深度学习理论初探

Deep Learning Theory

generalization
approximation
optimization !

一个重要特征：generalization theory
intertwined with optimization theory !

深度学习理论初探 Exploring Deep Learning Theory

Note：
1. 这是一个理论系列!!!
2. 该系列和"机器学习理论基础(FML)"系列相互独立
3. 部分内容(可能)已经过时，(甚至)被认为是错误的。

Part 1. 经典统计机器学习理论

ERM. Concentration. Uniform Convergence. Rademacher Complexity. Margin ...

Part 2 泛化理论：从线性模型到神经网络

Linear Models. 2-layer NN. Deep Neural Networks. PAC Bayes...

Part 3 逼近理论：神经网络与万有逼近

Universal Approximation. Barron's Theorem. Neural Tangent Kernel ...

Part 4 优化理论：非凸优化和隐式正则化

Non-convex Optimization. Implicit Regularization. Landscape Analysis ...

其他话题：

Double Descent. Benign Overfitting. Mean-Field Perspective. LLM theory ...

References.

Matus Telgarsky. Deep Learning Theory Lecture Notes

Anthony & Bartlett. Neural Network Learning: Theoretical Foundations.

Sanjeev Arora. Theory of Deep Learning

Simons Prince. Understanding Deep Learning.

Stanford CS229M / STATS214, Tengyu Ma

UCB CS281B / STAT 241B, Peter Bartlett.

Part 1 Statistical Learning Theory

Outline:

1) Supervised Learning Settings
2) Asymptotic Analysis
3) Concentration Inequality
4) Uniform Convergence
5) Rademacher Complexity
6) VC Dimension

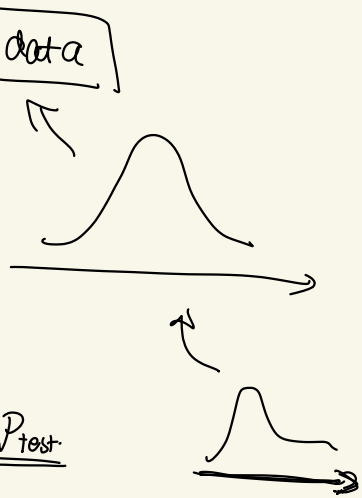
1.1. Supervised Learning Settings

$\mathcal{X}$: input space

$\mathcal{Y}$: output space

$\Rightarrow \mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$

distribution (unknown)



① 数据从哪来?

★ i.i.d 独立同分布 $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

"domain shift" $\mathbb{P}_{train} \neq \mathbb{P}_{test}$

Goal: learn $h: \mathcal{X} \to \mathcal{Y}$ (predictor / hypothesis / model)

$$x \longrightarrow \boxed{h} \longrightarrow \underset{\text{prediction}}{h(x)} \sim \underset{\text{true lable}}{y}$$

② 如何评估 h?

loss function $\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ $\quad \ell(\hat{y}, y) \geq 0$

e.g. Regression $\mathcal{Y} = \mathbb{R}$ $\quad \ell(\hat{y}, y) = (\hat{y} - y)^2$

Classification $\mathcal{Y} = \{-1, +1\}$ $\quad \ell(\hat{y}, y) = \mathbb{1}\{\hat{y} \neq y\} = \mathbb{1}\{\underbrace{y \cdot \hat{y}}_{\text{confidence margin}} \leq 0\}$

Risk: $\boxed{L(h) = \mathbb{E}_{(x,y)\sim \mathbb{P}}[\ell(h(x), y)]}$

或 $R(h)$ （population loss / expected loss）

③ h从哪里来？ $h \in \mathcal{H}$ hypothesis space

parametric.

e.g. $\mathcal{H} = \{ h : x \mapsto w^T x + b \}$ → w, b → Θ

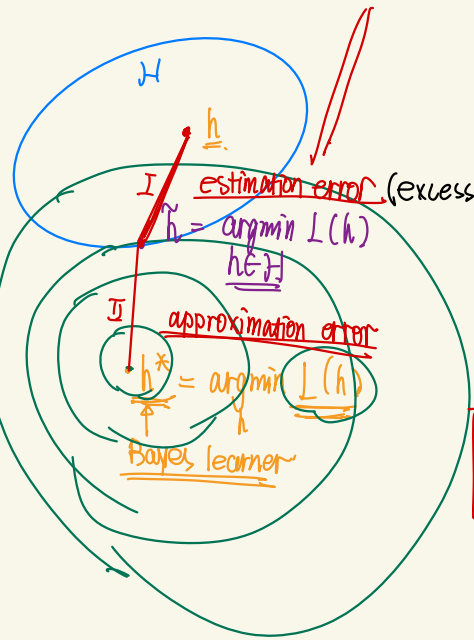
$\mathcal{H} = \{ h : h$ 是二叉决策树 $\}$

non-parametric.

④ h怎么学得？给定假设空间 $\mathcal{H}$ 后

def (excess risk) $E(h) = \boxed{L(h)} - \inf_{g \in \mathcal{H}} L(g)$

minimize



(excess risk) $L(h) = \mathbb{E}_{(x,y)\sim P}[\ell(h(x), y)]$ P unknown

sample: $\{(x_i, y_i)\}_{i=1}^n$

$\boxed{\hat{L}(h) = \frac{1}{n}\sum_{i=1}^n \ell(h(x_i), y_i)}$: empirical risk

h_θ $\quad \hookrightarrow \frac{1}{n}\sum_{i=1}^n \ell((x^{(i)}, y^{(i)}), \theta)$

empirical distribution: $\hat{P} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{(x,y) = (x^{(i)}, y^{(i)})\} \longrightarrow P$

minimize $\boxed{\hat{L}(h_\theta) = \mathbb{E}_{(x,y)\sim\hat{P}}[\ell((x^{(i)}, y^{(i)}), \theta)]} \longrightarrow L(h_\theta)$

ERM (empirical risk minimization 经验风险最小化)

$$\hat{\theta}_{ERM} = \arg\min_{\theta \in \Theta} \hat{L}(h_\theta)$$

generalization theorey : $\overset{P}{L(h_\theta)} \sim \overset{\hat{P}}{\hat{L}(h_\theta)}$

$\mathbb{E}[\hat{L}(h_\theta)]$

ERM is an unbiased estimator :

$$\mathbb{E}_{(x^{(i)}, y^{(i)}) \overset{i.i.d}{\sim} P}\left[\hat{L}(h_\theta)\right] = \mathbb{E}_P\left[\frac{1}{n}\sum_{i=1}^{n} \ell((x^{(i)}, y^{(i)}), \theta)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_P\left[\ell((x^{(i)}, y^{(i)}), \theta)\right]$$

$$= \mathbb{E}_{(x,y)\sim P}\left[\ell((x^{(i)}, y^{(i)}), \theta)\right] = L(h_\theta)$$

population loss

empirical loss & population risk are equal in expectation (over randomness of the training set)

1.2. Asymptotic Analysis

What guarantees do we have on the excess risk for the parameters learned by ERM ?

$L(\hat{\theta}) - \inf_{\theta \in \Theta} L(\theta) \to$ bound

$L(\hat{\theta}) - L(\theta^*)$

渐近分析: $n \to \infty$

Takeaway: $L(\hat{\theta}) - \inf_{\theta \in \Theta} L(\theta) \leq \frac{c}{n} + o\left(\frac{1}{n}\right)$

training data: $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

hypothesis class: $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}^p\}$

$\hat{\theta}$: (ERM) $= \arg\min_\theta \hat{L}(\theta)$

$\theta^* = \arg\min L(\theta)$

population risk

Thm 1 Suppose. (a) $\hat{\theta} \xrightarrow{P} \theta^*$, $n \to \infty$

consistency

$\mathbb{P}(|\hat{\theta} - \theta^*| > \varepsilon) \to 0$

(b) $\nabla^2 L(\theta^*)$ is full rank

(c) appropriate regularity conditions . then :

1. $\sqrt{n}(\hat{\theta} - \theta^*) = O_p(1)$ "bounded in probability"

$\{\sqrt{n}(\hat{\theta} - \theta^*)\}_{n=1}^{\infty}$

$\forall \varepsilon, \exists M,$

$\sup_n \mathbb{P}(\|\sqrt{n}(\hat{\theta} - \theta^*)\|_2 > M) < \varepsilon$

2. $\sqrt{n}(\hat{\theta}-\theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-1} \boxed{Cov[\nabla \ell((x,y),\theta^*)]} (\nabla^2 L(\theta^*))^{-1}\right)$

3. $n(L(\hat{\theta})-L(\theta^*)) = O_p(1)$

4. $n(L(\hat{\theta})-L(\theta^*)) \xrightarrow{d} \frac{1}{2}\|S\|_2^2$, $S \sim \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-\frac{1}{2}} Cov(\nabla \ell((x,y),\theta^*)) (\nabla^2 L(\theta^*))^{-\frac{1}{2}}\right)$

5. $\lim_{n\to\infty} \mathbb{E}[n(L(\hat{\theta})-L(\theta^*))] = \frac{1}{2} tr\left((\nabla^2 L(\theta^*))^{-1} Cov(\nabla \ell((x,y),\theta^*))\right)$

1, 3 → order of convergence. $(\frac{1}{\sqrt{n}}, \frac{1}{n})$

2, 4 → convergence distribution (Gaussian)

$X_1, \cdots, X_n$ (i.i.d) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

↳ mean: $\mathbb{E}[X]$

variance: Σ

Lemma 1 (Law of Large Numbers) $\bar{X} \xrightarrow{p} \mathbb{E}[X]$

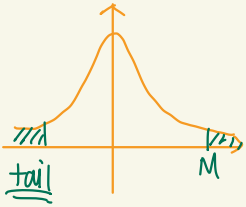
Lemma 2 (Central Limit Theorem) $\sqrt{n}(\bar{X}-\mathbb{E}[X]) \xrightarrow{d} \mathcal{N}(0,\Sigma)$

by LLN, $\hat{L}(\theta) \xrightarrow{p} L(\theta)$, $\nabla \hat{L}(\theta) \xrightarrow{p} \nabla L(\theta)$, $\nabla^2 \hat{L}(\theta) \xrightarrow{p} \nabla^2 L(\theta)$

by CLT, $\sqrt{n}(\bar{X}-\mathbb{E}[X]) = O_p(1)$

$\mathbb{P}(|\bar{X}-\mathbb{E}[X]| \geq a\sigma) \leq 2e^{-\frac{a^2}{2}} = \varepsilon$

↳ $2\sigma\sqrt{2\log\frac{2}{\varepsilon}} = M$



Lemma 3

1) $Z \sim \mathcal{N}(0,\Sigma)$ A deterministic matrix.

