Lecturer: Prof. Homrighausen Scribe: Chuck Vollmer

1 Concentration of Measure (cont...)

An Example: Uniform Risk

Recall... that we are trying to minimize $\hat{\mathbb{P}}\ell_f = n^{-1}\sum_{i=1}^n |f(X_i) - Y_i|^2$ over a set of functions $\mathcal{F}_n$.

The main tool in showing consistency/rates of convergence for such estimators is to show that the *empirical risk* looks like the *true risk*, uniformly over $\mathcal{F}_n$:

$$\sup_{f\in\mathcal{F}_n} \left|\hat{\mathbb{P}}\ell_f - \mathbb{P}\ell_f\right| \to 0 \text{ a.s.}$$

BUT...

We really just care about the loss and how it affects our empirical process, so we just generate an easier (read: less complex) set, $\mathcal{L}_n = \{\ell_f : f \in \mathcal{F}_n\}$, and pay attention to:

$$\sup_{\ell\in\mathcal{L}_n} \left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right|$$

Now, we bound *this* guy using Hoeffding's:

If $\ell(Z) \in [0, b]$ a.s., then

$$\mathbb{P}\left(\left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| > \epsilon\right) \leq 2\exp\left\{-\frac{2n\epsilon^2}{b^2}\right\}$$

and, subsequently, by using a union bound (with $\mathcal{L}$ finite)

$$\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| > \epsilon\right) \leq 2|\mathcal{L}_n|\exp\left\{-\frac{2n\epsilon^2}{b^2}\right\}$$

since $\mathbb{P}\cup_{j=1}^J A_j \leq \sum_{j=1}^J \mathbb{P}(A_j)$.

Therefore, we have convergence in probability.

We can strengthen this to include almost sure convergence using Borel-Cantelli:

If $|\mathcal{L}_n|$ grows slowly enough for

$$\sum_{n=1}^{\infty}|\mathcal{L}_n|\exp\left\{-\frac{2n\epsilon^2}{b^2}\right\} < \infty$$

Then

$$\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| \to 0 \text{ a.s.}$$

Though this will work for some situations, $|\mathcal{L}_n|$ is really infinite. In this case, the goal becomes to find a finite set $\mathcal{L}_{n,\epsilon}$ such that

$$\left\{\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell-\mathbb{P}\ell\right|>\epsilon\right\}\subseteq\left\{\sup_{\ell\in\mathcal{L}_{n,\epsilon}}\left|\hat{\mathbb{P}}\ell-\mathbb{P}\ell\right|>\epsilon'\right\}$$

Letting ϵ' be a function of ϵ but not n, we construct $\mathcal{L}_{n,\epsilon}$ using covering numbers with respect to different metrics $||\cdot||$ (which corresponds to $(T,d)\leftrightarrow(\mathcal{L}_n,||\cdot||)$.)

The most basic choice is

$$d_\infty(\ell,\ell')=||\ell-\ell'||_\infty=\sup_z|\ell(z)-\ell'(z)|$$

An ϵ-cover of $\mathcal{L}_n$ with respect to d_∞ is such that for every $\ell\in\mathcal{L}_n$, $\exists\ell_\epsilon\in\mathcal{L}_{n,\epsilon}$ such that

$$d_\infty(\ell,\ell_\epsilon)=||\ell-\ell_\epsilon||_\infty=\sup_z|\ell(z)-\ell_\epsilon(z)|<\epsilon$$

And we can define the L_∞ *covering number*

$$N_\infty(\epsilon,\mathcal{L}_n)=N_\infty(\epsilon,\mathcal{L}_n,||\cdot||_\infty)$$

to be the minimal ϵ-cover (alternatively known as the uniform covering number –not to be confused with uniform convergence).

Putting this all together...

$$\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell-\mathbb{P}\ell\right|>\epsilon\right)\leq 2N_\infty(\epsilon/3,\mathcal{L}_n)\exp\left\{-\frac{2n\epsilon^2}{9b^2}\right\}$$

Proof Sketch...

