Theoretical Stat/Machine Learning (Part-II)

1. Metric Entropy

Covering number: (T,d) metric space, $K \subset T$

$$\mathcal{N}(K,d,\varepsilon) = \min\{|\mathcal{N}| : \mathcal{N} \subset K \text{ is an } \varepsilon\text{-net of } K\}$$

Metric entropy: of K at scale ε,

$$H(K,d,\varepsilon) = \log \mathcal{N}(K,d,\varepsilon)$$

Packing number: (T,d) metric space, $K \subset T$

$$\mathcal{P}(K,d,\varepsilon) = \max\{|\mathcal{M}| : \mathcal{M} \subset K \text{ is } \varepsilon\text{-separated}\}$$

Covering vs Packing: Any $K \subset T$ and $\varepsilon > 0$,

$$\mathcal{P}(K,d,2\varepsilon) \leqslant \mathcal{N}(K,d,\varepsilon) \leqslant \mathcal{P}(K,d,\varepsilon)$$

ε-neighborhood of A:

$$A + \varepsilon B_2^n = \{x \ : \ d(x,A) \leqslant \varepsilon\}$$

Covering and Packing vs Volumes:

$$\frac{\mathrm{vol}(K)}{\mathrm{vol}(\varepsilon B_2^n)} \leqslant \mathcal{N}(K,\varepsilon) \leqslant \mathcal{P}(K,\varepsilon) \leqslant \frac{\mathrm{vol}(K + (\varepsilon/2)B_2^n)}{\mathrm{vol}((\varepsilon/2)B_2^n)}$$

Covering numbers of Euclidean ball:

$$\left(\frac{1}{\varepsilon}\right)^n \leqslant \mathcal{N}(B_2^n,\varepsilon) \leqslant \left(1 + \frac{2}{\varepsilon}\right)^n$$

Same upper bound holds for **sphere** S^{n-1}, and for $\varepsilon \in (0,1]$,

$$n\log(1/\varepsilon) \leqslant \log \mathcal{N}(B_2^n,\varepsilon) \leqslant n\log(3/\varepsilon)$$

Convex Hull: $T \subset \mathbb{R}^n$

$$\mathrm{conv}(T) = \left\{\sum_{i=1}^m \lambda_i t_i : m \in \mathbb{N}, t_i \in T, \lambda_i \geqslant 0, \sum_{i=1}^m \lambda_i = 1\right\}$$

Entropy of convex hulls: $P = \mathrm{conv}(v_1,\dots,v_n) \subset B_2^n$

$$H(P,\varepsilon) = \log \mathcal{N}(P,\varepsilon) \leqslant C\frac{\log N}{\varepsilon^2}$$

2. Covariance Estimation

Operator norm net reduction: $A \in \mathbb{R}^{m\times n}$, $\mathcal{N} \subset S^{n-1}$

$$\|A\|_{\mathrm{op}} \leqslant \frac{1}{1-\varepsilon}\sup_{x\in\mathcal{N}}\|Ax\|_2, \quad \varepsilon \in (0,1)$$

Symmetric quadratic-form net reduction:

$$\|A\|_{\mathrm{op}} \leqslant \frac{1}{1-2\varepsilon}\sup_{x\in\mathcal{N}}|x^\top Ax|, \quad \varepsilon \in (0,1/2)$$

Isotropic sub-Gaussian data: $X \in R^d$ is isotropic if

$$\mathbb{E}X = 0 \text{ and } \mathbb{E}[XX^\top] = I_d \implies \mathbb{E}\langle X,u\rangle^2 = 1, \ \forall u \in S^{d-1}$$

Sub-Gaussian norm for a vector: $X \in \mathbb{R}^d$,

$$\|X\|_{\psi_2} = \sup_{u\in S^{d-1}}\|\langle X,u\rangle\|_{\psi_2}$$

Useful result: $X_1,\dots,X_d$ are independent, mean-zero, sub-Gaussian r.v., and let $X = (X_1,\dots,X_d) \in \mathbb{R}^d$,

$$\max_{1\leqslant i\leqslant d}\|X_i\|_{\psi_2} \ \leqslant \ \|X\|_{\psi_2} \ \leqslant \ C\max_{1\leqslant i\leqslant d}\|X_i\|_{\psi_2}.$$

