Algorithmic Foundations of Learning **Lecture 7**

Bernstein's Concentration Inequalities. Fast Rates

Lecturer: Patrick Rebeschini *Version: November 6th 2019*

7.1 Introduction

In the last lecture we derived the following bound in probability, which says something about the *distribution* of the fluctuations of the excess risk $r(A^\star) - r(a^\star)$ around its mean, where $A^\star$ is the empirical risk minimizer (Theorem 6.13): if the loss function ℓ is bounded in the interval $[0, c]$, then

$$\mathbf{P}\left(r(A^\star) - r(a^\star) < 4\,\mathbf{E}\,\texttt{Rad}(\mathcal{L} \circ \{Z_1, \dots, Z_n\}) + c\sqrt{2\frac{\log(1/\delta)}{n}}\right) \geq 1 - \delta$$

Prior to that, we analyzed different settings in regression and classification and derived bounds on the expected value that take the following form

$$\mathbf{E}\,\texttt{Rad}(\mathcal{L} \circ \{Z_1, \dots, Z_n\}) \leq \frac{f(\text{dimension of the data, complexity of } \mathcal{A})}{\sqrt{n}}.$$

Plugging the two results together we get

$$\mathbf{P}\left(r(A^\star) - r(a^\star) < \frac{f(\text{dimension of the data, complexity of } \mathcal{A})}{\sqrt{n}} + c\sqrt{2\frac{\log(1/\delta)}{n}}\right) \geq 1 - \delta,$$

which seems to indicate that the excess risk of the empirical risk minimizer goes to zero with the "slow" rate $1/\sqrt{n}$ both in probability and in expectation.

Today we will see that under additional assumptions on the structure of the problem we can establish the "fast" rate $1/n$, and, more generally, a rate of the form $1/n^\alpha$ for $\alpha \in [1/2, 1]$. To achieve this, we first need to introduce a new type of concentration inequalities known as Bernstein's inequalities, which can exploit the variance of random variables. The good news is that this new type of concentration inequalities relies on the exact same general framework described in the previous lecture, where results are phrased in term of the convex conjugate $\psi^\star$ of a function ψ that upper bounds the moment generating function: $\mathbf{E}\,e^{\lambda(X - \mathbf{E}X)} \leq e^{\psi(\lambda)}$ for all $\lambda \geq 0$. The only difference is that today we will consider a class of random variables that is more general than sub-Gaussian, yielding better decay for $(\psi^\star)^{-1}$ and hence better confidence bounds.

7.2 Bernstein's Inequality

Definition 7.1 (One-sided Bernstein's condition) *A random variable X is said to satisfy the* one-sided Bernstein's condition *with parameter $b > 0$ if*

$$\boxed{\mathbf{E}\,e^{\lambda(X - \mathbf{E}X)} \leq \exp\left(\frac{(\mathbf{Var}X)\lambda^2/2}{1 - b\lambda}\right) \qquad \textit{for any } \lambda \in [0, 1/b)}$$

A direct application of the optimal Chernoff's bound (Proposition 6.3) immediately yields the upper-tail and upper-confidence bounds given in the proposition below. Henceforth, we define the function

$$u \in \mathbb{R}_+ \longrightarrow h(u) := 1 + u - \sqrt{1 + 2u},$$

which has inverse

$$u \in \mathbb{R}_+ \longrightarrow h^{-1}(u) = u + \sqrt{2u}.$$

Proposition 7.2 (Bernstein's upper-tail bound) *Let X be a random variable that satisfies the one-sided Bernstein's condition with parameter $b > 0$. Then, for any $\varepsilon \geq 0$ and any $\delta \in (0, 1)$ we have*

$$\boxed{\mathbf{P}(X - \mathbf{E}X \geq \varepsilon) \leq \exp\bigg(- \frac{\mathbf{Var}X}{b^2} h\bigg(\frac{b\varepsilon}{\mathbf{Var}X}\bigg)\bigg) \leq \exp\bigg(- \frac{\varepsilon^2/2}{\mathbf{Var}X + b\varepsilon}\bigg)}$$

$$\boxed{\mathbf{P}\Big(X - \mathbf{E}X < b\log(1/\delta) + \sqrt{2(\mathbf{Var}X)\log(1/\delta)}\Big) \geq 1 - \delta}$$

