Chapter 7: Recent Advances in Empirical Process Theory: Application to Statistical Learning

Empirical Processes (MATH-522)

Myrto Limnios

May, 2025

In this Chapter we will see how empirical process theory is used to explain recent achievements of particular Machine Learning algorithms with interest in classification. Without any further discussion, we place ourselves in highdimensional settings, wherein classical statistical theory might be applicable only at the price of some strong requisites. We will particularly focus on how to go beyond the rate $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$, and will give an insight on model selection procedures based on complexity penalization.

1 Framework and Basic Results

1.1 Framework

When considering an observation (image, text, etc.), the basic goal of classification is to predict its class. Suppose $x \in \mathcal{X}$ with $\mathcal{X}$ being the feature space equipped with its σ-algebra. In binary classification, the class of x is encoded in a variable y valued in $\{-1, +1\}$, and the goal is to build a measurable function $h : \mathcal{X} \to \{-1, +1\}$ to predict the class of x. We say that the *classifier* h commits an error if its prediction differs from its true class, i.e., if $h(x) \neq y$.

The learning problem is the following. Let (X, Y) be a $\mathcal{X} \times \{-1, 1\}$ pair of r.v.s, of p.d. described by the pair (P, η), where P is the marginal of X, and η is the *posterior distribution* of Y given X, i.e., $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

The *performance of the classifier* h is measured by its probability of error:

$$L(h) = \mathbb{P}(h(X) \neq Y) \ . \tag{1}$$

We formulate a straightforward result, yet fundamental.

Lemma 1.1. *The oracle classifier minimizing the probability of error is given by*

$$h^* : x \in \mathcal{X} \mapsto 2 1\{\eta(x) > 1/2\} - 1 \ , \tag{2}$$

and it is called the Bayes classifier. It satisfies

$$L(h) \geq L(h^*) =: L^*, \quad \forall h \ ,$$

and it holds true for any classifier h

$$L(h) - L^* = \mathbb{E}[1\{h(x) \neq h^*(x)\}|2 1\{\eta(x) > 1/2\} - 1|] \ . \tag{3}$$

In practice as h^* is unknown we estimate it based on a sample of observations. Let $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} =: \mathcal{D}_n$ composed of n-i.i.d. copies of (X, Y) that we use to build an empirical classifier $h_n(X) = h_n(X, \mathcal{D}_n)$. We measure the performance of h_n based on its conditional probability of error $L(h_n) = \mathbb{P}(h_n(X) \neq Y | \mathcal{D}_n)$.

The goal of Empirical Risk Minimization (ERM) is to find the *best* (see how) classificer h_n achieving a performance tending to that of h^*.

1.2 Empirical Risk Minimization

Definition 1.2. Let a class $\mathcal{H}$ composed of classifiers being measurable functions $h : \mathcal{X} \to \{-1, +1\}$. Define the *empirical error* of the classifier h based on $\mathcal{D}_n$ to be:

$$L_n(h) = \frac{1}{n} \sum_{i=1}^{n} 1\{h(X_i) \neq Y_i\} \ .$$

It is the most natural estimator of the probability of error $L(h)$.

We seek to estimate the empirical minimizer of L_n being

$$h_n^* \in \arg\min_{h \in \mathcal{H}} L_n(h) \ ,$$

corresponding to minimizing the average of mislabellings we commit on $\mathcal{D}_n$ using an element of the class $\mathcal{H}$. We list some basic properties that highlight the necessity to prove uniform control of the worst estimation error within the class $\mathcal{H}$.

Proposition 1.3. *Consider the previous notations, then the following assertions hold true with probability one, and for all $h \in \mathcal{H}$.*

1. $L_n(h_n^*) \leq L_n(h)$
2. $L(h_n^*) - \inf_{h \in \mathcal{H}} L(h) \leq 2 \sup_{h \in \mathcal{H}} |L_n(h) - L(h)|$
3. $L(h_n^*) \leq L_n(h_n^*) + 2 \sup_{h \in \mathcal{H}} |L_n(h) - L(h)|$

Proof. Exercise. □

Remark 1.1 (Interpretation).

