CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 16: October 26

Lecturer: Csaba Szepesvári *Scribes: Dávid Szepesvári*

Lecture 16 video

16.1 Outline

- Recap of Rademacher Complexity and "Expected Maximum Deviation".
- Bounding the Rademacher complexity of a class and sample by some measures of the size of the function class only.
- Introducing McDiarmid's Inequality, a concentration inequality that generalizes Hoeffding's,
- Applying this to Pg to see how its empirical estimate, $P_n g$, concentrates around it.
- Making the *Chaining* argument to arrive at a tighter bound on the Rademacher complexity of a class, which will allows us to remove the $\log n$ factor in our uniform deviation bounds.

16.2 Recap and Notation

We have been working towards removing the $\log n$ factor from the uniform deviation bounds for VC-classes. We do this by using Rademacher complexity and the so-called Chaining argument.

We have a space $\mathcal{Z}$, and $z_{1:n} \in \mathcal{Z}^n$ n-tuple. A function class $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ and $P \in \mathcal{M}_1(\mathcal{Z})$, a probability distribution on $\mathcal{Z}$. Let $\sigma \sim \mathrm{Rad}(n)$ be a random sign vector of length n. We defined the Rademacher complexity

$$R(\mathcal{G}, z_{1:n}) = \mathbb{E}[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i)],$$

which may sometimes be denoted by $R_n(\mathcal{G}, z_{1:n})$. Then we can define $R_n(\mathcal{G}, P)$ through samples $Z_{1:n} \sim P^{\otimes n}$:

$$R_n(\mathcal{G}, P) = \mathbb{E}[R(\mathcal{G}, Z_{1:n})].$$

We also defined the *Expected Maximum Deviation*,

$$\varepsilon_n(\mathcal{G}, P) = \mathbb{E}[\sup_{g \in \mathcal{G}} Pg - P_n g].$$

We, then, have two propositions.

Proposition 16.1. *Let* $g_n \in \arg\min_{g \in \mathcal{G}} P_n g$*, then* $Pg_n \leq \inf_{g \in \mathcal{G}} Pg + \varepsilon_n(\mathcal{G}, P)$.

Proposition 16.2. $\varepsilon_n(\mathcal{G}, P) \leq 2R_n(\mathcal{G}, P)$.

16.3 Relating the size of a class to Rademacher complexity

Definition 16.3. Let $z_{1:n} \in Z^n$. Define the *empirical norm* of it, $L_2(z_{1:n})$:

$$\|g\|^2_{L_2(z_{1:n})} = \frac{1}{n}\sum_{i=1}^{n} g^2(z_i).$$

We will use the shorthand $\|g\|^2_n$ to mean the same. Notice $\|g\|^2_{L_2(z_{1:n})} \to \|g\|^2_{L_2(P)}$ as $n \to \infty$.

Definition 16.4. We can, then, define the size of the norm of the class $\mathcal{G}$ through the $\sup$ of norms as before:

$$\|\mathcal{G}\|_n = \sup_{g\in\mathcal{G}} \|g\|_n.$$

Proposition 16.5. *Let $N = |\mathcal{G}(z_{1:n})|$, the number of behaviours of the function class when projected through $z_{1:n}$. Then*

$$R_n(\mathcal{G}, z_{1:n}) \le \|\mathcal{G}\|_n \sqrt{\frac{2\ln N}{n}}.$$

Discussion: while there is no general bound on $\|\mathcal{G}\|_n$, in many applications there will be natural bounds, such as in the case of binary functions. Otherwise extra work might be required, or this inequality might not be the most useful one.

Proof. We'll start from the definition.

$$R_n(\mathcal{G}, z_{1:n}) = \mathbb{E}[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)] = \mathbb{E}[\max_{g\in\tilde{\mathcal{G}}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)],$$

where we simply noticed that projecting through the samples, there are really only finitely many items we are taking the $\sup$ over. We set $\tilde{\mathcal{G}} \subseteq \mathcal{G}$ finite with $\tilde{\mathcal{G}}(z_{1:n}) = \mathcal{G}(z_{1:n})$. Then $|\tilde{\mathcal{G}}| = N$, too.