Proof: Let $\psi(\lambda) = \frac{a\lambda^2/2}{1-b\lambda}$ for $\lambda \in [0, 1/b)$, with $a = \mathbf{Var}X$. It can be showed that for every $\varepsilon > 0$ we have

$$\psi^\star(\varepsilon) = \sup_{\lambda\in[0,1/b)} \bigg(\varepsilon\lambda - \frac{a\lambda^2/2}{1 - b\lambda}\bigg) = \frac{a}{b^2} h\bigg(\frac{b\varepsilon}{a}\bigg) \geq \frac{\varepsilon^2/2}{a + b\varepsilon},$$

where $h(u) = 1 + u - \sqrt{1 + 2u}$ for $u > 0$ and we used that $h(u) \geq \frac{u^2}{2(1+u)}$ for $u > 0$. The upper-tail bound follows by Proposition 6.3, as

$$e^{-\psi^\star(\varepsilon)} \leq \exp\bigg(- \frac{a}{b^2} h\bigg(\frac{b\varepsilon}{a}\bigg)\bigg) \leq \exp\bigg(- \frac{\varepsilon^2/2}{a + b\varepsilon}\bigg).$$

The upper-confidence bound follows as $h^{-1}(u) = u+\sqrt{2u}$ for $u > 0$, so that setting $e^{-\psi^\star(\varepsilon)} = \exp(-\frac{a}{b^2}h(\frac{b\varepsilon}{a})) = \delta$ yields $\varepsilon = b\log(1/\delta) + \sqrt{2a\log(1/\delta)}$. ■

The one-sided Bernstein's condition is preserved by weighted sums of independent random variables when the weights are non-negative. In particular, if $X_1, \ldots, X_n$ are independent random variables that satisfy the one-sided Bernstein's condition with the same parameter b, then $\frac{1}{n}(X_1 + \cdots + X_n)$ satisfies the one-sided Bernstein's condition with parameter b/n:

$$\mathbf{E}\, e^{\lambda\frac{1}{n}\sum_{i=1}^n(X_i-\mathbf{E}X_i)} = \prod_{i=1}^n \mathbf{E}\, e^{\frac{\lambda}{n}(X_i-\mathbf{E}X_i)} \leq \prod_{i=1}^n \exp\bigg(\frac{(\mathbf{Var}X_i/n^2)\lambda^2/2}{1 - (b/n)\lambda}\bigg) = \exp\bigg(\frac{\mathbf{Var}(\frac{1}{n}\sum_{i=1}^n X_i)\lambda^2/2}{1 - (b/n)\lambda}\bigg).$$

This fact immediately yields the following concentration inequality, which is a corollary of Lemma 6.4 in the case of random variables that satisfy the one-sided Bernstein's condition.

Corollary 7.3 (Bernstein's inequality) *Let $X_1, \ldots, X_n \sim X$ be i.i.d. random variables that satisfy the one-sided Bernstein's condition with parameters $b > 0$. Then, for any $\varepsilon \geq 0$ and any $\delta \in [0, 1]$ we have*

$$\boxed{\mathbf{P}\bigg(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X \geq \varepsilon\bigg) \leq \exp\bigg(- \frac{n\mathbf{Var}X}{b^2} h\bigg(\frac{b\varepsilon}{\mathbf{Var}X}\bigg)\bigg) \leq \exp\bigg(- \frac{n\varepsilon^2/2}{\mathbf{Var}X + b\varepsilon}\bigg)}$$

$$\boxed{\mathbf{P}\bigg(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X < \frac{b}{n}\log(1/\delta) + \sqrt{\frac{2(\mathbf{Var}X)\log(1/\delta)}{n}}\bigg) \geq 1 - \delta}$$

The upper-confidence bound in Bernstein's inequality captures the trade-off between noise and rates and illustrates the main property that we need to exploit to get fast rates: in the extreme case that the variance is zero, i.e., $\mathbf{Var}X = 0$, the upper confidence interval decreases with the fast rate $1/n$. Otherwise, the slow rate $1/\sqrt{n}$ will eventually dominate as n grows. Later on we will see that the fast rate can be achieved even when $\mathbf{Var}X$ is not strictly zero, but when it decays fast enough.

The relevance of the one-sided Bernstein's condition is given by the fact that this condition is satisfied by any upper-bounded random variable.

