<u>Topics in Statistical Learning Theory</u>

Summarized by Ido Greenberg in 2021, based on various sources as mentioned below.

Contents

Background

PAC-learning & VC-dimension: see a brief summary here (P.13-14).

Glivenko-Cantelli Theorem: see here (P.7). Note that the proof relies on symmetrization argument.

Symmetrization

- ***Symmetrization lemma***: let ϕ convex, $E[Z]=0$, $\epsilon \sim unif\{\pm 1\}$ independently of Z, then:

$$\boldsymbol{E\phi(Z) \le E\phi(2\epsilon Z)}$$

  - I guess it's called symmetrization because ϵZ has a symmetric distribution.
  - That's a useful technique for proving various inequalities in statistics (e.g. bounding Orlicz norm of a sum using Orlicz norms of the elements; and VC-dimension bounds [1,2]).
- Examples:
  - $\phi(Z)=Z$ ➔ $0 \le 0$
  - $\phi(Z)=Z^2$ ➔ $Var(Z) \le Var(2Z)$
- The proof relies on 2 independent copies of Z, Jensen inequality and convexity of ϕ:

$$E_{Z_1}\phi(Z_1) = E_{Z_1}\phi\big(Z_1 - E_{Z_2}Z_2\big) \le E_{Z_1}E_{Z_2}\phi(Z_1 - Z_2) = E_\epsilon E_{Z_1}E_{Z_2}\phi\big(\epsilon(Z_1 - Z_2)\big)$$
$$\le E_\epsilon E_{Z_1}E_{Z_2}\frac{1}{2}\big(\phi(2\epsilon Z_1) + \phi(2\epsilon Z_2)\big) = E\phi(2\epsilon Z)$$

- Generalized ***Symmetrization Theorem***: for i.i.d $\{X_i\}$,$\epsilon_i \sim unif\{\pm 1\}$, and global $C_{1,2}$:

$$\boldsymbol{E}\sup_{\mathrm{f}\in\mathrm{F}}\left|\frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}}\boldsymbol{f}(\boldsymbol{X}_{\boldsymbol{i}}) - \boldsymbol{Ef}\right| \le C_1\boldsymbol{E}\sup_{\mathrm{f}\in\mathrm{F}}\left|\frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}}\boldsymbol{\epsilon_i f}(\boldsymbol{X}_{\boldsymbol{i}})\right| \le C_2 E\sup_{\mathrm{f}\in\mathrm{F}}\left|\frac{1}{N}\sum_{i=1}^{N}f(X_i) - Ef\right| + \frac{\sup_{\mathrm{f}\in\mathrm{F}}|Ef|}{\sqrt{N}}$$

  - I.e. the expected deviation $\left|\overline{f(X)} - Ef\right|$ remains similar after symmetrization $\left|\overline{\epsilon f(X)}\right|$.
- Another variant of the symmetrization theorem bounds the probability $Pr\left(\left|\overline{f(X)} - Ef\right| > \rho\right)$.

Rademacher Complexity

- Main source: [lecture notes of Clayton Scott by Deng & Moon]
- VC-dimension quantifies the expressiveness of a class of hypotheses of binary functions. Rademacher generalizes this notion for real-valued functions.
- Definition: $\boldsymbol{\tilde{R}_Z(G) \coloneqq E_\sigma\left[\sup_{g\in G}\frac{1}{n}\sum_{i=1}^n \sigma_i g(Z_i)\right]}$, $R_n(G) \coloneqq E_Z[\tilde{R}_Z(G)]$
  - These are ***Empirical Rademacher Complexity*** and ***Rademacher Complexity***, respectively.
  - G is a class of hypotheses (bounded functions); σ_i are iid $unif\{\pm 1\}$; Z are data samples.
- Interpretations:
  - G can fit different sign combinations of σ (to achieve large value, $g(Z_i)$ has to be very positive if $\sigma_i = 1$ and very negative if -1).
  - G can fit different directions of the vector σ.
- ***Rademacher complexity bound***: W.p. $1-\delta$:

$$\forall g \in G: \quad \boldsymbol{E[g(Z)] \le \frac{1}{n}\sum_{i=1}^n g(Z_i) + 2R_n(G) + B\sqrt{\frac{\log 1/\delta}{2n}}}$$

  - Z_i are iid data samples (the probability $1-\delta$ is wrt them); B is the uniform bound on G.
  - The proof relies on a symmetrization argument.
  - Interpretation: the expected loss is probably close to the empirical loss, up to the Rademacher complexity; if G is very expressive, we may choose g that overfits, but then the small empirical error does not necessarily represent the expected error.
  - Similar versions exist for $\tilde{R}_Z$ (instead of R_n) and for two-sided bound.

