Lecture Notes:
Beyond Empirical Risk minimization
Local Averages

Alessandro Rudi, Pierre Gaillard

May 5th

In this lecture we start from a different characterization of the target function f^*. We have seen that it is defined globally as $f^* : X \to Y$ satisfying

$$R(f^*) = \inf_{f:X\to Y} R(f),$$

(the infimum is over the measurable functions from X to Y) that is

$$f^* \in \arg\min_{f:X\to Y} R(f).$$

Assume without loss of generality that

$$f^* = \arg\min_{f:X\to Y} R(f). \tag{1}$$

We can provide a pointwise characterization of f^* as follows

Theorem 1. *When Y is a compact set and ℓ is continuous, then*

$$f^*(x) = \arg\min_{y'\in Y} \mathbb{E}[\ell(y', y) \mid x], \tag{2}$$

almost everywhere, where $\mathbb{E}[q(y) \mid x]$ denotes the conditional expectation of $q(y)$ given x, with $q : Y \to \mathbb{R}$.

Proof. We sketch the proof as follows. Denote by $\widetilde{f}$ the function in Eq. 2. Note that by definition

$$\mathbb{E}[\ell(\widetilde{f}(x), y) \mid x] = \inf_{y'\in Y} \mathbb{E}[\ell(y', y) \mid x],$$

almost everywhere. Then, by noting that for any function $f : X \to Y$

$$\mathbb{E}[\ell(f(x), y) \mid x] \geq \inf_{y'\in Y} \mathbb{E}[\ell(y', y) \mid x] = \mathbb{E}[\ell(\widetilde{f}(x), y) \mid x],$$

we have for any $f : X \to Y$

$$R(\widetilde{f}) = \mathbb{E}[\ell(\widetilde{f}(x), y)] = \mathbb{E}_x[\mathbb{E}[\ell(\widetilde{f}(x), y) \mid x]] = \mathbb{E}_x[\inf_{y'\in Y} \mathbb{E}[\ell(y', y) \mid x]] \tag{3}$$

$$\leq \mathbb{E}_x[\inf_{y'\in Y} \mathbb{E}[\ell(y', y) \mid x]] \leq \mathbb{E}_x[\mathbb{E}[\ell(\widetilde{f}(x), y) \mid x]] = R(f). \tag{4}$$

So $R(\widetilde{f}) = \inf_{f:X\to Y} R(f)$. To conclude the proof we need to prove that $\widetilde{f}$ is measurable, which is rather technical and out of the scope of the lecture (see [1]). □

1 Learning via Local Averages

While the characterization in Eq. (1) suggested approaches like empirical risk minimization (we have seen it in the previous lecture), the characterization in terms of Eq. (2) gave rise to the so called *local average methods*. Denoting by $\rho(y|x)$ the conditional probability of y given x, and by $\widehat{\rho}(y|x)$ an estimator for $\rho(y|x)$, local averages estimators are of the form

$$\widehat{f}(x) = \arg\min_{y' \in Y} \int \ell(y', y) d\widehat{\rho}(y|x).$$

To study the excess risk for this estimator we perform the following analysis. Denote by $E(y', x)$ the function $E(y', x) = \int \ell(y', y) d\rho(y|x)$ and by $\widehat{E}(y', x)$ the function $\widehat{E}(y', x) = \int \ell(y', y) d\widehat{\rho}(y|x)$, then

$$\begin{aligned} R(\widehat{f}) - R(f^*) &= \mathbb{E}_x \left[E(\widehat{f}(x), x) - E(f^*(x), x) \right] \\ &= \mathbb{E}_x \left[E(\widehat{f}(x), x) - \widehat{E}(\widehat{f}(x), x) \right] + \mathbb{E}_x \left[\widehat{E}(\widehat{f}(x), x) - E(f^*(x), x) \right]. \end{aligned}$$