Proposition 7.4 *If $X - \mathbf{E}X \leq c$ for a given $c > 0$, then X satisfies the one-sided Bernstein's condition with parameter $b = c/3$.*

Proof: Let $\mu = \mathbf{E}X$ and $\sigma^2 = \mathbf{Var}X$. By the series expansion of the exponential function we have, for any $\lambda \in \mathbb{R}$,

$$\mathbf{E}e^{\lambda(X-\mu)} = 1 + \mathbf{E}\sum_{k=2}^{\infty} \lambda^k \frac{(X-\mu)^k}{k!} = 1 + \lambda^2 \mathbf{E}[(X-\mu)^2 g(\lambda(X-\mu))]$$

where $g : \mathbb{R} \to \mathbb{R}$ is defined as

$$g(u) = \sum_{k=2}^{\infty} \frac{u^{k-2}}{k!} = \frac{e^u - u - 1}{u^2}.$$

Note that the function g is monotonically non-decreasing, as its first derivative is non-negative:

$$g'(u) = \frac{e^u(u-2) + u + 2}{u^3} \geq 0 \qquad \text{for any } u \in \mathbb{R}.$$

Therefore, using the assumption $X - \mu < c$ we find, using that $\lambda \geq 0$,

$$\mathbf{E}e^{\lambda(X-\mu)} \leq 1 + \lambda^2 \mathbf{E}[(X-\mu)^2 g(\lambda c)] = 1 + \lambda^2 g(\lambda c)\mathbf{E}[(X-\mu)^2] = 1 + \lambda^2\sigma^2 g(\lambda c).$$

Using the inequality $k!/2 \geq 3^{k-2}$ for any $k \geq 2$ and the bound $1 + x \leq e^x$, we obtain, for any $\lambda \in [0, 3/c)$,

$$\mathbf{E}e^{\lambda(X-\mu)} \leq 1 + \frac{\lambda^2\sigma^2}{2}\sum_{k=2}^{\infty}\left(\frac{\lambda c}{3}\right)^{k-2} = 1 + \frac{\lambda^2\sigma^2/2}{1-\lambda c/3} \leq \exp\left(\frac{\lambda^2\sigma^2/2}{1-\lambda c/3}\right).$$

■

Hence, a bounded random variable $|X - \mathbf{E}X| \leq c$ is both sub-Gaussian (with variance proxy $\sigma^2 = c^2$, by Hoeffding's Lemma 2.1) and it satisfies the one-sided Bernstein's condition (with parameter $c/3$, by Proposition 7.4; in fact, it satisfies the two-sided Bernstein's condition as we will see in the section below). Let us now compare Hoeffding's inequality (Corollary 6.8) and Bernstein's inequality (Corollary 7.3) in the case of bounded random variables. Let $X_1, \ldots, X_n \sim X$ be i.i.d. random variables bounded in the interval $[-c, c]$ for $c > 0$. Then, for any $\varepsilon \geq 0$,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X \geq \varepsilon\right) \leq e^{-n\varepsilon^2/(2c^2)} \qquad \text{Hoeffding's inequality}$$

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2/2}{\mathbf{Var}X + c\varepsilon/3}\right) \qquad \text{Bernstein's inequality}$$

and, for any $\delta \in [0, 1]$,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X < \sqrt{\frac{2c^2\log(1/\delta)}{n}}\right) \geq 1 - \delta \qquad \text{Hoeffding's inequality}$$

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X < \frac{c}{3n}\log(1/\delta) + \sqrt{\frac{2(\mathbf{Var}X)\log(1/\delta)}{n}}\right) \geq 1 - \delta \qquad \text{Bernstein's inequality}$$

We see that Bernstein's inequality provides much sharper control when $\mathbf{Var} X \ll c^2$. See **Problem 2.7** in the Problem Sheets for further insights about comparing different concentration bounds.

7.3 Two-sided Bernstein's condition and sub-exponential r.v.'s

The theory developed so far in this lecture holds only for upper-tail and upper-confidence bounds. A similar theory can be developed to yield only lower bounds, or both upper and lower bounds. In the case of upper and lower bound, we need the two-sided Bernstein's condition (with parameter $b > 0$), which reads

$$\boxed{\mathbf{E}\, e^{\lambda(X - \mathbf{E} X)} \le \exp\left(\frac{(\mathbf{Var} X)\lambda^2/2}{1 - b|\lambda|}\right) \qquad \text{for any } \lambda \in (-1/b, 1/b)} \tag{7.1}$$

If $|X| \le c$, then X satisfies the two-sided Bernstein's condition with parameter $b = c/3$ as the following proposition shows.