ψ_2-norm vs spectral norm: $X \sim \mathcal{N}(0,\Sigma) \in \mathbb{R}^d$, where $\Sigma \succeq 0$,

$$\|X\|_{\psi_2} \leqslant C\sqrt{\|\Sigma\|_{\mathrm{op}}}.$$

Sample Covariance for Isotropic Sub-Gaussian: $s \geqslant 0$,

$$\mathbb{P}\left\{\left\|\widehat{\Sigma} - I_d\right\|_{\mathrm{op}} \geqslant CK^2\left(\sqrt{\frac{d+s}{N}} + \frac{d+s}{N}\right)\right\} \leqslant 2e^{-cs}$$

Expectation bound:

$$\mathbb{E}\left\|\widehat{\Sigma} - I_d\right\|_{\mathrm{op}} \lesssim K^2\left(\sqrt{\frac{d}{N}} + \frac{d}{N}\right)$$

Equivalently, when $N \gg d$, the leading behavior is

$$\left\|\widehat{\Sigma} - I_d\right\|_{\mathrm{op}} \approx \sqrt{d/N}$$

General covariance by whitening: If $\Sigma \succ 0$, define

$$Y = \Sigma^{-1/2}X.$$

Then Y is isotropic with $\widehat{\Sigma}_Y = \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$. This gives

$$\widehat{\Sigma} - \Sigma = \Sigma^{1/2}(\widehat{\Sigma}_Y - I_d)\Sigma^{1/2},$$

and

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \leqslant \|\Sigma\|\,\|\widehat{\Sigma}_Y - I_d\|_{\mathrm{op}}$$

Hence, given $N \geqslant d$,

$$\mathbb{P}\left\{\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{op}} \geqslant CK^2\|\Sigma\|_{\mathrm{op}}\sqrt{\frac{d}{N}}\right\} \leqslant 2e^{-d}$$

Weyl: For symmetric $d\times d$ matrices A, B:

$$\max_{1\leqslant i\leqslant d}|\lambda_i(A) - \lambda_i(B)| \leqslant \|A - B\|_{\mathrm{op}}$$

Davis-Kahan: Let A, B be symmetric $d\times d$ matrices. Let P_A be the orthogonal projector onto $\mathrm{span}\{v_1(A),\dots,v_k(A)\}$. Assume there is an eigengap at k for A and that $\|A - B\| \leqslant \delta/2$. Let P_B be the projector onto $\mathrm{span}\{v_1(B),\dots,v_k(B)\}$. Then

$$\|P_A - P_B\|_{\mathrm{op}} \leqslant \frac{2\,\|A-B\|_{\mathrm{op}}}{\delta}$$

Projection matrices: $u, v \in S^{d-1}$ and define $P_u = uu^\top$, then there exists a sign $s \in \{-1,1\}$ such that

$$\frac{1}{2}\,\|u - sv\|_2 \ \leqslant \ \|P_u - P_v\|_{\mathrm{op}} \ \leqslant \ 2\,\|u - sv\|_2.$$

Perturbation of top eigenvectors: $A, B \in \mathbb{R}^{d\times d}$ symmetric matrices with $\delta := \lambda_1(A) - \lambda_2(A) > 0$. Then there exists a sign $s \in \{-1,1\}$ such that

$$\|v_1(A) - s\,v_1(B)\|_2 \ \leqslant \ C\,\frac{\|A-B\|_{\mathrm{op}}}{\delta}.$$

3. Random Processes

Random process: Let T be a nonempty set. A *random process* indexed by T is a family of real-valued random variables

$$(X_t)_{t\in T}.$$

Canonical pseudo-metric: Let $(X_t)_{t\in T}$ be a square-integrable random process. The *Canonical pseudo-metric* is:

$$d_X(s,t) := \|X_t - X_s\|_{L_2} = \left(\mathbb{E}|X_t - X_s|^2\right)^{1/2}, \qquad s,t \in T.$$

