Learning at the Turing

The Alan Turing Institute

Mathematics of Machine Learning - Summer School

Lecture 4
Maximal Inequalities and Rademacher Complexity

June 29, 2021

Patrick Rebeschini
Department of Statistics, University of Oxford

Maximum of finitely many bounded random variables (Proposition 2.2)

Let $X_1, \ldots, X_n$ be n centered random variables bounded in the interval $[a, b]$.

$$\boxed{\mathbf{E} \max_{i \in [n]} X_i \leq \frac{b-a}{\sqrt{2}} \sqrt{\log n}}$$

Proof

- $X = \max_{i \in [n]} X_i$. Exponentiate. Jensen's ineq. as $x \to e^{\lambda x}$ ($\lambda > 0$) is convex:

$$\mathbf{E} X = \frac{1}{\lambda} \log e^{\lambda \mathbf{E} X} \leq \frac{1}{\lambda} \log \mathbf{E}\, e^{\lambda X}$$

- Bound maximum of non-negative numbers by the sum:

$$\mathbf{E}\, e^{\lambda X} = \mathbf{E}\, e^{\lambda \max_{i \in [n]} X_i} = \mathbf{E} \max_{i \in [n]} e^{\lambda X_i} \leq \mathbf{E} \sum_{i=1}^{n} e^{\lambda X_i} = \sum_{i=1}^{n} \mathbf{E}\, e^{\lambda X_i}$$

- Put everything together and use Hoeffding's lemma ($\mathbf{E}\, e^{\lambda X_i} \leq e^{\lambda^2 (b-a)^2/8}$):

$$\mathbf{E} \max_{i \in [n]} X_i \leq \frac{1}{\lambda} \log \sum_{i=1}^{n} e^{\lambda^2 (b-a)^2/8} = \frac{1}{\lambda} \log n + \frac{\lambda (b-a)^2}{8}$$

- Optimizing the bound $\alpha/\lambda + \lambda\beta$ over $\lambda > 0$ yields the minimum is at $\lambda = \sqrt{\alpha/\beta}$ and the optimal value $2\sqrt{\alpha\beta} = (b-a)\sqrt{\log n / 2}$ $\square$

Bound in expectation for finitely-many actions

Bound in expectation (Proposition 2.3)

If the loss function ℓ is bounded by c, we have

$$\boxed{\mathbf{E} \max_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq c \frac{\sqrt{2 \log |\mathcal{A}|}}{\sqrt{n}}}$$

Proof: Same as above, using the independence of the data $Z_1, \dots, Z_n$
(note that for each $a \in \mathcal{A}$, $r(a) - R(a)$ is a centered random variable as $\mathbf{E} R(a) = r(a)$)

- Recall wish: $\boxed{\mathbf{E} \sup_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq \frac{f(\text{dimension}, \text{complexity of } \mathcal{A})}{n^{\alpha}}}$
- The dimension of the data is superseded by the boundedness assumption
- $\alpha = 1/2$, slow rate
- When $|\mathcal{A}| < \infty$, $\log |\mathcal{A}|$ is a valid notion of complexity of the problem
- When $|\mathcal{A}| = \infty$, upper bound is trivial and we need another notion of complexity

Rademacher complexity

Rademacher complexity (Definition 2.5)

The Rademacher complexity of a set $\mathcal{T} \subseteq \mathbb{R}^n$ is defined as

$$\boxed{\mathtt{Rad}(\mathcal{T}) := \mathbf{E} \sup_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i t_i}$$

where $\Omega_1, \ldots, \Omega_n \in \{-1, 1\}$ are i.i.d. uniform random variables (Rademacher)

- Measures of complexity: describes how well elements in $\mathcal{T}$ can replicate the sign pattern of a uniform random signal in $\mathbb{R}^n$ (see **Problem 1.5**)
- Useful properties:
  - $\boxed{\mathtt{Rad}(c\mathcal{T} + v) = |c| \, \mathtt{Rad}(\mathcal{T})}$ (Proposition 2.6)
  - $\boxed{\mathtt{Rad}(\mathcal{T} + \mathcal{T}') = \mathtt{Rad}(\mathcal{T}) + \mathtt{Rad}(\mathcal{T}')}$ (Proposition 2.7)
  - $\boxed{\mathtt{Rad}(\mathrm{conv}(\mathcal{T})) = \mathtt{Rad}(\mathcal{T})}$ (Proposition 2.8)

    with $\mathrm{conv}(\mathcal{T}) = \{\sum_{j=1}^{m} w_j t_j : w \in \Delta_m, t_1, \ldots, t_m \in \mathcal{T}, m \in \mathbb{N}\}$

Rademacher complexity

Massart's Lemma (Lemma 2.9)

