CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 17**

Glivenko-Cantelli Theorem

Lecturer: Peter Bartlett *Scribe: Percy Liang*

In the previous lecture, we introduced the notion of a Glivenko-Cantelli class of functions, for which we could get uniform convergence of empirical averages to their true expectations. We are interested in such a class because if a loss class is Glivenko-Cantelli, then the excess risk of the empirical risk minimizer converges to 0 a.s.

In this lecture, we take our first step towards proving uniform convergence. We will prove the Glivenko-Cantelli theorem, which gives us uniform convergence for the class of intervals functions on the real line. Although this theorem is a very classical result which can be proved by more bare-hands techniques, we will use a more sophisticated proof in this lecture. The main point of the proof will be to introduce new ideas that we can later apply to richer classes of functions.

First, the setup:

- Let $X_1, X_2, \ldots$ be i.i.d. data points drawn from some distribution F.
- Let $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1[X_i \le x]$ be the empirical distribution function with respect to the first n data points.

Now, the main theorem:

Theorem 1.1. For any probability measure P on $\mathbb{R}$, F_n converges uniformly on $\mathbb{R}$ to F a.s. In symbols,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

First, note that we easily have pointwise convergence of distribution functions just by the law of large numbers. That is, for each fixed x, $F_n(x) \xrightarrow{a.s.} F(x)$. However, we are after uniform convergence, which is much stronger.

Let us first rewrite the theorem to take the supremum over a class of functions rather than points in $\mathbb{R}$. This new notation will readily extend to richer function classes. Observe the following equivalences:

- $F(t) = \mathbb{E}1[X \le t]$
- $F_n(t) = \mathbb{E}_n 1[X \le t]$

Define

$$G = \{x \mapsto 1[x \le \theta] : \theta \in \mathbb{R}\}$$

There is a one-to-one mapping between G and $\mathbb{R}$. Theorem 1.1 is equivalent to the following: For any probability measure P,

$$\sup_{g\in G} |\mathbb{E}_n g - \mathbb{E} g| \xrightarrow{a.s.} 0.$$

Now we are ready to tackle the proof. The proof of the theorem is divided into the following three steps:

1. We will show that the quantity that we want to converge, $\sup_{g\in G} |\mathbb{E}_n g - \mathbb{E}g|$, is asymptotically concentrated around its mean.

2. Using a symmetrization trick, we turn the mean $\mathbb{E}\sup_{g\in G} |\mathbb{E}_n g - \mathbb{E}g|$ into an upper bound

$$\mathbb{E}\sup_{g\in G} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i g(X_i)|$$

   that is easier to work with (σ_i's are certain independent random variables defined in the proof below).

3. Finally, we show that this new quantity $\mathbb{E}\sup_{g\in G} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i g(X_i)|$ converges to 0 because G is "simple" in some sense.

Step 1: Concentration

Recall that McDiarmid's inequality states that if each argument of a function f has bounded influence on the value of the function, then we have concentration. In symbols, if

$$|f(x_1,\ldots,x_i,\ldots,x_n) - f(x_1,\ldots,x_i',\ldots,x_n)| \le c_i,$$

then

$$P(f(X_1,\ldots,X_n) > \mathbb{E}f(X_1,\ldots,X_n) + \epsilon) \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

We will apply McDiarmid's inequality with the following function f:

$$f(X_1,\ldots,X_n) = \sup_{g\in G} |\mathbb{E}_n g - \mathbb{E}g|.$$

This f satisfies the bounded differences property: if we change one of the data points X_i arbitrarily, then the empirical distribution function can change by at most $c_i = 1/n$ at any point on $\mathbb{R}$. Simple calculation shows that

$$\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) = \exp(-2\epsilon^2 n).$$

Therefore, with probability at least $1 - e^{-2\epsilon^2 n}$,

$$\sup_{g\in G} |\mathbb{E}_n g - \mathbb{E}g| \le \mathbb{E}\sup_{g\in G} |\mathbb{E}_n g - \mathbb{E}g| + \epsilon.$$

Step 2: Symmetrization

Now we would like

$$\mathbb{E}\sup_{g\in G} |\mathbb{E}_n g - \mathbb{E}g| \tag{1}$$

to converge to 0. However, this quantity is hard to get a handle on because $\mathbb{E}g$ involves the true distribution, which we do not know. Using symmetrization, we will upper bound (1) with something that does not depend on $\mathbb{E}g$.