Chaining

- Main sources: [David Pollard, Yale]; [Talagrand]; [Rakhlin]
- *Stochastic process*: $X = \{X_t\}_{t\in T}$
  - Example: $Z \sim N(0, I_d)$, $T = R^d$, $X_t = Z^\top t$.
- *Process control*:
  - Relies on assumptions about the increments of the process, e.g. $\left||X_s - X_t|\right| \le C\left||s-t|\right|$ or $P\{\left||X_s - X_t|\right| \ge \eta\left||s-t|\right|\} \le \beta(\eta)$.
    - E.g. ***sub-Gaussian process***: $\forall \lambda \in R, \forall t,s \in T$: $E\left[e^{\lambda(X_t - X_s)}\right] \le e^{\frac{\lambda^2\left||t-s|\right|^2}{2}}$
      - Equivalent definition: $\forall t,s \in T$: $X_t - X_s \sim subG\left(\left||t-s|\right|^2\right)$
      - For convenience, the process is usually assumed to be centered $EX_t = 0$.
  - Aims to find global tail-bounds, e.g. on $\sup_{t\in T}|X_t|$ or $OSC(\delta, X, T) \coloneqq \sup_{|s-t|<\delta}|X_s - X_t|$ (*oscillation*).
    - Note: for a symmetric process $E\sup|X_t - X_s| = 2E\sup X_t$ so both goals are essentially the same.
    - Also note that $\forall t_0$: $E\sup X_t = E\sup X_t - X_{t_0}$ (since $EX_{t_0} = 0$), which is sometimes more convenient to work with.
- A common approach to prove global bounds over an infinite set T:

- (1) Prove for finite subsets $T_n \subset T$; (2) take the limit $n \to \infty$ for a countable subset that is dense in T; (3) generalize the bound for T itself.
- For (1) to be effective, the bound must not diverge when $n \to \infty$. For example, naively taking the union bound $P\{\max|X_t| > \eta\} \le \sum P\{X_t > \eta\}$ would usually diverge with n. However, the union bound is clearly sub-optimal for positively-correlated variables.

- ***Chaining***:
  - We would like to find a subset $T_1 \subset T$ that is rather uncorrelated (hence a union bound is effective) and that covers T reasonably, so that a good mapping $\pi_1: T \to T_1$ would allow us to bound $X_t - X_{t_0} = \left(X_t - X_{\pi_1(t)}\right) + \left(X_{\pi_1(t)} - X_{t_0}\right)$ effectively.
  - More generally, we consider the subsets $\{t_0\} = T_0 \subset T_1 \subset T_2 \subset \cdots$, which decompose $X_t - X_{t_0}$ into increments along the ***chain*** $\{\pi_n\}_n$: $X_t - X_{t_0} = \sum_{n\ge0} X_{\pi_{n+1}(t)} - X_{\pi_n(t)}$ (the equality holds as is only if $\pi_n(t) = t$ for sufficiently large n).
  - We constraint $|T_n| = N_n = 2^{2^n}$ (except for $|T_0| = 1$). Note that $\sqrt{\log N_n} = 2^{n/2}$ ($\sqrt{\log x}$ will arise later as the inverse of e^{x^2}). Also $N_n^2 \le N_{n+1}$.
  - One can show that for a sub-Gaussian process,

$$\boldsymbol{P\left\{\sup_{t\in T}\left|X_t - X_{t_0}\right| > uS\right\} \le Ce^{-u^2/2}} \qquad\qquad \boldsymbol{E\sup X_t \le C\cdot S}$$

    - $S \coloneqq \sup_t \sum_{n\ge0} 2^{\frac{n+1}{2}}|\pi_{n+1}(t) - \pi_n(t)| \le 3\sup_t \sum_{n\ge0} 2^{\frac{n}{2}} d(t, T_n)$
  - ***Dudley's entropy bound***: as this holds for any chain $T_0 \subset T_1 \subset T_2 \subset \cdots$ (assuming $|T_n| \le N_n$), we obtain the bound: $\quad E\sup_t X_t \le C\sum_{n\ge0} 2^{n/2} \inf_{T_n\subset T}\sup_t d(t, T_n)$

- ***Entropy Integral***: $\quad \boldsymbol{J(D) \coloneqq \int_0^D \sqrt{\log N(\epsilon; T, \rho)}\, d\epsilon}$
  - N is the covering number of the set T by ϵ-balls wrt the metric ρ.
  - $\log N$ is also called the *metric entropy* of (T, ρ), not sure why (I guess N is kind of the number of bits in T up to ϵ-resolution, but then why log?).

