Sub-Gaussian Processes and Chaining

John Duchi

Stats 300b – Winter Quarter 2021

Outline

- Sub-gaussian processes
- Rademacher complexities
- Chaining and Dudley's entropy integral
- Comparison inequalities

Reading:

- Wainwright, *High Dimensional Statistics*, Chapters 5.1–5.3, 5.4–5.6 for extra perspective
- Vershynin, *High Dimensional Probability*, Chapters 8.1–8.4.

Motivation

- multiple examples of bounded supremum with expectation
- always have

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \le 2\mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| \mid X_1^n\right]\right]$$

- question today: bound processes like $f \mapsto \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)$ (fixing x_i)
- naive idea: just discretize $\mathcal{F}$, then use maxima

Sub-Gaussian Processes

Definition

Any collection $\{X_t\}_{t \in T}$ of $\mathbb{R}$-valued random variables is a *stochastic process*.

- we always assume the process is *separable*, so there exists a countable $T' \subset T$ such that

$$\sup_{s,t \in T'} |X_t - X_s| = \sup_{s,t \in T} |X_t - X_s|$$

Definition

The process $\{X_t\}_{t \in T}$ is a *sub-Gaussian-process* for a metric ρ on T if

$$\mathbb{E}[\exp(\lambda(X_s - X_t))] \leq \exp\left(\frac{\lambda^2 \rho(s,t)^2}{2}\right) \quad \text{for } \lambda \in \mathbb{R},\ s,t \in T.$$

Examples of sub-Gaussian processes

Example (Gaussian process)

Let $T = \mathbb{R}^d$ and $Z \sim \mathcal{N}(0, I_{d\times d})$. Then $X_t := \langle Z, t\rangle$ satisfies

$$\mathbb{E}[\exp(\lambda(X_s - X_t))] = \mathbb{E}\left[\exp(\lambda\langle Z, s-t\rangle)\right] = \exp\left(\frac{\lambda^2 \left\|s-t\right\|_2^2}{2}\right)$$

so $\rho(s,t) = \left\|s-t\right\|_2$

Example (Rademacher processes)

Let $T \subset \mathbb{V}$, vector space with norm $\left\|\cdot\right\|$, and $\ell : T \times \mathcal{X} \to \mathbb{R}$ be $M(x)$-Lipschitz in its first argument. For $x_1^n \in \mathcal{X}^n$ and $\varepsilon_i \overset{\text{iid}}{\sim} \{\pm 1\}$,

$$Z_t := \sum_{i=1}^n \varepsilon_i \ell(t, x_i)$$

is sub-Gaussian with $\rho(s,t)^2 = \sum_{i=1}^n M(x_i)^2 \left\|s-t\right\|^2$.

Another example: the symmetrized process

Example (Symmetrized process)

Fix x_1^n. The process $f \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(x_i)$ is $\|\cdot\|_{L^2(P_n)}$ sub-Gaussian.

Chaining

- saw covering numbers allowed "one" discretization
- chaining: a multi-scale (all scales) discretization of set

Let $\{X_t\}_{t \in T}$ be a mean-zero separable ρ-sub-Gaussian process

Idea: approximate $\sup_{t \in T} X_t$ by finer and finer approximations

- let $\operatorname{diam}(T) = \sup_{s,t \in T} \rho(s,t)$
- take increasing sequence of covers

$$T_0 \subset T_1 \subset T_2 \subset \cdots \subset T$$

- T_k is the minimal $2^{-k} \operatorname{diam}(T)$ cover of T

Chaining sequences

- assume w.l.o.g. that T is finite
- for $t \in T$ define

$$\pi_i(t) := \operatorname*{argmin}_{t_i \in T_i} \rho(t_i, t)$$

- for fixed $k \in \mathbb{N}$ also define composed "projection"

$$\pi^{(i)}(t) := \pi_i \circ \pi_{i+1} \circ \cdots \circ \pi_{k-1}(t)$$

Observation

For any k and $t \in T_k$, we have

$$X_t = \sum_{i=1}^{k} (X_{\pi^{(i)}(t)} - X_{\pi^{(i-1)}(t)}) + X_{\pi^{(0)}(t)}$$

