Note: These lecture notes are still rough, and have only have been mildly proofread.

Outline

- Consistency in classification
- Bounded difference inequality

Recap

Last time we proved that

$$
\Pr[|R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f)| > \epsilon] \leq 8s(\mathcal{F}, n) \exp\left(-n\epsilon^2 \frac{1}{128}\right),
$$

where $R(\hat{f}_n)$ is the population risk of empirically optimal $\hat{f}_n \in \mathcal{F}$, and $s(\mathcal{F}, n)$ is the $n$th shatter coefficient of the class $\mathcal{F}$.

We now state a consequence for the expected value:

**Corollary 15.1.** Under the same conditions, we have

$$
0 \leq \mathbb{E}[R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f)] \leq 16\sqrt{\frac{\log(8s(\mathcal{F}, n))}{2n}} = \mathcal{O}\left(\sqrt{\frac{V_{\mathcal{F}} \log(n)}{n}}\right),
$$

where $V_{\mathcal{F}}$ is the VC dimension of the class.

**Proof:** This corollary is a consequence of the bound (15.1) and the following lemma.

**Lemma 15.2.** If a non-negative random variable satisfies $\Pr[Z > t] \leq ce^{-2nt^2}$ for some constant $c < \infty$, then

$$
\mathbb{E}[Z] \leq \sqrt{\frac{\log(c\epsilon)}{2n}}.
$$

(15.2)
Proof: We begin by noting that since the variance of a random variable is non-negative, we have $\mathbb{E}[Z]^2 \leq \mathbb{E}[Z^2]$. Therefore, we have

$$
\mathbb{E}[Z^2] = \int_0^\infty \mathbb{P}(Z^2 > t)dt
= \int_0^u \mathbb{P}(Z^2 > s)ds + \int_u^\infty \mathbb{P}(Z^2 > s)ds
\leq u + \int_u^\infty c \exp(-2ns)ds
= u + \frac{c}{2n} e^{-2nu}.
$$

Optimizing over $u > 0$ concludes the proof. \(\square\)

15.1 Bounded difference inequality (McDiarmid, 1989)

It is often the case that we are given an i.i.d. sequence $X_1, X_2, \ldots$ of random variables, and some function $f : X^n \to \mathbb{R}$, and we would like to control the difference

$$
V := f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]
$$

of $f(X_1, \ldots, X_n)$ from its expectation. Let us say that a function $f$ satisfies the bounded difference condition with parameters $(c_1, \ldots, c_n)$ if for each $j = 1, \ldots, n$, we have

$$
\sup_{x_j, y_j \in X} |f(x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_n) - f(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n)| \leq c_j. \quad (15.3)
$$

This condition means that by changing the $j$th co-ordinate of the function, we can change the function value by at most $c_j$.

15.1.1 Statement of result and some examples

We now state a useful concentration result for i.i.d. RVs and functions that satisfy a bounded difference property.

Theorem 15.3. Suppose that a function $f$ satisfies the bounded difference property with parameters $(c_1, \ldots, c_n)$. Then for an i.i.d. sequence $X_1, X_2, \ldots, X_n$ and for all $\epsilon > 0$, we have:

$$
\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| > \epsilon) \leq 2 \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right). \quad (15.4)
$$

Let us consider some examples to illustrate:

Examples:

(a) Suppose that

$$
f(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i,
$$

corresponding to the set-up for the usual law of large numbers. If the \( X_i \) are bounded RVs (i.e., \( |X_i| \leq M/2 \) for all \( i \)), then the bounded difference property holds with \( c_i = \frac{M}{n} \) for all \( i \) and hence

\[
\mathbb{P}[\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X] \right| > \epsilon] \leq 2 \exp\left(\frac{-n\epsilon^2}{M^2}\right).
\]

Thus, we recover the Hoeffding bound for i.i.d. variables as a special case.

(b) Now suppose that \( f \) is differentiable with \( \sup_{x \in \mathbb{R}^n} |\frac{\partial f(x)}{\partial x}| = L_j < \infty \), corresponding to imposing a Lipschitz condition on each partial derivative of \( f \). This condition means that \( f \) does not change very fast in any particular direction. If \( |X_j| \leq \frac{M}{2} \) then the bounded difference condition holds with \( c_j = L_j M \). Example (a) is a special case of this more general condition.