Now note that

$$\mathbb{E}_x \left[E(\widehat{f}(x), x) - \widehat{E}(\widehat{f}(x), x) \right] \leq \mathbb{E}_x \left[\sup_{y' \in Y} |E(y', x) - \widehat{E}(y', x)| \right].$$

Moreover, since $\widehat{E}(\widehat{f}(x), x) = \inf_{y' \in Y} \widehat{E}(y', x)$ and $E(f^*(x), x) = \inf_{y' \in Y} E(y', x)$, then

$$\mathbb{E}_x \left[\widehat{E}(\widehat{f}(x), x) - E(f^*(x), x) \right] = \mathbb{E}_x \left[\inf_{y' \in Y} \widehat{E}(y', x) - \inf_{y' \in Y} E(y', x) \right] \leq \mathbb{E}_x \left[\sup_{y' \in Y} |E(y', x) - \widehat{E}(y', x)| \right].$$

So finally

$$R(\widehat{f}) - R(f^*) \leq 2\mathbb{E}_x \left[\sup_{y' \in Y} |E(y', x) - \widehat{E}(y', x)| \right].$$

2 Density estimation

A classical way to estimate probability density is to approximate it via convolutions of the empirical distribution. Let q be a probability density (i.e. $q(x) = e^{-\|x\|^2}$) and $\tau^{-d} q_\tau(x) = q(x/\tau)$, for $\tau > 0$. Let moreover $x_1, \ldots, x_n$ sampled i.i.d. from ρ. We define the estimator as

$$\widehat{\rho}(x) = \frac{1}{n} \sum_{i=1}^{n} q_\tau(x - x_i).$$

By denoting by $\widehat{\rho}_n$ the probability $\widetilde{\rho}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ (where δ is the Dirac's delta) and by $\star$ the convolution operator (i.e. $(f \star g)(x) = \int f(y) g(x - y) dy$) we have

$$\rho \quad \approx \quad \rho \star q_\tau \quad \approx \quad \widetilde{\rho}_n \star q_\tau \quad = \quad \widehat{\rho}(x).$$

In particular

Lemma 1 (Bias)**.** *Let $|\rho(x) - \rho(y)| \leq C\|x - y\|$ for any x, y, then for any $v \in \mathbb{R}^d$*

$$\sup_x |\rho(x) - (\rho \star q_\tau)(x)| \leq CT\tau,$$

where $T := \int \|z\| q(z) dz$. (The integrals are assumed on $\mathbb{R}^d$).

Proof. Since $\int q_\tau(x - y)dy = \int q_\tau(y)dy = 1$, we have $\rho(x) = \int \rho(x) q_\tau(x - y)dy$ and so

$$\begin{aligned}|\rho(x) - (\rho \star q_\tau)(x)| &= |\tau^{-d} \int (\rho(x) - \rho(y)) q((x - y)/\tau)dy| \leq \tau^{-d} \int |\rho(x) - \rho(y)| q((x - y)/\tau)dy \\ &\leq C\tau^{-d+1} \int \|x - y\|/\tau \; q((x - y)/\tau)dy = C\tau^{-d+1} \int \|u/\tau\| q(u/\tau)du = C\tau \int \|z\| q(z)dz,\end{aligned}$$

where the last step is due to the change of variable $u/\tau \in \mathbb{R}^d \mapsto z \in \mathbb{R}^d$. □

Lemma 2 (Variance)**.** *For any $v \in X$, we have, for any $v \in \mathbb{R}^d$*

$$\mathbb{E}|(\rho \star q_\tau)(v) - \hat{\rho}(v)|^2 \leq \frac{Q\tau^{-d}}{n},$$

where $Q = \max_t q(t) \max_t \rho(t)$.