$\Rightarrow AZ \sim \mathcal{N}(0, A\Sigma A^T)$

2) $Z \sim \mathcal{N}(0,\Sigma^{-1}) \Rightarrow Z^T\Sigma Z \sim \chi^2(p)$ (Proof: $W = \Sigma^{\frac{1}{2}} Z \sim \mathcal{N}(0, I)$

$\Rightarrow Z^T\Sigma Z = W^T W = \sum_{i=1}^{p} W_i^2 \sim \chi^2(p)$)

Proof of Thm 1: By Taylor Expansion

$0 = \nabla \hat{L}(\hat{\theta}) = \nabla \hat{L}(\theta^*) + \nabla^2 \hat{L}(\theta^*) \cdot (\hat{\theta}-\theta^*) + O(\|\hat{\theta}-\theta^*\|_2^2)$

ERM: $\arg\min \hat{L}$

first-order optimality

$\Rightarrow \hat{\theta}-\theta^* = -(\nabla^2 \hat{L}(\theta^*))^{-1} \cdot \nabla \hat{L}(\theta^*)$ + h.o.term.

$$\Rightarrow \boxed{\sqrt{n}(\hat{\theta}-\theta^*) \doteq -\left(\nabla^2 \hat{L}(\theta^*)\right)^{-1} \sqrt{n}\, \nabla \hat{L}(\theta^*)}$$

Let $X_i = \nabla \ell\left((x^{(i)}, y^{(i)}), \theta^*\right)$ $\quad \bar{X} = \nabla \hat{L}(\theta^*)$ $\quad \mathbb{E}[\bar{X}] = \nabla L(\theta^*) = 0$

minimizer of L

CLT: $\sqrt{n}\left(\nabla \hat{L}(\theta^*) - \nabla L(\theta^*)\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Cov}\left(\nabla \ell((x,y), \theta^*)\right)\right)$

$\Rightarrow \sqrt{n} \cdot \nabla \hat{L}(\theta^*) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Cov}\left(\nabla \ell((x,y), \theta^*)\right)\right)$

LLN: $\nabla^2 \hat{L}(\theta^*) \xrightarrow{P} \nabla^2 L(\theta^*)$

$\Rightarrow$ (Slutsky's Thm) $\boxed{\sqrt{n}(\hat{\theta}-\theta^*)} \xrightarrow{d} \left(\nabla^2 L(\theta^*)\right)^{-1} \cdot \mathcal{N}\left(0, \mathrm{Cov}\left(\nabla \ell((x,y), \theta^*)\right)\right)$

$\overset{d}{=} \mathcal{N}\left(0, \left(\nabla^2 L(\theta^*)\right)^{-1} \mathrm{Cov}\left(\nabla \ell((x,y), \theta^*)\right) \left(\nabla^2 L(\theta^*)\right)^{-1}\right)$ 2 ✓

If $X_n \xrightarrow{d} P$ for some probability distribution P. $X_n = O_p(1)$ 1 ✓

By Taylor's Expansion

$$L(\hat{\theta}) = L(\theta^*) + \langle \underbrace{\nabla L(\theta^*)}_{\to 0}, \hat{\theta}-\theta^* \rangle + \frac{1}{2} \langle \hat{\theta}-\theta^*, \nabla^2 L(\theta^*)(\hat{\theta}-\theta^*) \rangle + \text{rh.o}$$

$\langle v, Av \rangle = v^T A v = \|A^{\frac{1}{2}} v\|_2^2$

$$\Rightarrow n\left(L(\hat{\theta}) - L(\theta^*)\right) \doteq \frac{1}{2} \left\| \nabla^2 L(\theta^*)^{\frac{1}{2}} \sqrt{n}(\hat{\theta}-\theta^*) \right\|_2^2$$

Let $S = \left(\nabla^2 L(\theta^*)\right)^{\frac{1}{2}} \sqrt{n}(\hat{\theta}-\theta^*)$ by 2

By (2) $\quad S \xrightarrow{d} \mathcal{N}\left(0, \left(\nabla^2 L(\theta^*)\right)^{-\frac{1}{2}} \mathrm{Cov}\left(\nabla \ell((x,y), \theta^*)\right) \left(\nabla^2 L(\theta^*)\right)^{-\frac{1}{2}}\right)$

$\Rightarrow n\left(L(\hat{\theta}) - L(\theta^*)\right) \xrightarrow{d} \frac{1}{2} \|S\|_2^2$ 4 ✓ 3 ✓

(5): $\lim_{n\to\infty} \mathbb{E}\left[n\left(L(\hat{\theta}) - L(\theta^*)\right)\right] = \frac{1}{2} \mathbb{E}\left[\|S\|_2^2\right]$ $\quad S \in \mathbb{R}^p$

$= \frac{1}{2} \mathbb{E}\left[\mathrm{tr}(S^T S)\right] = \frac{1}{2} \mathbb{E}\left[\mathrm{tr}(S S^T)\right] = \frac{1}{2} \mathrm{tr}\left(\mathbb{E}[S S^T]\right)$

covariance matrix

$$= \frac{1}{2}\mathrm{tr}\left(\mathrm{Cov}(S)\right) = \frac{1}{2}\mathrm{tr}\left(\left(\nabla^2 L(\theta^*)\right)^{-\frac{1}{2}}\mathrm{Cov}\left(\nabla \ell((x,y),\theta^*)\right)\left(\nabla^2 L(\theta^*)\right)^{-\frac{1}{2}}\right)$$

$$= \frac{1}{2}\mathrm{tr}\left(\left(\nabla^2 L(\theta^*)\right)^{-1}\cdot\mathrm{Cov}\left(\nabla \ell((x,y),\theta^*)\right)\right)$$

5V

$\square$.

θ^* θ_*

Well-Specified Case 假设模型的结构和真实数据的生成过程是一致的

$$\mathbb{P}(y|x;\theta) \quad ; \quad \{y^{(i)}|x^{(i)}\}_{i=1}^{n} \sim \mathbb{P}(y^{(i)}|x^{(i)};\theta_*)$$

P_{θ_*}: true data-generating distribution

Thm2 MLE: $\ell((x^{(i)},y^{(i)}),\theta) = -\log \mathbb{P}(y^{(i)}|x^{(i)};\theta)$

$\Rightarrow$ 1. $\theta^* = \theta_*$ (well-specified)

2. $\mathbb{E}[\nabla \ell((x,y),\theta^*)] = 0$ ✓ $\nabla \mathbb{E}[\ell((x,y),\theta^*)] = \nabla L(\theta^*) = 0$

3. $\mathrm{Cov}(\nabla \ell((x,y),\theta^*)) = \nabla^2 L(\theta^*)$ Bartlett's indentity

$$\sqrt{n}(\hat{\theta}-\theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-1}\cdot\mathrm{Cov}(\nabla \ell((x,y),\theta^*))\cdot(\nabla^2 L(\theta^*))^{-1}\right)$$

4. $\sqrt{n}(\hat{\theta}-\theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla^2 L(\theta^*))^{-1}\right)$.

