Lecture 10 – February 7

Lecturer: Stephen Bates  Scribe: Zhihan Xiong

Warning: these notes may contain factual errors

Reading:  VDV Chapter 10, HDP Chapter 1, 2, 8

Outline:
• Concentration Inequalities:
  – sub-Gaussian random variables
  – symmetrization
• Rademacher Complexity, Metric Entropy and ULLNs

1 Recap

In last lecture, we defined Uniform Law of Large Numbers (ULLNs), which is
\[ \| \mathbb{P}_n - \mathbb{P} \|_F = \sup_{f \in \mathcal{F}} | \mathbb{P}_n f - \mathbb{P} f | \overset{p}{\to} 0. \]

Further, we developed a condition that leads to ULLNs, based on \( \epsilon \)-covering number of parameter space and the bracketing number of the corresponding function space. The \( \epsilon \)-covering number is defined as
\[ N(\Theta, d, \epsilon) = \min \{ N \mid \exists \{ \theta_i \}_{i=1}^N \text{ such that } \forall \theta \in \Theta, d(\theta, \theta_i) < \epsilon \text{ for some } i \} \]

In this lecture, we will develop another condition that leads to ULLNs, based on the metric entropy and Rademacher complexity. Before doing that, we will first build some tools about concentration inequalities based on sub-Gaussian random variables.

2 Sub-Gaussian R.V.’s and Concentration Inequalities

We start by defining the sub-Gaussian random variable.

**Definition 2.1.** A random variable \( X \) is \( \sigma^2 \)-sub-Gaussian if
\[ \mathbb{E} \left[ e^{\lambda (X - \mathbb{E}[X])} \right] \leq e^{\lambda^2 \sigma^2/2} \]

**Example 1:** If \( X \sim \mathcal{N}(\theta, \sigma^2) \), then it is a \( \sigma^2 \)-sub-Gaussian.

**Example 2:** If \( X \in [a, b] \) almost surely, then it is possible to show that
\[ \mathbb{E} \left[ e^{\lambda (X - \mathbb{E}[X])} \right] \leq e^{\frac{\lambda^2 (b-a)^2}{4}} \]

This result is called Hoeffding’s lemma. We will not prove it here, but people who are interested in its proof may refer to section B.4 of Shalev-Shwartz’s textbook[1].

It indicates that such an \( X \) is \( \frac{(b-a)^2}{4} \)-sub-Gaussian.
we will now introduce some useful properties of sub-Gaussian random variables so that we can build concentration inequalities based on them.

**Proposition 1.** Let $X_i$’s be independent $\sigma_i^2$-sub-Gaussian random variables. Then, $\sum_{i=1}^n X_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian.

**Proof**

\[
\mathbb{E} \left[ e^{\lambda \sum_{i=1}^n (X_i - \mathbb{E}[X_i])} \right] = \prod_{i=1}^n \mathbb{E} \left[ e^{\lambda (X_i - \mathbb{E}[X_i])} \right] \\
\leq \prod_{i=1}^n e^{\lambda^2 \sigma_i^2 / 2} \quad \text{(By definition of sub-Gaussian)}
\]

\[= e^{\frac{\lambda^2}{2} \sum_{i=1}^n \sigma_i^2} \]

\[\square\]

**Proposition 2.** Let $X$ be $\sigma^2$-sub-Gaussian. Then, $\mathbb{P}(X - \mathbb{E}[X] > t) \leq e^{-\frac{t^2}{2\sigma^2}}$.

**Proof** Without loss of generality, we assume that $\mathbb{E}[X] = 0$. Then, it will be sufficient to prove that $\mathbb{P}(X > t) \leq e^{-\frac{t^2}{2\sigma^2}}$. To show this, we can have

\[
\mathbb{P}(X > t) = \mathbb{P}\left( e^{\lambda X} > e^{\lambda t} \right) \quad \text{(By monotonicity of } \lambda \mapsto e^{\lambda x} \text{)}
\]

\[\leq \frac{1}{e^{\lambda t}} \mathbb{E}[e^{\lambda X}] \quad \text{(By Markov’s inequality)}
\]

\[\leq e^{-\lambda t + \frac{\lambda^2 \sigma^2}{2}} \quad \text{(By definition of sub-Gaussian)}
\]

Since the above bound will be true for any $\lambda$, we can minimize this upper bound over $\lambda$. This can be done by simply taking derivative and setting it to 0. Specifically, when we choose $\lambda^* = \frac{t}{\sigma^2}$, we will get the desired bound since

\[e^{-\lambda^* t + \frac{\lambda^2 \sigma^2}{2}} = e^{-\frac{t^2}{2\sigma^2}}\]

\[\square\]

**Corollary 3.** Let $X_i$’s be independent $\sigma_i^2$-sub-Gaussian. Then, by combining above two propositions, we can get

\[
\mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > t \right) \leq \exp \left( -\frac{n^2 t^2}{2 \sum_{i=1}^n \sigma_i^2} \right)
\]

In particular, when all $\sigma_i^2$’s are equal, we can have

\[
\mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > t \right) \leq \exp \left( -\frac{nt^2}{2\sigma^2} \right)
\]
**Remark** It is possible to show that if \( P(|X| > t) \leq 2e^{-ct^2} \) for some constant \( c > 0 \), then \( X \) is a \( \sigma^2 \)-sub-Gaussian for some \( \sigma^2 \).