15.1.2 Proof of Theorem 15.3

We begin by noting that \( V = f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \) can be decomposed as the sum \( V = \sum_{i=1}^{n} Z_i \), where \( Z_1 = \mathbb{E}[V | X_1] - \mathbb{E}[V] \), and

\[
Z_{i+1} = \mathbb{E}[V | X_1, \ldots, X_{i+1}] - \mathbb{E}[V | X_1, \ldots, X_i] \quad \text{for} \quad i = 1, \ldots, n - 1. \tag{15.5}
\]

(Note that \( \mathbb{E}[V | X_1, \ldots, X_n] = V \) since \( V \) is a function of the \( \{X_i\} \).) Moreover, note that \( \mathbb{E}[Z_1] = 0 \) and \( \mathbb{E}[Z_i | X_1, \ldots, X_{i-1}] = 0 \), which implies that \( \{Z_i\} \) is a martingale difference sequence with respect to \( \{X_i\} \). With this decomposition, our strategy is to show that \( Z_i \in [A_i, B_i] \) for suitable random variables that are functions of \( (X_1, \ldots, X_{i-1}) \). We will then show that \( B_i - A_i \) is bounded almost surely by some constant \( c \). Under this condition, the conditioned random variable \( (Z_i | X_1, \ldots, X_{i-1}) \) is a bounded random variable, so that we upper bound the moment-generating function \( \mathbb{E}[\exp(tZ_i) | X_1, \ldots, X_{i-1}] \) by \( \exp(t^2c^2/2) \), and then proceed as in the proof of the Azuma-Hoeffding inequality.

It remains to show that \( Z_i \in [A_i, B_i] \) for suitable random variables that are functions of \( (X_1, \ldots, X_{i-1}) \). Define the function \( H_i(x_1, \ldots, x_i) = \mathbb{E}[f(X_1, \ldots, X_n) | X_1 = x_1, \ldots, X_i = x_i] \), and let us write

\[
Z_i = H_i(X_1, \ldots, X_i) - \int H_i(X_1, \ldots, X_{i-1}, t) dF_i(t)
\]

where \( F_i \) is the distribution function of \( X_i \). Next define the random variables

\[
B_i := \sup_{u \leq X_i} [H_i(X_1, \ldots, X_{i-1}, u) - \int H_i(X_1, \ldots, X_{i-1}, t) dF_i(t)], \quad \text{and} \quad \tag{15.6a}
\]

\[
A_i := \inf_{v \leq X_i} [H_i(X_1, \ldots, X_{i-1}, v) - \int H_i(X_1, \ldots, X_{i-1}, t) dF_i(t)]. \tag{15.6b}
\]

By construction, we have the inclusion \( Z_i \in [A_i, B_i] \).

Finally, let us show that \( B_i - A_i \) is bounded. We have

\[
B_i - A_i = \sup_{u \leq X_i} H_i(X_1, \ldots, X_{i-1}, u) - \inf_{v \leq X_i} H_i(X_1, \ldots, X_{i-1}, v)
\]

\[
\leq \sup_{u, v} [H_i(X_1, \ldots, X_{i-1}, u) - H_i(X_1, \ldots, X_{i-1}, v)]
\]

15-3
But since $f$ satisfies the bounded difference property with coefficients $(c_1, \ldots, c_n)$, and using the definition of $H_i$, we have

$$\sup_{u,v}[H_i(X_1, \cdots, X_{i-1}, u) - H_i(X_1, \cdots, X_{i-1}, v)] \leq c_i,$$

so that we have shown that $0 \leq B_i - A_i \leq c_i$, which completes the proof.

### 15.2 Covering and Packing (Metric Entropy)

Metric entropy is another way to measure the richness of classes of functions/sets; it is a concept that dates back to the seminal paper of Kolmogorov & Tikhomirov (1961). Intuitively, the covering number measures the volume of a class in terms of the number of balls of radius $\varepsilon$ required to cover it. In particular, let us define

$$B_{\varepsilon}(y) := \{x \mid \rho(x, y) \leq \varepsilon\}$$

to be the ball of radius $\varepsilon$ centered at $y$ measured in some metric $\rho$. With this notation, we have:

**Definition:** Given a set $S$ and a metric $\rho : S \times S \to \mathbb{R}^+$, the $\varepsilon$-covering number of $S$ with respect to $\rho$ is the smallest integer $N = N(\varepsilon; S, \rho)$ such that

$$S \subseteq \bigcup_{i=1}^{N} B_{\varepsilon}(x^i). \quad (15.7)$$

The collection of points $\{x^1, \ldots, x^N\}$ is called an $\varepsilon$-cover of $S$. Finally, the quantity $H(\varepsilon; S, \rho) = \log N(\varepsilon; S, \rho)$ is called the metric entropy of the set $S$ in the $\rho$-metric.

Next time we consider some examples of metric entropy for different classes of functions. Clearly, the metric entropy grows as $\varepsilon \to 0$, and we will see that for statistical problems, the growth rate of this entropy affects the rates (in terms of sample size $n$) at which we can perform classification or regression over that class of sets/functions.