Proposition 7.5 *If $|X - \mathbf{E} X| \le c$ for a given $c > 0$, then X satisfies the two-sided Bernstein's condition with parameter $b = c/3$.*

Proof: Let $\mu = \mathbf{E} X$ and $\sigma^2 = \mathbf{Var} X$. By the series expansion of the exponential function we have, for any $\lambda \ge 0$,

$$\mathbf{E} e^{\lambda(X-\mu)} = 1 + \sum_{k=2}^{\infty} \lambda^k \frac{\mathbf{E}[(X-\mu)^k]}{k!} \le 1 + \lambda^2 \sigma^2 \sum_{k=2}^{\infty} \frac{(|\lambda| c)^{k-2}}{k!},$$

where we used that

$$\mathbf{E}[(X-\mu)^k] = \mathbf{E}[(X-\mu)^2 (X-\mu)^{k-2}] \le \mathbf{E}[(X-\mu)^2 |X-\mu|^{k-2}] \le \mathbf{E}[(X-\mu)^2 c^{k-2}] = c^{k-2}\sigma^2.$$

Using the inequality $k!/2 \ge 3^{k-2}$ for any $k \ge 2$ and the bound $1 + x \le e^x$, we obtain, for any $|\lambda| < 3/c$,

$$\mathbf{E} e^{\lambda(X-\mu)} \le 1 + \frac{\lambda^2\sigma^2}{2} \sum_{k=2}^{\infty} \left(\frac{|\lambda| c}{3}\right)^{k-2} = 1 + \frac{\lambda^2\sigma^2/2}{1 - |\lambda| c/3} \le \exp\left(\frac{\lambda^2\sigma^2/2}{1 - |\lambda| c/3}\right).$$

■

If a random variable satisfies the two-sided Bernstein's condition, then it belongs to the class of random variables called *sub-exponential*, which is characterized by the bound $\mathbf{E}\, e^{\lambda(X - \mathbf{E} X)} \le e^{a\lambda^2}$ for any $\lambda \in (-1/c, 1/c)$, for given $a, c \ge 0$. See **Problem 3.3** in the Problem Sheets. Note that if $c = 0$ we recover the sub-Gaussian case, so this class is more general. Recall that for an exponential random variable with mean $b > 0$ we have

$$\mathbf{E}\, e^{\lambda(X - \mathbf{E} X)} = \begin{cases} \frac{1}{1-b\lambda} e^{-b\lambda} & \text{for } \lambda < 1/b \\ +\infty & \text{otherwise} \end{cases}$$

and there exists $a, c \ge 0$ such that $\mathbf{E}\, e^{\lambda(X - \mathbf{E} X)} \le e^{a\lambda^2}$ for any $\lambda \in (-1/c, 1/c)$. It can be shown that the sub-exponential property is equivalent to the fact that the moment generating function is finite in a neighbourhood of 0, which gives an indication of the importance of this class.

7.4 Fast Rates: Back to Learning Part III

To explore the trade-off between noise and rate embodied by Bernstein's inequality, we consider the binary classification case when $|\mathcal{A}| < \infty$. Let us first recall the setting of binary classification:

- Training data $Z_1, \dots, Z_n$ such that $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$;
- Admissible action set $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \to \{-1, 1\}\}$;
- Loss function $\ell(a, (x, y)) = \phi(a(x), y)$, for $\phi : \{-1, 1\}^2 \to \mathbb{R}_+$.

We consider the "true" zero-one loss function $\phi(\hat{y}, y) = \mathbf{1}_{\hat{y} \neq y}$. In this case, for each $a \in \mathcal{B}$ we have

$$\begin{aligned} r(a) &= \mathbf{E}\phi(a(X), Y) = \mathbf{P}(a(X) \neq Y), \\ R(a) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{a(X_i) \neq Y_i}. \end{aligned}$$

Recall the definitions:

$$a^\star = \operatorname*{argmin}_{a \in \mathcal{A}} r(a), \qquad a^{\star\star} = \operatorname*{argmin}_{a \in \mathcal{B}} r(a), \qquad A^\star = \operatorname*{argmin}_{a \in \mathcal{A}} R(a).$$

By Example 1.6, the Bayes decision rule $a^{\star\star}$ reads

$$a^{\star\star}(x) = \operatorname*{argmax}_{\hat{y} \in \mathcal{Y}} \mathbf{P}(Y = \hat{y} | X = x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ -1 & \text{if } \eta(x) \leq 1/2 \end{cases} \tag{7.2}$$

where the *regression function* $\eta : \mathbb{R}^d \to [0, 1]$ is defined as

$$\eta(x) := \mathbf{P}(Y = 1 | X = x).$$

The regression function measures the amount of noise in the problem. If $\eta \in \{0, 1\}$, then Y has to be a deterministic (i.e., noiseless) function of X. On the other hand, if $\eta = 1/2$, then Y does not contain any information on X. The following result makes precise the link between the regression function η, the excess risk $r(a) - r(a^{\star\star})$, and the optimal Bayes error $r(a^{\star\star})$.