MLE is asymptotically efficient (achieves the CRLB: Cramer-Rao Lower Bound)

Fisher Information Matrix

Proof: $L(\theta) = \mathbb{E}[\ell((x,y),\theta)]$

$$= \mathbb{E}[-\log \mathbb{P}(y|x;\theta)]$$

$$= \mathbb{E}[-\log \mathbb{P}(y|x;\theta) + \log \mathbb{P}(y|x;\theta_*)] + \mathbb{E}[-\log \mathbb{P}(y|x;\theta_*)]$$

$$= \mathbb{E}\left[\log \frac{\mathbb{P}(y|x;\theta_*)}{\mathbb{P}(y|x;\theta)}\right] + \mathbb{E}[-\log \mathbb{P}(y|x;\theta_*)]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\log \frac{P(y|x;\theta_*)}{P(y|x;\theta)} \,\middle|\, x\right]\right] + H(y|x;\theta_*)$$

(tower rule)

$$= \mathbb{E}\left[D_{KL}\left(y|x;\theta_* \,\|\, y|x;\theta\right)\right] + H(y|x;\theta_*)$$

$$\geq H(y|x;\theta_*)$$

(KL divergence: non-negative)

$$\Rightarrow L(\theta^*) \geq H(y|x;\theta_*) = L(\theta_*)$$

$$\theta^* \in \arg\min_\theta L(\theta) \quad (\nabla L(\theta^*) = 0)$$

$$\hat{\theta} \xrightarrow{p} \theta^* \quad \text{(consistency)} \Rightarrow \text{unique limit.}$$

$$\Rightarrow \theta^* = \theta_*$$

1 ✓

假设模型的结构

真实数据的生成过程

$$0 = \nabla L(\theta^*) = \nabla \mathbb{E}\left[\ell((x,y),\theta^*)\right] = \mathbb{E}\left[\nabla \ell((x,y),\theta^*)\right]$$

2 ✓

$$\text{LHS} = \mathrm{Cov}\left(\nabla \ell((x,y),\theta^*)\right) = \mathbb{E}\left[\nabla \ell((x,y),\theta^*) \cdot \nabla \ell((x,y),\theta^*)^T\right]$$

矢量 → 标量

$$= \int p(x) \left(\int p(y|x;\theta^*) \cdot \nabla \log P(y|x;\theta^*) \cdot \nabla \log P(y|x;\theta^*)^T \, dy \right) dx.$$

→ 标量

$$= \int p(x) \cdot \left(\int \frac{\nabla P(y|x;\theta^*) \, \nabla P(y|x;\theta^*)^T}{P(y|x;\theta^*)} \, dy \right) dx.$$

$$\text{RHS} = \nabla^2 L(\theta^*) = \mathbb{E}\left[-\nabla^2 \log P(y|x;\theta^*)\right]$$

$$= \int p(x) \left(\int -\nabla^2 P(y|x;\theta^*) + \frac{\nabla P(y|x;\theta^*) \nabla P(y|x;\theta^*)^T}{P(y|x;\theta^*)} \, dy \right) dx$$

$$\text{WTS: } \int \nabla^2 P(y|x;\theta^*) \, dy = \nabla^2 \int P(y|x;\theta^*) \, dy = \nabla^2 1 = 0$$

$$\Rightarrow \nabla^2 L(\theta^*) = \mathrm{Cov}\left(\nabla \ell((x,y),\theta^*)\right)$$

3 ✓ 4 ✓ □

probability

Similarly, $\begin{cases} \sqrt{n}(\hat{\theta}-\theta^*) \xrightarrow{d} N(0, (\nabla^2 L(\theta^*))^{-1}) \\ n(L(\hat{\theta})-L(\theta^*)) \xrightarrow{d} \frac{1}{2}\|S\|_2^2 \end{cases}$ $\quad S \sim N(0, (\nabla^2 L(\theta^*))^{-1} Cov(\nabla l(x,y,\theta^*))) = N(0, I)$

$$\Rightarrow 2n(\underbrace{L(\hat{\theta})-L(\theta^*)}_{\text{excess risk}}) \sim \chi^2(p)$$

by Thm1(5) $\Rightarrow$ $\boxed{\lim_{n\to\infty} \mathbb{E}[L(\hat{\theta})-L(\theta^*)] = \frac{p}{2n}}$ $(n\to\infty)$: asymptotic analysis

1.3. Concentration Inequality

SLT

ERM $\quad \hat{\theta} = \arg\min_{\theta} \hat{L}(\theta)$

asymptotic: $(n\to\infty)$ $\quad \boxed{\text{excess risk } L(\hat{\theta}) - \inf_{\theta\in\Theta} L(\theta)} = \frac{p}{2n} + o\left(\frac{1}{n}\right)$ "obscure"

#features : constant

① Taylor Expansion

② LLN, CLT

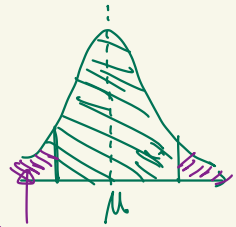
大样本定理

$\frac{p}{2n} + \frac{1}{n^2}$ v.s. $\frac{1}{2n} + \frac{p^{100}}{n^2}$ $\quad (n > p^{50})$

通常 $L(\theta) \in [0,1]$ "informative"

Non-Asymptotic Analysis: concentration inequality 集中不等式

concentration $\begin{cases} X_1+\cdots+X_n \text{ concentrates around } \mathbb{E}[X_1+\cdots+X_n] \\ f(X_1,\cdots,X_n) \text{ concentrates around } \mathbb{E}[f(X_1,\cdots,X_n)] \end{cases}$



tail bound

$\hat{L} = \frac{1}{n}\sum_{i=1}^{n} l_i \longleftrightarrow L = \mathbb{E}[l]$

$\mathbb{P}(|f(X_1,\cdots,X_n) - \mathbb{E}[f(X_1,\cdots,X_n)]| \geq \varepsilon) \leq \delta$

deviation

High-Dimensional Statistics: A Non-Asymptotic View. by Martin Wainwright

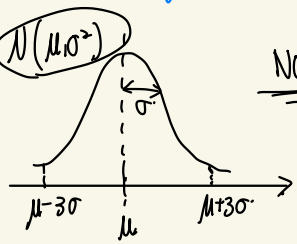
Chebyshev inequality $\quad \mathbb{P}(|Z-\mathbb{E}[Z]| \geq t) \leq \frac{Var[Z]}{t^2} \quad \forall t>0$

Markov inequality: $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$

Proof: $\mathbb{P}(|Z-\mathbb{E}[Z]| \geq t) = \mathbb{P}((Z-\mathbb{E}[Z])^2 \geq t^2)$

$\leq Var[Z]/t^2 \quad \square$

Convergence rate: $O\left(\frac{1}{t^2}\right)$ (quadratic rate) (exponential rate)



Normal Large Deviation Bound: $\boxed{\mathbb{P}(|X-\mu| \geq a\sigma) \leq 2e^{-a^2/2}}$

$Z \sim N(0,1) \Rightarrow$ MGF: $M_Z(t) = \mathbb{E}[e^{tZ}] = e^{t^2/2}$

$$\left(\mathbb{E}[e^{tZ}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tz} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \cdot e^{+\frac{t^2}{2}} \underbrace{\int_{-\infty}^{+\infty} e^{-\frac{(z-t)^2}{2}} d(z-t)}_{1} = e^{\frac{t^2}{2}} \right)$$

$$\mathbb{P}(X-\mu \geq a\sigma) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \geq a\right) = \mathbb{P}(Z \geq a) = \mathbb{P}(e^{tZ} \geq e^{ta}) \quad (t>0)$$

$$\leq \frac{\mathbb{E}[e^{tZ}]}{e^{ta}} = e^{\frac{t^2}{2} - ta}$$ (Markov's ineq.)

$$\leq e^{-a^2/2} \quad (t=a)$$

$$\mathbb{P}(X-\mu \leq -a\sigma) \leq e^{-a^2/2}$$

Bounded Random Variable

Hoeffding's inequality (Thm) $X_1, \cdots, X_n$ i.i.d. r.v $\boxed{X_i \in [a_i, b_i]}$ $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\mu = \mathbb{E}[\bar{X}]$.

$$\Rightarrow \mathbb{P}(|\bar{X}-\mu| \geq \varepsilon) \leq 2\exp\left(-\frac{2n^2\varepsilon^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)$$

exponential rate

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \leq \boxed{\frac{1}{n^2}\sum_{i=1}^n (b_i-a_i)^2 = \sigma^2}$$

upper bound of the variance (proxy)

Let $\varepsilon = O(\sigma\sqrt{\log n}) = \sigma\sqrt{c\log n}$.

$$\Rightarrow \mathbb{P}(|\bar{X}-\mu| \geq \varepsilon) \leq 2\exp\left(-\frac{2\cancel{n^2}\cancel{\sigma^2}c\log n}{\cancel{\sum_{i=1}^n (b_i-a_i)^2}}\right) = 2\exp(-2c\log n) = 2n^{-2c}$$

$\Rightarrow$ for sufficiently large n, $|\bar{X}-\mu| \leq O(\sigma\sqrt{\log n})$ w.p. $\geq 1 - O\left(\frac{1}{n^c}\right)$

With high probability (whp.)

若进一步知道 $a_i = -O(1)$, $b_i = O(1)$

$\sigma^2 = O\left(\frac{1}{n}\right) \Rightarrow |\bar{X}-\mu| \leq O\left(\sqrt{\frac{\log n}{n}}\right) \doteq \tilde{O}\left(\frac{1}{\sqrt{n}}\right)$ w.h.p

(only under exponential rate)

Sub-Gaussian Random Variables

def A r.v. X with finite mean μ is $\boxed{\sigma\text{-sub-Gaussian}}$ if $\boxed{\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\sigma^2\lambda^2/2}}$ $\forall \lambda \in \mathbb{R}$

variance proxy

MGF of $X-\mu$.

MGF of $\mathcal{N}(0,\sigma^2)$

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}[X^k] = \mathbb{E}\left[\sum_{k=0}^{\infty}\frac{(\lambda X)^k}{k!}\right] \Rightarrow \boxed{\mathbb{E}[e^{\lambda X}] \le e^{\frac{\sigma^2\lambda^2}{2}+\mu\lambda}}$$

infinitely many moments of X exists!

MGF of X — MGF of $\mathcal{N}(\mu,\sigma^2)$

(An equivalent definition:
$\mathbb{E}[|X|^p]^{1/p} \lesssim \sqrt{p}$. $\forall p \ge 1$.)