Let $\mathcal{T} \subseteq \mathbb{R}^n$ and $\bar{t} := \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} t$. We have

$$\boxed{\mathtt{Rad}(\mathcal{T}) \leq \max_{t \in \mathcal{T}} \|t - \bar{t}\|_2 \frac{\sqrt{2 \log |\mathcal{T}|}}{n}}$$

Proof: Similar to ones given above. **Problem 1.6**

Contraction property - Talagrand's Lemma (Lemma 2.10)

Let $\mathcal{T} \subseteq \mathbb{R}^n$. For each $i \in \{1, \ldots, n\}$, let $f_i : \mathbb{R} \to \mathbb{R}$ be a γ-Lipschitz function. Then,

$$\boxed{\mathtt{Rad}((f_1, \ldots, f_n) \circ \mathcal{T}) \leq \gamma \, \mathtt{Rad}(\mathcal{T})}$$

with $(f_1, \ldots, f_n) \circ \mathcal{T} := \{(f_1(t_1), \ldots, f_n(t_n)) \in \mathbb{R}^n : t \in \mathcal{T}\}$

Proof: **Problem 1.7**

Recap

- Goal: $$\boxed{\mathbf{E}\,\underbrace{r(A^\star) - r(a^\star)}_{\text{estimation error for ERM}} \lesssim \frac{f(\text{dimension})}{n^\alpha}}$$
- Sufficient: $$\boxed{\mathbf{E}\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} \le \frac{f(\text{dimension}, \text{complexity of } \mathcal{A})}{n^\alpha}}$$

Bound in expectation (Proposition 2.3)

If the loss function ℓ is bounded by c, we have

$$\boxed{\mathbf{E}\max_{a\in\mathcal{A}}\{r(a) - R(a)\} \le c\frac{\sqrt{2\log|\mathcal{A}|}}{\sqrt{n}}}$$

Bound in expectation via Rademacher complexity (Proposition 2.11)

$$\boxed{\mathbf{E}\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} \le 2\,\mathbf{E}\,\mathtt{Rad}(\mathcal{L}\circ\{Z_1,\dots,Z_n\})}$$

with $\mathcal{L}\circ\{Z_1,\dots,Z_n\} := \{(\ell(a,Z_1),\dots,\ell(a,Z_n)) \in \mathbb{R}^n : a\in\mathcal{A}\}$

Note

If $|\mathcal{A}| < \infty$, Massart's Lemma recovers previous result (modulo constant)

Massart's Lemma (Lemma 2.9)

Let $\mathcal{T} \subseteq \mathbb{R}^n$ and $\bar{t} := \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} t$. We have

$$\boxed{\mathtt{Rad}(\mathcal{T}) \leq \max_{t \in \mathcal{T}} \|t - \bar{t}\|_2 \frac{\sqrt{2 \log |\mathcal{T}|}}{n}}$$

We have

- $\mathcal{T} = \mathcal{L} \circ \{Z_1, \ldots, Z_n\} = \{(\ell(a, Z_1), \ldots, \ell(a, Z_n)) \in \mathbb{R}^n : a \in \mathcal{A}\}$
- $|\mathcal{T}| \leq |\mathcal{A}|$
- $\|t - \bar{t}\|_2 \leq \|t\|_2 = \sqrt{\sum_{i=1}^n \ell(a, Z_i)^2} \leq \sqrt{nc^2} = c\sqrt{n}$

so we obtain

$$\boxed{\mathbf{E} \max_{a \in \mathcal{A}} \{r(a) - R(a)\} \leq 2c \frac{\sqrt{2 \log |\mathcal{A}|}}{\sqrt{n}}}$$

For the proof, let's review some basic properties of conditional expectations...

Properties of conditional expectations

Let X, Y be real-valued random variables. The following can be made precise:

- $\mathbf{E}X$ is the "best" estimate of X with no information. It is a **constant**
- $\mathbf{E}[X|Y]$ is the "best" estimate of X if we know Y. It is a **random variable**
- If X and Y are independent, Y does not contain any information on X and

  $\mathbf{E}[X|Y] = \mathbf{E}X$ independence property (a)

- If f is a deterministic function, if we know Y we also know $f(Y)$ and

  $\mathbf{E}[Xf(Y)|Y] = f(Y)\mathbf{E}[X|Y]$ "taking out what is known" property (b)

- Law of total expectation ("Ignorants win in life" phenomenon)

  $\mathbf{E}\mathbf{E}[X|Y] = \mathbf{E}X$ "tower" property (c)

Remark: the above holds with $\mathbf{E} \to \mathbf{E}[\,\cdot\,|Z]$, $\mathbf{E}[\,\cdot\,|Y] \to \mathbf{E}[\,\cdot\,|Y, Z]$ possibly using the notion of conditional independence.