Intuitively, sub-Gaussian represents distributions with tail not fatter than Gaussian. As an counterexample, a Cauchy random variable is not sub-Gaussian. We will then give a result that is essentially the same as problem 1.7 in homework, so we will not prove it again.

**Proposition 4.** If \( X_1, \ldots, X_n \) are \( \sigma^2 \)-sub-Gaussian with zero mean, then we can have

\[
E \left[ \max_{1 \leq i \leq n} X_i \right] \leq \sqrt{2 \sigma^2 \log n}
\]

When we try to show a collection of function \( F \) satisfies ULLNs under certain distribution, we can instead have

\[
P \left( \sup_{f \in F} |P_n f - Pf| > t \right) \leq \frac{1}{t} E \left[ \sup_{f \in F} |P_n f - Pf| \right]
\]

If we can show that \( E \left[ \sup_{f \in F} |P_n f - Pf| \right] \) converges to 0 as \( n \to \infty \), then we are done. However, doing this directly is difficult. Thus, we need to resort to a technique called symmetrization, which will be introduced next.

3 Symmetrization and Rademacher Complexity

**Definition 3.1.** \( \epsilon \) is Rademacher random variable if it is uniformly distributed over \( \{-1, 1\} \).

When doing symmetrization, we need to introduce Rademacher random variable, which leads to the following theorem.

**Theorem 5.** Let \( X_1, \ldots, X_n \) be random vectors in some Banach space (complete normed space) with norm \( \|\cdot\| \). Let \( \epsilon_1, \ldots, \epsilon_n \) be iid Rademacher random variables that are independent of \( X_i \)'s. Then, for \( p \geq 1 \), we have

\[
E \left[ \left\| \sum_{i=1}^n (X_i - E[X_i]) \right\|_p \right] \leq 2^p E \left[ \left\| \sum_{i=1}^n \epsilon_i X_i \right\|_p \right]
\]

**Proof** Let \( X'_i \) be an independent copy of \( X_i \) for convenience. Then, we can have

\[
E \left[ \left\| \sum_{i=1}^n (X_i - E[X_i]) \right\|_p \right] = E \left[ \left\| \sum_{i=1}^n (X_i - E[X'_i]) \right\|_p \right] 
\]

\[
\leq E \left[ \left\| \sum_{i=1}^n (X_i - X'_i) \right\|_p \right] 
\]

(By Jensen’s inequality and convexity of \( x \mapsto \|x\|_p \))

\[
= E \left[ \left\| \sum_{i=1}^n \epsilon_i (X_i - X'_i) \right\|_p \right] 
\]

(Since \( \epsilon_i (X_i - X'_i) \) and \( (X_i - X'_i) \) are identically distributed)

\[
= 2^p E \left[ \left\| \frac{1}{2} \sum_{i=1}^n (X_i - X'_i) \right\|_p \right]
\]
\[ \leq 2^p \cdot \frac{1}{2} \mathbb{E} \left[ \left\| \sum_{i=1}^{n} \varepsilon_i X_i \right\|^p \right] + 2^p \cdot \frac{1}{2} \mathbb{E} \left[ \left\| \sum_{i=1}^{n} \varepsilon_i X_i' \right\|^p \right] \]

(By triangle inequality)

\[ = 2^p \mathbb{E} \left[ \left\| \sum_{i=1}^{n} \varepsilon_i X_i \right\|^p \right] \]

\[ \square \]

**Remark**  By using the above theorem, we can have

\[ \mathbb{E} \left[ \| \mathbb{P}_n f - \mathbb{P} f \|_\mathcal{F} \right] = \mathbb{E} \left[ \sup_{f \in \mathcal{F}} | \mathbb{P}_n f - \mathbb{P} f | \right] \leq 2 \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left\| \frac{1}{n} \sum_{i=1}^{n} f (X_i) \right\| \right] \]

Therefore, now our goal becomes to bound \( \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left\| \frac{1}{n} \sum_{i=1}^{n} f (X_i) \right\| \right] = R_n (\mathcal{F}) \), which is called the Rademacher complexity of \( \mathcal{F} \).