Proof. Define the random variable $z = q_\tau(v - x)$, with x distributed according to ρ. Now note that

$$\mathbb{E}z = \int q_\tau(v - x)d\rho(x) = \int q_\tau(v - x)\rho(x)dx = \rho \star q_\tau.$$

Let $z_1, \ldots, z_n$ defined as $z_i = q_\tau(v - x_i)$, since $x_1, \ldots, x_n$ are independently and identically distributed according to ρ, then $z_1, \ldots, z_n$ are independent copies of z and

$$\mathbb{E}|\frac{1}{n}\sum_{i=1}^n (z_i - \mathbb{E}z)|^2 = \frac{1}{n}\mathbb{E}(z_1 - \mathbb{E}z)^2$$

Now

$$\begin{aligned}\mathbb{E}(z - \mathbb{E}z)^2 \leq \mathbb{E}z^2 &= \int q_\tau(v - x)^2 \rho(x)dx \leq (\max_t q_\tau(t)) \int q_\tau(v - x)\rho(x)dx \\ &= (\max_t q_\tau(t))(q_\tau \star \rho)(v).\end{aligned}$$

To conclude note that $q_\tau \star \rho \leq \max_t \rho(t)$. □

Finally

Theorem 2. *Let ρ such that $|\rho(x) - \rho(y)| \leq C\|x - y\|$, then for any $v \in \mathbb{R}^d$*

$$\left(\mathbb{E}_v \mathbb{E}_{\mathcal{D}} |\rho(v) - \hat{\rho}(v)|^2\right)^{1/2} \;\leq\; CT\,\tau \;+\; \sqrt{\frac{Q\tau^{-d}}{n}}.$$

Proof. The result is obtained combining the two lemmas above □

Finally the result is obtained by optimizing τ to minimize the trade off between bias and variance. By choosing $\tau = n^{-\frac{1}{2+d}}$ we obtain for any $v \in \mathbb{R}^d$

$$\mathbb{E}\,|\rho(v) - \hat{\rho}(v)|^2 \leq (CT + \sqrt{Q})^2 \; n^{-\frac{2}{2+d}}.$$

3 Digression. Controlling the supremum of a non-discrete set of random variables

Note that the result above is in expectation with respect to the observed points $x_1, \dots, x_n$. Here we provide the tools to study the problem in high probability. To this aim we need to use some results from last class, in particular Lemma 2 that allows us to control the supremum of random variables, and the Bernstein inequality to control bounded random variables, in Lemma 4 from last class. We first introduce the concept of covering. The result we are going to obtain can be used to control the excess risk of an estimator when $\mathcal{F}$ is not discrete but continuous (compare with the last exercise of the previous class).

Definition 1 (Coverings of a set and covering numbers)**.** *Given a metric space S equipped with metric d and a compact subset X, we say that the set of points $\mathcal{C}(X, d, \varepsilon)$ is an ε-covering of X if*

$$X \subseteq \bigcup_{x \in \mathcal{C}(X,d,\varepsilon)} B_\varepsilon(x, d),$$

where $B_\varepsilon(x, d) = \{z \in S \mid d(x, z) \leq \varepsilon\}$. We denote with covering number *$\mathcal{N}(X, \varepsilon, d)$ the number*

$$\mathcal{N}(X, \varepsilon, d) = \min\{|\mathcal{C}(X, d, \varepsilon)| \mid \mathcal{C}(X, d, \varepsilon) \text{ is an } \varepsilon\text{-covering of } X\}.$$

Example. (Covering numbers of $[-R, R]^d$ with the $\|\cdot\|_\infty$ metric) Let $X = [-R, R]^d$, with $d \in \mathbb{N}$ and let $\|\cdot\|_\infty$ the ℓ_∞ metric defined as

$$\|x\|_\infty = \max_{j \in \{1,\dots,d\}} |x_j|,$$

for $x \in \mathbb{R}^d$. The ε-ball $B_\varepsilon(x_0, \|x\|_\infty)$ corresponds to a cube of side 2ε as follows

$$B_\varepsilon(x_0, \|x\|_\infty) = [x_0 - \varepsilon, x_0 + \varepsilon]^d.$$

Note that we can cover X with at most $\lceil \frac{R}{\varepsilon} \rceil^d$ cubes of sides 2ε, by disposing them on a grid of step 2ε. Then

$$\mathcal{N}([-R, R]^d, \|\cdot\|_\infty, \varepsilon) \leq \lceil \frac{R}{\varepsilon} \rceil^d.$$

More generally it is possible to prove the following theorem that holds for any metric space.