Let $\mathcal{L}_{n,\epsilon/3}$ be a minimal $\epsilon/3$-cover of $\mathcal{L}_n$. Fix a $\ell\in\mathcal{L}_n$, then $\exists\ell'\in\mathcal{L}_{n,\epsilon/3}$ such that $||\ell-\ell'||_\infty<\epsilon/3$

$$\begin{aligned}\left|\hat{\mathbb{P}}\ell-\mathbb{P}\ell\right|&\leq\left|\hat{\mathbb{P}}\ell-\hat{\mathbb{P}}\ell'\right|+\left|\hat{\mathbb{P}}\ell'-\mathbb{P}\ell'\right|+|\mathbb{P}\ell'-\mathbb{P}\ell|\\&\leq||\ell-\ell'||_\infty+\left|\hat{\mathbb{P}}\ell'-\mathbb{P}\ell'\right|+||\ell-\ell'||_\infty\\&\leq 2\epsilon/3+\left|\hat{\mathbb{P}}\ell'-\mathbb{P}\ell'\right|\end{aligned}$$

Now, we do as before, but with $N_\infty(\epsilon/3,\mathcal{L}_n)=|\mathcal{L}_{n,\epsilon/3}|$

BUT...

Unfortunately, $N_\infty(\epsilon/3, \mathcal{L}_n)$ is often too large for

$$\sum_{n=1}^{\infty} N_\infty(\epsilon/3, \mathcal{L}_n) \exp\left\{-\frac{2n\epsilon^2}{9b^2}\right\} < \infty$$

(The uniform norm is quite strong)

At issue is that

$$\left|\hat{\mathbb{P}}\ell - \hat{\mathbb{P}}\ell'\right| + |\mathbb{P}\ell' - \mathbb{P}\ell| \leq ||\ell - \ell'||_\infty + ||\ell - \ell'||_\infty$$

is quite coarse.

So...

Doing the above turns out to be a very crude thing to do... and we really need something better.

Yeah? Well... then how do we do it?

We do this by adding randomness t our indexing set.

mmhmm....

And we do *that* by introducing "fictive" data.

Really?!? ok... tell me more...

We appeal to *random L_1-norm covers*.

2 Introducing randomness to our indexing set

The idea is based around fictively creating a *ghost sample* $Z_1, \ldots, Z_{2n}$ such that $Z_i \overset{i.i.d}{\sim} \mathbb{P}$ for $i = 1, \ldots, 2n$.

Now, we will think of $\mathbb{P}\ell \approx \hat{\mathbb{P}}_{n+1}^{2n}\ell$ and form

$$\left\{\sup_{\ell \in \mathcal{L}_n} \left|\hat{\mathbb{P}}\ell - \hat{\mathbb{P}}_{n+1}^{2n}\ell\right| > \epsilon\right\} \subseteq \left\{\sup_{\ell \in \mathcal{L}_{n,\epsilon}} \left|\hat{\mathbb{P}}\ell - \hat{\mathbb{P}}_{n+1}^{2n}\ell\right| > \epsilon'\right\}$$

where now $\mathcal{L}_{n,\epsilon}$ is going to be a data-dependent set (i.e. a random variable).

Now, we cover $\mathcal{L}_n$ with ϵ-covers with respect to

$$||\ell - \ell'||_n = \int |\ell - \ell'| d\hat{\mathbb{P}}_1^n$$

with covering number $N_1(\epsilon, \mathcal{L}_n) = N(\epsilon, \mathcal{L}_n, ||\cdot||_n)$.

2.1 Introducing the randomness through a *ghost sample*

The main idea and set-up is shown above, and the following important result can be shown:

for $n \geq 2b^2/\epsilon^2$:

$$\mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} \left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| > \epsilon\right) \leq 2\mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} \left|\hat{\mathbb{P}}\ell - \hat{\mathbb{P}}^{2n}_{n+1}\ell\right| > \epsilon/2\right)$$

(See pages 136-138 of Györfi et al. (2002) for details)

2.2 Introduction of additional randomness through a Rademacher random variables

As all the R.V.s are iid, their difference is invariant to a random sign change with Rademacher R.V.s

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} \left|\hat{\mathbb{P}}\ell - \hat{\mathbb{P}}^{2n}_{n+1}\ell\right| > \frac{\epsilon}{2}\right) \\
&= \mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} \left|\sum_{i=1}^{n} \epsilon_i \left(\ell(Z_i) - \ell(Z_{i+n})\right)\right| > \frac{n\epsilon}{2}\right) \\
&\leq \mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} \left|\sum_{i=1}^{n} \epsilon_i \ell(Z_i)\right| > \frac{n\epsilon}{4}\right) + \mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} \left|\sum_{i=1}^{n} \epsilon_i \ell(Z_{i+n})\right| > \frac{n\epsilon}{4}\right) \\
&= 2\mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} \left|\sum_{i=1}^{n} \epsilon_i \ell(Z_i)\right| > \frac{n\epsilon}{4}\right)
\end{aligned}$$

(This is known as *symmetrization*.)