- [***Dudley's Theorem***]: $\{X_t\}\sim subG$ ➔ $\boldsymbol{E\left[\sup_{t\in T} X_t\right] \lesssim J(\infty)}$ (similarly: $E\left[\sup_{t,s\in T}(X_t - X_s)\right] \lesssim J(\infty)$).
  - This bounds a process using only its subG property and the geometry of its indices.
  - Note: for a meaningful bound the integral must be finite. We usually assume that T is bounded, hence $J(\infty) = J\big(diam(T)\big)$.

- [Application – Rademacher]:
  - $\forall f, g \in G:\ \ \rho_n(f,g) \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^n\big(f(x_i) - g(x_i)\big)^2}; \qquad \forall f \in G:\ X_f \coloneqq \frac{1}{\sqrt{n}}\sum\sigma_i f(x_i)$
  - $\{X_f\}_{f\in G}$ is clearly a sub-Gaussian process, thus (for a corresponding D):

$$\boldsymbol{R_n(G) = E\left[\sup_{f\in G}\frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right] \lesssim \frac{1}{\sqrt{n}}\int_0^D \sqrt{\log N(\epsilon; G, \rho_n)}\, d\epsilon}$$

  - Using another bound on N, Rademacher complexity can be further bounded by $\lesssim \boldsymbol{\sqrt{\nu/n}}$, where ν is the **VC-dimension** of the domain X of the function-class G.

Martingales

- Main source: James Aspnes, Peter Morters
- ***Martingale***: a stochastic process $\{X_t\}_{t\in N}$ with $E[X_{t+1}|X_1 \dots X_t] = X_t$.
  - Example: share price in an efficient market (or any random walk).
  - Note: the definition is local (X_t vs. X_{t+1}), but the consequences on the process are global.
  - The conditioned variables $X_1 \dots X_t$ are often replaced with observed information F_t named ***filtration***.
- By induction: $\forall k \geq 1$: $\boldsymbol{E[X_{t+k}|X_1 \dots X_t] = X_t}$ (in particular: $\forall t$: $E[X_t] = E[X_0]$)
  - Example: let X_t random walk; then (by direct calculation) $Y_t := X_t^2 - t$ is martingale, hence $E[X_t^2] = E[Y_t] + t = E[Y_0] + t = t$ (which is indeed the variance of a r.w.).
- A martingale (with $E[X_0] = 0$) as a **sum of uncorrelated random variables $\boldsymbol{\Delta_t := X_t - X_{t-1}}$**:
  - $E[\Delta_{t+1}|\Delta_1 \dots \Delta_t] = E[X_{t+1} - X_t|X_1 \dots X_t] = 0$ ➔ $\{\Delta_t\}$ are uncorrelated.
  - In particular: $Var(X_t) = \sum_{s\leq t} E[\Delta_s^2]$.
  - ***Azuma-Hoeffding inequality***: $|\Delta_t| \leq c_t$ a.s. ➔ $\boldsymbol{P(|X_t| \geq \epsilon) \leq 2e^{-\frac{\epsilon^2}{2\sum_t c_t^2}}}$
- **Stopping time**:
  - The martingale property does not hold in general if t is replaced by a random variable T.
    - Example: T is the first step with $X_t > 1$ (thus clearly $E[X_T] \geq 1 > 0 = E[X_0]$).
    - Example: stopping time of the (infinite) double-or-nothing strategy.
  - ***Optional Stopping Theorem***: $E[X_T] = E[X_0]$ for a martingale $\{X_t\}$, if (1) $P(T < \infty) = 1$; (2) $E[|X_T|] < \infty$; and (3) $\lim_{t\to\infty} E[X_t \cdot \chi_{T>t}] = 0$.
- ***Sub-martingale***: $X_t \leq E[X_{t+1}|X_1 \dots X_t]$ (hence $X_t \leq E[X_{t+k}|X_1 \dots X_t]$ and $E[X_0] \leq E[X_t]$).
  - Terminology: sub = current value is below future expectation = increasing expectations.
  - Azuma-Hoeffding inequality holds in a one-sided variant ($P(X_t \leq -\epsilon) \leq \cdots$).
  - ***Super-martingale***: same with opposite inequalities.