Theorem 7.6 *For any classifier $a \in \mathcal{B}$ we have*

$$\boxed{r(a) - r(a^{\star\star}) = \mathbf{E}[|2\eta(X) - 1| \mathbf{1}_{a(X) \neq a^{\star\star}(X)}]}$$

Moreover,

$$\boxed{r(a^{\star\star}) = \mathbf{E}\min\{\eta(X), 1 - \eta(X)\} \leq \frac{1}{2}}$$

Proof: By the tower property and the "take out what is known" property of conditional expectations we have, for any $a \in \mathcal{B}$,

$$\begin{aligned} r(a) = \mathbf{P}(a(X) \neq Y) = \mathbf{E}\,\mathbf{1}_{a(X) \neq Y} &= \mathbf{E}[\mathbf{1}_{a(X) \neq Y}(\mathbf{1}_{a(X)=1} + \mathbf{1}_{a(X)=-1})] \\ &= \mathbf{E}[\mathbf{1}_{Y=-1}\mathbf{1}_{a(X)=1}] + \mathbf{E}[\mathbf{1}_{Y=1}\mathbf{1}_{a(X)=-1}] \\ &= \mathbf{E}\,\mathbf{E}[\mathbf{1}_{Y=-1}\mathbf{1}_{a(X)=1}|X] + \mathbf{E}\,\mathbf{E}[\mathbf{1}_{Y=1}\mathbf{1}_{a(X)=-1}|X] \\ &= \mathbf{E}[\mathbf{1}_{a(X)=1}\mathbf{E}[\mathbf{1}_{Y=-1}|X]] + \mathbf{E}[\mathbf{1}_{a(X)=-1}\mathbf{E}[\mathbf{1}_{Y=1}|X]] \\ &= \mathbf{E}[\mathbf{1}_{a(X)=1}(1 - \eta(X)) + \mathbf{1}_{a(X)=-1}\eta(X)]. \end{aligned}$$

Hence,

$$r(a) - r(a^{\star\star}) = \mathbf{E}[(\mathbf{1}_{a(X)=1} - \mathbf{1}_{a^{\star\star}(X)=1})(1 - \eta(X)) + (\mathbf{1}_{a(X)=-1} - \mathbf{1}_{a^{\star\star}(X)=-1})\eta(X)].$$

As each $a \in \mathcal{B}$ takes values in $\{-1, 1\}$ we have $\mathbf{1}_{a(X)=1} = 1 - \mathbf{1}_{a(X)=-1}$ and

$$\begin{aligned} r(a) - r(a^{\star\star}) &= \mathbf{E}[-(\mathbf{1}_{a(X)=-1} - \mathbf{1}_{a^{\star\star}(X)=-1})(1-\eta(X)) + (\mathbf{1}_{a(X)=-1} - \mathbf{1}_{a^{\star\star}(X)=-1})\eta(X)] \\ &= \mathbf{E}[(2\eta(X)-1)(\mathbf{1}_{a(X)=-1} - \mathbf{1}_{a^{\star\star}(X)=-1})]. \end{aligned}$$

The term $(\mathbf{1}_{a(X)=-1} - \mathbf{1}_{a^{\star\star}(X)=-1})$ takes the value 0 when a and $a^{\star\star}$ agree, the value -1 when they disagree and $a^{\star\star}(X) = -1$ and the value 1 when they disagree and $a^{\star\star}(X) = 1$. Recalling definition (7.2) of $a^{\star\star}$, we above can be rewritten as

$$r(a) - r(a^{\star\star}) = \mathbf{E}[(2\eta(X)-1)\operatorname{sign}(\eta(X)-1/2)\mathbf{1}_{a(X)\neq a^{\star\star}(X)}] = \mathbf{E}[|2\eta(X)-1|\mathbf{1}_{a(X)\neq a^{\star\star}(X)}],$$

where

$$\operatorname{sign}(u) := \begin{cases} 1 & \text{if } u > 0 \\ -1 & \text{if } u \leq 0 \end{cases}$$