Thm (Tail Bound for sub-Gaussian r.v.) $\boxed{\mathbb{P}(|X-\mu| \ge t) \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right)}$

Proof: $\mathbb{P}(X-\mu \ge t) = \mathbb{P}(e^{\lambda(X-\mu)} \ge e^{\lambda t})$ $\quad (\lambda > 0)$

$\le e^{-\lambda t}\cdot\mathbb{E}[e^{\lambda(X-\mu)}]$ (Markov inequality)

$\le e^{-\lambda t}\cdot e^{\sigma^2\lambda^2/2}$ (sub-Gaussian)

$= e^{-\lambda t + \frac{\sigma^2\lambda^2}{2}} = e^{-\frac{t^2}{2\sigma^2}}$ $\quad (\lambda = t/\sigma^2)$

Similarly. $\mathbb{P}(X-\mu \le -t) \le e^{-\frac{t^2}{2\sigma^2}}$ $\Rightarrow$ Q.E.D. $\square$

Note: Tail Bound also implies sub-Gaussianity.

Thm (Sum of sub-Gaussian r.v's is sub-Gaussian) $X_1, \cdots, X_n$ is $\sigma_1, \cdots, \sigma_n$ sub-Gaussian.

$Z = \sum_{i=1}^n X_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian. $\left(\mathbb{P}(|Z-\mathbb{E}[Z]| \ge t) \le 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n\sigma_i^2}\right)\right)$ $\forall t > 0$

Proof: $\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda(X_i-\mathbb{E}[X_i])}\right]$

$$= \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \le \prod_{i=1}^{n} e^{\lambda^2\sigma_i^2/2} = e^{\frac{\lambda^2 \sum_{i=1}^{n}\sigma_i^2}{2}} \qquad \square$$

Examples of Sub-Gaussian Random Variables.

1) $X \sim N(\mu, \sigma^2)$ $\quad \mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\lambda^2\sigma^2/2}$. $\Rightarrow \mathbb{P}(|X-\mu| \ge t) \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$

2) X satisfying $|X - \mathbb{E}[X]| \le M$ a.s. for $M > 0$ $\Rightarrow X$ is $O(M)$-sub-Gaussian.

$\Rightarrow$ (Bounded Random Variable) $X \in [a,b]$ $\Rightarrow \mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \le \frac{(b-a)^2}{4} = \sigma^2$.

variance proxy

$$\Leftrightarrow \boxed{\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)}$$

Hoeffding's Lemma

$+$ Thm 1 $+$ Thm 2 : $\mathbb{P}(|\bar{X} - \mu| \ge t) \le 2\exp\left(-\frac{2t^2n^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$

Hoeffding's inequality.

3) (Rademacher r.v.) $\varepsilon \sim \{\pm 1\}$ $\quad \begin{cases} \text{w.p. } \frac{1}{2} & \varepsilon = 1 \\ \text{w.p. } \frac{1}{2} & \varepsilon = -1 \end{cases}$ "random noise"

Claim: ε is 1-sub-Gaussian.

Proof: $\mathbb{E}[e^{\lambda\varepsilon}] = \frac{1}{2}(e^{\lambda} + e^{-\lambda}) = \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} + \sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}\right)$

$$= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \le 1 + \sum_{k=0}^{\infty}\frac{(\lambda^2)^k}{2^k \cdot k!} = \exp\left(\frac{1}{2}\lambda^2\right) \qquad \square$$

McDiarmid's Inequality

f is not overly sensitive to arbitrary changes in a single coordinate

$f: \mathbb{R}^n \to \mathbb{R}$, satisfying bounded difference condition.

$\exists\, c_1, \cdots, c_n \in \mathbb{R}$, s.t. $\forall\, x_1, \cdots, x_n, x_i'$, s.t.

$$|f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \le c_i \qquad i \in [n].$$

r.v. $X_1, \cdots, X_n$.

$$\mathbb{P}\left(\left|f(X_1,\cdots,X_n) - \mathbb{E}[f(X_1,\cdots,X_n)]\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

Note: 1) Hoeffding's inequality is when $f = \sum_{i=1}^{n} X_i$

2) $f(X_1,\cdots,X_n)$ is $\mathcal{O}\left(\sqrt{\sum_{i=1}^{n} c_i^2}\right)$-sub-Gaussian.

1.4. Uniform Convergence

一致收敛

Generalization: bound excess error $L(\hat{\theta}) - L(\theta^*)$ ✓

$\Leftrightarrow \mathbb{P}\left(\left|L(\hat{\theta}) - L(\theta^*)\right| < \varepsilon\right) \geq 1-\delta$

ERM $= \arg\min \hat{L}(\theta)$

PAC Learning (Probably Approximately correct)

$$L(\hat{\theta}) - L(\theta^*) = L(\hat{\theta}) - \hat{L}(\hat{\theta})$$
$$\boxed{+\hat{L}(\hat{\theta}) - \hat{L}(\theta^*)} \leq 0$$
$$+\hat{L}(\theta^*) - L(\theta^*)$$

$\hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}, \theta)$

optimization

$$\leq \left|\hat{L}(\hat{\theta}) - L(\hat{\theta})\right| + \left|\hat{L}(\theta^*) - L(\theta^*)\right| \leq 2\boxed{\sup_{\theta\in\Theta}\left|\hat{L}(\theta) - L(\theta)\right|}$$

经验风险和泛化风险之间的差距 (gap)

generalization gap (optimism)

Uniform Convergence: $\boxed{\mathbb{P}\left(\left|\hat{L}(\theta) - L(\theta)\right| \geq \varepsilon\right) \leq \delta.}$

$\forall \theta \in \Theta$

$\Rightarrow$ Uniform Convergence implies generalization!

Recall: concentration inequality

$\hat{L}$ L

Hoeffding's inequality: $\mathbb{P}\left(\left|\bar{X} - \mu\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$

McDiarmid's inequality: $\mathbb{P}\left(\left|f(X_1,\cdots,X_n) - \mathbb{E}[f(X_1,\cdots,X_n)]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right)$

High-level ideas about deriving uniform convergence bound:

一个θ 1) fixed choice of θ: $\mathbb{P}\left(|\hat{L}(\theta)-L(\theta)|\geq\varepsilon'\right)\leq\delta'_\theta$ by concentration inequality

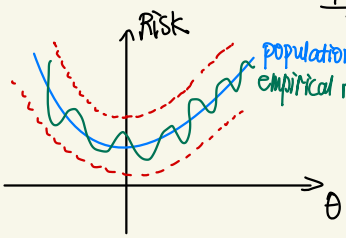
所有θ 2) all values of θ: union-bound $\mathbb{P}\left(\bigcup_i A_i\right)\leq\sum_i\mathbb{P}(A_i)$

uniform

$$\mathbb{P}\left(\exists\,\theta\in\Theta:\ |\hat{L}(\theta)-L(\theta)|\geq\varepsilon'\right)=\mathbb{P}\left(\bigcup_{\theta\in\Theta}|\hat{L}(\theta)-L(\theta)|\geq\varepsilon'\right)$$

$$\leq\sum_{\theta\in\Theta}\mathbb{P}\left(|\hat{L}(\theta)-L(\theta)|\geq\varepsilon'\right)\leq\sum_{\theta\in\Theta}\delta'_\theta=\delta$$

$$\Rightarrow\mathbb{P}\left(\forall\,\theta\in\Theta:\ |\hat{L}(\theta)-L(\theta)|\leq\varepsilon\right)\geq1-\delta:\ \text{uniform convergence}$$



population risk (L)
empirical risk ($\hat{L}$) 落在红线($L\pm\varepsilon$)内的概率 $\geq1-\delta$. ($\forall\theta\in\Theta$)

1. Finite Hypothesis Space ($|\mathcal{H}|$ finite)

Thm 1 $\mathcal{H}$ finite hypothesis space. ℓ: bounded in $[0,1]$ ($0\leq\ell((x,y),h)\leq1$)

$\forall\delta\in(0,\frac{1}{2})$ $\boxed{\text{w.p.}\geq1-\delta:\ |L(h)-\hat{L}(h)|\leq\sqrt{\frac{\log|\mathcal{H}|+\log 2/\delta}{2n}}}$ ($\forall h\in\mathcal{H}$) worst case

ε'

样本数(n)大: gap 小
假设空间($\mathcal{H}$)太大: gap 大 → $|\mathcal{H}|$ 不宜过大

Occam's Razor 奥卡姆剃刀

Proof: 1) fixed h

$\bar{X}\leftarrow\hat{L}(h)=\frac{1}{n}\sum_{i=1}^{n}\ell((x^{(i)},y^{(i)}),h)$

$\mu\leftarrow L(h)=\mathbb{E}[\ell((x,y),h)]=\mathbb{E}[\hat{L}(h)]$

Let $X_i=\ell((x^{(i)},y^{(i)}),h)\in[0,1]$ $i=1,\cdots,n$.

