Recap on Probably Approximately Correct learning theory

Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
https://mathematical-tours.github.io
www.numerical-tours.com

December 31, 2020

Abstract

This document is a short presentation of some important results of Probably Approximately Correct (PAC) learning theory. Its goal is to asses the generalization performance of learning methods. The main reference (and in particular the proofs of the mentioned results) is the fantastic book "Foundations of Machine Learning" by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar https://cs.nyu.edu/~mohri/mlbook/ and the cristal clear course notes "Learning Theory from First Principles" https://www.di.ens.fr/~fbach/learning_theory_class/index.html of Francis Bach.

The underlying assumption is that the data $\mathcal{D} \overset{\text{def.}}{=} (x_i, y_i)_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ are independent realizations of a random vector (X, Y). The goal is to learn from $\mathcal{D}$ alone a function $f : \mathcal{X} \to \mathcal{Y}$ which is as close a possible of minimizing the risk

$$L(f) \overset{\text{def.}}{=} \mathbb{E}(\ell(Y, f(X))) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, f(x)) \mathrm{d}\mathbb{P}_{X,Y}(x, y)$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is some loss function. In order to achieve this goal, the method selects a class of functions $\mathcal{F}$ and minimizes the empirical risk

$$\hat{f} \overset{\text{def.}}{=} \operatorname*{argmin}_{f\in\mathcal{F}} \hat{L}(f) \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)).$$

This is thus a random function (since it depends on $\mathcal{D}$).

Example 1 (Regression and classification). In regression, $\mathcal{Y} = \mathbb{R}$ and the most common choice of loss is $\ell(y, z) = (y - z)^2$. For binary classification, $\mathcal{Y} = \{-1, +1\}$ and the ideal choice is the 0-1 loss $\ell(y, z) = 1_{y\neq z}$. Minimizing $\hat{L}$ for the 0-1 loss is often NP hardÂ for most choice of $\mathcal{F}$. So one uses other loss functions of the form $\ell(y, z) = \Gamma(-yz)$ where φ is a convex function upper-bounding $1_{\mathbb{R}^+}$, which makes $\min_f \hat{L}(f)$ a convex optimization problem.

The goal of PAC learning is to derive, with probability at least $1 - \delta$ (intended to be close to 1), a bound on the generalization error $L(\hat{f}) - \inf(L) \geqslant 0$ (also called excess risk), and this bound depends on n and δ. In order for this generalization error to go to zero, one needs to put some hypothesis on the distribution of (X, Y).

1 Non parametric setup and calibration

If $\mathcal{F}$ is the whole set of measurable functions, the minimizer $f^\star$ of L is often called "Bayes estimator" and is the best possible estimator.

Risk decomposition Denoting

$$\alpha(z|x) \stackrel{\text{def.}}{=} \mathbb{E}_Y(\ell(Y,z)|X=x) = \int_{\mathcal{Y}} \ell(y,f(x))\mathrm{d}\mathbb{P}_{Y|X}(y|x)$$

the average error associate to the choice of some predicted value $z \in \mathcal{Y}$ at location $x \in \mathcal{X}$, one has the decomposition of the risk

$$L(f) = \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} \ell(y,f(x))\mathrm{d}\mathbb{P}_{Y|X}(y|x) \right] \mathrm{d}\mathbb{P}_X(x) = \int_{\mathcal{X}} \alpha(f(x)|x)\mathrm{d}\mathbb{P}_X(x)$$

so that computing $f^\star$ can be done independently at each location x by solving

$$f^\star(x) = \operatorname*{argmin}_z \alpha(z|x).$$

Example 2 (Regression). For regression applications, where $\mathcal{Y} = \mathbb{R}$, $\ell(y,z) = (y-z)^2$, one has $f^\star(x) = \mathbb{E}(Y|X=x) = \int_{\mathcal{Y}} y\mathrm{d}\mathbb{P}_{Y|X}(y|x)$, which is the conditional expectation.