Finally, if we take $a = a^{\star\star}$ given in (7.2), this yields

$$\begin{aligned} r(a^{\star\star}) &= \mathbf{E}[\mathbf{1}_{\eta(x)>1/2}(1-\eta(X)) + \mathbf{1}_{\eta(x)\leq 1/2}\eta(X)] \\ &= \mathbf{E}[(\mathbf{1}_{\eta(x)>1/2} + \mathbf{1}_{\eta(x)\leq 1/2})\min\{\eta(X), 1-\eta(X)\}] \\ &= \mathbf{E}\min\{\eta(X), 1-\eta(X)\} \leq \frac{1}{2}. \end{aligned}$$

■

Theorem 7.6 shows that $r(a^{\star\star}) = 1/2$ if and only if $\eta(X) = 1/2$ almost surely, which is when Y contains no information on X, i.e., when Y and X are independent. To see this, note that if $\eta(X) = \mathbf{P}(Y=1|X) = 1/2$ almost surely then

$$\mathbf{P}(Y=1) = \mathbf{E}\mathbf{P}(Y=1|X) = \frac{1}{2},$$

so that $\mathbf{P}(Y=1) = \mathbf{P}(Y=1|X) = 1/2$ (i.e., X and Y are independent).

When η is close to 1/2, the optimal Bayes risk is large; no classifier can do well and the excess risk is small. When η is far from 1/2, either close to 0 or to 1, the Bayes classifier is smaller and the excess risk is bigger.

7.4.1 Massart's noise condition

Definition 7.7 (Massart's noise condition) *There exists $\gamma \in (0, 1/2]$ such that for any $x \in \mathbb{R}^d$ we have*

$$\boxed{\left|\eta(x) - \frac{1}{2}\right| \geq \gamma}$$

Remark 7.8 *Massart's noise condition originally reads as the following weaker version: there exists $\gamma \in (0, 1/2]$ such that*

$$\boxed{\mathbf{P}\left(\left|\eta(X) - \frac{1}{2}\right| \geq \gamma\right) = 1}$$

Here the requirement $|\eta(x) - \frac{1}{2}| \geq \gamma$ is assumed to hold for $\mathbf{P}$-almost-every *x, i.e., for a set of x's with probability 1 under the distribution $\mathbf{P}$. We will see later on how to weaken this type of assumption even further by requiring a general (non-trivial) control on the probability of the event $\{|\eta(X) - \frac{1}{2}| \geq \gamma\}$. This will lead to Tsybakov's noise condition.*

Remark 7.9 (Noiseless case) *The noiseless case in binary classification is represented by the case $\eta \in \{0,1\}$, which corresponds to the choice $\gamma = 1/2$ in Massart's noise condition.*

Massart's noise condition yields the fast rate $1/n$ when we assume that $a^{\star\star} \in \mathcal{A}$ and $|\mathcal{A}| < \infty$.

Theorem 7.10 (Fast rate in binary classification) *Let $a^{\star\star} \in \mathcal{A}$ so that $a^\star = a^{\star\star}$. Let Massart's noise condition holds for a given $\gamma \in (0,1/2]$. Then,*

$$\boxed{\mathbf{P}\left(r(A^\star) - r(a^\star) \le \frac{\log(|\mathcal{A}|/\delta)}{\gamma n}\right) \ge 1-\delta}$$

Proof: Note that as $a^{\star\star} \in \mathcal{A}$, then $a^\star = a^{\star\star}$. Let us rearrange the decomposition of the error for the empirical risk minimizer $A^\star$ given in Section 1.2.1 as follows:

$$r(A^\star) - r(a^\star) = r(A^\star) - R(A^\star) + \underbrace{R(A^\star) - R(a^\star)}_{\le 0} + R(a^\star) - r(a^\star) \le R(a^\star) - R(A^\star) - (r(a^\star) - r(A^\star)).$$