Theorem 3. *Let $R \geq \varepsilon > 0$. Let X be a subset of $\mathbb{R}^D$ such that $X \subseteq B_R(x_0, d)$ with $x_0 \in \mathbb{R}^D, R > 0$ and d a metric for $\mathbb{R}^D$. Then there exists a covering $\mathcal{C}(X, d, \varepsilon)$ with covering numbers*

$$\mathcal{N}(X, \|\cdot\|, \varepsilon) \leq \left(\frac{6R}{\varepsilon}\right)^D.$$

Now we are able to state the concentration result

Theorem 4. *Let $\sigma, B > 0$. Let $\mathcal{F}$ be a compact set with respect to the metric d and for any $\theta \in \mathcal{F}$ let $z_\theta^1, \dots, z_\theta^n$ be real independent random variables, such that*

$$\mathbb{E} z_\theta^i = 0, \quad \mathbb{E}|z_\theta^i|^2 \leq \sigma^2, \quad |z_\theta^i| \leq B,$$

for any $i = 1, \dots, n$. Moreover assume that there exists $L > 0$, such that for any $\theta, \theta' \in \mathcal{F}$ we have

$$|z_\theta - z_{\theta'}| \leq L d(\theta, \theta'),$$

Then the following holds for any $t > 0$

$$\mathbb{P}\left(\sup_{\theta \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} z_\theta^i \right| \; > \; L\varepsilon + t \right) \quad \leq \quad \mathcal{N}(\mathcal{F}, d, \varepsilon) e^{-\frac{t^2 n/2}{\sigma^2 + Bt}}.$$

Proof. Let $\mathcal{C}(\mathcal{F}, d, \varepsilon)$ be covering of $\mathcal{F}$ with cardinality $\mathcal{N}(\mathcal{F}, d, \varepsilon)$. Denote by u_θ the random variable

$$u_\theta = \left| \frac{1}{n} \sum_{i=1}^{n} z_\theta^i \right|.$$

Since

$$\mathcal{F} = \bigcup_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} B_\varepsilon(\bar{\theta}, d) \cap \mathcal{F},$$

then we have

$$\sup_{\theta \in \mathcal{F}} u_\theta = \sup_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} \sup_{\theta \in B_\varepsilon(\bar{\theta}, d) \cap \mathcal{F}} u_\theta \leq \sup_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} \sup_{\theta \in B_\varepsilon(\bar{\theta}, d)} u_\theta.$$

Now note that for any $\bar{\theta} \in \mathcal{F}$ and any $\theta \in B_\varepsilon(\bar{\theta}, d)$, by construction and the Lipschitzianity of z, we have

$$u_\theta = \left| \frac{1}{n} \sum_{i=1}^{n} (z_\theta^i - z_{\bar{\theta}}^i) + \frac{1}{n} \sum_{i=1}^{n} z_{\bar{\theta}}^i \right| \leq \frac{1}{n} \sum_{i=1}^{n} \left| z_\theta^i - z_{\bar{\theta}}^i \right| + u_{\bar{\theta}} \leq L d(\theta, \bar{\theta}) + u_{\bar{\theta}} \leq L\varepsilon + u_{\bar{\theta}}.$$

Then

$$\sup_{\theta \in \mathcal{F}} u_\theta \leq \sup_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} \sup_{\theta \in B_\varepsilon(\bar{\theta}, d)} u_\theta \leq \sup_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} \sup_{\theta \in B_\varepsilon(\bar{\theta}, d)} u_{\bar{\theta}} + L\varepsilon = \sup_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} u_{\bar{\theta}} + L\varepsilon.$$