Let $X_1',\ldots,X_n'$ be independent copies of $X_1,\ldots,X_n$.

$$
\begin{aligned}
\mathbb{E}\sup_{g\in G}|\mathbb{E}_n g-\mathbb{E}g| &= \mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i g(X_i)-\mathbb{E}g\right| \quad \text{[expand definition of } \mathbb{E}_n\text{]}\\
&= \mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i g(X_i)-\mathbb{E}\frac{1}{n}\sum_i g(X_i')\right| \quad \text{[write } \mathbb{E}g \text{ verbosely]}\\
&= \mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i \left(g(X_i)-\mathbb{E}g(X_i')\right)\right| \quad \text{[linearity of expectation]}\\
&= \mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i \left(g(X_i)-\mathbb{E}(g(X_i')\mid \mathbf{X})\right)\right| \quad [\mathbf{X}=X_1,\dots,X_n \text{ independent of } \mathbf{X}']\\
&= \mathbb{E}\sup_{g\in G}\left|\mathbb{E}\left(\frac{1}{n}\sum_i \left(g(X_i)-g(X_i')\right)\mid \mathbf{X}\right)\right| \quad \text{[properties of conditional expectation]}\\
&\le \mathbb{E}\,\mathbb{E}\left(\sup_{g\in G}\left|\frac{1}{n}\sum_i \left(g(X_i)-g(X_i')\right)\right|\mid \mathbf{X}\right) \quad \text{[convexity of sup and } |\cdot|\text{]}\\
&= \mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i \left(g(X_i)-g(X_i')\right)\right| \quad \text{[iterated conditional expectation]}
\end{aligned}
$$

We are half way there. We have gotten rid of $\mathbb{E}g$ like we wanted, but now we have these extra data points $X_1',\dots,X_n'$. Let's get rid of them too. To that end, introduce *Rademacher variables* $\sigma_1,\dots,\sigma_n$ which are independent of each other and everything else, and are each drawn uniformly from $\{-1,+1\}$.

First, note that the distribution of $g(X_i)-g(X_i')$ is symmetric around 0, so it has the same distribution as $\sigma_i(g(X_i)-g(X_i'))$ for any fixed value of $\sigma_i\in\{-1,+1\}$. So, the above quantity does not change if multiply any term in the summation by σ_i. Since this is true for all fixed values of σ_i, we can take an expectation over the σ_i's as well, with σ_i's having the distribution specified above. With the understanding that the outer expectation is with respect to $X_1,\dots,X_n$, $X_1',\dots,X_n'$ and $\sigma_1,\dots,\sigma_n$, we continue with our string of equalities and inequalities:

$$
\begin{aligned}
&= \mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i \sigma_i\left(g(X_i)-g(X_i')\right)\right| \quad \text{[by the above argument]}\\
&= \mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i \sigma_i g(X_i)-\frac{1}{n}\sum_i \sigma_i g(X_i')\right| \quad \text{[simple algebra]}\\
&\le \mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i \sigma_i g(X_i)\right|+\left|\frac{1}{n}\sum_i \sigma_i g(X_i')\right| \quad \text{[triangle inequality]}\\
&\le \mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i \sigma_i g(X_i)\right|+\sup_{g\in G}\left|\frac{1}{n}\sum_i \sigma_i g(X_i')\right| \quad [\sup_i(a_i+b_i)\le\sup_i a_i+\sup_i b_i]\\
&= 2\mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i \sigma_i g(X_i)\right| \quad \text{[linearity of expectation, combine terms]}
\end{aligned}
$$

Note that the "ghost sample" $X_1',\dots,X_n'$ has disappeared from the final bound. Let us try to interpret that last quantity. Inside the expectation and sup, we have $\sum_i \sigma_i g(X_i)$, which looks kind of like an inner product between the vectors $(\sigma_1, \ldots, \sigma_n)'$ and $(g(X_1), \ldots, g(X_n))'$. Then we consider the inner product with the largest magnitude as we range over functions in G.

Think of this maximum inner product as an indication of how well the class $\{(g(X_1), \ldots, g(X_n))' : g \in G\}$ can align (by having a big inner product) with random directions $(\sigma_1, \ldots, \sigma_n)$, which are distributed uniformly over the 2^n corners of an n-cube.

We shall see that because G is just the class of one-sided intervals, we can't even get close to hitting all the corners of the n-cube. This is good, because then

$$\mathbb{E} \sup_{g \in G} \left| \frac{1}{n} \sum_i \sigma_i g(X_i) \right|$$

can go to zero quickly. We will also see later that this quantity is like the Rademacher complexity of the class G (the only difference being the extra absolute value inside the sup).