For any $a \in \mathcal{A}$, define

$$G(a) := R(a^\star) - R(a) - (r(a^\star) - r(a)) = R(a^\star) - R(a) - \mathbf{E}[R(a^\star) - R(a)] = \frac{1}{n}\sum_{i=1}^n (g(a,Z_i) - \mathbf{E}g(a,Z_i)),$$

with

$$g(a,z) := \mathbf{1}_{a^\star(x)\neq y} - \mathbf{1}_{a(x)\neq y}$$

for $a \in \mathcal{A}$ and $z = (x,y) \in \mathbb{R}^d \times \{-1,1\}$. Then the above yields

$$r(A^\star) - r(a^\star) \le G(A^\star).$$

Bernstein's inequality for bounded random variables (note that $g(a,z) \in \{-1,0,1\}$ so $g(a,Z) - \mathbf{E}g(a,Z) \le 2$ and $g(a,Z)$ satisfy the one-sided Bernstein's condition with parameter $b = 2/3$) yields, for any $a \in \mathcal{A}$,

$$\mathbf{P}(G(a) \ge \varepsilon) \le \exp\left(-\frac{n\mathbf{Var}\, g(a,Z)}{b^2} h\left(\frac{b\varepsilon}{\mathbf{Var}\, g(a,Z)}\right)\right).$$

Setting the right-hand side of this bound equal to $\delta/|\mathcal{A}|$, and using that $h^{-1}(u) = u + \sqrt{2u}$ for $u > 0$, we obtain

$$\begin{aligned}
&\mathbf{P}\left(G(A^\star) < \frac{b}{n}\log(|\mathcal{A}|/\delta) + \sqrt{\frac{2(\mathbf{Var}\, g(A^\star,Z))\log(|\mathcal{A}|/\delta)}{n}}\right) \\
&\ge \mathbf{P}\left(\bigcap_{a\in\mathcal{A}} \left\{G(a) < \frac{b}{n}\log(|\mathcal{A}|/\delta) + \sqrt{\frac{2(\mathbf{Var}\, g(a,Z))\log(|\mathcal{A}|/\delta)}{n}}\right\}\right) \\
&\ge 1-\delta,
\end{aligned}$$

where the first inequality comes as $A^\star \in \mathcal{A}$ and the second inequality comes by the union bound. As for any $a \in \mathcal{A}$ we have $|g(a,Z)| = \mathbf{1}_{a(X)\neq a^\star(X)}$, then

$$\mathbf{Var}\, g(a,Z) \le \mathbf{E}[g(a,Z)^2] = \mathbf{P}(a(X) \neq a^\star(X))$$

and from Theorem 7.6 and Massart's noise condition we have

$$r(a) - r(a^\star) = \mathbf{E}[|2\eta(X) - 1|\mathbf{1}_{a(X)\neq a^\star(X)}] \ge 2\gamma\mathbf{P}[a(X) \neq a^\star(X)], \tag{7.3}$$

which yields $\mathbf{Var}\, g(a,Z) \le \frac{1}{2\gamma}(r(a) - r(a^\star))$. Using that $r(A^\star) - r(a^\star) \le G(A^\star)$, we can conclude

$$\mathbf{P}\biggl(r(A^\star) - r(a^\star) < \frac{2}{3n}\log(|\mathcal{A}|/\delta) + \sqrt{\frac{(r(A^\star) - r(a^\star))\log(|\mathcal{A}|/\delta)}{\gamma n}} \biggr) \ge 1 - \delta.$$

The proof follows by solving the expression in the event with respect to the excess risk $r(A^\star) - r(a^\star)$, using that $x < 2\alpha/3 + \sqrt{x\alpha/\gamma}$ for $x \in [0,1]$, with $\alpha > 0$ and $\gamma \in (0, 1/2]$, implies $x < \alpha/\gamma$. ■

7.4.2 Tsybakov's noise condition

Massart's noise condition is strong (indeed, it yields the fast rate!), as it requires the regression function η be uniformly bounded away from 1/2. We will now explore a weaker condition that allows η to be arbitrarily close to 1/2, but with small probability.

Definition 7.11 (Tsybakov's noise condition) *There exist $\alpha \in (0,1)$, $\beta > 0$, and $\gamma \in (0,1/2]$ such that, for all $t \in [0,\gamma]$, we have*

$$\boxed{\mathbf{P}\biggl(\Bigl| \eta(X) - \frac{1}{2} \Bigr| \le t \biggr) \le \beta t^{\alpha/(1-\alpha)}}$$

As $\alpha \to 0$, we have that $t^{\alpha/(1-\alpha)} \to 1$ for all $t \in [0,\gamma]$ and so Tsybakov's condition is void and we expect the slow rate $1/\sqrt{n}$. As $\alpha \to 1$, we have that $t^{\alpha/(1-\alpha)} \to 0$ for all $t \in [0,\gamma]$ and so Tsybakov's condition recovers Massart's condition and we expect the fast rate $1/n$. For values of α between 0 and 1, we expect Tsybakov's condition to yield a rate that interpolates between $1/\sqrt{n}$ and $1/n$. To prove this, we first need the following proposition.