Next, we'll use the *log-sum-exp* inequality. This states that for $A \subseteq \mathbb{R}$, $|A| < \infty$, if we order the elements $a_1 \ge \ldots \ge a_n$, then $\forall \eta > 0$

$$e^{\eta a_1} \le \sum_{j=1}^{n} e^{\eta a_j}.$$

This is trivially true as the LHS is in the positive sum on the RHS. By taking the log of both sides and rearranging we get $\max A \le \frac{1}{\eta}\log\sum_{j=1}^{n} e^{\eta a_j}$. Back to the main inequality:

$$\begin{aligned}
\ldots &\le \mathbb{E}[\frac{1}{\eta}\log\sum_{g\in\tilde{\mathcal{G}}} \exp\left(\frac{\eta}{n}\sum_{i=1}^{n}\sigma_i g(z_i)\right)] \\
&\overset{\text{Jensen}}{\le} \frac{1}{\eta}\log\sum_{g\in\tilde{\mathcal{G}}} \mathbb{E}[\exp\left(\frac{\eta}{n}\sum_{i=1}^{n}\sigma_i g(z_i)\right)] \\
&= \frac{1}{\eta}\log\sum_{g\in\tilde{\mathcal{G}}} \mathbb{E}[\prod_{i=1}^{n}\exp\left(\frac{\eta}{n}\sigma_i g(z_i)\right)] \\
&\overset{\text{Indep}}{=} \frac{1}{\eta}\log\sum_{g\in\tilde{\mathcal{G}}} \prod_{i=1}^{n}\mathbb{E}[\exp\left(\frac{\eta}{n}\sigma_i g(z_i)\right)].
\end{aligned}$$

Next, we use $\mathbb{E}[\exp(\sigma x)] \leq \exp(x^2/2)$ for $x \in \mathbb{R}$, $\sigma \sim \mathrm{Rad}(1)$:

$$\begin{aligned}
\dots &\leq \frac{1}{\eta} \log \sum_{g \in \tilde{\mathcal{G}}} \prod_{i=1}^{n} \exp\left(\frac{\eta^2}{2n^2} g^2(z_i)\right) \\
&\leq \frac{1}{\eta} \log \sum_{g \in \tilde{\mathcal{G}}} \exp\left(\frac{\eta^2}{2n} \underbrace{\frac{1}{n} \sum_{i=1}^{n} g^2(z_i)}_{\|g\|_n^2}\right) \\
&\leq \frac{1}{\eta} \log \left[N \exp\left(\frac{\eta^2}{2n} \frac{1}{n} \|\mathcal{G}\|_n^2\right)\right],
\end{aligned}$$

where in the last step we bounded each term in the sum by $\|\mathcal{G}\|_n^2$. Next, we optimize η. The last line is equal to $\frac{1}{\eta} \log N + \frac{1}{\eta} \frac{\eta^2}{2n} \|\mathcal{G}\|_n^2$, so by setting these two terms to be equal, we get $\eta = \sqrt{\frac{2n \log N}{\|\mathcal{G}\|_n^2}}$. Then

$$\dots = 2\|\mathcal{G}\|_n \sqrt{\frac{\log N}{2n}} \leq \|\mathcal{G}\|_n \sqrt{\frac{2 \log N}{n}}. \qquad \square$$

The essence of this argument is an upper bound on the expected maximum of a bunch of random variables that concentrate at a rate of $1/\sqrt{n}$. You could e.g. make the same style argument for (centered) sub-Gaussians (of the same constant). The maximum will yield a $\sqrt{(\log N)}$ boost to the expected value. In fact, there is a lower bound that says that you can't do much better, so this upper bound is pretty tight. Here, the common sub-Gaussian constant was the scale of $\mathcal{G}$.

Note the connection to VC-classes: the number of behaviours, $N = |\mathcal{G}(z_{1:n})|$, is limited for VC-classes. It is bounded by a polynomial, n^d, so $logN \leq d \log n$. It gets even better; we will see refined bounds based on Haussler's result that instead of the $\log n$ term use a $\log \frac{1}{\varepsilon}$ term instead. We will then, through a careful analysis, the chaining argument, will be able to remove the $\log n$ factor. But first, we need to work towards a high probability oracle inequality for P_n.

