{R}_n(\mathcal{H})$, and the same with the two terms exchanged.

Rademacher complexity is computable. Finite classes give $\ell_\infty\sqrt{2\log|\mathcal{H}|/n}$ by Massart, and so does any convex hull of $|\mathcal{H}|$ functions, so a polytope costs the logarithm of its number of vertices. Constrained linear models give $D\mathbb{E}[\Omega^*(\Phi^\top\epsilon)]/n$, a dual norm computation.

Contraction removes the loss. A G-Lipschitz loss costs a factor G and nothing else, which is why the complexity of the predictor class is the only geometric quantity in the final bound.

The final bound. Excess risk at most $2GD\sqrt{\mathbb{E}_x\|\varphi(x)\|^2}/\sqrt{n}$, in expectation, plus $2\ell_\infty\sqrt{\log(2/\delta)/(2n)}$ for the high probability version.