By Hoeffding's inequality. $\mathbb{P}\left(|\hat{L}(h)-L(h)|\geq\varepsilon\right)$ ("bad event") $\leq2\exp\left(-\frac{2n^2\varepsilon^2}{\sum_{i=1}^{n}1}\right)=\boxed{2\exp(-2n\varepsilon^2)}$ δ'

2) for all $h \in \mathcal{H}$:

$$\mathbb{P}(\exists h \in \mathcal{H}: |\hat{L}(h) - L(h)| \geq \varepsilon) \overset{\text{Union Bound}}{\leq} \sum_{h \in \mathcal{H}} \mathbb{P}(|\hat{L}(h) - L(h)| \geq \varepsilon)$$

$$\leq \sum_{h \in \mathcal{H}} 2\exp(-2n\varepsilon^2) = 2|\mathcal{H}| \cdot \exp(-2n\varepsilon^2) = \delta$$

$$\Rightarrow \varepsilon = \sqrt{\frac{\log|2\mathcal{H}| + \log\frac{1}{\delta}}{2n}}$$

$$\Rightarrow \text{w.p.} \geq 1-\delta: \underbrace{\forall h \in \mathcal{H}}_{\text{uniform}}, |\hat{L}(h) - L(h)| \leq \sqrt{\frac{\log|2\mathcal{H}| + \log\frac{1}{\delta}}{2n}} \quad \square$$

concentration inequality: $\forall h \in \mathcal{H}$ w.h.p $|\hat{L}(h) - L(h)| \leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right)$

uniform convergence: w.h.p $\forall h \in \mathcal{H}$ $|\hat{L}(h) - L(h)| \leq \tilde{O}\left(\frac{\log|\mathcal{H}|}{\sqrt{n}}\right)$ *

$\Rightarrow$ excess error: $L(\hat{h}) - L(h^*) \leq \tilde{O}\left(\frac{\log|\mathcal{H}|}{\sqrt{n}}\right)$

trade-off in order to make the bound uniform

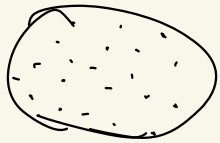
$O\left(\frac{1}{\sqrt{n}}\right)$: slow rate

* asymptotic: $L(\hat{h}) - L(h^*) \leq \frac{c}{n} + o\left(\frac{1}{n}\right)$

$O\left(\frac{1}{n}\right)$: fast rate

2. Infinite Hypothesis Space

- brute-force discretization "离散化"
- complexity measure → (下节)



$\mathcal{H} = \{h: x \mapsto w^T x + b,\ w \in \mathbb{R}^p, b \in \mathbb{R}\}$

Assume hypothesis space is bounded: $\boxed{\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}^p, \|\theta\|_2 \leq B\}}$

Bound → regularization.

$\min f(\theta)$ s.t. $\|\theta\|_2 \leq B$ $\Rightarrow \min f(\theta) + \lambda \cdot \|\theta\|_2$

Key Idea: $E_\theta = \{|\hat{L}(\theta) - L(\theta)| \geq \varepsilon\}$ "bad" event

bound the probability of $\bigcup_{\theta \in \Theta} E_\theta$ → 不能用 Union Bound.

无限

有限

Observation: these bad events can overlap with each other significantly

$\Rightarrow$ prototype / cover
"原型" "覆盖"

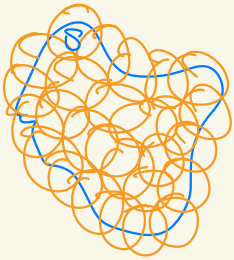
$\bigcup_{\theta\in\Theta} E_\theta \approx \bigcup_{i=1}^{N} E_{\theta_i}$ 有限 $\rightarrow$ Union Bound.

$\{E_{\theta_1}, \cdots, E_{\theta_N}\}$

def (ε-cover / ε-net) Let $\varepsilon>0$. An ε-cover of a set S w.r.t. distance metric ρ is a subset $C \subseteq S$ s.t. $\forall x\in S.\ \exists x'\in C$ s.t. $\rho(x,x')\le\varepsilon$

$$\boxed{S \subseteq \bigcup_{x\in C} \mathrm{Ball}(x,\varepsilon,\rho)}$$

"有限覆盖定理" (Heine-Borel)



$\mathrm{Ball}(x,\varepsilon,\rho) = \{x' : \rho(x,x')\le\varepsilon\}$

$S = \{x\in\mathbb{R}^p : \|x\|_2 \le B\} \rightarrow \ell_2$-ball

Lemma (ε-cover of ℓ_2-ball) Let $B.\ \varepsilon>0$. then there exist an ε-cover of S w.r.t. ℓ_2 norm with at most $\boxed{\max\left(\left(\frac{3B\sqrt{p}}{\varepsilon}\right)^p, 1\right)}$ elements

Proof: If $\varepsilon > B\sqrt{p}$. trivially holds ε-cover has size 1.

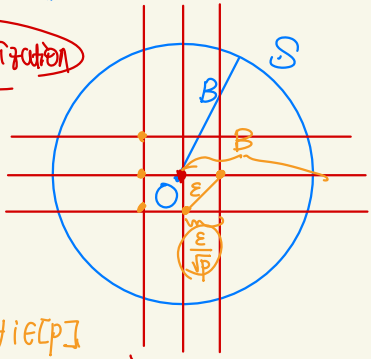
Assume $\varepsilon \le B\sqrt{p}$

grid point

length of unit grid

discretization

$$C = \left\{x\in S : x_i = k_i\cdot\frac{\varepsilon}{\sqrt{p}},\ k_i\in\mathbb{Z},\ |k_i|\le\frac{B\sqrt{p}}{\varepsilon}\right\}$$



Claim: C is an ε-cover of S w.r.t. ℓ_2-norm

$\forall x\in S\ \exists$ grid point $x'\in C$. $|x_i - x_i'| \le \frac{\varepsilon}{\sqrt{p}}$ $\forall i\in[p]$

$\Rightarrow \|x-x'\|_2 = \sqrt{\sum_{i=1}^{p}|x_i-x_i'|^2} \le \sqrt{p\cdot\frac{\varepsilon^2}{p}} = \varepsilon.$ ✓

只考虑网格的格点构成 ε-cover.

size of C? each k_i has at most $\frac{2B\sqrt{p}}{\varepsilon}+1$ choices

$$\Rightarrow |C| \le \left(\frac{2B\sqrt{p}}{\varepsilon}+1\right)^p \le \left(\frac{3B\sqrt{p}}{\varepsilon}\right)^p$$

$\Rightarrow$ ε-cover of an ℓ_2-ball has size $\max\left\{\left(\frac{3B\sqrt{p}}{\varepsilon}\right)^p, 1\right\}$. $\square$

Note: A strong version $\max\left(\left(\frac{3B\sqrt{p}}{\varepsilon}\right)^p, 1\right) \to \boxed{\left(\frac{3B}{\varepsilon}\right)^p}$

def (K-Lipschitz function) Let $K>0$ and $\|\cdot\|$ be a norm on the domain D.

A function $L: D \to \mathbb{R}$ is called K-Lipschitz w.r.t. $\|\cdot\|$ if $\forall \theta, \theta' \in D$:

$$|L(\theta) - L(\theta')| \le K \|\theta - \theta'\|.$$

(arbitrary normed space)

Thm2. $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}^p, \|\theta\|_2 \le B\}$. $\ell((x,y),\theta) \in [0,1]$. K-Lipschitz in θ w.r.t. ℓ_2-norm,

then: w.p. $\ge 1 - O(\exp(-\Omega(p)))$: $\boxed{\forall \theta \in \Theta: |\hat{L}(\theta) - L(\theta)| \le O\left(\sqrt{\frac{p \max(\log KBn, 1)}{n}}\right)}$

Proof: fix params $\delta, \varepsilon > 0$ 待定(TBD)

Let C be an ε-cover of our parameter space $S = \{\theta \in \mathbb{R}^p : \|\theta\|_2 \le B\}$ w.r.t. ℓ_2-norm

(C是一个有限的设空间)

$E = \{\forall \theta \in C: |L(\hat{\theta}) - L(\theta)| \le \delta\}$ "good" events

$\mathbb{P}(E) \ge 1 - 2 \cdot |C| \cdot \exp(-2n\delta^2)$ (by Thm1)

C 有限

S 无限

$\forall \theta \in S: \exists \theta_0 \in C$ s.t. $\|\theta - \theta_0\|_2 \le \varepsilon$

$\theta_0 \in C$

bound $|\hat{L}(\theta) - L(\theta)| \le |\hat{L}(\theta) - \hat{L}(\theta_0)| + |\hat{L}(\theta_0) - L(\theta_0)| + |L(\theta_0) - L(\theta)|$

ℓ: K-Lipschitz $\Rightarrow \begin{cases} \hat{L}: K\text{-Lipschitz} \Rightarrow |\hat{L}(\theta) - \hat{L}(\theta_0)| \le K \|\theta - \theta_0\|_2 \le K \cdot \varepsilon \\ L: K\text{-Lipschitz} \Rightarrow |L(\theta) - L(\theta_0)| \le K \|\theta - \theta_0\|_2 \le K \cdot \varepsilon \end{cases}$

$$\Rightarrow \left|\hat{L}(\theta) - L(\theta)\right| \le \boxed{2K\varepsilon + \delta} \quad \text{w.p.} \ge 1 - 2\cdot|C|\cdot\exp(-2n\delta^2)$$

ε: cover bound

δ: gap bound

$\Rightarrow$ Let $\varepsilon = \frac{\delta}{2K}$ $\Rightarrow$ $\left|\hat{L}(\theta) - L(\theta)\right| \le 2\delta$

How to choose proper δ ?

probability: $1 - 2|C|\cdot\exp(-2n\delta^2) = 1 - 2\cdot\exp(\log|C| - 2n\delta^2)$

by Lemma, $|C| \le \left(\frac{3B}{\varepsilon}\right)^p = \left(\frac{3B}{\delta}\cdot 2K\right)^p = \left(\frac{6BK}{\delta}\right)^p$

$$\Rightarrow \log|C| \le \log\left(\frac{6BK}{\delta}\right)^p = p\cdot\boxed{\log\frac{6BK}{\delta}}$$

If we drop the log term, $\log|C| \le p$

$\Rightarrow$ prob. of "bad" events: $2|C|\cdot\exp(-2n\delta^2) \le 2\exp(p - 2n\delta^2)$ $\quad \left(\delta = \sqrt{\frac{p}{n}}\right)$