1. *If we can guarantee that* $\sup_{h \in \mathcal{H}} |L_n(h) - L(h)|$ *is small with high probability, then the probability of error using h_n^* is not much larger than the probability of error within the class $\mathcal{H}$.*
2. *We recognize a quantity for which we proved finite sample control of its expectation, taking the form of maximal inequalities (Chapter 6), namely* $\sup_{h \in \mathcal{H}} |L_n(h) - L(h)|$. *It essentially corresponds to the uniform deviation of the empirical measure L_n w.r.t. its mean L.*

Theorem 1.4. *Suppose the class $\mathcal{H}$ to be a* VC*-class of functions, with finite* VC*-dimension $\mathcal{V}$. Let $\delta \in (0, 1)$, then the estimation error of h_n^* is upperbounded as follows, with probability at least $1 - \delta$,*

$$L(h_n^*) - \inf_{h \in \mathcal{H}} L(h) \leq C \sqrt{\frac{\mathcal{V}}{n}} + \sqrt{\frac{2 \log(1/\delta)}{n}} \ ,$$

with $C > 0$ being a universal constant.

Proof. Exercise. □

In the following section, we will focus on explaining how to obtain faster convergence rates than $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$.

2 Fast Rates of Convergence

We intend to explain some phenomena witnessed in experiments wherein algorithms converge to the oracle at rate $\mathcal{O}_{\mathbb{P}}(1/n^{\alpha})$, with $\alpha \in [1/2, 1]$. The key ingredient for these recent results take advantage on second-order control of the second moments, being natural for binary-valued functions in $\{0,1\}$.

2.1 Relative Deviation

We begin with some analysis highlighting the importance of controlling the variance of empirical processes.

Consider a generic measurable function $g : \mathcal{Z} \to \{0,1\}$ and consider an n-i.i.d. random sample $Z_1, \ldots, Z_n$. Notice that the empirical measure $P_n g$ based on the Z_i-s is just an average of Bernoulli r.v.s with parameter the mean Pg. Then, for any fixed function g, because the images $g(Z_1), \ldots, g(Z_n)$ are a.s. bounded, Hoeffding's inequality yields w.p. at least $1-\delta$

$$Pg - P_n g \leq \sqrt{\frac{2\log(1/\delta)}{n}} \ . \tag{4}$$

Notice that the most difficult case is when $Pg = 1/2$. Also, the variance is also bounded, and Bernstein's inequality yields

$$Pg - P_n g \leq \sqrt{\frac{2\mathrm{Var}_P(g)\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n} \ . \tag{5}$$

Proof. Exercise. □

Indeed, because the function g is valued in $\{0,1\}$, then $Var_P(g) \leq Pg(1-Pg) \leq Pg$. Again, the smaller the expectation, the tighter the bound. We already see some hope for improvement from a rate of $1/\sqrt{n}$ requiring some further assumptions and analysis of the variance.

We now state a Theorem going back to Vapnik and Chervonenkis (around 1980). We are essentially interested in proving advanced results for the estimation error of the empirical classifier defined in the previous section.

We want to tackle the eventuality of having candidate functions in $\mathcal{G}$ that potentially could present very large values but with small probability (e.g. unbounded functions) and that of functions with Pg close to $1/2$, i.e., having high variance.

As first important idea, still studied now, is to rescale by $\sqrt{Pg}$ the empirical deviation, ending up with quantity having similar behavior. Thus we want for bound

$$\sup_{g\in\mathcal{G}} \frac{P_n g - Pg}{\sqrt{Pg}} \ .$$

Theorem 2.1 (Vapnik-Chervonenkis, 1982). *Let $\mathcal{G}$ be a class of measurable functions valued in $\{0,1\}$, and denote by $m_{2n}(\mathcal{G})$ the shattering coefficient based on the dataset $\mathcal{D} = \{Z_1, \ldots, Z_{2n}\}$ of i.i.d. r.v.s. Let $\delta \in (0,1)$, then w.p. at least $1-\delta$, it holds true for all $g \in \mathcal{G}$*

$$\begin{aligned} \frac{P_n g - Pg}{\sqrt{Pg}} &\leq 2\sqrt{\frac{\log m_{2n}(\mathcal{G}) + \log(4/\delta)}{n}} \ , \\ \frac{P_n g - Pg}{\sqrt{P_n g}} &\leq 2\sqrt{\frac{\log m_{2n}(\mathcal{G}) + \log(4/\delta)}{n}} \ . \end{aligned}$$

Remark 2.1. *If $P_n g \leq (1-t)Pg$ for all g then (Exercise)*

$$Pg \leq 4\frac{\log m_{2n}(\mathcal{G}) + \log(4/\delta)}{t^2 n} .$$

We now apply it in classification.

