36-755: Advanced Statistical Theory 1 **Fall 2016**

Lecture 17: October 31

Lecturer: Alessandro Rinaldo *Scribes: Arun Sai Suggala*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

17.1 Sub Gaussian Processes, Metric Entropy and Chaining

17.1.1 Overview

In this class and the next few classes we will look at techniques to control collection of random variables indexed by sets with infinite number of elements. This collection of random variables is called a *stochastic process*. To be more specific, given random variables $\{X_\theta, \theta \in \mathbb{T}\}$ indexed by $\mathbb{T}$, we are interested in bounding

$$\mathbb{E}\left[\sup_{\theta \in \mathbb{T}} X_\theta\right], \tag{17.1}$$

where $\mathbb{T} \subseteq \mathbb{R}^n$. Lets now look at two important stochastic processes that we often come across:

- When $X_\theta = \langle \theta, X \rangle$, X is rademacher (i.e., $X_1, X_2 \dots X_n$ are i.i.d rademacher random variables) then the expression in (17.1) is called the *Rademacher Complexity* of $\mathbb{T}$.

  Note that, we came across *Rademacher Complexity* when we discussed Uniform Law of Large Numbers. When $\mathbb{T} = \mathcal{F}(Z_1^n) = \{(f(z_1), f(z_2), \dots f(z_n)) \in \mathbb{R}^n, f \in \mathcal{F}\}$ and X_θ is as defined above then (17.1) is the conditional *Rademacher Complexity* of $\mathcal{F}$ given Z_1^n.

- If $X_\theta = \langle \theta, X \rangle$ and $X \sim \mathcal{N}(0, I)$ is gaussian, then the expression in (17.1) is called the *Gaussian Complexity* of $\mathbb{T}$.

Example 17.1 *Let $\mathbb{T}$ be the unit euclidean ball in $\mathbb{R}^n$. And let $\mathcal{R}_n(\mathbb{T})$ be the Rademacher Complexity of $\mathbb{T}$ and $\mathcal{G}_n(\mathbb{T})$ be its Gaussian Complexity. Then*

$$\mathcal{R}_n(\mathbb{T}) = \sqrt{n}, \quad \mathcal{G}_n(\mathbb{T}) \leq \sqrt{n}. \tag{17.2}$$

Example 17.2 *Let $\mathbb{T} = \{\theta \in \mathbb{R}^n, \|\theta\|_1 \leq 1\}$. Then*

$$\sup_{\theta \in \mathbb{T}} \langle \theta, X \rangle = \|X\|_\infty = \max_i |X_i|.$$

Then $\mathcal{R}_n(\mathbb{T}) = 1, \quad \mathcal{G}_n(\mathbb{T}) \leq 2\sqrt{\log n}$.

17.1.2 Sub-Gaussian Stocahstic Process

Definition 17.3 *A zero mean stochastic process $\{X_\theta, \theta \in \mathbb{T}\}$ is a sub-gaussian process w.r.t d, a metric on $\mathbb{T}$, if for every $\lambda \in \mathbb{R}$, the following is true:*

$$\mathbb{E}\left[\exp \lambda(X_\theta - X_{\theta'})\right] \leq \exp \frac{\lambda^2 d^2(\theta, \theta')}{2} \quad \forall \theta, \theta' \in \mathbb{T}.$$

or equivalently $(X_\theta - X_{\theta'}) \in SG(d^2(\theta, \theta'))$.

Note that in many cases $d(\theta, \theta') = \Theta(\|\theta - \theta'\|)$. This is the case with Rademacher and Gaussian complexities. When $d^2(\theta, \theta') = \mathbb{E}[(X_\theta - X_{\theta'})^2]$, then d is called the canonical distance.

17.1.3 1-Step Discretization Method

We now present the 1-step discretization method to bound (17.1) when $\{X_\theta, \theta \in \mathbb{T}\}$ is a sub-gaussian process. This method uses the *metric entropy* of a set, which is a measure of the size of a set with infinitely many elements. Before we present the theorem, we review the definition of covering number.