$= 2\exp(-p) = O(\exp(-\Omega(p)))$ $\quad 2n\delta^2 = 2p$

$$\Rightarrow \left|L(\hat{\theta}) - L(\theta)\right| \le 2\delta = 2\sqrt{\frac{p}{n}} = O\left(\sqrt{\frac{p\cdot\max(\log KBn, 1)}{n}}\right)$$

However, we CANNOT drop the log term! $\quad$ δ bigger

$$\delta = \sqrt{\frac{C_0\cdot p\cdot\max(\log KBn, 1)}{n}}. \qquad C_0 = 36 \ (\text{TBD}) \text{ is a constant.}$$

$$\log|C| - 2n\delta^2 \le p\cdot\log\frac{6BK}{\delta} - 2n\delta^2$$

$\max(a,b) \ge a$
$\max(a,b) \ge b$

$$\le p\cdot\log\frac{6BK\cdot\sqrt{n}}{\sqrt{C_0}\,p\cdot\max(\log KBn, 1)} - 2n\cdot\frac{C_0 p}{n}\cdot\max(\log KBn, 1)$$

$$\le p\cdot\log\frac{BK\sqrt{n}}{\sqrt{p}} - 72\cdot p\cdot\log KBn \qquad \sqrt{\frac{n}{p}} \le \sqrt{n} \le n.$$

$$\le p\cdot\log BKn - 72\cdot p\cdot\log KBn = -71p\log KBn \le -p \qquad \left(n \ge \frac{1}{KB}\cdot e^{\frac{1}{71}}\right)$$

$\Rightarrow$ w.p. $\geq 1 - 2|C|\exp(-2n\delta^2) \geq 1 - \exp(-p) = 1 - O(\exp(-\Omega(p)))$

we have $|\hat{L}(\theta) - L(\theta)| \leq 2\delta = O\left(\sqrt{\frac{p \cdot \max(\log \frac{p}{\delta n}, 1)}{n}}\right)$ $\square$

$$|\hat{L}(\theta) - L(\theta)| \leq \delta + 2\varepsilon K \leq \sqrt{\frac{\log|C|}{n}} + 2\varepsilon K$$

有限假设空间

error from bf discretization

$|C| \to |S|$ trade off

- 假设空间简单. ← Occam's Razor SLT
- 假设空间复杂 $\Rightarrow$ 更多样本 overparameterization DLT

1.5. Rademacher Complexity

Recall: bound excess risk $L(\hat{h}) - L(h^*) \xrightarrow{\text{uniform convergence}} \sup_{h \in H} |L(h) - \hat{L}(h)|$

generalization gap

1) Finite Hypothesis Space: $L(h) - \hat{L}(h) \leq \tilde{O}\left(\sqrt{\frac{\log|H|}{n}}\right)$

2) Bounded Hypothesis Space (p-dimensional): $L(h) - \hat{L}(h) \leq \tilde{O}\left(\sqrt{\frac{p}{n}}\right)$ slow rate

$\log|H|$ p: complexity measure of H

dimension-free

$\{\theta : \|\theta\|_1 \leq B\}$

$\{\theta : \|\theta\|_2 \leq B\}$ $\Rightarrow \tilde{O}\left(\sqrt{\frac{p}{n}}\right)$

$$\Rightarrow |L(h) - \hat{L}(h)| \leq \tilde{O}\left(\sqrt{\frac{\text{Complexity}(H)}{n}}\right)$$

① This complexity measure will depend on the distribution P over $\mathcal{X} \times \mathcal{Y}$

$\Rightarrow$ how easy it is to learn P

if P is easy to learn $\Rightarrow$ complexity measure is small (even if H large)

② Restrict the hypothesis space to a subspace with small complexity

Regularization $\boxed{\hat{L}(\theta) + \lambda \cdot \text{Complexity}(\theta)}$

empirical risk — Structural risk minimization (SRM)

def (Rademacher Complexity) Let $\mathcal{F}$ be a family of mapping $\mathcal{Z} \to \mathbb{R}$. Let P be a distribution over $\mathcal{Z}$. The (average) Rademacher complexity of $\mathcal{F}$:

$$\boxed{\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{Z_1,\cdots,Z_n \overset{iid}{\sim} P}\left[\mathbb{E}_{\sigma_1,\cdots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right]\right]}$$

定义时抹

where $\sigma_1, \cdots, \sigma_n$ are independent Rademacher r.v. $\begin{cases} +1: \text{ w.p } \frac{1}{2} \\ -1: \text{ w.p } \frac{1}{2} \end{cases}$ ⇒ 随机噪声 (random noise)

$\sup_{h\in\mathcal{H}} |\mathcal{L}(h) - \hat{\mathcal{L}}(h)|$ $\xrightarrow{\text{weak}}$ $\mathbb{E}_{Z_1,\cdots,Z_n}\left[\sup_{h\in\mathcal{H}} |\mathcal{L}(h) - \hat{\mathcal{L}}(h)|\right] \le$ upper bound

randomness

in expectation!

(not with high probability!)

$\sigma = (\sigma_1, \cdots, \sigma_n)$

$f(Z) = (f(Z_1), \cdots, f(Z_n))$

$\sum_{i=1}^{n} \sigma_i f(Z_i) = \langle \sigma, f(Z) \rangle$ inner product 内积

⇒ 相似性 / 关联程度

⇒ $\mathcal{R}_n(\mathcal{F})$: maximal possible correlation between outputs of some $f \in \mathcal{F}$ and random $(\sigma_1, \cdots, \sigma_n)$

$f(Z)$ σ

Takeaway: Rademacher 复杂度测量的是一个函数族在随机噪声下的拟合能力

In machine learning $\begin{cases} \mathcal{F}: \text{假设空间 } \mathcal{H} \text{ 损失函数的空间} \\ \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \quad Z_i = (x_i, y_i) \end{cases}$

Thm1 $$\mathbb{E}_{Z_1,\cdots,Z_n \overset{iid}{\sim} P}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}_{Z\sim P}[f(Z)]\right)\right] \le 2\cdot\mathcal{R}_n(\mathcal{F})$$ ☆

Observation: $\mathcal{F} = \{f: Z = (x,y) \in \mathcal{Z} \mapsto \ell((x,y), h) \mid h \in \mathcal{H}\}$

uniform convergence $\mathbb{E}\left[\sup_{h\in H}\left|L(h)-\hat{L}(h)\right|\right]$ (upper bound in expectation)

$$= \mathbb{E}_{\{(x^{(i)},y^{(i)})\}\sim p^n}\left[\sup_{h\in H}\left|\mathbb{E}_{(x,y)\sim p}[\ell((x,y),h)] - \frac{1}{n}\sum_{i=1}^{n}\ell((x^{(i)},y^{(i)}),h)\right|\right]$$

$$= \mathbb{E}_{\{z_i\}\sim p^n}\left[\sup_{f\in F}\left|\mathbb{E}_{z}[f(z)] - \frac{1}{n}\sum_{i=1}^{n}f(z_i)\right|\right]$$

$$= \mathbb{E}_{\{z_i\}\sim p^n}\left[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}f(z_i) - \mathbb{E}_{z}[f(z_i)]\right]$$

$$\le 2\cdot R_n(-F) = 2\cdot R_n(F) \quad \text{(by symmetry of } \sigma\text{)}$$

$\Rightarrow 2R_n(F)$ is an upper bound for the generalization gap

↳ how well $(\ell((x^{(1)},y^{(1)}),h), \cdots, \ell((x^{(n)},y^{(n)}),h))$ correlates with $(\sigma_1, \cdots, \sigma_n)$

Note: $R_n(F)$: 模类函数空间的 Rademacher 复杂度

$R_n(H)$: 假设空间的 Rademacher 复杂度

$R_n(F)$ is NOT always the same as $R_n(H)$

Talagrand Lemma

$$\boxed{R_n(F) \le K\cdot R_n(H)}$$

Example. Binary Classification $y\in\mathcal{Y}=\{\pm 1\}$ $\quad \ell_{0-1}((x,y),h) = \mathbb{1}\{h(x)\neq y\} = \frac{1-yh(x)}{2}$

$$\Rightarrow R_n(F) = \mathbb{E}_{\{x^{(i)},y^{(i)}\}}\left[\mathbb{E}_{\{\sigma_i\}}\left[\sup_{h\in H}\frac{1}{n}\sum_{i=1}^{n}\ell_{0-1}((x^{(i)},y^{(i)}),h)\cdot\sigma_i\right]\right]$$

同号: 0
异号: 1

$$= \mathbb{E}_{\{x^{(i)},y^{(i)}\}}\left[\mathbb{E}_{\{\sigma_i\}}\left[\sup_{h\in H}\frac{1}{n}\sum_{i=1}^{n}\frac{-y^{(i)}h(x^{(i)})+1}{2}\cdot\sigma_i\right]\right]$$

$$\sigma_i \begin{cases} +1 \text{ w.p. } \frac{1}{2} \\ -1 \text{ w.p. } \frac{1}{2}\end{cases}$$

$$= \frac{1}{2}\mathbb{E}_{\{x^{(i)},y^{(i)}\}}\left[\mathbb{E}_{\{\sigma_i\}}\left[\sup_{h\in H}\frac{1}{n}\sum_{i=1}^{n}y^{(i)}h(x^{(i)})\cdot(-\sigma_i) + \boxed{\frac{1}{n}\sum_{i=1}^{n}\sigma_i}\right]\right]$$

$\mathbb{E}=0$

$$= \frac{1}{2}\mathbb{E}_{\{x^{(i)},y^{(i)}\}}\left[\mathbb{E}_{\{\sigma_i\}}\left[\sup_{h\in H}\frac{1}{n}\sum_{i=1}^{n}h(x^{(i)})\sigma_i\right]\right]$$