2.3 Conditioning to introduce a Covering Number

We observe that:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} \left|\sum_{i=1}^{n} \epsilon_i \ell(Z_i)\right| > \frac{n\epsilon}{4}\right) \\
&= \mathbb{E}_Z \mathbb{P}_\epsilon \left(\exists \ell \in \mathcal{L}_n : \left|\sum_{i=1}^{n} \epsilon_i \ell(z_i)\right| > \frac{n\epsilon}{4} \middle| (Z_i)_{i=1}^n = (z_i)_{i=1}^n\right)
\end{aligned}$$

Now, we can apply the complexity part:

Let $\mathcal{L}_{n,\epsilon/8}$ be an $\epsilon/8$-cover of $\mathcal{L}_n$ with respect to $||\cdot||_n$:

$$\frac{1}{n}\sum_{i=1}^{n} |\ell(z_i) - \ell'(z_i)| < \epsilon/8$$

Now... fix an ℓ and find its $\ell' \in \mathcal{L}_{n,\epsilon/8}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(z_i)\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell'(z_i)\right| + \epsilon/8$$

Showing (conditional on $(Z_i)_{i=1}^n = (z_i)_{i=1}^n$)

$$\begin{aligned}&\mathbb{P}\left(\exists \ell \in \mathcal{L}_{n,\epsilon/8} : \left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(z_i)\right| + \epsilon/8 > \frac{\epsilon}{4}\right)\\ &\leq N_1(\epsilon,\mathcal{L}_n)\max_{\ell\in\mathcal{L}_{n,\epsilon/8}}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(z_i)\right| > \frac{\epsilon}{8}\right)\end{aligned}$$

2.4 And, finally, bound using Hoeffding's Inequality

Recall... that we want to bound the following object:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(z_i)\right| > \frac{\epsilon}{8}\right)$$

Now...this can be done with Hoeffding:

(Recall, though, that Bernstein's tends to give better bounds.)

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(z_i)\right| > \frac{\epsilon}{8}\right) \leq 2\exp\left\{-\frac{2n(\epsilon/8)^2}{(2b)^2}\right\} \leq 2\exp\left\{-\frac{2n\epsilon^2}{128b^2}\right\}$$

2.5 And putting this all together we find that

$$\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| > \epsilon\right) \leq 8\mathbb{E}[N]_1(\epsilon,\mathcal{L}_n)\exp\left\{-\frac{2n\epsilon^2}{128b^2}\right\}$$

■

BUT...

We need to know the expectation of the covering number... and we know that covering numbers are essentially packing numbers. So... we use those!

2.6 Quantifying the complexity using VC dimension

Recall that the VC dimension of a class of sets $\mathcal{A}$ is the largest number of points $\mathcal{G}_n$ which we can shatter.

(That is, the largest n such that $|\{G \subset \mathcal{G}_n : G = \mathcal{G}_n \cap A, A \in \mathcal{A}\}| = 2^n$.)

Let

$$\mathcal{G} = \{(z,t) \in \mathbb{R}^{p+1} \times \mathbb{R} : t \leq \ell(z), \ell \in \mathcal{L}_n\}$$

(This is the set of all *subgraphs* of functions in $\mathcal{L}_n$.)

We can bound the L_q packing numbers by $VC_{\mathcal{G}}$ = VC dimension of $\mathcal{G}$.

Let ν be a probability measure on $\mathbb{R}^{p+1}$ and let $\sup_{\ell \in \mathcal{L}_n} ||\ell||_\infty \leq b$.

Then, for any $\epsilon \in (0, b/4)$

$$M(\epsilon, \mathcal{L}_n, ||\cdot||_{L_q(\nu)}) \leq 3\left(\frac{2eb^q}{\epsilon^q}\log\left(\frac{3eb^q}{\epsilon^q}\right)\right)^{VC_{\mathcal{G}}}$$

(Remember: packing numbers are essentially covering numbers for our purposes. Note that the bound on the RHS doesn't depend on ν.)

Hence, we can leverage the "agreement" between the empirical L_q norm and packing numbers to bound the random quantity with the nonrandom VC dimension.

Unfortunately,

- the gap between these bounds can be huge
- VC dimension can frequently only be upper-bounded itself

So... what's the solution???

Bracketing numbers!!!