Example 3 (Classification). For classification applications, where $\mathcal{Y} = \{-1,1\}$, it is convenient to introduce $\eta(x) \stackrel{\text{def.}}{=} \mathbb{P}_{Y|X}(y=1|x) \in [0,1]$. If the two classes are separable, then $\eta(x) \in \{0,1\}$ on the support of X (it is not defined elsewhere). For the 0-1 loss $\ell(y,z) = 1_{y=z} = 1_{\mathbb{R}^+}(-yz)$, one has $f^\star = \mathrm{sign}(2\eta-1)$ and $L(f^\star) = \mathbb{E}_X(\min(\eta(X), 1-\eta(X)))$. In practice, computing this η is not possible from the data $\mathcal{D}$ alone, and minimizing the 0-1 loss is NP hard for most $\mathcal{F}$. Considering a loss of the form $\ell(y,z) = \Gamma(-yz)$, one has that the Bayes estimator then reads in the fully non-parametric setup

$$f^\star(x) = \operatorname*{argmin}_z \alpha(z|x) = \eta(x)\Gamma(-z) + (1-\eta(x))\Gamma(z),$$

so that it is non-linear function of $\eta(x)$.

Calibration in the classification setup A natural question is to ensure that in this (non realistic ...) setup, the final binary classifier $\mathrm{sign}(f^\star)$ is equal to $\mathrm{sign}(2\eta-1)$, which is the Bayes classifier of the (non-convex) 0-1 loss. In this case, the loss is said to be calibrated. Note that this does not mean that $f^\star$ is itself equal to $2\eta-1$ of course. One has the following result.

Proposition 1. *A loss loss ℓ associated to a convex Γ is calibrated if and only if Γ is differentiable at 0 and $\Gamma'(0) > 0$.*

In particular, the hinge and logistic loss are thus calibrated. Denoting L_Γ the loss associated to $\ell(y,z) = \Gamma(-yz)$, and denoting $\Gamma_0 = 1_{\mathbb{R}^+}$ the 0-1 loss, stronger quantitative controls are of the form

$$L_{\Gamma_0}(f) - \inf L_{\Gamma_0} \leqslant \Psi(L_\Gamma(f) - \inf L_\Gamma) \tag{1}$$

for some increasing function $\Psi : \mathbb{R}^+ \to \mathbb{R}^+$. Such a control ensures in particular that if $f^\star$ minimize L_Γ, it also minimizes L_{Γ_0} and hence $\mathrm{sign}(f^\star) = \mathrm{sign}(2\eta-1)$ and the loss is calibrated. One can show that the hinge loss enjoys such a quantitative control with $\Psi(r) = r$ and that the logistic loss has a worse control since it requires $\Psi(s) = \sqrt{s}$.

2 PAC bounds

Bias-variance decomposition. For a class $\mathcal{F}$ of functions, the excess risk of the empirical estimator

$$\hat{f} \stackrel{\text{def.}}{=} \operatorname*{argmin}_{f\in\mathcal{F}} \hat{L}(f)$$

is decomposed as the sum of the estimation (random) error and the approximation (deterministic) error

$$L(\hat{f}) - \inf L = \Big[L(\hat{f}) - \inf_{\mathcal{F}} L\Big] + \mathcal{A}(\mathcal{F}) \quad \text{where} \quad \mathcal{A}(\mathcal{F}) \stackrel{\text{def.}}{=} \Big[\inf_{\mathcal{F}} L - \inf L\Big]. \tag{2}$$

This splitting is a form of variance/bias separation.

As the size of $\mathcal{F}$ increases, the estimation error increases but the approximation error decreases, and the goal is to optimize this trade-off. This size should thus depend on n to avoid over-fitting (selecting a too large class $\mathcal{F}$).

Approximation error. Bounding the approximation error fundamentally requires some hypothesis on $f^\star$. This is somehow the take home message of "no free lunch" results, which shows that learning is not possible without regularity assumption on the distribution of (X, Y). We only give here a simple example.

Example 4 (Linearly parameterized functionals). A popular class of functions are linearly parameterized maps of the form

$$f(x) = f_w(x) = \langle \varphi(x),\, w\rangle_{\mathcal{H}}$$

where $\varphi : \mathcal{X} \to \mathcal{H}$ somehow "lifts" the data features to a Hilbert space $\mathcal{H}$. In the particular case where $\mathcal{X} = \mathbb{R}^p$ is already a (finite dimensional) Hilbert space, one can use $\varphi(x) = x$ and recovers usual linear methods. One can also consider for instance polynomial features, $\varphi(x) = (1, x_1, \ldots, x_p, x_1^2, x_1x_2, \ldots)$, giving rise to polynomial regression and polynomial classification boundaries. One can then use a restricted class of functions of the form $\mathcal{F} = \{f_w \;;\; \|w\|_{\mathcal{H}} \leqslant R\}$ for some radius R, and if one assumes for simplicity that $f^\star = f_{w^\star}$ is of this form, and that the loss $\ell(y, \cdot)$ is Q-Lipschitz, then the approximation error is bounded by an orthogonal projection on this ball

$$\mathcal{A}(\mathcal{F}) \leqslant Q\mathbb{E}(\|\varphi(x)\|_{\mathcal{H}}) \max(\|w^\star\|_{\mathcal{H}} - R, 0).$$