Step 3: G is simple

To bound

$$\mathbb{E} \sup_{g \in G} \left| \frac{1}{n} \sum_i \sigma_i g(X_i) \right|,$$

first let us rewrite the expression by first conditioning on the data $\mathbf{X}$:

$$\mathbb{E}\, \mathbb{E} \left(\sup_{g \in G} \left| \frac{1}{n} \sum_i \sigma_i g(X_i) \right| \mid \mathbf{X} \right).$$

We will show that, quite remarkably, we can bound the inside conditional expectation (which is a function of the data $\mathbf{X}$) by something that does not depend on $\mathbf{X}$. This is due to the fact that G is not a very rich class, so there aren't that many values that $(g(X_1), \ldots, g(X_n))'$ can take. From now on, think of the data $\mathbf{X} = (X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ as fixed.

Consider the order statistics $x_{(1)} \leq \cdots \leq x_{(n)}$ of the n data points. Note that $x_{(i)} \leq x_{(i+1)}$ implies $g(x_{(i)}) \geq g(x_{(i+1)})$, and remember that $g(x) \in \{0, 1\}$. Therefore, the vector $(g(x_{(1)}), \ldots, g(x_{(n)}))'$ can take at most $n + 1$ different values as g ranges over G. In particular, these values are a subset of

$$\{(0, \ldots, 0)', (1, 0, \ldots, 0)', \ldots, (1, \ldots, 1, 0, \ldots, 0)', \ldots, (1 \ldots, 1, 0)', (1, \ldots, 1)'\}.$$

Since there is a bijection between $(x_1, \ldots, x_n)$ and $(x_{(1)}, \ldots, x_{(n)})$, the vectors $\{(g(x_1), \ldots, g(x_n))' : g \in G\}$ also has cardinality at most $n + 1$.

Now, we will bound

$$\mathbb{E} \left(\sup_{g \in G} \left| \frac{1}{n} \sum_i \sigma_i g(X_i) \right| \mid \mathbf{X} \right)$$

using Massart's finite class lemma, which requires that $\{(g(X_1), \ldots, g(X_n))' : g \in G\}$ is not too big.

Lemma 1.2 (Massart's finite class lemma). Let A be a finite subset of $\mathbb{R}^n$ with each vector in $a \in A$ bounded by $R = \max_{a \in A} ||a|| = \max_{a \in A} (\sum_i a_i^2)^{1/2}$.

Then

$$\mathbb{E} \max_{a \in A} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i a_i \right) \leq \frac{R\sqrt{2 \log |A|}}{n}.$$

Proof. Let $s > 0$ and define $Z_a = \sum_{i=1}^n \sigma_i a_i$.

$$\begin{aligned}
\exp(s\mathbb{E}\max_{a\in A} Z_a) &\leq \mathbb{E}\exp(s\max_{a\in A} Z_a) \quad \text{[convexity of } \exp(\cdot)\text{]} \\
&= \mathbb{E}\max_{a\in A}\exp(sZ_a) \quad [x \mapsto \exp(sx) \text{ is increasing}] \\
&\leq \mathbb{E}\sum_{a\in A}\exp(sZ_a) \quad \text{[replacing max by sum, exp is positive]} \\
&= \sum_{a\in A} \underbrace{\mathbb{E}\exp(sZ_a)}_{\text{moment generating function}} \qquad \text{[linearity of expectation]} \\
&\leq \sum_{a\in A}\exp(s^2R^2/2) \quad \text{[Hoeffding's lemma]} \\
&= |A|\exp(s^2R^2/2) \quad \text{[collect terms]}
\end{aligned}$$

The third inequality above is an application of the generalized Hoeffding's lemma as applied to bounded martingale differences (see notes for lecture 15). The constant c_i here is $2a_i$ since $\sigma_i a_i$ lies in $[-a_i, a_i]$. Thus the term $(s^2\sum_i c_i^2)/8$ becomes $(s^2\sum_i a_i^2)/2$, which is bounded above by $s^2R^2/2$.

Rearranging terms gives

$$\mathbb{E}\max_{a\in A} Z_a \leq \inf_{s>0}\left(\frac{\log|A|}{s} + \frac{sR^2}{2}\right).$$

If we optimize s to make the RHS as small as possible by setting derivatives to 0, we find that the optimum is $s^* = \sqrt{\frac{2\log|A|}{R^2}}$, so

$$\mathbb{E}\max_{a\in A} Z_a \leq R\sqrt{2\log|A|}.$$

The claim in the lemma follows by dividing both sides by n. $\square$

Note that R usually grows as $\sqrt{n}$, so this lemma is only useful when $|A|$ is $o(\exp(n))$. Otherwise, the bound does not converge to 0 as $n \to \infty$.