16.4 High Probability Bounds

We want a high probability oracle inequality for the deviations, so far we only have one for the expected maximum deviation in Prop 16.2, $\varepsilon_n(\mathcal{G}, P) \leq 2R_n(\mathcal{G}, P)$. We bring out a "big cannon" to help us achieve this goal: McDiarmid's concentration inequality. While the concentration inequalities we have seen so far were for the average of r.v.s, McDiarmid's inequality is a concentration inequality for any function with limited "sensitivity" to its arguments. The key observation is that the concentration inequalities worked, because the "average function" has limited sensitivity to any one of its inputs: if you swapped out any input for any another value, the average can only change by 1/n at most. McDiarmid's inequality captures this more general result.

Theorem 16.6 (McDiarmid)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}$. Define the sensitivity to the i^{th} input of f as*

$$\Delta_i = \sup_{x \in \mathcal{X}^n} \sup_{x_i' \in \mathcal{X}} f(x) - f(x_1, \dots, x_{i-1}, x_{i'}, x_{i+1}, \dots, x_n).$$

Then, for $X \in P^{\otimes n} \in \mathcal{M}_1(\mathcal{X})$ and $0 < \delta < 1$,

(a) w.p. $1 - \delta$: $f(X) \leq \mathbb{E}f(X) + \sqrt{\frac{1}{2} \sum_i^n \Delta_i^2 \log \frac{1}{\delta}}$.

(b) w.p. $1 - \delta$: $f(X) \geq \mathbb{E}f(X) - \sqrt{\frac{1}{2} \sum_i^n \Delta_i^2 \log \frac{1}{\delta}}$.

Proof. We will not prove this here. The proof can be found in lots of place. The argument is Chernoff-like and employs techniques from martingales. Not too different from the proof of Hoeffding's. $\square$

Example. If you apply this theorem to the average function, you will get Hoeffding's Inequality.

We will now use this inequality to prove our first high probability result relating Pg with its empirical estimate $P_n g$.

Theorem 16.7. *$\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$, $P \in \mathcal{M}_1(\mathcal{Z})$. Let*

$$M = \sup_{g \in \mathcal{G}} \sup_{z, z' \in \mathcal{Z}} |g(z) - g(z')|.$$

Then for $0 < \delta < 1$ we have w.p. $1 - \delta$ that $\forall g \in \mathcal{G}$:

$$\begin{aligned} Pg &\leq P_n g + \varepsilon_n(\mathcal{G}, P) + M\sqrt{\frac{\ln 1/\delta}{2n}} \\ &\leq P_n g + 2R_n(\mathcal{G}, P) + M\sqrt{\frac{\ln 1/\delta}{2n}}. \end{aligned}$$

Proof. Denote $z_{1:n} \in \mathcal{Z}^n$ simply by z for brevity. Further, denote

$$f(z) = \sup_{g \in \mathcal{G}} \left[Pg - \frac{1}{n} \sum_{j=1}^{n} g(z_j) \right] = \sup_{g \in \mathcal{G}} u(g, z).$$

We will bound the sensitivity of f to its inputs. Towards this, for $1 \leq i \leq n$, $z_i' \in \mathcal{Z}$ denote $z' = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)$, where we swapped out the i^{th} entry in z for z_i'. Then we can write

$$\begin{aligned} f(z) - f(z') &= \sup_{g \in \mathcal{G}} u(g, z) - \sup_{g \in \mathcal{G}} u(g, z') \\ &\leq \sup_{g \in \mathcal{G}} [u(g, z) - u(g, z')] \\ &= \sup_{g \in \mathcal{G}} \left[-\frac{1}{n} \Big(g(z_i) + g(z_i') \Big) \right] \\ &\leq \frac{M}{n}, \end{aligned}$$

where we used that $\forall g' \in \mathcal{G}$ $\sup_{g \in \mathcal{G}} u(g, z') \geq u(g', z')$, that most terms in $u(g, z) - u(g, z')$ cancel, and the definition of M, respectively. Of course, we get the same bound for $f(z') - f(z)$, therefore $\Delta_i \leq M/n \; \forall i$. Then $\sum_i^n \Delta_i^2 = n\frac{M^2}{n^2} = \frac{M^2}{n}$, so by McDiarmid, w.p. $1 - \delta$, $Z_{1:n} \sim P^{\otimes n}$

$$f(Z_{1:n}) \leq \underbrace{\mathbb{E}[\sup_{g \in \mathcal{G}} Pg - P_n g]}_{\varepsilon_n(\mathcal{G}, P)} + M\sqrt{\frac{\log(1/\delta)}{2n}}.$$

By writing out the definition of $f(Z_{1:n})$ and rearranging, we complete the proof. □

Discussion about Fast Rates. McDiarmid gave us a Hoeffding-type bound – can we get a Bernstein-type, small-risk bound, as well? The answer is yes. The arguments rely on Talagrand's concentration inequality, but we do not do this here. The book works out a number of results in this setting.

