COMS 4773: Uniform convergence

Daniel Hsu

February 23, 2024

1 Uniform convergence

Let μ be a probability distribution on $\mathcal{X}$, and let $X_1, \ldots, X_n$ be an iid sample from μ. For a function class $\mathcal{F} \subset \{-1, 1\}^{\mathcal{X}}$ and probability distribution μ on $\mathcal{X}$, we are interested in

$$\sup_{f \in \mathcal{F}} \mu_n f - \mu f \tag{1}$$

or

$$\sup_{f \in \mathcal{F}} |\mu_n f - \mu f|, \tag{2}$$

where we use μf as shorthand for $\mathbb{E}_{X \sim \mu}[f(X)]$ and $\mu_n f$ as shorthand for $\frac{1}{n} \sum_{i=1}^n f(X_i)$. Here, μ_n is the *empirical distribution* based on the iid sample from μ of size n.

Note that (2) is the same as (1) if $\mathcal{F}$ is "closed under negation", i.e., $-f \in \mathcal{F}$ iff $f \in \mathcal{F}$. For technical reasons, it is sometimes easier to handle (1) instead of (2).

Suppose $\mathcal{F}$ has VC dimension $d < \infty$. Then, by Sauer's lemma, $\mathcal{F}$ only has $\binom{n}{\leq d}$ many different "behaviors" on n points. So we hope to be able to replace the supremum in (1) by a maximum over the different "behaviors". If this were the case, we could use Hoeffding's inequality and a union bound to bound the probability that (1) is larger than $\epsilon > 0$ by

$$\binom{n}{\leq d} \exp\bigl(-\Omega(n\epsilon^2)\bigr).$$

Since $\binom{n}{\leq d} = O(n^d)$, this probability bound would be at most δ for

$$\epsilon = O\left(\sqrt{\frac{d \log n + \log(1/\delta)}{n}}\right). \tag{3}$$

This is essentially what we will be able to prove.

The supremum in (1) considers differences between empirical averages and population means. It is true that the empirical average can take only $O(n^d)$ different values as we range over $f \in \mathcal{F}$, but it is possible that the population mean μf will take infinitely-many different values. This is a technical obstacle to analyzing (1) directly. Instead, we will use two "symmetrization" tricks to bypass this obstacle, before concluding with a concentration argument. Here is the three-step plan:

1. Symmetrization by ghost sample
2. Symmetrization by random signs
3. Conditioning and concentration

1.1 Symmetrization by ghost sample

The first trick is to instead analyze a variant of (1), where μf is replaced by an empirical average over an independent iid sample $X'_1, \dots, X'_n$ from μ, called the *ghost sample*:

$$\sup_{f \in \mathcal{F}} \mu_n f - \mu'_n f \tag{4}$$

(where we use $\mu'_n f$ as shorthand for $\frac{1}{n}\sum_{i=1}^n f(X'_i)$). The high-level idea is that if two iid random variables are far from each other, then at least one of them must be far from their common mean. Differences of empirical averages are entirely determined by effective behaviors of $\mathcal{F}$ on finite sets.

Here is a simple way to use this symmetrization trick. For any $\epsilon_1, \epsilon_2 > 0$, define the following notations for events:

- $\mathcal{E}[f] = \{\mu_n f - \mu f \geq \epsilon_1 + \epsilon_2\}$;
- $\mathcal{G}[f] = \{\mu'_n f - \mu f \geq \epsilon_1\}$;
- $\mathcal{S}[f] = \{\mu_n f - \mu'_n f \geq \epsilon_2\}$;
- $\mathcal{S}^* = \{\sup_{f \in \mathcal{F}} \mu_n f - \mu'_n f \geq \epsilon_2\}$.

Observe that $\mathcal{E}[f] \Rightarrow \mathcal{G}[f] \vee \mathcal{S}[f]$, and $\mathcal{S}[f] \Rightarrow \mathcal{S}^*$ for any $f \in \mathcal{F}$, even if f depends on the sample.