Gaussian process: A random process $(X_t)_{t\in T}$ is a *Gaussian process* if every finite subcollection has a multivariate Gaussian distribution, i.e.,

$$(X_{t_1},\dots,X_{t_m}) \sim \mathrm{MVN}(\,\cdot\,,\,\cdot\,)$$

Canonical Gaussian process on $\mathbb{R}^n$: Let $g \sim \mathcal{N}(0,I_n)$, and let $T \subset \mathbb{R}^n$. Define

$$X_t := \langle g,t\rangle, \qquad t \in T.$$

Then $(X_t)_{t\in T}$ is a centered Gaussian process, and

$$X_t - X_s = \langle g, t-s\rangle \sim \mathcal{N}(0,\|t-s\|_2^2).$$

Therefore, the canonical metric is

$$d_X(s,t) = \|t-s\|_2.$$

Finite maximal inequality: Let $Z_1,\dots,Z_M$ be centered random variables s.t. $\|Z_j\|_{\psi_2} \leqslant L$ for all j. Then

$$\mathbb{E}\max_{1\leqslant j\leqslant M} Z_j \leqslant CL\sqrt{\log M}$$

4. Gaussian Width

Definition: For a bounded set $T \subset \mathbb{R}^d$

$$w(T) := \mathbb{E}\sup_{t\in T}\langle g,t\rangle, \qquad g \sim \mathcal{N}(0,I_d).$$

Euclidean ball:

$$w(B_2^d) = \mathbb{E}\|g\|_2 \asymp \sqrt{d}.$$

ℓ_1-ball:

$$w(B_1^d) = \mathbb{E}\|g\|_\infty \asymp \sqrt{\log d}.$$

Gaussian tail upper bound:

$$G \sim \mathcal{N}(0,1) \implies \mathbb{P}\{|G| > t\} \geqslant c_0\frac{e^{-t^2/2}}{t}, \quad t \geqslant 1$$

Finite sets: If $T = \{t_1,\dots,t_M\}$, then

$$w(T) \leqslant \left(\max_{1\leqslant j\leqslant M}\|t_j\|_2\right)\sqrt{2\log M}$$

Sparse unit vectors: If $T_k := \{x \in S^{d-1} : \|x\|_0 \leq k\}$, then

$$w(T_k) \lesssim \sqrt{k\log(ed/k)}$$

5. Chaining

Dudley's discrete inequality:

$$\mathbb{E}\sup_{t\in T} X_t \leqslant CK\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}(T,d,2^{-k})}$$

Dudley's entropy integral inequality: Let $(X_t)_{t\in T}$ be a random process on metric space (T,d) s.t. $\|X_t - X_s\|_{\psi_2} \leqslant Kd(t,s)$ for all $s,t\in T$. Then for all fixed $t_0\in T$,

$$\mathbb{E}\sup_{t\in T}(X_t - X_{t_0}) \leqslant CK\int_0^{\operatorname{diam}(T,d)} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon$$

Supremum of Increments:

$$\mathbb{E}\sup_{s,t\in T}|X_t - X_s| \leqslant 2CK\int_0^{\operatorname{diam}(T,d)} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon$$

High-Probability Dudley Bound: Let $\mathrm{D_T} = \operatorname{diam}(T,d)$, then with probability at least $1-2e^{-cu^2}$

$$\sup_{s,t\in T}|X_t - X_s| \leqslant CK\left[\int_0^{\mathrm{D_T}} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon + u\cdot \mathrm{D_T}\right]$$

Sub-exponential increments: Suppose $(X_t)_{t\in T}$ satisfies

$$\|X_t - X_s\|_{\psi_1} \leqslant Kd(t,s) \qquad \text{for all } s,t\in T.$$

Then for every $t_0\in T$,

$$\mathbb{E}\sup_{t\in T}(X_t - X_{t_0}) \leqslant CK\int_0^{\operatorname{diam}(T,d)} \log \mathcal{N}(T,d,\varepsilon)\, d\varepsilon$$