Now we can use the bound on the supremum of a discrete set of random variables in Lemma 2 from previous class obtaining for $t > 0$,

$$\mathbb{P}(\sup_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} u_{\bar{\theta}} > t) \leq \sum_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} \mathbb{P}(u_{\bar{\theta}} > t).$$

Moreover since we have a bound on the absolute value and the variance of the random variables $z_{\bar{\theta}}^i$ for any $\bar{\theta}$, then we can use Bernstein inequality from Lemma 4 of previous class, obtaining

$$\mathbb{P}(u_{\bar{\theta}} > t) \leq e^{-\frac{t^2 n/2}{\sigma^2 + Bt}}.$$

Then, we obtain

$$\sum_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} \mathbb{P}(u_{\bar{\theta}} > t) \leq \mathcal{N}(\mathcal{F}, d, \varepsilon) e^{-\frac{t^2 n/2}{\sigma^2 + Bt}},$$

and so, by Lemma 1 of the previous class

$$\mathbb{P}(\sup_{\theta \in \mathcal{F}} u_\theta > L\varepsilon + t) \leq \mathbb{P}(\sup_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} u_{\bar{\theta}} > t) \leq \sum_{\bar{\theta} \in \mathcal{C}(\mathcal{F}, d, \varepsilon)} \mathbb{P}(u_{\bar{\theta}} > t) \leq \mathcal{N}(\mathcal{F}, d, \varepsilon) e^{-\frac{t^2/2}{\sigma^2 + Bt}}.$$

□

4 Consistency of density estimation

We are now using the theorem above to derive a proper consistency analysis for density estimation. The goal is to obtain a bound for the error of the estimator $\widehat{\rho}$ of the form

$$\sup_{v\in X} |\widehat{\rho}(v) - \rho(v)|,$$

that holds in high probability. As we did in Section 2 we split the error in bias and variance

$$\sup_{v\in X} |\widehat{\rho}(v) - (\rho \star q_\tau)(v)| \leq \sup_{v\in X} |\widehat{\rho}(v) - \rho(v)| \; + \; \sup_{v\in X} |\widehat{\rho}(v) - (\rho \star q_\tau)(v)|.$$

The bias is controlled by Lemma 1. For the variance we are going to apply Thm. 4. Let $X = [-R, R]^d$ with $R > 0$, Let $x_1, \ldots, x_n$ be the sampled independently from ρ. Define the random variable z_v^i with $v \in X$ as

$$z_v^i = q_\tau(v - x_i) - (\rho \star q_\tau)(v).$$

In particular note, that since x_i is sampled from ρ, we have

$$\mathbb{E} q_\tau(v - x_i) = \int q_\tau(v - x_i)\rho(x)dx = \rho \star q_\tau.$$

Then by definition

$$\widehat{\rho}(v) - (\rho \star q_\tau)(v) = \frac{1}{n}\sum_{i=1}^{n} z_v^i.$$

Moreover

$$\mathbb{E} z_v^i = 0,$$

by construction, and we have

$$\sup_{v\in X} |z_v^i| \leq \sup_{v\in X} |q_\tau(v - x_i)| + \sup_{v\in X} \int q_\tau(v - x)\rho(x)dx \leq 2 \sup_{z\in\mathbb{R}^d} q_\tau(z) \leq Q\tau^{-d},$$

moreover analogously to Lemma 2

$$\mathbb{E}|z_v^i|^2 = \int (q(v - x) - (\rho \star q_\tau)(v))^2 \rho(x) \leq \int q(v - x)^2 \rho(x) \leq Q\tau^{-d}.$$

To apply the theorem of previous section we assume that q is Lipschitz with constant L. Then

$$|q_\tau(z) - q_\tau(z')| = \tau^{-d}|q(z/\tau) - q(z'/\tau)| \leq \tau^{-d-1}\|z - z'\|,$$

so we have

$$|z_v^i - z_{v'}^i| \leq |q_\tau(v - x_i) - q_\tau(v' - x_i)| + \int |q_\tau(v - x) - q_\tau(v' - x)|\rho(x)dx \leq 2L\tau^{-d-1}\|v - v'\|.$$