Let $f_n \in \arg\max_{f \in \mathcal{F}} \mu_n f - \mu f$, breaking ties in some arbitrary but fixed manner. Then

$$\begin{aligned}
\Pr(\mathcal{E}[f_n]) &\leq \Pr(\mathcal{G}[f_n] \vee \mathcal{S}[f_n]) && (\text{since } \mathcal{E}[f_n] \Rightarrow \mathcal{G}[f_n] \vee \mathcal{S}[f_n]) \\
&\leq \Pr(\mathcal{G}[f_n]) + \Pr(\mathcal{S}[f_n]) && (\text{union bound}) \\
&\leq \sup_{f \in \mathcal{F}} \Pr(\mathcal{G}[f]) + \Pr(\mathcal{S}[f_n]) && (\text{since } f_n \perp\!\!\!\perp \mu'_n) \\
&\leq \sup_{f \in \mathcal{F}} \Pr(\mathcal{G}[f]) + \Pr(\mathcal{S}^*) && (\text{since } \mathcal{S}[f_n] \Rightarrow \mathcal{S}^*).
\end{aligned}$$

We can bound $\Pr(\mathcal{G}[f])$ using Hoeffding's inequality, but we should choose ϵ_1 in a way so that $\Pr(\mathcal{G}[f])$ is dominated by $\Pr(\mathcal{S}^*)$.[1] After this is done, the main task is to bound $\Pr(\mathcal{S}^*)$.

1.2 Symmetrization by random signs

We defined (4) by considering two independent iid samples $X_1, \dots, X_n$ and $X'_1, \dots, X'_n$ from μ. But we can equivalently arrive at the same stochastic process by starting from a single iid sample from μ of size $2n$, and then partitioning it into two samples, each of size n. One way to do this is to pair up the $2n$ points arbitrarily—say, $(X_1^{\mathrm{a}}, X_1^{\mathrm{b}}), \dots, (X_n^{\mathrm{a}}, X_n^{\mathrm{b}})$—and then for each pair $(X_i^{\mathrm{a}}, X_i^{\mathrm{b}})$, put one in the first sample and the other in the second sample. We record our choices using a collection of signs $\sigma_1, \dots, \sigma_n \in \{-1, +1\}$:

- if $\sigma_i = +1$, then we put X_i^{a} in the first sample and X_i^{b} in the second sample;
- if $\sigma_i = -1$, then we put X_i^{a} in the second sample and X_i^{b} in the first sample.

[1] A different argument shows that $\Pr(\mathcal{E}[f_n]) \leq \Pr(\mathcal{S}^*)/(1 - \sup_{f \in \mathcal{F}} \Pr(\mathcal{G}[f]))$. In this case, it suffices to choose ϵ_1 so that $\Pr(\mathcal{G}[f]) \leq 1/2$ (say) for all $f \in \mathcal{F}$.

With a particular choice $\sigma_1, \ldots, \sigma_n$, we have

$$\mu_n f - \mu'_n f = \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Big(f(X_i^{\mathrm{a}}) - f(X_i^{\mathrm{b}}) \Big).$$

This works for any choice of $\sigma_1, \ldots, \sigma_n$, and hence it also works for a *random* choice.

So let us regard $\sigma_1, \ldots, \sigma_n$ as random variables—in fact, iid random variables with $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = 1/2$. We call such random variables *Rademacher random variables*. Then, again with $X_1^{\mathrm{a}}, X_1^{\mathrm{b}}, \ldots, X_n^{\mathrm{a}}, X_n^{\mathrm{b}}$ being an iid sample from μ of size $2n$, (4) has the same distribution as

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Big(f(X_i^{\mathrm{a}}) - f(X_i^{\mathrm{b}}) \Big).$$

We are already in position to move on to the final step, but there is one more simplification we can make. Note that

$$\begin{aligned} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Big(f(X_i^{\mathrm{a}}) - f(X_i^{\mathrm{b}}) \Big) &= \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i^{\mathrm{a}}) + \frac{1}{n} \sum_{i=1}^{n} (-\sigma_i) f(X_i^{\mathrm{b}}), \\ &\leq \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i^{\mathrm{a}}) + \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (-\sigma_i) f(X_i^{\mathrm{b}}). \end{aligned}$$

Therefore, if the supremum on the left-hand side is at least $\epsilon > 0$, then at least one of the two suprema on the right-hand side must be at least $\epsilon/2$. Hence,

$$\begin{aligned} &\Pr\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Big(f(X_i^{\mathrm{a}}) - f(X_i^{\mathrm{b}}) \Big) \geq \epsilon \right) \\ &\leq \Pr\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i^{\mathrm{a}}) \geq \epsilon/2 \right) + \Pr\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (-\sigma_i) f(X_i^{\mathrm{b}}) \geq \epsilon/2 \right) \\ &= 2 \Pr\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \geq \epsilon/2 \right), \end{aligned}$$

where the inequality follows from a union bound, and the last step uses the symmetry of the Rademacher random variables. (We have also gone back to expressing the events in terms of the original sample $X_1, \ldots, X_n$.)