Theorem 2.2 (Classification)**.** *Let $h_n^* \in \mathcal{H}$ be the empirical minimizer of the empirical probability of error L_n based on $\mathcal{D}_n$. Suppose $\mathcal{G}$ to be* VC*-class with finite* VC*-dimension $\mathcal{V}$. Let $\delta \in (0,1)$, then w.p. at least $1-\delta$,*

$$L(h_n^*) \leq L_n(h_n^*) + 2\sqrt{L_n(h_n^*)\frac{\mathcal{V}\log(n+1) + \log(4/\delta)}{n}} + 4\frac{\mathcal{V}\log(n+1) + \log(4/\delta)}{n} .$$

Proof. Exercise. □

Remark 2.2.

1. *Notice that all the quantities on the right-hand-side are known and only depend on $\mathcal{G}$ and on the dataset $\mathcal{D}_n$.*
2. *Relative deviation yields generalization guarantees interpolating between rates $\mathcal{O}_{\mathbb{P}}(\sqrt{\mathcal{V}\log(n)/n})$ and $\mathcal{O}_{\mathbb{P}}(\mathcal{V}\log(n)/n)$ with multiplier being the empirical classification error of h_n^*.*
3. *Suppose a quite strong assumption, that the minimizer of $L(h)$ in $\mathcal{H}$ achieves no error, i.e., $\exists h' \in \mathcal{H}$, $L(h') = 0$, i.e., $h'(X) = Y$ a.s. then $L_n(h_n^*) = 0$ and we get*

$$L(h_n^*) - \inf_{h\in\mathcal{H}} L(h) \leq 4\frac{\mathcal{V}\log(n+1) + \log(4/\delta)}{n} .$$

So now the question is, how to interpolate between the two terms appearing in the upperbound on Thereom 2.2?

Corollary 2.3. *Suppose that the infimum of $L(h)$ over $\mathcal{H}$ is achieved, i.e., that there exists $h' \in \mathcal{H}$ such that $L(h') = \inf_{h\in\mathcal{H}} L(h)$. Let $\delta \in (0,1)$, then w.p. at least $1-\delta$,*

$$L(h_n^*) - L(h') \leq C\sqrt{L(h')\frac{\mathcal{V}\log(n+1) + \log(4/\delta)}{n}} + C\frac{\mathcal{V}\log(n+1) + \log(4/\delta)}{n} ,$$

where $C > 0$ is a universal constant.

Proof. Exercise. □

2.2 Noise Conditions

The key ingredient is the relation between the variance and the mean for functions valued in $\{0,1\}$ to obtain fast rates of convergence for $L(h_n^*) - \inf_{h\in\mathcal{H}} L(h)$. In particular, if we are able to choose the class of candidate functions $\mathcal{H}$ such the the infimum equals to 0 (ideal setting), then we expect the upperbound using $\sup_{h\in\mathcal{H}} |L_n(h) - L(h)|$ to be loose.

To use this intuition, consider $\mathcal{G}$ to be composed of functions based on $\mathcal{H}$ defined as follows $g(x,y) \mapsto 1\{h(x) \neq y\} - 1\{h'(x) \neq y\}$, with $L(h') = \inf_{h\in\mathcal{H}} L(h)$.

Massart and Mammen-Tsybakov's noise conditions. Suppose the class $\mathcal{H}$ is composed of N functions fulfilling for some $s > 0$ and $\alpha \in (0, 1]$,

$$\mathrm{Var}(g) \leq \left(\frac{Pg}{s}\right)^{\alpha} . \tag{6}$$

Theorem 2.4. *Consider the framework above, and suppose that there exists a pair $(s, \alpha) \in (0, \infty) \times (0, 1]$, such that for all $g \in \mathcal{G}$, the condition (6) is fulfilled. Let $\delta \in (0, 1)$, then w.p. at least $1 - \delta$,*

$$L(h_n^*) - \inf_{h \in \mathcal{H}} L(h) \leq \left(2\frac{\log(N/\delta)}{ns^{\alpha}}\right)^{1/(2-\alpha)} .$$