Proof: Symmetrization

Proof

- $S = \{Z_1, \dots, Z_n\}$ and $\widetilde{S} = \{\widetilde{Z}_1, \dots, \widetilde{Z}_n\}$ be independent samples with same distribution

$$r(a) = \mathbf{E}\,\ell(a, Z) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\,\ell(a, \widetilde{Z}_i) \overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[\ell(a, \widetilde{Z}_i)|S]$$

- By properties of conditional expectations (tower property and others) we get

$$\begin{aligned}
\mathbf{E}\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} &= \mathbf{E}\sup_{a\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{E}[\ell(a, \widetilde{Z}_i)|S] - \ell(a, Z_i)\Big) \\
&\overset{(b)}{=} \mathbf{E}\sup_{a\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[\ell(a, \widetilde{Z}_i) - \ell(a, Z_i)|S] \\
&\le \mathbf{E}\mathbf{E}\Big[\sup_{a\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\{\ell(a, \widetilde{Z}_i) - \ell(a, Z_i)\}\Big| S\Big] \\
&\overset{(c)}{=} \mathbf{E}\sup_{a\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\{\ell(a, \widetilde{Z}_i) - \ell(a, Z_i)\} \\
&= \mathbf{E}\sup_{a\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\Omega_i\{\ell(a, \widetilde{Z}_i) - \ell(a, Z_i)\} \\
&\le 2\,\mathbf{E}\sup_{a\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\Omega_i \ell(a, Z_i) = 2\,\mathbf{E}\,\mathtt{Rad}(\mathcal{L}\circ\{Z_1, \dots, Z_n\})
\end{aligned}$$

$\square$

Supervised Learning. Regression

Today, we consider the setting of regression:

- $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$
- Admissible action set $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \to \mathbb{R}\}$
- Loss function is of the form $\ell(a, (x, y)) = \phi(a(x), y)$, for $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$

(Proposition 3.1)

If the function $\hat{y} \to \phi(\hat{y}, y)$ is γ-Lipschitz for any $y \in \mathcal{Y}$, then

$$\boxed{\mathtt{Rad}(\mathcal{L} \circ \{z_1, \ldots, z_n\}) \leq \gamma\, \mathtt{Rad}(\mathcal{A} \circ \{x_1, \ldots, x_n\})}$$

with $\mathcal{A} \circ \{x_1, \ldots, x_n\} := \{(a(x_1), \ldots, a(x_n)) \in \mathbb{R}^n : a \in \mathcal{A}\}$

- New goal: $\boxed{\mathtt{Rad}(\mathcal{A} \circ \{x_1, \ldots, x_n\}) \leq \dfrac{f(\text{dimension}, \text{complexity of } \mathcal{A})}{n^\alpha}}$

Linear predictors ℓ_2/ℓ_2 constraints (SVM)

(Proposition 3.2)

Let $\mathcal{A}_2 := \{x \in \mathbb{R}^d \to w^\top x : w \in \mathbb{R}^d, \|w\|_2 \le c\}$. Then,

$$\boxed{\texttt{Rad}(\mathcal{A}_2 \circ \{x_1, \ldots, x_n\}) \le c \frac{\max_i \|x_i\|_2}{\sqrt{n}}}$$

Note: typically, $\max_i \|x_i\|_2 \sim \sqrt{d}$ as

$$\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2} \le \sqrt{d} \max_{i \in [d]} |x_i|$$

Proof

$$
\begin{aligned}
& n\,\mathtt{Rad}(\mathcal{A}_2 \circ \{x_1, \dots, x_n\}) \\
& = \mathbf{E} \sup_{w \in \mathbb{R}^d : \|w\|_2 \le 1} \sum_{i=1}^n \Omega_i w^\top x_i = \mathbf{E} \sup_{w \in \mathbb{R}^d : \|w\|_2 \le 1} w^\top \Big(\sum_{i=1}^n \Omega_i x_i \Big) \\
& \le \sup_{w \in \mathbb{R}^d : \|w\|_2 \le 1} \|w\|_2 \, \mathbf{E} \Big\| \sum_{i=1}^n \Omega_i x_i \Big\|_2 \quad \text{by Cauchy-Schwarz's ineq. } x^\top y \le \|x\|_2 \|y\|_2 \\
& \le \mathbf{E} \sqrt{\Big\| \sum_{i=1}^n \Omega_i x_i \Big\|_2^2} \le \sqrt{\mathbf{E} \Big\| \sum_{i=1}^n \Omega_i x_i \Big\|_2^2} \quad \text{by Jensen's, as } x \to \sqrt{x} \text{ is concave} \\
& = \sqrt{\mathbf{E} \sum_{j=1}^d \Big(\sum_{i=1}^n \Omega_i x_{i,j} \Big)^2} \\
& = \sqrt{\mathbf{E} \sum_{j=1}^d \sum_{i=1}^n (\Omega_i x_{i,j})^2} \quad \text{as the } \Omega_i\text{'s are independent and } \mathbf{E}\Omega_i = 0 \\
& = \sqrt{\mathbf{E} \sum_{i=1}^n \|x_i\|_2^2} \le \sqrt{n} \max_i \|x_i\|_2 \quad \text{as } \Omega_i^2 = 1.
\end{aligned}
$$