Definition 17.4 ***(Covering Number)*** *A δ-cover of a set $\mathbb{T}$ w.r.t metric d is a set $\{\theta_1, \theta_2 \dots \theta_N\} \subseteq \mathbb{T}$ such that for each $\theta \in \mathbb{T}$, there exists some $i \in \{1, 2, \dots N\}$ such that $d(\theta, \theta_i) \leq \delta$. The δ-covering number $N(\delta, \mathbb{T})$ is the cardinality of the smallest δ-cover.*

Theorem 17.5 *Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a sub-gaussian stochastic process w.r.t metric d. Let $D = \sup_{\theta, \theta' \in \mathbb{T}} d(\theta, \theta')$ be the diameter of $\mathbb{T}$ w.r.t d (assume $D < \infty$). Then $\forall \delta \in [0, D]$*

$$\mathbb{E}\left[\sup_{\theta, \theta' \in \mathbb{T}} X_\theta - X_{\theta'}\right] \leq 2\mathbb{E}\left[\sup_{\gamma, \gamma' \in \mathbb{T}, d(\gamma, \gamma') \leq \delta} X_\gamma - X_{\gamma'}\right] + 4D\sqrt{\log N(\delta, \mathbb{T})}. \tag{17.3}$$

Proof: Fix $\delta \in [0, D]$. Let $\{\theta_1, \theta_2, \dots \theta_N\}$ be a δ-covering of $\mathbb{T}$ where $N = N(\delta, \mathbb{T})$. For any $\theta \in \mathbb{T}$, let θ_i be a point in $\delta-$covering such that $d(\theta, \theta_i) \leq \delta$. Then:

$$X_\theta - X_{\theta_1} = X_\theta - X_{\theta_i} + X_{\theta_i} - X_{\theta_1} \leq \left[\sup_{\gamma, \gamma' \in \mathbb{T}, d(\gamma, \gamma') \leq \delta} X_\gamma - X_{\gamma'}\right] + \left[\max_{i \in [N]} |X_{\theta_i} - X_{\theta_1}|\right].$$

For any other point $\theta' \in \mathbb{T}$, the same bound applies to $X_{\theta_1} - X_{\theta'}$. Adding these two, we get:

$$\sup_{\theta, \theta' \in \mathbb{T}} X_\theta - X_{\theta'} \leq 2\left[\sup_{\gamma, \gamma' \in \mathbb{T}, d(\gamma, \gamma') \leq \delta} X_\gamma - X_{\gamma'}\right] + 2\max_{i \in [N]} |X_{\theta_i} - X_{\theta_1}|.$$

To complete the proof we take expectations on both sides of the above equation and use the fact that $\{X_\theta, \theta \in \mathbb{T}\}$ is a sub-gaussian process:

$$\mathbb{E}\left[\sup_{\theta, \theta' \in \mathbb{T}} X_\theta - X_{\theta'}\right] \leq 2\mathbb{E}\left[\sup_{\gamma, \gamma' \in \mathbb{T}, d(\gamma, \gamma') \leq \delta} X_\gamma - X_{\gamma'}\right] + 4D\sqrt{\log N(\delta, \mathbb{T})}.$$

■

Remark 17.6 *Note that* (17.3) *also gives a bound for* $\mathbb{E}\left[\sup_{\theta\in\mathbb{T}} X_\theta\right]$*:*

$$\mathbb{E}\left[\sup_{\theta\in\mathbb{T}} X_\theta\right] = \mathbb{E}\left[\sup_{\theta\in\mathbb{T}} X_\theta - X_{\theta_0}\right] \quad \text{for any } \theta_0 \in \mathbb{T}$$

$$\leq \mathbb{E}\left[\sup_{\theta,\theta'\in\mathbb{T}} X_\theta - X_{\theta'}\right]$$

Remark 17.7 *This is the same discretization bound we derived before in the class when we talked about covering and packing numbers.*

Example 17.8 ***(Gaussian and Rademacher complexities)***

Let $\mathbb{T} \subset \mathbb{R}^n$ *be bounded and let* $X_\theta = \langle \theta, X\rangle$ *where* X *is a rademacher or gaussian random vector. Note that* X_θ *is a sub-gaussian stochastic process w.r.t* $\|.\|_2$*. We now apply the 1-step discretization theorem to get an upper bound for Gaussian and Rademacher complexities of* $\mathbb{T}$*. From to* (17.2) *we have:*

$$\mathbb{E}\left[\sup_{\gamma,\gamma'\in\mathbb{T}, d(\gamma,\gamma')\leq\delta} X_\gamma - X_{\gamma'}\right] \leq \delta\sqrt{n}.$$

Then the 1-step discretization bound gives us:

$$\mathbb{E}\left[\sup_{\theta\in\mathbb{T}} X_\theta\right] \leq \min_{\delta\in[0,D]} 2\delta\sqrt{n} + 4D\sqrt{\log N(\delta,\mathbb{T})}.$$