We apply the lemma with $A = \{(g(X_1), \ldots, g(X_n))' : g \in G\}$. Simple calculation shows that $R \leq \sqrt{n}$ and $|A| \leq n + 1$. The lemma yields the following result:

$$\mathbb{E}\left(\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right| \mid \mathbf{X}\right) \leq \sqrt{\frac{2\log(n+1)}{n}}$$

Now we combine the three steps of the proof:

$$\begin{aligned}
& P\left(\sup_{g\in G}|\mathbb{E}_n g - \mathbb{E}g| > \epsilon + 2\sqrt{\frac{2\log(n+1)}{n}}\right) \\
\leq\ & P\left(\sup_{g\in G}|\mathbb{E}_n g - \mathbb{E}g| > \epsilon + 2\mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum_i \sigma_i g(X_i)\right|\right) \quad \text{[Step 3]} \\
\leq\ & P\left(\sup_{g\in G}|\mathbb{E}_n g - \mathbb{E}g| > \epsilon + \mathbb{E}\sup_{g\in G}|\mathbb{E}_n g - \mathbb{E}g|\right) \quad \text{[Step 2]} \\
\leq\ & e^{-2\epsilon^2 n} \quad \text{[Step 1]}
\end{aligned}$$

From this, we conclude that $\sup_{g\in G} |\mathbb{E}_n g - \mathbb{E}g| \xrightarrow{P} 0$ because for any $\epsilon > 0$, letting $n \to \infty$ causes both $2\sqrt{\frac{2\log(n+1)}{n}}$ and $e^{-2\epsilon^2 n}$ to converge to zero.

Furthermore, we can upgrade convergence in probability to a.s. convergence by using the first Borel-Cantelli lemma, which states that $P(A_n \text{ i.o.}) = 0$ if $\sum_n P(A_n) < \infty$. If we let $A_n = \{\sup_{g\in G} |\mathbb{E}_n g - \mathbb{E}g| > \epsilon\}$ and let $\epsilon \downarrow 0$, we can conclude that we have convergence to 0 with probability 1.

$\square$

In the future, we would like to prove uniform convergence for more complicated G. When can we use this basic approach? Step 1 required that the functions G be uniformly bounded. Step 2 did not require any assumptions. Step 3 required the most, that given any fixed data points $\mathbf{X}$, the set of vectors $\{(g(X_1), \ldots, g(X_n))' : g \in G\}$ is not too big.

However, for richer classes G, we might have an infinite set of vectors. For example, this is the case with bounded monotonic functions and functions with bounded variations. These classes are in fact Glivenko-Cantelli, so we will need more powerful machinery to prove those results.

1.1 Rademacher averages

Now we turn to a way of measuring the complexity of a function class G.

Definition. For a class $F \subset \mathbb{R}^{\mathbf{X}}$ and a distribution on $\mathbf{X}$, the Rademacher average of F is defined as

$$R_n(F) = \mathbb{E} \sup_{f\in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i)$$

and the empirical Rademacher average is defined as

$$\hat{R}_n(F) = \mathbb{E}\left(\sup_{f\in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \mid X_1, \ldots, X_n \right),$$

where $X_1, \ldots, X_n \in \mathbf{X}$ are i.i.d. and $\sigma_1, \ldots, \sigma_n$ are Rademacher variables.

First, some intuition about $R_n(F)$. Think of F consisting of binary classifier functions of the form $f : \mathbf{X} \mapsto \{\pm 1\}$. Think of $\sigma_1, \ldots, \sigma_n$ as the labels on the data points.

We can interpret the quantity inside the sup of $R_n(F)$ in terms of a loss of a classifier f:

$$\begin{aligned} \frac{1}{n}\sum_i \sigma_i f(x_i) &= \frac{1}{n}\sum_i (1 - 2\cdot 1[\sigma_i \neq f(x_i)]) \\ &= 1 - 2\cdot \frac{1}{n}\sum_i 1[\sigma_i \neq f(x_i)] \\ &= 1 - 2\hat{L}_n(f) \end{aligned}$$

Taking the sup finds the minimum loss classifier over $f \in F$. By taking the expectation over the labels σ_i's, we are measuring the empirical loss of the best classifier in the class against random labels.

Also, note that $R_n(F)$ integrates over the distribution of $\mathcal{X}$. In step 3 of the Glivenko-Cantelli proof, we were able to get a bound on an $\hat{R}_n(F)$-like quantity but for any data set. But in some cases, integrating over the distribution of $\mathcal{X}$ could give us a tighter control over the complexity of the function class, since we could allow $\hat{R}_n(F)$ to be high, but just on portions of the space $\mathcal{X}$ that have low probability. This way, we could have a very complex function class, and if the distribution is simple, we might still be able to prove uniform convergence.