1.3 Conditioning and concentration

Define the *empirical inner product* of two n-vectors $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ by

$$\langle u, v \rangle_n = \frac{1}{n} \sum_{i=1}^{n} u_i v_i.$$

Also define the *empirical norm* of an n-vector $v = (v_1, \ldots, v_n)$ by $\|v\|_n = \sqrt{\langle v, v \rangle_n}$.

The previous step left us with analyzing the random variable

$$\sup_{v \in \mathcal{F}(X_{1:n})} \langle \sigma, v \rangle_n$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a Rademacher random vector (i.e., vector of Rademacher random variables), and $\mathcal{F}(X_{1:n}) = \{f(X_1), \ldots, f(X_n) : f \in \mathcal{F}\}$ is the set of behaviors of $\mathcal{F}$ on the sample $X_1, \ldots, X_n$. For any fixed $v \in \mathbb{R}^n$, we call $\langle \sigma, v \rangle_n$ a *Rademacher average*, so we have a supremum of Rademacher averages for the set of (random) vectors $\mathcal{F}(X_{1:n})$.

The final (simple) trick is to condition on the sample $X_{1:n}$, and then to apply a concentration inequality for the supremum of Rademacher averages. Upon conditioning on $X_{1:n}$, the effective behaviors $\mathcal{F}(X_{1:n})$ becomes a deterministic set of vectors. In general, it could be a set of cardinality 2^n; but if $\mathcal{F}$ has VC dimension $d < \infty$, then it has cardinality at most $O(n^d)$ by Sauer's lemma.

Since $\sigma_1, \ldots, \sigma_n$ are iid mean-zero 1-subgaussian random variables, we have for any $v \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$,

$$\mathbb{E} \exp(\lambda \langle \sigma, v \rangle_n) = \mathbb{E} \exp\left(\frac{\lambda}{n} \sum_{i=1}^n \sigma_i v_i \right) \leq \prod_{i=1}^n \mathbb{E} \exp\left(\frac{\lambda \sigma_i v_i}{n} \right) \leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 v_i^2}{2n^2} \right) = \exp\left(\frac{\lambda^2 \|v\|_n^2}{2n} \right),$$

which implies $\langle \sigma, v \rangle_n$ is a mean-zero $\frac{1}{n}\|v\|_n^2$-subgaussian random variable. So for any finite set of vectors $V \subset \mathbb{R}^n$ and any $\epsilon > 0$,

$$\Pr\left(\max_{v \in V} \langle \sigma, v \rangle_n \geq \epsilon \right) \leq \sum_{v \in V} \exp\left(-\frac{n\epsilon^2}{2\|v\|_n^2} \right).$$

Since the range of each $f \in \mathcal{F}$ is $\{-1, 1\}$, we have $\|v\|_n = 1$ for all $v \in \mathcal{F}(X_{1:n})$. So, when $\mathcal{F}$ has VC dimension $d < \infty$, we have

$$\Pr\left(\sup_{v \in \mathcal{F}(X_{1:n})} \langle \sigma, v \rangle_n \geq \epsilon \mid X_{1:n} \right) \leq \binom{n}{\leq d} \exp\left(-\frac{n\epsilon^2}{2} \right).$$

1.4 Putting it all together

Theorem 1. *Let $\mathcal{F} \subset \{-1, 1\}^{\mathcal{X}}$ have VC dimension $d < \infty$. Let μ be a probability distribution on $\mathcal{X}$, and let μ_n be the empirical distribution based on an iid sample from μ of size n. Then for any $\epsilon > 0$,*

$$\Pr\left(\sup_{f \in \mathcal{F}} \mu_n f - \mu f \geq \epsilon \right) \leq \exp\left(-\frac{n\epsilon^2}{8} \right) + 2\binom{n}{\leq d} \exp\left(-\frac{n\epsilon^2}{32} \right).$$

This implies that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \mu_n f - \mu f \leq O\left(\sqrt{\frac{d \log n + \log(1/\delta)}{n}} \right).$$

This is proved by combining the previous steps with appropriate choices for ϵ, ϵ_1, ϵ_2, etc.