Further, we can notice that when conditional on \( X_1, \ldots, X_n \), \( \sum_{i=1}^{n} \varepsilon_i X_i \) is \( \sum_{i=1}^{n} X_i^2 \)-sub-Gaussian, which can potentially help us to bound the Rademacher complexity.

### 4 Metric Entropy and ULLNs

Now, with all tools needed, we will develop a condition for ULLNs based on metric entropy. It is stated as the following theorem.

**Theorem 6.** For function family \( \mathcal{F} \), assume that \( \exists \: F \in L^1 (\mathbb{P}) \) such that \( |f (x)| \leq F (x) \) for all \( x \in \mathcal{X} \) and \( f \in \mathcal{F} \). Further, define

\[ f_M (x) = \begin{cases} f (x) & \text{if } |f (x)| \leq M \\ 0 & \text{otherwise} \end{cases} \]

and let \( \mathcal{F}_M = \{ f_M : f \in \mathcal{F} \} \).

Now, if \( \log N (\mathcal{F}_M, L^1 (\mathbb{P}_n), \varepsilon) = o_p (n) \) for all \( M, \varepsilon > 0 \), we can have \( \| \mathbb{P}_n - \mathbb{P} \|_\mathcal{F} \xrightarrow{p} 0 \).

**Proof**  By previous remarks, it suffices to show that \( \lim_{n \to \infty} \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left\| \frac{1}{n} \sum_{i=1}^{n} f (X_i) \right\| \right] = 0 \).

First, we fix some \( M, \varepsilon > 0 \). Then, by triangle inequality, we can have

\[ \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left\| \frac{1}{n} \sum_{i=1}^{n} f (X_i) \right\| \right] \leq \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f (X_i) - f_M (X_i)) \right\| \right] + \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_M (X_i) \right\| \right] \]

By noticing that \( f (x) - f_M (x) \leq F (x) \mathbb{I}_{\{F (x) > M\}} \) for any \( x \in \mathcal{X} \) and any \( f \in \mathcal{F} \), we can have

\[ \text{first term} \leq \mathbb{E} \left[ F (X) \mathbb{I}_{\{F (X) > M\}} \right] \]

For second term, let \( \mathcal{G} \subseteq \mathcal{F}_M \) be a minimum \( \varepsilon \)-cover of \( \mathcal{F}_M \) under \( L^1 (\mathbb{P}_n) \). Then, condition on \( X_1, \ldots, X_n \), with probability 1, we can have

\[ \sup_{f \in \mathcal{F}_M} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f (X_i) \right\| \leq \max_{g \in \mathcal{G}} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g (X_i) \right\| + \varepsilon \]


Condition on $X_1, \ldots, X_n$, $\sum_{i=1}^n \varepsilon_i g(X_i)$ is $\sum_{i=1}^n g(X_i)^2$-sub-Gaussian as we mentioned in last section. Then, since $g \in \mathcal{F}_M$, $\sum_{i=1}^n \varepsilon_i g(X_i)$ is also $nM^2$-sub-Gaussian. Therefore, we can have

\[
E \left[ \max_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(X_i) \right| \right] = \frac{1}{\sqrt{n}} E \left[ \max_{g \in \mathcal{G}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i g(X_i) \right| \right] \\
\leq \frac{1}{\sqrt{n}} \cdot \sqrt{2M^2 \log (2|\mathcal{G}|)} \\
= M \sqrt{\frac{2 \log 2 + 2 \log |\mathcal{G}|}{n}} \\
= M \sqrt{o_P \left( 1 + \frac{1}{n} \right)} \\
= Mo_P (1)
\]

The inequality above comes from the fact that $\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i g(X_i)$ condition on $X_i$ is $M^2$-sub-Gaussian and Proposition 4 in section 2. The extra factor 2 before $|\mathcal{G}|$ is because $|x| = \max \{x, -x\}$. This bound indicates that

second term $\leq M E \left[ o_P (1) \right]$

With all above, we can now have

\[
E \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \right] \leq E \left[ F(X) \mathbb{I}_{\{F(X) > M\}} \right] + \epsilon + M E \left[ o_P (1) \right] \\
\rightarrow E \left[ F(X) \mathbb{I}_{\{F(X) > M\}} \right] + \epsilon \quad \text{as } n \to \infty
\]

Since this bound is true for any $M, \epsilon > 0$, by noticing that $\inf_{M>0} E \left[ F(X) \mathbb{I}_{\{F(X) > M\}} \right] = 0$, we can have

\[
\lim_{n \to \infty} E \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \right] \leq \inf_{M, \epsilon > 0} \left\{ E \left[ F(X) \mathbb{I}_{\{F(X) > M\}} \right] + \epsilon \right\} = 0
\]

\[
\implies \lim_{n \to \infty} E \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \right] = 0
\]

This completes the proof. \qed

References