$= \frac{1}{2} R_n(\mathcal{H})$

↳ how well $h \in \mathcal{H}$ can fit random patterns

$-y_i \sigma_i \overset{d}{=} \sigma_i$ ("Bayes y_i")

Recall (Thm 1) $\mathbb{E}_{Z_1,\dots,Z_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}_{Z\sim P}[f(Z)]\right] \le 2\cdot R_n(\mathcal{F})$

Proof: (symmetrization) fix $Z_1,\dots,Z_n$ and draw $Z_1',\dots,Z_n' \sim P$

$$\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}_{Z\sim P}[f(Z)]\right) = \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}_{Z_1',\dots,Z_n'}\left[\frac{1}{n}\sum_{i=1}^n f(Z_i')\right]\right)$$

$$= \sup_{f\in\mathcal{F}} \mathbb{E}_{Z_1',\dots,Z_n'}\left[\frac{1}{n}\sum_{i=1}^n f(Z_i) - \frac{1}{n}\sum_{i=1}^n f(Z_i')\right]$$

$$\le \mathbb{E}_{Z_1',\dots,Z_n'}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(Z_i) - f(Z_i'))\right],$$ by sub-additivity of sup

$$\sup_u \mathbb{E}_v[g(u,v)] \le \sup_u \mathbb{E}_v\left[\sup_{u'} g(u',v)\right] = \mathbb{E}_v\left[\sup_{u'} g(u',v)\right]$$

u: f

v: Z_i'

$$\Rightarrow \text{LHS} \le \mathbb{E}_{Z_1,\dots,Z_n}\left[\mathbb{E}_{Z_1',\dots,Z_n'}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(Z_i) - f(Z_i'))\right]\right]$$

$$= \mathbb{E}_{Z_i,Z_i'\sim P}\left[\mathbb{E}_{\sigma_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i (f(Z_i) - f(Z_i'))\right]\right]$$

$f(Z_i) - f(Z_i')$: symmetric

$\Rightarrow \sigma_i(f(Z_i) - f(Z_i')) \overset{d}{=} f(Z_i) - f(Z_i')$

$$\le \mathbb{E}_{Z_i,\sigma_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right] + \mathbb{E}_{Z_i',\sigma_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i')\right]$$

$\sup(A-B) \le \sup A + \sup B$

$$= R_n(\mathcal{F}) + R_n(\mathcal{F}) = 2R_n(\mathcal{F}) \qquad \square$$

LHS: $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)]$

RHS: $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)$

① Remove unknown $\mathbb{E}[f(Z)]$ (depend on P)

② add more randomness via σ_i

Rademacher Complexity depends on distribution P! 分布相关

e.g. P point mass ($Z = Z_0$ a.s.) Assume $-1 \le f(Z_0) \le 1$. $\forall f \in \mathcal{F}$

$$\mathbb{E}_{Z_1,\cdots,Z_n \sim P, \sigma_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right] = \mathbb{E}_{\sigma_1,\cdots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n} f(Z_0)\sum_{i=1}^n \sigma_i\right]$$

$$= \mathbb{E}_{\sigma_i}\left[\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\right|\right] \le \mathbb{E}_{\sigma_i}\left[\left(\frac{1}{n}\sum_{i=1}^n \sigma_i\right)^2\right]^{\frac{1}{2}}$$

Jenson's inequality

$$= \frac{1}{n}\left(\mathbb{E}_{\sigma_i}\left[\sum_{i=1}^n \sigma_i^2\right]\right)^{\frac{1}{2}} = \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}$$

$\mathbb{E}[\sigma_i] = 0$

$\Rightarrow \begin{cases} \mathbb{E}[\sigma_i^2] = 1 \\ \mathbb{E}[\sigma_i \sigma_j] = 0 \end{cases}$

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} L(h) - \hat{L}(h)\right] \le O(R_n(\mathcal{H})) \quad \text{(in expectation)}$$

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} L(h) - \hat{L}(h) \le \varepsilon\right) \ge 1-\delta \quad \text{(with high probability)}$$

def (Empirical Rademacher Complexity) Given a dataset $S = \{Z_1, \cdots, Z_n\}$

$$R_S(\mathcal{F}) = \mathbb{E}_{\sigma_1,\cdots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right]$$

function class $\mathcal{F}$

dataset S

$$R_n(\mathcal{F}) = \mathbb{E}_{\substack{Z_1,\cdots,Z_n \overset{iid}{\sim} P \\ S = \{Z_1,\cdots,Z_n\}}}\left[R_S(\mathcal{F})\right]$$

Thm 2 Suppose for all $f \in \mathcal{F}$: $0 \le f(Z) \le 1$,

then w.p. $\ge 1-\delta$

$$\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)]\right) \le \boxed{2R_S(\mathcal{F})} + 3\sqrt{\frac{\log 2/\delta}{2n}}$$

Proof: $g(Z_1,\cdots,Z_n) = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)]$

bounded difference property $\Rightarrow$ McDiarmid's inequality

$$g(Z_1,\cdots,Z_n) - g(Z_1,\cdots,Z_i',\cdots,Z_n) \le \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(Z_i)\right) - \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{j=1,j\neq i}^n f(Z_j) + \frac{f(Z_i')}{n}\right)$$

$$\le \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\left(f(Z_i) - f(Z_i')\right)\right) \le \frac{1}{n} \triangleq c_i$$

$\sup A - \sup B \le \sup(A-B)$

By McDiarmid's ineq: $\mathbb{P}\left(g(Z_1,\cdots,Z_n) \ge \underbrace{\mathbb{E}_{Z_1,\cdots,Z_n \overset{iid}{\sim} P}[g(Z_1,\cdots,Z_n)]}_{} + \varepsilon\right) \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$

$$= \exp(-2n\varepsilon^2) = \frac{\delta}{2}$$

$$\varepsilon = \sqrt{\frac{\log 1/\delta}{2n}}$$

By Thm1. $\mathbb{E}_{Z_1,\cdots,Z_n \overset{iid}{\sim} P}[g(Z_1,\cdots,Z_n)] = \mathbb{E}_{Z_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}_Z[f(Z)]\right]$

$$\le 2\cdot R_n(\mathcal{F})$$

$$(*_1) \Rightarrow \boxed{\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)]\right) \le 2\cdot R_n(\mathcal{F}) + \sqrt{\frac{\log 2/\delta}{2n}}}$$

$$\varepsilon = \sqrt{\frac{\log 2/\delta}{2n}}$$

w.p. $\ge 1 - \frac{\delta}{2}$

Similarly, $\tilde{g}(Z_1,\cdots,Z_n) = R_S(\mathcal{F}) = \mathbb{E}_{\sigma_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right]$

$$\tilde{g}(Z_1,\cdots,Z_n) - \tilde{g}(Z_1,\cdots,Z_i',\cdots,Z_n) \le \mathbb{E}_{\sigma_i}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sigma_i\left(f(Z_i) - f(Z_i')\right)\right)\right] \le \frac{1}{n}$$

$\Rightarrow$ Bounded Difference.

$\Rightarrow$ McDiarmid's:

$$(*_2)\quad \mathbb{P}\left(\underbrace{\tilde{g}}_{R_S(\mathcal{F})} - \underbrace{\mathbb{E}[\tilde{g}]}_{R_n(\mathcal{F})} \ge \varepsilon'\right) \le \exp(-2n\varepsilon'^2) = \frac{\delta}{2}$$

$$\Rightarrow \varepsilon' = \sqrt{\frac{\log 2/\delta}{2n}}$$

Union Bound of $(*_1)(*_2)$,

w.p. $\ge 1-\delta$: $\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)]\right) = g \le \mathbb{E}[g] + \varepsilon$ ①

$$\le 2R_n(\mathcal{F}) + \varepsilon \quad ②$$

$$\le 2\cdot(R_S(\mathcal{F}) + \varepsilon') + \varepsilon \quad ③$$

$$= 2R_S(\mathcal{F}) + 3\cdot\sqrt{\frac{\log 2/\delta}{2n}} \quad \square$$

①: McDiarmid's ineq. $\left(\frac{\delta}{2}\right)$

②: Thm1 (1)

③: McDiarmid's ineq. $\left(\frac{\delta}{2}\right)$

Cor Let $\mathcal{F} = \{ f : (x,y) \mapsto \ell((x,y),h) \mid h \in \mathcal{H} \}$ $\quad \ell((x,y),h) \in [0,1]$

$\Rightarrow$ w.p. $\geq 1-\delta$: $\quad \hat{L}(h) - L(h) \leq 2\,\boxed{R_S(\mathcal{F})} + 3\cdot\sqrt{\frac{\log 2/\delta}{2n}} \quad \forall h \in \mathcal{H}$

Typically, $O\left(\sqrt{\frac{\log 2/\delta}{n}}\right) \ll R_S(\mathcal{F})$

$\left(\frac{c}{\sqrt{n}}\right)$

$\left(2R_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}\right)$

dominating term. $\tilde{O}\left(\frac{1}{\sqrt{n}}\right)$

(dimension free)

1.6 Bounding Rademacher Complexity

Recall:

$$R_n(\mathcal{F}) = \mathbb{E}_{Z_1,\cdots,Z_n \overset{iid}{\sim} P}\left[\mathbb{E}_{\sigma_1,\cdots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right]\right].$$

$$\sup_{h\in\mathcal{H}}\left(L(h) - \hat{L}(h)\right) \longrightarrow \mathbb{E}_Z\left[\sup_{h\in\mathcal{H}}\left(L(h) - \hat{L}(h)\right)\right] \leq \text{upper bound}$$