Remark 1 (Connexions with RKHS). Note that this lifting actually corresponds to using functions f in a reproducing Hilbert space, denoting

$$\|f\|_k \stackrel{\text{def.}}{=} \inf_{w \in \mathcal{H}} \{\|w\|_{\mathcal{H}} \;;\; f = f_w\}$$

and the associated kernel is $k(x, x') \stackrel{\text{def.}}{=} \langle \varphi(x),\, \varphi(x')\rangle_{\mathcal{H}}$. But this is not important for our discussion here.

Estimation error. The estimation error can be bounded for arbitrary distributions by leveraging concentration inequalities (to controls the impact of the noise) and using some bound on the size of $\mathcal{F}$.

The first simple but fundamental inequality bounds the estimator error by some uniform distance between L and $\hat{L}$. Denoting $g \in \mathcal{F}$ an optimal estimator such that $L(g) = \inf_{\mathcal{F}} L$ (assuming for simplicity it exists) one has

$$L(\hat{f}) - \inf_{\mathcal{F}} L = \Big[L(\hat{f}) - \hat{L}(\hat{f})\Big] + \Big[\hat{L}(\hat{f}) - \hat{L}(g)\Big] + \Big[\hat{L}(g) - L(\hat{g})\Big] \leqslant 2 \sup_{f \in \mathcal{F}} |\hat{L}(f) - L(f)| \tag{3}$$

since $\hat{L}(\hat{f}) - \hat{L}(g) \geqslant 0$. So the goal is "simply" to control $\Delta(\mathcal{D}) \stackrel{\text{def.}}{=} \sup_{\mathcal{F}} |\hat{L} - L|$, which is a random value (depending on the data $\mathcal{D}$).

The following proposition, which is a corollary of McDiarmid inequality, bounds with high probability the deviation of $\Delta(\mathcal{D})$ from its mean.

Proposition 2 (McDiarmid control to the mean)**.** *If $\ell(Y, f(X))$ is almost surely bounded by ℓ_∞ for any $f \in \mathcal{F}$, then with probability $1 - \delta$,*

$$\Delta(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}(\Delta(\mathcal{D})) \leqslant \ell_\infty \sqrt{\frac{2\log(1/\delta)}{n}}. \tag{4}$$

Now we need to control $\mathbb{E}_{\mathcal{D}}(\Delta(\mathcal{D}))$ which requires to somehow bound the size of $\mathcal{F}$. This can be achieved by the so-called Vapnik-Chervonenkis (VC) dimension, but this leads to overly pessimistic (dimension-dependent) bounds for linear models. A more refined analysis makes use of the so-called Rademacher complexity of a set of functions $\mathcal{G}$ from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$

$$\mathcal{R}_n(\mathcal{G}) \stackrel{\text{def.}}{=} \mathbb{E}_{\varepsilon,\mathcal{D}}\Big[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(x_i, y_i)\Big]$$

where ε_i are independent Bernoulli random variable (i.e. $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$). Note that $\mathcal{R}_n(\mathcal{G})$ actually depends on the distribution of (X, Y) ...

Here, one needs to apply this notion of complexity to the functions $\mathcal{G} = \ell[\mathcal{F}]$ defined as

$$\ell[\mathcal{F}] \stackrel{\text{def.}}{=} \{(x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto \ell(y, f(x)) \; ; \; f \in \mathcal{F}\},$$

and that one has the following control, which can be proved using a simple but powerful symmetrization trick.

Proposition 3. *One has*

$$\mathbb{E}(\Delta(\mathcal{D})) \leqslant 2\mathcal{R}_n(\ell[\mathcal{F}]).$$

If ℓ is Q-Lipschitz, one furthermore has $\mathcal{R}_n(\ell[\mathcal{F}]) \leqslant Q\mathcal{R}_n(\mathcal{F})$ (here the class of functions only depends on x).

Putting (2), (3), Propositions 2 and 3 together, one obtains the following final result.