A little counting

$$\max_{t \in T_k} X_t \le \sum_{i=1}^{k} \max_{t \in T_k} \underbrace{\left(X_{\pi^{(i)}(t)} - X_{\pi^{(i-1)}(t)} \right)}_{\text{How many points?}} + X_{t_0}$$

Lemma

For $D = \operatorname{diam}(T)$ *and* $k \in \mathbb{N}$,

$$\mathbb{E}\left[\max_{t \in T_k} X_t\right] \le \sum_{i=1}^{k} \sqrt{8 \cdot 2^{-2i} D^2 \log N(T, \rho, 2^{-i} D)}$$

Dudley's entropy integral

Theorem (Dudley)

For any ρ-sub-Gaussian process $\{X_t\}_{t\in T}$,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le C \int_0^\infty \sqrt{\log N(T,\rho,u)}\,du.$$

A refined entropy integral bound

Corollary

For any ρ-sub-Gaussian process $\{X_t\}_{t\in T}$ and $\delta > 0$,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq C\left(\mathbb{E}\left[\sup_{\rho(s,t)\leq\delta}(X_t - X_s)\right] + J(\delta, T)\right)$$

where

$$J(\delta, T) := \int_{\delta}^{\infty} \sqrt{\log N(T, \rho, u)}du$$

Absolute values in suprema

- need to recenter process to work out
- for any $t_0 \in T$, obtain

$$\mathbb{E}[\sup_{t \in T} |X_t|] \leq \mathbb{E}[\sup_{t \in T} X_t] + \mathbb{E}[\sup_{t \in T}(-X_t)] + \mathbb{E}[|X_{t_0}|]$$

- for a symmetric process,

$$\mathbb{E}[\sup_{t \in T} |X_t|] \leq 2\mathbb{E}[\sup_{t \in T} X_t] + \inf_{t_0 \in T} \mathbb{E}[|X_{t_0}|].$$

Finite sample bounds for Lipschitz functions

- function class $\mathcal{F} = \{\ell(\theta, \cdot)\}_{\theta \in \Theta}$
- $t \mapsto \ell(t, x)$ is $M(x)$-Lipschitz
- know that $\log N(\Theta, \|\cdot\|_2, \epsilon) \lesssim d \log \frac{\mathrm{diam}(\Theta)}{\epsilon}$

Proposition

For this class,

$$\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \lesssim \frac{1}{\sqrt{n}} \mathrm{diam}(\Theta) \sqrt{PM^2} \sqrt{d}.$$

A uniform concentration bound for Lipschitz functions

- as in previous slide, except $M(x) \leq M < \infty$ for all x

Corollary

There exists a (numerical) constant C such that for all $t \geq 0$,

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |P_n \ell(\theta, X) - P\ell(\theta, X)| \geq CM \,\mathrm{diam}(\Theta) \sqrt{\frac{d+t}{n}}\right) \leq \exp(-t).$$

Comparison inequalities

- sometimes nice to compare expectation of complicated quantity to a simpler one
- e.g. compare function class $\phi \circ \mathcal{F} = \{\phi \circ f\}_{f \in \mathcal{F}}$ to $\mathcal{F}$

Example (Rademacher complexities of norm balls)

Say $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \le r\}$, $\mathcal{X} \subset \mathbb{R}^d$. Then function class $\mathcal{F} = \{f(x) = \theta^T x\}_{\theta \in \Theta}$ satisfies

$$\frac{1}{n} R_n(\mathcal{F} \mid x_1^n) \le \frac{r}{\sqrt{n}} \sqrt{\frac{1}{n} \sum_{i=1}^n \|x_i\|_2^2}$$

An ordering inequality

- mean-zero Gaussian vectors $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^n$, with

$$\mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2], \quad i = 1, \dots, n$$

  and disjoint index sets $A, B \subset [n] \times [n]$ with

$$\begin{aligned} \mathbb{E}[X_i X_j] &\le \mathbb{E}[Y_i Y_j] \ \text{ for } (i,j) \in A \\ \mathbb{E}[X_i X_j] &\ge \mathbb{E}[Y_i Y_j] \ \text{ for } (i,j) \in B \\ \mathbb{E}[X_i X_j] &= \mathbb{E}[Y_i Y_j] \ \text{ for } (i,j) \notin A \cup B \end{aligned}$$

