Gliveako-Cantelli theorem and Rademacher averages

Lecturer: Peter L. Bartlett *Scribe: Tao Shi*

Recall that Hoeffding's inequality shows that a sum of bounded independent random variables is concentrated about its expectation:
Hoeffding's Inequality: For $f : \mathcal{X} \mapsto [a, b]$, $x_1, \cdots, x_n \in \mathcal{X}$, *i.i.d.*, define $Z = \sum_{i=1}^n f(x_i)$

$$Pr\{Z - \mathbf{E}Z \geq t\} \leq \exp \frac{-2t^2}{n(b-a)^2}$$

We have also seen that, as a consequence of McDiarmid's inequality, the same is true for the supremum of a sum of bounded independent random variables:
Theorem 1: For $f \in F$, $f : \mathcal{X} \mapsto [a, b]$; $x_1, \cdots, x_n$ *i.i.d.*, define

$$Z = \sup_{f \in F} \sum_{i=1}^n (Ef - f(x_i))$$

then

$$Pr\{Z - \mathbf{E}Z \geq t\} \leq \exp \frac{-2t^2}{n(b-a)^2}.$$

We are interested in learning algorithms that choose $\hat{f} \in F$ to minimize the empirical risk $\hat{R}$, defined by

$$\hat{\mathbf{R}}(f) = \hat{\mathbf{E}}(l_f) = \frac{1}{n} \sum_{i=1}^n l(Y_i f(x_i))$$

so that $\hat{f}$ is defined by

$$\hat{R}(\hat{f}) = \min_{f \in F} \hat{R}(f) \;\; (\text{or } \; \leq \inf_{f \in F} \hat{R}(f) + \epsilon)$$

We want to know what we can say about $R(\hat{f})$. For example, is it true that

$$\hat{R}(f) \leq \epsilon \Rightarrow R(f) \leq 2\epsilon?$$

(perhaps for most samples $\{X_1, \ldots, X_n\}$) Or if $\hat{f}$ minimizes $\hat{R}$, do we have

$$R(\hat{f}) \leq \inf_{f \in F} R(f) + \epsilon?$$

For a class F and loss function l, define the loss class

$$l_F = \{(x, y) \mapsto l(y, f(x)) : f \in F\} = \{l_F : f \in F\}$$

and the excess loss class

$$g_F = \{(x, y) \mapsto l(y, f(x)) - l(y, f^*(x) : f \in F\} = \{g_f : f \in F\}$$

with $f^*(x) = argmin_{f \in F} \mathbf{E}(l_f)$.

Now if the class l_F is such that

$$\sup_{g \in l_F} |\mathbf{E}g - \hat{\mathbf{E}}g| \leq \epsilon,$$

then this implies that, for any f, $\mathbf{E}l_{\hat{f}} \leq \hat{\mathbf{E}}l_{\hat{f}} + \epsilon$, and so an f with small empirical risk will have small risk. Also, if $\hat{f} \in F$ minimizes empirical risk, $\hat{\mathbf{E}}\ l_{\hat{f}}$, and $f^* \in F$ minimizes risk, $\mathbf{E}l_f$, then

$$\begin{aligned}
\mathbf{E}\ l_{\hat{f}} - \mathbf{E}\ l_{f^*} &= (\mathbf{E}\ l_{\hat{f}} - \hat{\mathbf{E}}\ l_{\hat{f}}) + (\hat{\mathbf{E}}\ l_{\hat{f}} - \hat{\mathbf{E}}\ l_{f^*}) + (\hat{\mathbf{E}}\ l_{f^*} - \mathbf{E}\ l_{f^*}) \\
&\leq 2\epsilon.
\end{aligned}$$

(The first and third difference are no more than ϵ because the supremum of these differences over any f is no more than ϵ, and the second difference is no more than 0 by definition of $\hat{f}$.)

Thus, if $\mathbb{E}g$ and $\hat{\mathbb{E}}g$ are close, uniformly over l_F, then the minimizer of the empirical risk has near minimal risk. We'll see that this is a rather loose sufficient condition, because it's really only the behavior of $\mathbb{E}g - \hat{\mathbb{E}}g$ for the empirical risk minimizer that matters. But for now, we'll study the uniform convergence of the empirical risk to the risk.

Definition 1: We say that F is a Glivenko-Cantelli class (G-C class), if for every $\epsilon > 0$,

$$\lim_{n \to \infty} \sup_p P^n \{\sup_{f \in F} |\mathbf{E}f - \hat{\mathbf{E}}_n f| > \epsilon\} = 0$$

For a G-C class F, define $\epsilon(n, \delta)$ as the smallest ϵ for which

$$\sup_p P^n \{\sup_{f \in F} |\mathbf{E}f - \hat{\mathbf{E}}_n f| > \epsilon\} < \delta$$

The name Glivenko-Cantelli arises from the following classical result, which can be viewed as a demonstration that the class of semi-infinite intervals is a G-C class.