Given a dataset $S = \{Z_1,\cdots,Z_n\}$. $\quad R_S(\mathcal{F}) = \mathbb{E}_{\sigma_1,\cdots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right]$

$\mathcal{F} = \{ (x,y) \mapsto \ell((x,y),h) : h \in \mathcal{H} \}$. $\quad \ell((x,y),h) \in [0,1]$

w.p. $\geq 1-\delta$:

$$\hat{L}(h) - L(h) \leq 2R_S(\mathcal{F}) + 3\sqrt{\frac{\log 2/\delta}{2n}} \quad \forall h \in \mathcal{H}$$

$$\hat{L}(h) - L(h) \leq 2R_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}} \quad \forall h \in \mathcal{H}$$

$$\boxed{R_S(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right]}$$

f_1, f_2

$\downarrow$

$\{f_1(Z_i)\}_{i=1}^n = \{f_2(Z_i)\}_{i=1}^n$

Output space : $Q = \{ (f(z_1), \cdots, f(z_n))^T : f \in F \} \subseteq \mathbb{R}^n$

$$|Q| \le |F| \quad \Rightarrow \quad R_S(F) = \mathbb{E}_\sigma\left[\sup_{v \in Q} \frac{1}{n} \langle \sigma, v \rangle \right]$$

$$\Updownarrow$$

$$\mathbb{E}_{g \sim \mathcal{N}(0,I)}\left[\sup_{v \in Q} \langle g, v \rangle \right]$$

$$F_1 = \{ f(x) = \sum_i \theta_i x_i \mid \theta \in \mathbb{R}^d \}$$

$$F_2 = \{ f(x) = \sum_i \theta_i^3 w_i x_i \mid \theta \in \mathbb{R}^d, w \in \mathbb{R}^d \}$$

$$Q_{F_1} = Q_{F_2} \Rightarrow R_S(F_1) = R_S(F_2)$$

Lemma (Maximal Inequality) $X_1, \cdots, X_n$ random variable (subGaussian)

$$\boxed{\mathbb{E}[\max_j X_j] \le r \cdot \sqrt{2 \log n}} \qquad \exists r > 0 \quad \mathbb{E}[e^{tX_j}] \le e^{\frac{t^2 r^2}{2}}$$

Proof: Jenson's inequality (convex function f: $\mathbb{E}[f(X)] \ge f(\mathbb{E}[X])$)

$$e^{t \cdot \mathbb{E}[\max_j X_j]} \le \mathbb{E}[e^{t \max_j X_j}]$$

$$= \mathbb{E}[\max_j e^{tX_j}] \le \mathbb{E}[\sum_{j=1}^n e^{tX_j}] \le n \cdot e^{\frac{t^2 r^2}{2}}$$

$$\Rightarrow \mathbb{E}[\max_j X_j] \le \frac{\log n}{t} + \frac{t r^2}{2} \qquad \forall t > 0$$

$$t = \sqrt{2 \log n} / r \Rightarrow \mathbb{E}[\max_j X_j] \le r \cdot \sqrt{2 \log n} \quad \square$$

Lemma' (Maximal Inequality) $X_1, \cdots, X_n$ r.v. $\forall j \in [n]$, $X_j = \sum_{i=1}^m Y_{ij}$

Fixed j. Y_{ij} independent. Zero mean. taking values in $[-r_i, r_i]$

$$\Rightarrow \mathbb{E}[\max_j X_j] \le r \cdot \sqrt{2 \log n}, \quad r = \sqrt{\sum_{i=1}^m r_i^2},$$

Y_{ij} : bounded variable

Proof: 先证 sub-Gaussianity.

$$\mathbb{E}[e^{tX_j}] = \mathbb{E}\left[\prod_{i=1}^{m} e^{tY_{ij}}\right] = \prod_{i=1}^{m} \mathbb{E}[e^{tY_{ij}}]$$

$$\leq \prod_{i=1}^{m} e^{t^2 \cdot \frac{r_i^2}{2}} = e^{\frac{t^2 r^2}{2}}.$$

Hoeffding's Lemma

$\Rightarrow X_j$ is r-sub-Gaussian. $\square$

$Z \in [a,b]$. $\mathbb{E}[Z]=0$

$$\mathbb{E}[e^{tZ}] \leq e^{\frac{t^2(b-a)^2}{8}}$$

Hoeffding's Lemma

Thm (Massart's Lemma) $A \in \mathbb{R}^m$ finite set,

$$r = \max_{x \in A} \|x\|_2 \Rightarrow \mathbb{E}_\sigma\left[\sup_{x \in A} \frac{1}{n}\sum_{i=1}^{n} \sigma_i x_i\right] \leq \frac{r \cdot \sqrt{2\log|A|}}{n}.$$

Proof: By Maximal inequality, $\sigma_i x_i \in [-|x_i|, |x_i|]$ $\qquad \sqrt{\sum_{i=1}^{n} x_i^2} \leq r.$

Y_{ij}. $\Leftrightarrow X_j : \sum_{i=1}^{n} \sigma_i x_i$

$$\Rightarrow \mathbb{E}_\sigma\left[\sup_{x \in A} \frac{1}{n}\sum_{i=1}^{n} \sigma_i x_i\right] \leq \frac{r \cdot \sqrt{2\log|A|}}{n} \quad !$$

$$\Rightarrow \boxed{R_S(\mathcal{F}) \leq \sqrt{\frac{2M^2 \log|Q|}{n}}}$$

$$M \geq \frac{1}{\sqrt{n}}\|v\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f(z_i)^2} \Rightarrow f(z) \leq M$$

Rademacher Complexity Bound of finite Hypothesis Class

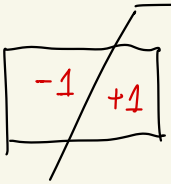
infinite $\sim$ $\rightarrow$ covering number (见第二部分)

二分类 (Binary Classification):

$$Q = \{(h(x_1), \cdots, h(x_n))^T : h \in \mathcal{H}\} \qquad h(x) \in \{-1, 1\}$$

Growth Function (增长函数) $\quad \Pi_{\mathcal{H}}(n) = \max_{\{x_1, \cdots, x_n\} \subseteq X} |\{(h(x_1), \cdots, h(x_n))^T : h \in \mathcal{H}\}|$

dichotomy 对分



n个样本最多可以产生多少种不同的对分

$$2^n \geq \Pi_{\mathcal{H}}(n) \geq |\mathcal{Q}|$$

$$\Rightarrow \mathcal{R}_S(\mathcal{F}) \leq \sqrt{\frac{2\log \Pi_{\mathcal{H}}(n)}{n}}$$

purely combinatorial (distribution-free)

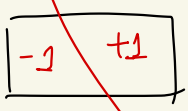
def (VC-Dimension) $\quad VCdim(\mathcal{H}) = \max\{n : \Pi_{\mathcal{H}}(n) = 2^n\}$

Vapnik Chervonenkis

打散 shattering: $\Pi_{\mathcal{H}}(n) = 2^n$

($\mathcal{H}$ realizes all possible dichotomies)

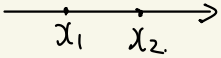
$$VCdim(\mathcal{H}) = d \Rightarrow \begin{cases} \exists (x_1, \cdots, x_d) \text{ 能被 } \mathcal{H} \text{ 打散} \\ \forall (x_1, \cdots, x_d, x_{d+1}) \text{ 不能被 } \mathcal{H} \text{ 打散} \end{cases}$$



"the size of the largest sample set that can be shattered by $\mathcal{H}$"

e.g. Intervals on the real line

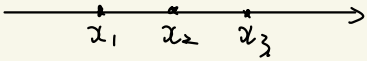
$$\mathcal{H} = \{h : h(x) = \mathbb{1}\{x \in (a,b)\},\ a \in \mathbb{R}, b \in \mathbb{R}, a < b\}$$



(+ +)
(+) −
− (+)
( ) − −

∃2个点能被打散

$\Rightarrow VCdim(\mathcal{H}) = 2.$



(+ + +)
(+ +) −
− (+ +)
(+) − −
− (+) −
− − (+)
− − − ( )
+ − +

∀3个点不能被打散

→ unrealizable!

Thm (Sauer's Lemma) $VCdim(\mathcal{H}) = d$; $\forall n \in \mathbb{N}$, $\Pi_{\mathcal{H}}(n) \le \sum_{i=0}^{d} \binom{n}{i}$

$\Rightarrow \Pi_{\mathcal{H}}(n) \le \left(\frac{en}{d}\right)^d$ (when $n \ge d$)

△ Limitation: VC dimension (typically) highly depends on the dimension of feature space.

$R_S(\mathcal{F}) \le \tilde{O}\left(\sqrt{\frac{d}{n}}\right)$

$\Downarrow$

$R_S(\mathcal{F}) \le \tilde{O}\left(\sqrt{\frac{\#\text{ of params}}{n}}\right)$

deep NN
- ① 参数量太大
- ② 参数维数过高

Rademacher Complexity ☆

Recall: Statistical Learning Theory bound excess risk

$L(\hat{\theta}) - L(\theta^*)$

ERM: $\hat{\theta} = \arg\min \hat{L}(\theta)$

① asymptotic analysis

② Concentration

③ uniform convergence $\sup_{\theta \in \Theta} |L(\theta) - \hat{L}(\theta)|$

↳
- finite hypothesis class
- infinite hypothesis class
  - BF discretization
  - complexity measure!