Linear predictors $simplex/\ell_\infty$ constraints (Boosting)

Define d-dimensional simplex: $\Delta_d := \{w \in \mathbb{R}^d : \|w\|_1 = 1, w_1, \ldots, w_d \geq 0\}$.

(Proposition 3.4)

Let $\mathcal{A}_\Delta := \{x \in \mathbb{R}^d \to w^\top x : w \in c\Delta_d\}$. Then

$$\boxed{\mathtt{Rad}(\mathcal{A}_\Delta \circ \{x_1, \ldots, x_n\} \leq c\frac{\max_i \|x_i\|_\infty}{\sqrt{n}}\sqrt{2\log d}}$$

Note: typically, $\max_i \|x_i\|_\infty \not\propto d$, so overall dependence is $\sim \sqrt{\log d}$

(Similar result for Proposition 3.3 for ℓ_1/ℓ_∞ constraints. In that case we present a different argument in the lecture notes, based on Hölder's inequality $x^\top y \leq \|x\|_1 \|y\|_\infty$. The same argument used for the $simplex/\ell_\infty$ case also works)

Remark: Difference between d and $\log d$ is ultimately linked with the different dependence with the dimension d for the ℓ_2 and ℓ_1 ball, respectively.

Proof

- We have

$$n\,\mathtt{Rad}(\mathcal{A}_\Delta \circ \{x_1, \ldots, x_n\}) = \mathbf{E} \sup_{w \in \Delta_d} \sum_{i=1}^{n} \Omega_i w^\top x_i = \mathbf{E} \sup_{w \in \Delta_d} w^\top \Big(\sum_{i=1}^{n} \Omega_i x_i \Big)$$

- Note that for any vector $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ we have

$$\sup_{w \in \Delta_d} w^\top v = \max_{j \in 1:d} v_j$$

- Then,

$$\mathbf{E} \sup_{w \in \Delta_d} w^\top \Big(\sum_{i=1}^{n} \Omega_i x_i \Big) = \mathbf{E} \max_{j \in 1:d} \sum_{i=1}^{n} \Omega_i x_{i,j} = n\,\mathtt{Rad}(\mathcal{T})$$

with $\mathcal{T} = \{t_1 \ldots, t_d\}$ with $t_j = (x_{1,j}, \ldots, x_{d,j})$ for any $j \in \{1, \ldots, d\}$

- The proof follows by Massart's lemma as

$$\mathtt{Rad}(\mathcal{T}) \leq \max_{t \in \mathcal{T}} \|t\|_2 \frac{\sqrt{2 \log |\mathcal{T}|}}{n} \leq \sqrt{n} \max_i \|x_i\|_\infty \frac{\sqrt{2 \log d}}{n}$$

Feed-forward neural networks

- A **layer** $l^{(k)} : \mathbb{R}^{d_{k-1}} \to \mathbb{R}^{d_k}$ consists of a coordinate-wise composition of an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and an affine map:

$$l^{(k)}(x) := \sigma(W^{(k)}x + b^{(k)})$$

- A **neural network** with **depth** ι is the function (with $d_0 = d$, $d_\iota = 1$)

$$\boxed{f_{nn}^{\iota} : x \in \mathbb{R}^d \longrightarrow f_{nn}^{(\iota)}(x) := l^{(\iota)}(\cdots l^{(2)}(l^{(1)}(x))\cdots)}$$

(Proposition 3.6)

Let $\mathcal{A}_{nn}^{(\iota)} := \{x \in \mathbb{R}^d \to f_{nn}^{(\iota)}(x) : \|\mathbf{w}^{(k)}\|_\infty \leq \omega, \|b^{(k)}\|_\infty \leq \beta \ \forall k\}$.

$$\boxed{\mathtt{Rad}(\mathcal{A}_{nn}^{(\iota)} \circ \{x_1, \ldots, x_n\} \leq \frac{1}{\sqrt{n}}\Big(\beta \sum_{k=0}^{\iota-2} \omega^k + \omega^{\iota-1} \max_i \|x_i\|_\infty \sqrt{2\log(2d)}\Big)}$$