Example 17.9 ***(Random Matrices)***

Let $W \in \mathbb{R}^{n\times d}$ *be a matrix with i.i.d* $SG(1)$ *entries. We are interested in bounding* $\|W\|_{op} = \sup_{v\in\mathbb{S}^{d-1}} \|Wv\|_2$*. We can write* $\|W\|_{op}$ *as*

$$\|W\|_{op} = \sup_{\Theta\in M^{n,d}(1)} X_\Theta,$$

where $M^{n,d}(1)$ *is the set of all* $n \times d$ *matrices of rank 1 and frobenius norm 1,* $X_\Theta = \langle W, \Theta\rangle_F$*. Note that* X_Θ *is a zero-mean, SG process w.r.t frobenius distance. So applying 1-step discretization bound we get:*

$$\mathbb{E}\left[\|W\|_{op}\right] \leq 2\mathbb{E}\left[\sup_{\Gamma,\Gamma'\in M^{n,d}(1), \|\Gamma-\Gamma'\|_F\leq\delta} X_\Gamma - X_{\Gamma'}\right] + 6\sqrt{\log N(\delta, M^{n,d}(1))}$$

$$\leq 2\sqrt{2}\delta\mathbb{E}[\|W\|_{op}] + 6\sqrt{(n+d)\log\left(1+\frac{2}{\delta}\right)}.$$

Setting $\delta = \frac{1}{4\sqrt{2}}$*, we get* $\frac{1}{\sqrt{n}}\mathbb{E}\left[\|W\|_{op}\right] \leq c\left(1+\sqrt{\frac{d}{n}}\right)$ *where* $c > 0$ *is a universal constant.*

Example 17.10 ***(Non-parametric Regression)***

Let $\mathcal{F}_L$ *be the set of lipschitz functions on* $[0,1]$ *defined as:*

$$\mathcal{F}_L = \{f : [0,1] \to \mathbb{R}, |f(x) - f(y)| \leq L|x-y|, x, y \in [0,1]\}.$$

Consider the $\|.\|_\infty$ metric on $\mathcal{F}_L$ defined as : $\|f-g\|_\infty = \sup_x |f(x)-g(x)|$. Then we have the following bound on covering number of $\mathcal{F}_L$ using $\|.\|_\infty$ metric: $\log N_\infty(\delta, \mathcal{F}_L) \asymp \frac{L}{\delta}$, $\forall \delta < \delta_0$.

In non-parametric settings we usually work with $\frac{\mathcal{F}_L(X_1^n)}{\sqrt{n}}$ because in this case $\|.\|_2$ on $\mathcal{F}_L(X_1^n)$ corresponds to empirical L_2 distance: $\|f-g\|_n = \frac{1}{\sqrt{n}}\sqrt{\sum_i (f(x_i)-g(x_i))^2}$. It is easy to see that $\|f-g\|_n \le \|f-g\|_\infty$. So $\log N_2\left(\delta, \frac{\mathcal{F}_L}{\sqrt{n}}\right) \le \log N_\infty(\delta, \mathcal{F}_L)$. So if we are interested in Gaussian complexity of $\frac{\mathcal{F}_L(X_1^n)}{n}$, we get the following upper bound from 1-step discretization method:

$$\mathcal{G}_n\left(\frac{\mathcal{F}_L(X_1^n)}{n}\right) \le \frac{1}{\sqrt{n}} \inf_{\delta \in (0,\delta_0)} \delta\sqrt{n} + c\sqrt{\frac{L}{\delta}}.$$

Setting $\delta = n^{-1/3}$ we get $\mathcal{G}_n\left(\frac{\mathcal{F}_L}{n}\right) \le cn^{-1/3}$, where $c > 0$ is a constant.

17.1.4 Chaining Method

In the next class we will look at chaining technique (by *Dudley*) that provides a better bound for $\mathbb{E}\left[\sup_{\theta\in\mathbb{T}} X_\theta\right]$. Specifically, chaining gives a better bound for $\max_{i\in[N]} |X_{\theta_i} - X_{\theta_1}|$ of the form:

$$\int_{\delta/4}^{D} \sqrt{\log N(\mu, \mathbb{T})} d\mu.$$

If we let $\delta \to 0$, we will recover Dudley's result.

References

[WM09] Wainwright, Martin J, High-dimensional statistics: A non-asymptotic viewpoint.