Theorem 1. *Assuming ℓ is Q-Lipschitz and bounded by ℓ_∞ on the support of $(Y, f(X))$, one has with probability $1 - \delta$*

$$L(\hat{f}) - \inf L \leqslant 2\ell_\infty \sqrt{\frac{2\log(1/\delta)}{n}} + 4Q\mathcal{R}_n(\mathcal{F}) + \mathcal{A}(\mathcal{F}). \tag{5}$$

Example 5 (Linear models). In the case where $\mathcal{F} = \{\langle \varphi(\cdot), w\rangle_{\mathcal{H}} \; ; \; \|w\| \leqslant R\}$ where $\|\cdot\|$ is some norm on $\mathcal{H}$, one has

$$\mathcal{R}_n(\mathcal{F}) \leqslant \frac{R}{n}\Big\|\sum_i \varepsilon_i \varphi(x_i)\Big\|_*$$

where $\|\cdot\|_*$ is the so-called dual norm

$$\|u\|_* \stackrel{\text{def.}}{=} \sup_{\|w\|\leqslant 1} \langle u, w\rangle_{\mathcal{H}}.$$

In the special case where $\|\cdot\| = \|\cdot\|_{\mathcal{H}}$ is Hilbertian, then one can further simplify this expression since $\|u\|_* = \|u\|_{\mathcal{H}}$ and

$$\mathcal{R}_n(\mathcal{F}) \leqslant \frac{R\sqrt{\mathbb{E}(\|\varphi(x)\|_{\mathcal{H}}^2)}}{\sqrt{n}}.$$

This result is powerful since the bound does not depend on the feature dimension (and can be even applied in the RKHS setting where $\mathcal{H}$ is infinite dimensional). In this case, one sees that the convergence speed in (5) is of the order $1/\sqrt{n}$ (plus the approximation error). One should keep in mind that one needs also to select the "regularization" parameter R to obtain the best possible trade-off. In practice, this is done by cross validation on the data themselves.

Example 6 (Application to SVM classification). One cannot use the result (5) in the case of the 0-1 loss $\ell(y, z) = 1_{z\neq y}$ since it is not Lipschitz (and also because minimizing $\hat{L}$ would be intractable). One can however applies it to a softer piece-wise affine upper-bounding loss $\ell_\rho(z, y) = \Gamma_\rho(-zy)$ for

$$\Gamma_\rho(s) \stackrel{\text{def.}}{=} \min\left(1, \max\left(0, 1 + s/\rho\right)\right).$$

This function is $1/\rho$ Lipschitz, and it is sandwiched between the 0-1 loss and a (scaled) hinge loss

$$\Gamma_0 \leqslant \Gamma_\rho \leqslant \Gamma_{\mathrm{SVM}}(\cdot/\rho) \quad \text{where} \quad \Gamma_{\mathrm{SVM}}(s) \stackrel{\text{def.}}{=} \max(1+s, 0).$$

This allows one, after a change of variable $w \mapsto w/\rho$, to bound with probability $1-\delta$ the 0-1 risk using a SVM risk by applying (5)

$$L_{\Gamma_0}(\hat{f}) - \inf_{f \in \mathcal{F}_\rho} L_{\Gamma_{\mathrm{SVM}}}(f) \leqslant 2\sqrt{\frac{2\log(1/\delta)}{n}} + 4\frac{\sqrt{\mathbb{E}(\|\varphi(x)\|_{\mathcal{H}}^2)}/\rho}{\sqrt{n}}$$

where $\mathcal{F}_\rho \stackrel{\text{def.}}{=} \{f_w = \langle \varphi(\cdot), w\rangle_{\mathcal{H}} \; ; \; \|w\| \leqslant 1/\rho\}$. In practice, one rather solves a penalized version of the above risk (in its empirical version)

$$\min_w \hat{L}_{\Gamma_{\mathrm{SVM}}}(f_w) + \lambda \|w\|_{\mathcal{H}}^2 \tag{6}$$

which corresponds to the so-called kernel-SVM method. The kernel trick allows one to solve this problem by recasting this possibly infinite dimensional problem into a problem of finite dimension n by observing that the solution necessarily has the form $w = \sum_{i=1}^n c_i \varphi(x_i)$ for some $c \in \mathbb{R}^n$ and plugging this expression in (6) only necessitates the evaluation of the empirical kernel matrix $(k(x_i, x_j) \stackrel{\text{def.}}{=} \langle \varphi(x_i), \varphi(x_j)\rangle)_{i,j=1}^n$.