Now we are ready to apply the Thm 4,

$$\mathbb{P}\left(\sup_{v\in X} |\widehat{\rho}(v) - (\rho \star q_\tau)(v)| > 2L\varepsilon + t\right) \leq \mathcal{N}(X, \|\cdot\|, \varepsilon) e^{-\frac{t^2 n}{Q\tau^{-d}(1+t)}}.$$

Considering that $\mathcal{N}([-R,R]^d, \|\cdot\|, \varepsilon) \leq (\frac{6R}{\varepsilon})^d$ by Thm. 3 and rewriting t in terms to a given confidence level δ, we have $\frac{t^2}{1+t} = Qt^{-d}(\log\frac{1}{\delta} + d\log\frac{6R}{\varepsilon})$ that implies

$$\mathbb{P}\left(\sup_{v\in X}|\widehat{\rho}(v) - (\rho\star q_\tau)(v)| > 2L\varepsilon + \frac{Qt^{-d}(\log\frac{1}{\delta} + d\log\frac{6R}{\varepsilon})}{n} + \sqrt{\frac{Qt^{-d}(\log\frac{1}{\delta} + d\log\frac{6R}{\varepsilon})}{n}}\right) \leq \delta.$$

In particular we can choose ε to minimize the bound in the equation above. Choosing $\varepsilon = 1/n$ we have

$$\mathbb{P}\left(\sup_{v\in X}|\widehat{\rho}(v) - (\rho\star q_\tau)(v)| > \frac{2L + Qt^{-d}(\log\frac{1}{\delta} + d\log 6Rn)}{n} + \sqrt{\frac{Qt^{-d}(\log\frac{1}{\delta} + d\log(6Rn))}{n}}\right) \leq \delta.$$

To conclude the analysis we combine this result with the bias term in Lemma 1 and we select $\tau = n^{-\frac{1}{2+d}}$ as before, obtaining

$$\mathbb{P}\left(\sup_{v\in X}|\widehat{\rho}(v) - \rho(v)| > CTn^{-\frac{1}{2+d}} \;+\; C_1(n)n^{-\frac{2}{2+d}} + C_2(n)^{1/2}n^{-\frac{1}{2+d}}\right). \leq \delta.$$

where

$$C_1(n) = 2L + C_2(n), \quad C_2(n) = Q(\log(1/\delta) + d\log 6Rn).$$

This leads to bound of the form

$$\sup_{v\in[-R,R]^d}|\widehat{\rho}(v) - \rho(v)| = O\left(n^{-\frac{1}{2+d}}\sqrt{\log(1/\delta) + d\log Rn}\right),$$

with probability $1-\delta$.

5 Estimators for the conditional expectation

Assume here to have $X \subseteq \mathbb{R}^d$, that $Y \subset \mathbb{R}$ and that $\rho(y|x), \rho(y,x), \rho(x)$ are probability densities. We characterize $\rho(y|x)$ as

$$\rho(y|x) = \frac{\rho(y,x)}{\rho(x)}.$$

Usually estimators for the conditional probability have the following form

$$\widehat{\rho}(y|x) = \frac{\widehat{\rho}(y,x)}{\widehat{\rho}(x)},$$

where $\widehat{\rho}(y,x)$ and $\widehat{\rho}(x)$ are estimators for $\rho(y,x)$ and $\rho(x)$. The estimator for $\rho(x,y)$ can be derived in the same way as we did for the one of $\rho(x)$,i.e., using $(x_1,y_1),\ldots,(x_n,y_n) \in \mathbb{R}^{d'}$ with $d' = d+p$ where d is the dimension of the euclidean space containing X and p the dimension of the space containing Y.

References

[1] D Aliprantis Charalambos and Kim C Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer, 2006.