2 Statistical learning via empirical risk minimization

Theorem 1 can be used in the analysis of *empirical risk minimization (ERM)* for statistical learning. In statistical learning (for binary classification), μ is a probability distribution on $\mathcal{X} \times \{-1, 1\}$, where $\mathcal{X}$ is the input space and $\{-1, 1\}$ is the output space. The learner is provided training data, i.e.,

an iid sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from μ, and the goal is to output a hypothesis $h\colon \mathcal{X} \to \{-1, 1\}$ that has low *error rate*

$$\operatorname{err}(h) = \mu(\{(x, y) \in \mathcal{X} \times \{-1, 1\} : h(x) \neq y\}).$$

Given a hypothesis class $\mathcal{H} \subseteq \{-1, 1\}^{\mathcal{X}}$, ERM selects a hypothesis $h_n \in \arg\min_{h \in \mathcal{H}} \operatorname{err}_n(h)$ that minimizes the *empirical error rate*

$$\operatorname{err}_n(h) = \mu_n(\{(x, y) \in \mathcal{X} \times \{-1, 1\} : h(x) \neq y\}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(X_i) \neq Y_i\},$$

where μ_n is the empirical distribution on the training data.

Let $h^* \in \arg\min_{h \in \mathcal{H}} \operatorname{err}(h)$ be a hypothesis in $\mathcal{H}$ of minimum error rate. The relevance of uniform convergence to the analysis of ERM comes from the following decomposition:

$$\begin{aligned}
\operatorname{err}(h_n) - \operatorname{err}(h^*) &= \operatorname{err}(h_n) - \operatorname{err}_n(h_n) + \operatorname{err}_n(h_n) - \operatorname{err}_n(h^*) + \operatorname{err}_n(h^*) - \operatorname{err}(h^*) \\
&\leq \operatorname{err}(h_n) - \operatorname{err}_n(h_n) + \operatorname{err}_n(h^*) - \operatorname{err}(h^*) \\
&\leq \underbrace{\sup_{h \in \mathcal{H}} \operatorname{err}(h) - \operatorname{err}_n(h)}_{A} + \underbrace{\operatorname{err}_n(h^*) - \operatorname{err}(h^*)}_{B}
\end{aligned}$$

since $\operatorname{err}_n(h_n) \leq \operatorname{err}(h)$ for all $h \in \mathcal{H}$. The probability that term marked B is large can be bounded using Hoeffding's inequality: for any $\epsilon > 0$,

$$\Pr(\operatorname{err}_n(h^*) - \operatorname{err}(h^*) \geq \epsilon) \leq \exp\bigl(-2n\epsilon^2\bigr).$$

Moreover, the term marked A can be written as

$$\sup_{h \in \mathcal{H}} \operatorname{err}(h) - \operatorname{err}_n(h) = \frac{1}{2}\left(\sup_{f \in \mathcal{F}^{\mathcal{H}}} \mu_n f - \mu f\right)$$

where

$$\mathcal{F}^{\mathcal{H}} = \{f^h : h \in \mathcal{H}\} \quad \text{and} \quad f^h(x, y) = y\, h(x).$$

We can apply Theorem 1 to bound the probability that this supremum is large, provided that $\mathcal{F}^{\mathcal{H}}$ has finite VC dimension. Consider any set of n points $z_1, \ldots, z_n \in \mathcal{X} \times \{-1, 1\}$ with $z_i = (x_i, y_i)$ for each i. Then it is easy to see that $\mathcal{F}^{\mathcal{H}}(z_{1:n})$ is in one-to-one correspondence with $\mathcal{H}(x_{1:n})$. This implies that $\mathcal{F}^{\mathcal{H}}$ has the same VC dimension as $\mathcal{H}$. So, by Theorem 1, for any $\epsilon > 0$,

$$\Pr\left(\sup_{h \in \mathcal{H}} \operatorname{err}(h) - \operatorname{err}_n(h) \geq 2\epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{8}\right) + 2\binom{n}{\leq d} \exp\left(-\frac{n\epsilon^2}{32}\right).$$

where d is the VC dimension of $\mathcal{H}$. Putting everything together, we obtain the following.

Theorem 2. *Let $\mathcal{H} \subset \{-1, 1\}^{\mathcal{X}}$ have VC dimension $d < \infty$. Let μ be a probability distribution on $\mathcal{X} \times \{-1, 1\}$, and let h_n be the ERM from $\mathcal{H}$ based on an iid sample from μ of size n. Then for any $\varepsilon, \delta \in (0, 1)$, if*

$$n \geq O\left(\frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon^2}\right),$$

then with probability at least $1 - \delta$,

$$\operatorname{err}(h_n) - \inf_{h \in \mathcal{H}} \operatorname{err}(h) \leq \varepsilon.$$