Proof. Exercise. □

Remark 2.3. *We may wonder whether (6) is a reasonable assumption to make. Let us consider classification framework, and let h^* be Bayes classifier and that η is bounded away from $1/2$, i.e., suppose that there exists $s > 0$ such that $|\eta(x) - 1/2| > s$, for all $x \in \mathcal{X}$. This is known as* **Massart's noise condition**. *Consider the class $\mathcal{G}$ based on the classifiers $\mathcal{H}$. We can conclude that*

$$\begin{aligned} \mathrm{Var} g &\leq \mathbb{E}[(1\{h(x) \neq y\} - 1\{h^*(x) \neq y\})^2] \leq \mathbb{E}[(1\{h(x) \neq h^*(x)\})^2] = \mathbb{E}[1\{h(x) \neq h^*(x)\}] \\ &\leq (1/s)\mathbb{E}[1\{h(x) \neq h^*(x)\}|\eta(x) - 1/2|] = (1/s)(L(h) - L^*) = Pg/s , \end{aligned}$$

where we used the assumption of the posterior distribution, and Lemma 1.1. Thus Eq. (6) is fulfilled with $\alpha = 1$, and thus

Theorem 2.4 yields

$$L(h_n^*) - L^* \leq 2\frac{\log(N/\delta)}{ns} .$$

It is clear that the larger s is, and the easier the problem is, yielding a tighter control of the generalization error of the empirical minimizer h_n^*.

We will see that we can relax the condition (6), or Massart's noise condition.

Definition 2.5 (Mammen-Tsybakov's noise condition)**.** The noise in binary classification satisfies *Mammen-Tsybakov's noise condition* if there exists constants $\alpha \in (0, 1)$, $B > 0$, $t_0 \in (0, 1/2]$, such that for all $t \in [0, t_0]$

$$\mathbb{P}(|\eta(X) - 1/2| \leq t) \leq Bt^{\alpha/(1-\alpha)} .$$

Remark 2.4.

1. *It relaxes Massart's condition by assuming that the posterior distribution is bounded aways from $1/2$ with large probability.*
2. *Notice that is the zero-error case, i.e., when $\eta \in \{0, 1\}$, then we obtain fast rates.*

We formulate an important consequence of Mammen-Tsybakov's noise condition, relating the variance to the mean, as initially wished.

Lemma 2.6. *Suppose Mammen-Tsybakov's noise condition to be fulfilled for a triplet of constants α, B, t_0. Then, it holds true that for any classifier $h \in \mathcal{H}$:*

$$\mathbb{P}(h(X) \neq h^*(X)) \leq C(L(h) - L^*)^{\alpha} ,$$

where the constant $C > 0$ depends on α, B, t_0.

Proof. Recall that

$$\begin{aligned} L(h) - L^* &= \mathbb{E}[1\{h(x) \neq h^*(x)\}|\eta(x) - 1/2|] \geq \mathbb{E}[1\{h(x) \neq h^*(x)\}|\eta(x) - 1/2|1\{|\eta(x) - 1/2| \geq t\}] \\ &\geq 2t\mathbb{P}[h(x) \neq h^*(x), |\eta(x) - 1/2| > t] \geq 2t\mathbb{P}[h(x) \neq h^*(x)] - 2t\mathbb{P}[|\eta(x) - 1/2| \leq t] \\ &\geq 2t\mathbb{P}[h(x) \neq h^*(x)] - 2Bt^{\alpha/(1-\alpha)} . \end{aligned}$$

We can choose t, such that $t \in [0, t_0]$. Take $t = b\mathbb{P}[h(x) \neq h^*(x)]^{(1-\alpha)/\alpha}$, with $b > 0$ constant, such that $b \leq t_0$, then the last quantity equals to $2b\mathbb{P}[h(x) \neq h^*(x)]^{1/\alpha} - 2Bb^{\alpha/(1-\alpha)}\mathbb{P}[h(x) \neq h^*(x)]$. Because the second term is positive, we can choose $C = 2b^{-\alpha}$ and the result is obtained. □

We conclude by proving fast rates for the excess of risk of the empirical minimizer h_n^* under Mammen-Tsybakov's noise condition.

Theorem 2.7. *Suppose the conditions of Mammen-Tsybakov's noise condition to be fulfilled. We consider the size of the class $\mathcal{H}$ to be finite and equal to $N \in \mathbb{N}^*$. Let $\delta \in (0,1)$, with probability at least $1-\delta$, it holds true that*

$$L(h_n^*) - \inf_{h\in\mathcal{H}} L(h) \leq C\left(\frac{\log(N/\delta)}{n}\right)^{1/(2-\alpha)} ,$$

where the constant $C > 0$ depends on α, B, t_0.