Theorem 2 (Glivenko-Cantelli):
For any P on $\Re$, with distribution function $F(t) = P\{(-\infty, t]\}$, almost surely, $F_n \to F$ uniformly on $\Re$ as $n \to \infty$, where F_n is the empirical distribution function, $F_n(t) = P_n\{(-\infty, t]\}$.

To see the relationship between this theorem and the definition of a GC class, notice that

$$\begin{aligned}
F(t) &= \mathbf{E}1_{[X \leq t]} \\
F_n(t) &= \hat{\mathbf{E}}1_{[X \leq t]}.
\end{aligned}$$

Define $G = \{x \longmapsto 1_{[x \leq \theta]} :\ \theta \in \Re\}$. Then the theorem is equivalent to:

For all probability distribution P, $\sup_{g \in G} |\mathbf{E}g - \hat{\mathbf{E}}_n g| \to 0\ a.s.$

Proof of G-C Theorem:

STEP 1: concentration.
McDiarmid's inequality implies that, with probability $\geq 1 - \exp(-2\epsilon^2 n)$,

$$\sup_{g \in G} |\hat{\mathbf{E}}_n g - \mathbf{E}g| \leq \mathbf{E} \sup_{g \in G} |\hat{\mathbf{E}}_n g - \mathbf{E}g| + \epsilon$$

STEP 2: symmetrization.

$$\begin{aligned}
\mathbf{E}\sup_{g\in G}|\hat{\mathbf{E}}_n g - \mathbf{E}g| &= \mathbf{E}\sup_{g\in G}\frac{1}{n}\sum_{i=1}^{n}(g(x_i) - \mathbf{E}'g(x_i')) \\
&\leq \mathbf{E}\sup_{g\in G}\frac{1}{n}\sum_{i=1}^{n}(g(x_i) - g(x_i')) \\
&= \mathbf{E}\sup_{g\in G}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(g(x_i) - g(x_i')) \\
&\quad \text{(choosing } \epsilon_i \in \{\pm 1\} \text{ uniformly deos not affect expectation)} \\
&\leq \mathbf{E}\sup_{g\in G}\frac{1}{n}(|\sum_{i=1}^{n}\epsilon_i(g(x_i)| + |\sum_{i=1}^{n}\epsilon_i(g(x_i)|) \\
&\leq 2\mathbf{E}\sup_{g\in G}\frac{1}{n}|\sum_{i=1}^{n}\epsilon_i(g(x_i)| \\
&= 2\mathbf{E}\{\mathbf{E}[\sup_{g\in G}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(x_i)|x_1,\cdots,x_n]\}
\end{aligned}$$

Notice the function class G is not rich. Consider any $x_1, \ldots, x_n$. Without loss of generality, assume that $x_1 \leq x_2 \leq \cdots \leq x_n$. Then it is clear that

$$|\{g(x_1), g(x_2), \cdots, g \in G\}| \leq |\{(0,0,\cdots,0), (1,0,\cdots,0), \cdots, (1,1,\cdots,1)\}| = n+1.$$

Here we need a Lemma to help the proof,
Lemma 1 [Massart's finite class lemma]: For a finite set $A \subseteq \Re^n$ and $R > 0$ satisfying $R^2 = \max_{a\in A}\|a\|_2^2 = \max_{a\in A}\sum_i a_i^2$,

$$\mathbf{E}\sup_{a\in A}\sum_{i=1}^{n}\epsilon_i a_i \leq R\sqrt{2\log|A|}$$

Proof of Lemma: For $s > 0$, $Z_a = \sum_{i=1}^{n}\epsilon_i a_i$,

$$\begin{aligned}
\exp(s\ \mathbf{E}\sup_{a\in A}Z_a) &\leq \mathbf{E}\exp(s\ \sup_{a\in A}Z_a) \\
&= \mathbf{E}\sup_{a\in A}\exp(sZ_a) \\
&\leq \sum_{a\in A}\mathbf{E}\exp(sZ_a) \\
&\leq \sum_{a\in A}\exp(\frac{s^2}{2}\sum_{i=1}^{n}a_i^2) \\
&\leq |A|\exp(\frac{s^2R^2}{2})
\end{aligned}$$

where the third inequality follows from Hoeffding's lemma (see Lemma 2.3 in Lecture 13). Thus, we have

$$\mathbf{E}\sup_{a\in A}Z_a \leq \inf_s(\frac{\log(|A|)}{s} + s^2R^2/2) = R\sqrt{2\log|A|}.$$