16.5 Removing $\log n$ from the upper bound

Recall, in our upper bounds so far we had a $\sqrt{\frac{n \ln n}{d}}$ term, while the lower bound is $\sqrt{\frac{n}{d}}$. We have been working towards an upper bound with the same rate. We are already on a promising trajectory, as the bounds we just proved in Thm 16.7 do not have the $\ln n$ term, so as long as it does not appear in R_n we are set! Let's work this out. First we will bound the Rademacher complexity by covering numbers through a *Chaining* argument. Then, we will show that that leads to a bound on the Rademacher complexity without the $\ln n$ term.

Theorem 16.8. *Let $z_{1:n} \in \mathcal{Z}$ and $N(s) = N(s, \mathcal{G}, L_2(z_{1:n}))$, the covering number. Then*

$$R(\mathcal{G}, z_{1:n}) \leq \inf_{\varepsilon > 0} 4\varepsilon + 12 \int_{\varepsilon}^{\infty} \sqrt{\frac{\ln N(s)}{n}} ds.$$

Note that the integral in the Theorem is actually a finite integral, as for bounded classes, beyond a certain scale you can cover with just 1 function, so the logarithm of that will be 0.

Fast-forwarding a bit, we will then have the following corollary, which will directly lead to a high probability upper bound with a rate matching the lower bound!

Corollary 16.9. *Let $\mathcal{G} \in \{0,1\}^{\mathcal{Z}}$, $d = \mathrm{VC}(\mathcal{G})$, then*

$$R_n(\mathcal{G}, z_{1:n}) \leq O(\sqrt{d/n}).$$

This combined with Thm 16.7 will yield the bound we were looking for without the $\log n$ term.

Proof of Corollary. Haussler has a result, which we saw earlier, that upper bounded the 2-norm empirical covering number for a VC-class like so:

$$\ln N_2(\varepsilon, G, n) \leq 1 + \ln(d+1) + d \ln \frac{2e}{\varepsilon^2}.$$

Then, plugging this into Thm 16.8 and noting that (a) we can pick ε basically 0, and (b) for a binary class, for a covering scale s larger than $1/2$ we can cover with a single function so the metric entropy will be 0:

$$R(\mathcal{G}, z_{1:n}) \leq \int_0^{\frac{1}{2}} \sqrt{\ln N(s)} \leq c\sqrt{d}$$

for some constant c. This is because $\ln N(s)$ is linear in d by Haussler – it is an exercise to show how to integrate out the $\ln \frac{2e}{\varepsilon^2}$ term. □

Proof of Thm 16.8. The idea for this argument is to consider coverings at multiple scales, from very big to small. Before, we always had this issue of trading off the covering number for scale ε under the square root, with the ε additive term. We always considered just one scale. Considering multiple scales, we'll arrive at a better bound. Let $B = \|G\|_n$, the empirical 2-norm. Define covering scales

$$\varepsilon_0 = B,\ \varepsilon_1 = B/2,\ \varepsilon_3 = B/4,\ \ldots, \varepsilon_l = 2^{-l}B,\ \ldots.$$

Let $\mathcal{G}_l$ be a min ε_l-cover of $\mathcal{G}$ w.r.t. $\|\cdot\|_n$, the empirical norm. Denote $N_l = |\mathcal{G}_l| = N(\varepsilon_l)$. Further, for convenience, set $\mathcal{G}_0 := \{0\}$ (even if it is not in $\mathcal{G}$).