Entropy integral bound for Gaussian width:

$$w(T) \leqslant C\int_0^{\operatorname{diam}(T,\|\cdot\|_2)} \sqrt{\log \mathcal{N}(T,\|\cdot\|_2,\varepsilon)}\, d\varepsilon$$

6. Empirical Processes

Definition: Given a class $\mathcal{F}$ of real-valued functions on Ω, define

$$Z_f := \mu_n(f) - \mu(f) = \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X), \qquad f\in\mathcal{F}.$$

The random family $(Z_f)_{f\in\mathcal{F}}$ is called the *empirical process* indexed by $\mathcal{F}$, and $\sup_{f\in\mathcal{F}}|Z_f|$ the uniform empirical error.

Uniform LLN for Lipschitz Functions: Let μ be any probability measure on $[0,1]$, and let $X_1,\dots,X_n$ be iid with law μ. For $L>0$, define the anchored Lipschitz class

$$\mathcal{F}_L = \{f:[0,1]\to\mathbb{R} : f(0)=0, \|f\|_{\mathrm{Lip}} \leqslant L\}.$$

Then

$$\mathbb{E}\sup_{f\in\mathcal{F}_L}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right| \leqslant \frac{CL}{\sqrt{n}}.$$

Covering Lipschitz functions: For $L>0$, define

$$\mathcal{F}_{L,d} := \left\{f:[0,1]^d\to[0,1] \;:\; \|f\|_{\mathrm{Lip}}\leqslant L\right\}.$$

Then

$$\log \mathcal{N}(\mathcal{F}_{L,1},\|\cdot\|_\infty,\varepsilon) \leqslant C\frac{L}{\varepsilon}, \qquad 0<\varepsilon\leqslant 1,$$

$$\log \mathcal{N}(\mathcal{F}_{L,d},\|\cdot\|_\infty,\varepsilon) \leqslant C_d\left(\frac{L}{\varepsilon}\right)^d, \qquad 0<\varepsilon\leqslant 1,$$

where C_d may depend on the ambient dimension d.

Giné-Zynn Symmetrization: Let $\mathcal{F}$ be a class of measurable real-valued functions on Ω, and let $\varepsilon_1,\dots,\varepsilon_n$ be iid Rademacher random variables, independent of everything else. Then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-\mathbb{E}f(X))\right| \leqslant 2\,\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|.$$

Empirical and expected Rademacher complexity:

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_\varepsilon\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| \implies \mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{X_{1:n}}\widehat{\mathfrak{R}}_n(\mathcal{F})$$

Empirical L_2 pseudometric:

$$\|f-g\|_{L_2(\mu_n)} := \left(\frac{1}{n}\sum_{i=1}^n \big(f(X_i)-g(X_i)\big)^2\right)^{1/2}.$$

Conditional chaining bound for Rademacher averages:

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) \leqslant \frac{C}{\sqrt{n}}\int_0^{\operatorname{diam}(\mathcal{F},L_2(\mu_n))} \sqrt{\log \mathcal{N}(\mathcal{F},L_2(\mu_n),\varepsilon)}\, d\varepsilon$$

Finite-class Massart-type bound: $\mathcal{F}$ is finite, and $\forall\, f\in\mathcal{F}$,

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)^2 \leqslant r^2.$$

Then

$$\widehat{\mathfrak{R}}_n(\mathcal{F};x_1,\dots,x_n) \leqslant r\sqrt{\frac{2\log|\mathcal{F}|}{n}}.$$

7. Boolean Classes/VC Bounds

Definition: Let $\mathcal{H}$ be a class of Boolean functions on Ω. Given sample points $x_1,\dots,x_n\in\Omega$, define the set of **traces of** $\mathcal{H}$ on the sample by

$$\mathcal{H}|_{x_{1:n}} := \{(h(x_1),\dots,h(x_n)) : h\in\mathcal{H}\} \subseteq \{0,1\}^n.$$

$\mathcal{H}|_{x_{1:n}}$ records all labelings that the class can realize on sample.