Theorem

Let $F : \mathbb{R}^n \to \mathbb{R}$ be $\mathcal{C}^2$ with $\nabla_{ij}^2 F(x) \ge 0$ for all $(i,j) \in A$ and $\nabla_{ij}^2 F(x) \le 0$ for $(i,j) \in B$. Then

$$\mathbb{E}[F(X)] \le \mathbb{E}[F(Y)].$$

Slepian's inequality

Corollary (Slepian's inequality)

Let X, Y be mean-zero Gaussian vectors with $\mathbb{E}[(X_i - X_j)^2] \le \mathbb{E}[(Y_i - Y_j)^2]$ and $\mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2]$. Then

$$\mathbb{E}[\max_{i \le n} X_i] \le \mathbb{E}[\max_{i \le n} Y_i].$$

Proof of Gaussian ordering inequality

- Starting point: rotating Gaussians,

$$Z(\theta) := X\cos\theta + Y\sin\theta$$

  so $Z(0) = X$ and $Z(\pi/2) = Y$
- show that function

$$h(\theta) := \mathbb{E}[F(Z(\theta))]$$

  satisfies $h'(\theta) \geq 0$, all $\theta \in [0, \pi/2]$
- notation: $X \sim \mathcal{N}(0, \Sigma)$ and $Y \sim \mathcal{N}(0, \Gamma)$,
  $\dot{Z}(\theta) = -X\sin\theta + Y\cos\theta$, $\rho_{ij}(\theta) = (\sin\theta\cos\theta)(\Gamma_{ij} - \Sigma_{ij})$

$$\begin{bmatrix} Z_i(\theta) \\ \dot{Z}_j(\theta) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_{ii} & \rho_{ij}(\theta) \\ \rho_{ij}(\theta) & \Sigma_{jj} \end{bmatrix}\right)$$

Proof of Gaussian ordering inequality continued

Lemma

If

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma^2 & \rho \\ \rho & \sigma^2 \end{bmatrix}\right)$$

then $V_2 = \frac{\rho}{\sigma^2} V_1 + W$ *for* $W \sim \mathcal{N}(0, (1 - \frac{\rho^2}{\sigma^4})\sigma^2)$.

Finalizing proof of Gaussian ordering inequality

Lemma

For random vectors $U(i) \in \mathbb{R}^n$ that may depend on θ and Z,

$$h'(\theta) = \sum_{j=1}^{n} \sum_{i=1}^{n} \mathbb{E}\left[\nabla_{ij}^2 F(U(i)) \frac{\rho_{ij}(\theta)}{\Sigma_{ii}} \dot{Z}_j(\theta)^2\right]$$

Gaussian contraction

Theorem (Sudakov-Fernique)

Let X, Y be mean-zero Gaussian vectors with $\mathbb{E}[(X_i - X_j)^2] \le \mathbb{E}[(Y_i - Y_j)^2]$ for all i, j. Then

$$\mathbb{E}[\max_{i \le n} X_i] \le \mathbb{E}[\max_{i \le n} Y_i].$$

Example (Gaussian complexity)

Gaussian complexity of a set $T \subset \mathbb{R}^n$ is

$$G_n(T) := \mathbb{E}\left[\sup_{t \in T} \langle t, g \rangle\right] \quad \text{for } g \sim \mathcal{N}(0, I_n).$$

Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be non-expansive. Then for $\phi(t) = (\phi_i(t_i))_{i=1}^n$

$$G_n(\phi(T)) \le G_n(T)$$

Rademacher contraction

Theorem (Ledoux-Talagrand contraction)

For a bounded set $T \subset \mathbb{R}^n$ with Rademacher complexity

$$R_n(T) := \mathbb{E}\left[\sup_{t \in T} |\langle \varepsilon, t \rangle|\right],$$

if $\phi_i : \mathbb{R} \to \mathbb{R}$ are M-Lipschitz and $\phi_i(0) = 0$, then

$$\mathbb{E}\left[\sup_{t \in T} |\langle \phi(t), \varepsilon \rangle|\right] \leq 2MR_n(T).$$

- some consequences in exercises
- important in generalization guarantees for machine learning