Proof. Board. □

Remark 2.5.
1. *This bound is of similar flavor as that of Theorem 2.4. It interpolates between the slow regime ($\alpha \to 0$) and the fast one ($\alpha \to 1$). Notice in addition that we loose the effect of parameter s: if the η is bounded away from $1/2$, then $s \to 1/2$, and thus the upperbound under Massart's condition will be sharper.*
2. *For both Theorems, the empirical minimizer h_n^* is estimated independently on the noise condition. It is thus independent on α, thanks to ERM, that is very important (because unknown in practice).*
3. *Again, we emphasize that those results hold true only if the oracle classifier is in the class $\mathcal{H}$!!*
4. *In statistical learning literature, the exposed noise assumptions are often referred to* **margin condition***.*
5. *More general classes of classifiers. There is a rich literature establishing generalization of the theory exposed, under entropic/VC conditions on $\mathcal{H}$, that rely on localization of the risk formulating a Bernstein-type of bound in terms of the modulus of continuity of the functions, initially proved by Talagrand. We will not study this case however. Recently, it was further extended by not considering that $h^* \in \mathcal{H}$ (e.g. Bousquet and Zhivotovskiy, 2020).*

A minimax lower bound. Lastly, we may wonder whether the stated rates are informative, i.e., if we could think of a better method to select a classifier than by ERM. We will give some intuition by stating a minimax lower bound. We consider $\mathcal{H}$ to have finite VC-dimension $\mathcal{V} > 0$. We want to show that for any picked classifier from $\mathcal{H}$, we can find a distribution on the data, performing worse than the bound. So, we want to establish a lower bound on the quantity

$$\inf_{h_n} \sup_{G(\mathrm{d}x,\mathrm{d}y)} \{L(h_n) - \inf_{h\in\mathcal{H}} L(h)\}$$

where the infimum is taken over all datadriven estimators in $\mathcal{H}$, and the supremum is taken over all the distributions G of the pair (X, Y), and h_n is a classifier depending on the dataset $\mathcal{D}_n$ (not restricted to ERM solutions).

Notice first a direct consequence of Corollary 2.3 is the following result.

Corollary 2.8. *Consider the assumptions of Corollary 2.3 to be fulfilled. Then, there exists a constant $C > 0$, such that*

$$\mathbb{E}L(h_n) - \inf_{h\in\mathcal{H}} L(h) \leq C\sqrt{\frac{\inf_{h\in\mathcal{H}} L(h)\mathcal{V}\log n}{n}} + C\frac{\mathcal{V}\log n}{n} .$$

Proof. Exercise. □

Before stating the main result, we formulate lower bounds for two extreme cases (no margin restriction and zero-error classifier), when the class of functions $\mathcal{H}$ is supposed to be VC-type, from Haussler, Littlestone, and Warmuth (1994) and Vapnik Chervonenkis (1971,1979) for instance.

Corollary 2.9. *Consider the assumptions of Corollary 2.3 to be fulfilled. And consider the class of joint distributions G of the pair (X, Y). The following two assertions hold true.*

1. *If $s = 0$, then*
$$\inf_{h_n} \sup_G \mathbb{E}L(h_n) - L^* \geq \frac{e^{-8}}{2\sqrt{6}} \sqrt{\frac{\mathcal{V}-1}{n}} \ , \quad \forall n \geq 5(\mathcal{V}-1) \tag{7}$$

2. *If $s = 1$, then*
$$\inf_{h_n} \sup_G \mathbb{E}L(h_n) - L^* \geq \frac{\mathcal{V}-1}{4n} \ , \quad \forall n \geq \max(2, \mathcal{V}-1) \ , \tag{8}$$

where the infimum is taken over all datadriven estimators.

Massart and Nédélec (2006) proved a lower bound interpolating between the two extreme cases.

Theorem 2.10. *Suppose the minimax risk over the set of distributions fulfilling Massart's noise condition for a given $s > 0$. Then, there exists a universal constant $C > 0$, such that if $\mathcal{V} \geq 2$, then*

$$\inf_{h_n} \sup_G \mathbb{E}L(h_n) - L^* \geq C \min(\frac{\mathcal{V}}{ns}, \sqrt{\frac{\mathcal{V}}{n}}) \ ,$$

for all $n \geq \mathcal{V}$.

It proves that there is a small gap between the lower and upperbounds, of rate $\log n$. Again, it has been improved under some additional conditions on $\mathcal{H}$. Notice that the margin condition $s > \sqrt{\frac{\mathcal{V}}{n}}$ implies fast rates but at the price of the constant C.