Proposition 7.12 *Let Tsybakov's noise condition holds for given $\alpha \in (0,1)$, $\beta > 0$, and $\gamma \in (0,1/2]$. Then, for any $a \in \mathcal{A}$ we have*

$$\boxed{\mathbf{P}[a(X) \ne a^{\star\star}(X)] \le c(r(a) - r(a^{\star\star}))^\alpha}$$

for a given constant c that depends on α, β, γ.

Proof: By Theorem 7.6 and Tsybakov's noise condition we obtain

$$\begin{aligned} r(a) - r(a^{\star\star}) &= \mathbf{E}[|2\eta(X) - 1|\mathbf{1}_{a(X)\ne a^{\star\star}(X)}] \\ &\ge \mathbf{E}[|2\eta(X) - 1|\mathbf{1}_{|\eta(X)-1/2|>t}\mathbf{1}_{a(X)\ne a^{\star\star}(X)}] \\ &> 2t\mathbf{P}(|\eta(X) - 1/2| > t, a(X) \ne a^{\star\star}(X)) \\ &\ge 2t\mathbf{P}(a(X) \ne a^{\star\star}(X)) - 2t\mathbf{P}(|\eta(X) - 1/2| \le t) \\ &\ge 2t\mathbf{P}(a(X) \ne a^{\star\star}(X)) - 2\beta t^{1/(1-\alpha)}. \end{aligned}$$

Take $t = \tilde{c}\mathbf{P}[a(X) \ne a^{\star\star}(X)]^{(1-\alpha)/\alpha}$ for some positive $\tilde{c} \le \gamma$ to be chosen later. We get

$$\begin{aligned} r(a) - r(a^{\star\star}) &\ge 2\tilde{c}\mathbf{P}(a(X) \ne a^{\star\star}(X))^{1/\alpha} - 2\beta\tilde{c}^{1/(1-\alpha)}\mathbf{P}(a(X) \ne a^{\star\star}(X))^{1/\alpha} \\ &= 2(\tilde{c} - \beta\tilde{c}^{1/(1-\alpha)})\mathbf{P}(a(X) \ne a^{\star\star}(X))^{1/\alpha} \\ &\ge \tilde{c}\mathbf{P}(a(X) \ne a^{\star\star}(X))^{1/\alpha}, \end{aligned}$$

where the last inequality follows by choosing $\tilde{c}$ sufficiently small. Hence,

$$\mathbf{P}(a(X) \ne a^{\star\star}(X)) \le \frac{(r(a) - r(a^{\star\star}))^\alpha}{\tilde{c}^\alpha},$$

and the proof follows by choosing $c = \tilde{c}^{-\alpha}$. ■

Theorem 7.13 (Interpolation between slow and fast rate in binary classification) *Let $a^{\star\star} \in \mathcal{A}$ so that $a^\star = a^{\star\star}$. Let Tsybakov's noise condition holds for given $\alpha \in (0,1)$, $\beta > 0$, and $\gamma \in (0, 1/2]$. Then,*

$$\boxed{\mathbf{P}\left(r(A^\star) - r(a^\star) \le c\left(\frac{\log(|\mathcal{A}|/\delta)}{n}\right)^{\frac{1}{2-\alpha}}\right) \ge 1 - \delta}$$

for a given constant c that depends on α, β, γ.

Proof: The proof follows the same steps as in the proof of Theorem 7.10, up to equation (7.3) as in this case we can use Proposition 7.12 instead. This yields

$$\mathbf{Var}\, g(a, Z) \le \mathbf{E}[g(a, Z)^2] = \mathbf{P}(a(X) \ne a^\star(X)) \le c(r(a) - r(a^{\star\star}))^\alpha$$

and

$$\mathbf{P}\left(r(A^\star) - r(a^\star) < \frac{2}{3n}\log(|\mathcal{A}|/\delta) + \sqrt{\frac{2c(r(A^\star) - r(a^\star))^\alpha \log(|\mathcal{A}|/\delta)}{n}}\right) \ge 1 - \delta,$$

which yields the result solving for $r(A^\star) - r(a^\star)$. ■

We note that when $\alpha \to 0$ we approach the slow rate, while when $\alpha \to 1$ we approach the fast rate. We also highlight that the empirical risk minimizer $A^\star$ does *not* depend on α: it automatically adjusts to the noise level!