For $l \geq 0$, let $g_l(g) = \arg\min_{g' \in \mathcal{G}_l} \|g' - g\|_n$. Then, by the construction of the cover

$$\|g - g_l(g)\|_n \leq \varepsilon_l.$$

Now, pick $L > 0$ large positive integer, corresponding to the smallest scale approximation we consider. Then we can write

$$g = \underbrace{g - g_L(g)}_{\text{very good approx}} + \underbrace{g_L(g) - g_{L-1}(g)}_{\substack{\text{trading off approx} \\ \text{and covering \#}}} + \ldots + g_1(g) - g_0(g) + \underbrace{g_0(g)}_{=0}. \tag{16.1}$$

The idea is that $g_L(g)$ is a very fine approximation, so we pay a small additive ε price for it, but its corresponding covering number would be too large. Therefore we add it back, moving to a less fine approximation. We will bound the empirical norm of each term:

$$\|g_l(g) - g_{l-1}(g)\|_n \leq \|g_l(g) - g\|_n + \|g_{l-1}(g) - g\|_n \leq \varepsilon_l + \varepsilon_{l-1} \leq \varepsilon_l + 2\varepsilon_l = 3\varepsilon_l.$$

We also have $|\{g_l(g) - g_{l-1}(g) : g \in \mathcal{G}\}| \leq N_l N_{l-1}$. We are ready to derive the promised bound on the Rademacher complexity! We will use the notation $[f]_z$ to mean f evaluated at z. An explanation of each step will come after.

$$
\begin{aligned}
R(\mathcal{G}, z_{1:n}) &= \mathbb{E} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \\
&\leq \mathbb{E} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i [g - g_L(g)]_{z_i} + \mathbb{E} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \sum_{l=1}^{L} [g_l - g_{l-1}(g)]_{z_i} && (16.2) \\
&\leq \varepsilon_L + \mathbb{E} \sum_{l=1}^{L} \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i [g_l - g_{l-1}(g)]_{z_i} && (16.3) \\
&\leq \varepsilon_L + \sum_{l=1}^{L} 3\varepsilon_l \sqrt{\frac{2 \ln N_l N_{l-1}}{n}} && (16.4) \\
&\leq \varepsilon_L + 6 \sum_{l=1}^{L} \varepsilon_l \sqrt{\frac{\ln N_l}{n}} && (16.5) \\
&= \varepsilon_L + 12 \sum_{l=1}^{L} \frac{1}{2} \varepsilon_l \sqrt{\frac{\ln N_l}{n}} && (16.6) \\
&= \varepsilon_L + 12 \sum_{l=1}^{L} (\varepsilon_l - \varepsilon_{l-1}) \sqrt{\frac{\ln N(\varepsilon_l)}{n}} && (16.7) \\
&\leq \varepsilon_L + 12 \int_{\varepsilon_L/2}^{\infty} \sqrt{\frac{\ln N(s)}{n}} ds. && (16.8) \\
&\leq \inf_{\varepsilon > 0} 4\varepsilon + 12 \int_{\varepsilon}^{\infty} \sqrt{\frac{\ln N(s)}{n}} ds. && (16.9)
\end{aligned}
$$

The steps carried out were:

- Eq 16.2: split g according to the sum in 16.1, take the $\sup$ of the first term (fine approximation) and the rest of the terms separately.
- Eq 16.3: In the first term, each $\sigma_i [g - g_L(g)]_{z_i}$ is upper bounded by ε_L, hence the whole expression is. In the second term we move the sum over L outside the $\sup$ (upper bound), and even the expectation.
- Eq 16.4: Notice that the expectation is the Rademacher complexity for a finite class, with number of elements $N_l N_{l-1}$ as we saw above. The scale is $3\varepsilon_l$, as shown above. The bound is from Prop 16.5.
- Eq 16.5: use $N_l \geq N_{l-1}$, also adjusting constants for convenience.
- Eq 16.6: setup for the next step, where..
- Eq 16.7: we use $\frac{1}{2}\varepsilon_l = \varepsilon_l - \varepsilon_{l+1}$ as $\varepsilon_{l+1} = \frac{1}{2}\varepsilon_l$ by construction. Also writing $N(\varepsilon_l) = N_l$.
- Eq 16.8: note that what we have is a Riemann sum, as $N(\varepsilon_l)$ is a decreasing function of ε_l, so we can bound it by the corresponding integral. We are generous with the upper limit of the integral here.
- Eq 16.9: first, use $\varepsilon' = \varepsilon_L/2$. Then note that the inequalities so far were true for any L, so we can take the $\inf$ over the corresponding discrete set of ε_L. However the last line has an $\inf$ over all continuous values. The best continuous ε is a factor of 2 off from the best discrete ε, hence the additional factor of 2 in this step.

Note, B is the upper end of the integral, otherwise it does not matter. □