♠

In our setting, $R \leq 1/\sqrt{n}$ and $|A| \leq n+1$, so

$$\Pr\left(\sup_{g\in G}\left|\hat{\mathbf{E}}g - \mathbf{E}g\right| > \epsilon + 2\sqrt{\frac{2\log(n+1)}{n}}\right) \leq e^{-2\epsilon^2 n},$$

and hence averages converge in probability to expectations uniformly. Also, we can apply the Borel-Cantelli theorem, since

$$\sum_{n=1}^{\infty} \Pr(\sup \cdots > \epsilon)$$

is finite. Thus, the supremum converges to zero almost surely.

♠

G-C class and Rademacher averages

Definition 2: For a class F of real-valued functions on $\mathcal{X}$, $x_1, \cdots, x_n \in \mathcal{X}$ *i.i.d.*, $\epsilon_1, \cdots, \epsilon_n$ are rademacher random variables, define Rademacher average of F as:

$$R_n(F) = \mathbf{E}\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)$$

$$\hat{R}_n(F) = \mathbf{E}[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)|x_1, \cdots, x_n]$$

We can think of the Rademacher averages as capturing how well functions in F can fit the noise sequence $\epsilon_1, \ldots, \epsilon_n$. Indeed, for $f : \mathcal{X} \longrightarrow \{\pm 1\}$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i) &= \frac{1}{n}\sum_{i=1}^{n}(2l(\epsilon_i, f(x_i)) - 1) \\
&= \frac{1}{n}\sum_{i=1}^{n}(1 - 2l(-\epsilon_i, f(x_i))) \\
&= 1 - \frac{2}{n}\sum_{i=1}^{n}l(-\epsilon_i, f(x_i))
\end{aligned}$$

Theorem 3 (Symmetrization) :: For $F \subseteq \Re^{\mathcal{X}}$,

$$\begin{aligned}
\mathbf{E}\sup_{f\in F}(\mathbf{E}f - \frac{1}{n}\sum_{i=1}^{n}f(x_i)) &\leq 2R_n(F) \\
&= \mathbf{E}\sup_{f\in F}\frac{2}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)
\end{aligned}$$

Proof:

$$
\begin{aligned}
LHS &= \mathbf{E} \sup_{f\in F} \mathbf{E}' \frac{1}{n}\sum (f(x_i') - f(x_i)) \\
&\leq \mathbf{E} \sup_{f\in F} \frac{1}{n}\sum (f(x_i') - f(x_i)) \\
&= \mathbf{E} \sup_{f\in F} \frac{1}{n}\sum \epsilon_i (f(x_i') - f(x_i)) \\
&\leq \mathbf{E} \sup_{f\in F} \frac{2}{n}\sum \epsilon_i f(x_i) \\
&= 2R_n(F)
\end{aligned}
$$

♠

By symmetry,

$$\mathbf{E} \sup_{f\in F}(\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbf{E}f) \leq 2R_n(F)$$

Also converse:

$$R_n(F) \leq \mathbf{E} \sup_{f\in F}(\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbf{E}f) + \mathbf{E} \sup_{f\in F}(\mathbf{E}f - \frac{1}{n}\sum_{i=1}^{n} f(x_i)) + \mathbf{E} \sup_{f\in F}(\mathbf{E}f + \frac{1}{n}\sum \epsilon_i)$$

Combining concentration (McDiarmid) and symmetrization, we get

Theorem 4: For $F \subseteq [0,1]^{\mathcal{X}}$, with probability at least $1-\delta$,

$$\sup_{f\in F}(\mathbf{E}f - \frac{1}{n}\sum_{i=1}^{n} f(x_i)) \leq 2R_n(F) + \sqrt{\frac{\log(1/\delta))}{2n}}$$

Note: we also have,

$$R_n(F) = \mathbf{E} \sup_{f\in F} \frac{1}{n}\sum \epsilon_i f(x_i),$$

and changing any ϵ_i or x_i changes this random variable by no more than $2\sup_{f\in F} ||f||_\infty / n$, so McDiarmid implies this quantity is also concentrated. This gives the following result.

Theorem 5: With probability at least $1-\delta$, over $x_1, \cdots, x_n, \epsilon_1, \cdots, \epsilon_n$,

$$\sup_{f\in F}(\mathbf{E}f - \frac{1}{n}\sum_{i=1}^{n} f(x_i)) \leq \frac{2}{n}\sum_{i=1}^{n} \epsilon_i f(x_i) + C\sqrt{\frac{\log(1/\delta)}{n}}.$$