Growth function: For a class $\mathcal{H}$ of Boolean functions,

$$\Pi_{\mathcal{H}}(n) := \sup_{x_1,\dots,x_n\in\Omega}\left|\mathcal{H}|_{x_{1:n}}\right|.$$

Pajor Lemma: $\mathcal{H}$ is finite class of Boolean functions on Ω,

$$|\mathcal{H}| \leqslant \#\{\Lambda\subseteq\Omega : \Lambda \text{ is shattered by } \mathcal{H}\}$$

Sauer-Shelah Lemma: Let $\mathrm{VC}(\mathcal{H}) = d < \infty$,

$$\Pi_{\mathcal{H}}(n) \leqslant \sum_{k=0}^d \binom{n}{k} \leqslant \left(\frac{en}{d}\right)^d \quad (n\geqslant d)$$

Preliminary VC bound for uniform empirical error: Let $\mathcal{H}$ be a class of Boolean functions on Ω, and $\mathrm{VC}(\mathcal{H}) = d$. Let $X_1,\dots,X_n$ be i.i.d. with law μ, and assume $n\geq d$. Then

$$\mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n h(X_i) - \mathbb{E}h(X)\right| \leqslant C\sqrt{\frac{d\log(en/d)}{n}},$$

Good to remember: Cond. on the sample $X_{1:n}$, the process

$$R_h := \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(X_i), \qquad h\in\mathcal{H},$$

has sub-Gaussian increments with respect to

$$\frac{1}{\sqrt{n}}\|h-g\|_{L_\infty(\mu_n)}.$$

8. Statistical Learning

Population and empirical risk:

$$R(h) := \mathbb{E}[\ell(h(X),Y)], \qquad R_n(h) := \frac{1}{n}\sum_{i=1}^n \ell(h(X_i),Y_i)$$

Population and empirical risk minimizer:

$$h^\star \in \arg\min_{h\in\mathcal{H}} R(h), \qquad \widehat{h}_n \in \arg\min_{h\in\mathcal{H}} R_n(h)$$

Excess Risk bound: For any hypothesis class $\mathcal{H}$,

$$R(\widehat{h}_n) - R(h^\star) \leqslant 2\sup_{h\in\mathcal{H}}|R_n(h) - R(h)|$$

Finite-class generalization bound: For every $\delta\in(0,1)$, with probability at least $1-\delta$,

$$R(\widehat{h}_n) - R(h^\star) \leqslant 2\sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2n}}$$

and using the tail integral identity gives

$$\mathbb{E}\big[R(\widehat{h}_n) - R(h^\star)\big] \lesssim \sqrt{\frac{\log|\mathcal{H}|}{n}}.$$

Loss class has the same trace complexity: The loss class:

$$\mathcal{L} := \left\{(x,y)\mapsto \mathbf{1}_{\{h(x)\neq y\}} : h\in\mathcal{H}\right\}.$$

Then for every sample $(x_1,y_1),\dots,(x_n,y_n)$,

$$\left|\mathcal{L}|_{(x_i,y_i)_{i=1}^n}\right| = \left|\mathcal{H}|_{x_{1:n}}\right|.$$

Consequently,

$$\Pi_{\mathcal{L}}(n) \leq \Pi_{\mathcal{H}}(n) \qquad \text{and} \qquad \mathrm{VC}(\mathcal{L}) \leq \mathrm{VC}(\mathcal{H}).$$

VC Generalization Bound: Assume $n\geqslant d$,

$$\mathbb{E}\big[R(\widehat{h}_n) - R(h^\star)\big] \lesssim \sqrt{\frac{\mathrm{VC}(\mathcal{L})\log(en/\mathrm{VC}(\mathcal{L}))}{n}}.$$

One-dimensional Lipschitz regression: with squared loss

$$\mathbb{E}\big[R(\widehat{h}_n) - R(h^\star)\big] \lesssim \sqrt{L/n}$$

Higher-dimensional Lipschitz regression: for all $\varepsilon\in(0,1)$

$$\mathbb{E}[R(\widehat{h}_n) - R(h^\star)] \lesssim \varepsilon + \sqrt{\frac{\log \mathcal{N}(\mathcal{F}_{L,d},\|\cdot\|_\infty,\varepsilon)}{n}}.$$