

Guarantees for Machine Learning: Fact Sheet

Tao SUN

taosun47@student.ethz.ch

September 12, 2021

Contents

1	Introduction (Lec. 1)	2
1.1	The Well-specified Case	2
1.2	The Mis-specification Case	2
2	Concentration bounds (Ch. 2)	2
2.1	Classical bounds	2
2.2	Sub-Gaussian and Hoeffding bounds	3
2.3	Sub-exponential and Bernstein bounds	3
2.4	Martingale-based Methods	4
2.5	Functional bounds	4
3	Uniform Laws of Large Numbers (Ch. 4)	5
3.1	Motivation	5
3.2	A uniform law via Rademacher complexity	5
3.3	Upper bounds on the Rademacher complexity	5
4	Non-uniform Learnability (Lec. 6-7)	6
4.1	Structural risk minimization (SRM)	6
4.2	Margin bound for linear classifiers	6
4.3	Margin bounds for SVM	7
5	Metric Entropy (Ch. 5)	7
5.1	Covering and Packing	7
5.2	Metric entropy and sub-Gaussian processes	7
6	Reproducing Kernel Hilbert Spaces (Ch. 12)	8
6.1	Basics of Hilbert space	8
6.2	Reproducing kernel Hilbert space	9
6.3	Mercer's theorem and its consequences	9
7	Non-parametric Least Squares (Ch. 13)	9
7.1	Fixed design	9
7.2	Error bounds for RKHS	11
7.3	Random design	11

Disclaimer

This note is mainly based on our slides and the textbook (High-Dimensional Statistics: A Non-Asymptotic Viewpoint by Martin J. Wainwright). The main purpose of writing this note is to familiarize me with the concepts and mathematical derivations in the course. Therefore, I do not guarantee the correctness and completeness of it.

1 Introduction (Lec. 1)

1.1 The Well-specified Case

Given a collection of n samples $\{x_i\}_{i=1}^n$ sampled from a fixed distribution \mathbb{P} . Let f^* be the “true” estimator that minimizes some loss $\ell(x; f)$, i.e., $f^* := \arg \min_{f \in \mathcal{F}} \mathbb{E}_x \ell(x; f)$.

Def. 1 (Risk) The estimation of losses is called *risk*:

- Empirical risk:

$$R_n(f) := \frac{1}{n} \sum_{i=1}^n \ell(x_i; f)$$

- Population risk:

$$R(f) := \mathbb{E}_{x \sim \mathbb{P}} \ell(x; f)$$

Remark: Please note that the notation in our course is slightly different from the MW Chapter 4, where it uses a more burdensome notation as $R(f, f^*)$. Here, the f^* is omitted since it is fixed in a problem.

Def. 2 (Excess risk) The excess risk is defined as

$$\mathcal{E}_n(\hat{f}_n, f^*) := R(\hat{f}_n) - R(f^*).$$

Thm. 1 (Risk decomposition) The excess risk $R(\hat{f}) - R(f^*)$ can be decomposed into

$$= \underbrace{R(\hat{f}) - R_n(\hat{f})}_{T_1} + \underbrace{R_n(\hat{f}) - R_n(f^*)}_{T_3} + \underbrace{R_n(f^*) - R(f^*)}_{T_2}$$

Remark: The following figure illustrate the risk decomposition. In the figure, the upper and lower layers represent the empirical risk and the population risk, respectively.

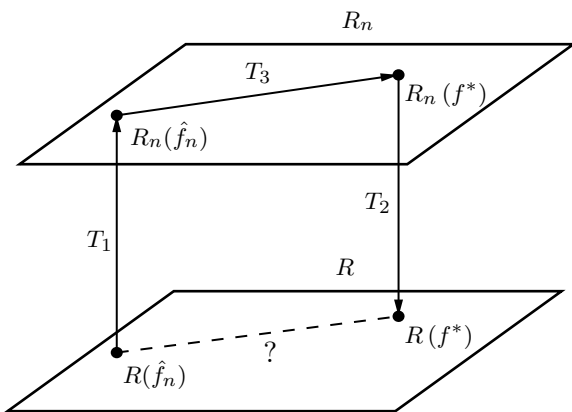


Figure: Risk decomposition (well-specified)

1.2 The Mis-specification Case

We call the problem mis-specification, when $f^* \notin \mathcal{F}$.

Thm. 2 (Bias-variance trade-off) The excess risk $R(\hat{f}) -$

$R(f^*)$ can be decomposed into

$$R(\hat{f}_n) - R(f^*) = \underbrace{R(\hat{f}_n) - R(f^{*, \mathcal{F}})}_{\text{"Variance"}} + \underbrace{R(f^{*, \mathcal{F}}) - R(f^*)}_{\text{"Bias"} = \text{"approx error"}}$$

where $f^{*, \mathcal{F}} = \arg \min_{f \in \mathcal{F}} R(f)$.

Remark: The following figure provides some insights.

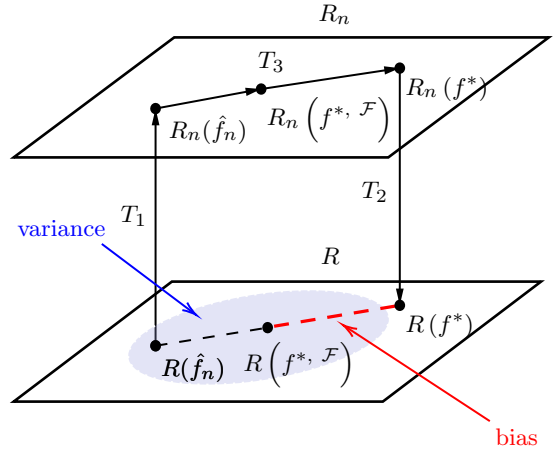


Figure: Bias and variance under mis-specification

2 Concentration bounds (Ch. 2)

2.1 Classical bounds

The most basic tail bound is Markov's inequality.

Thm. 3 (Markov's inequality) For a random non-negative variable X with finite mean,

$$P(X \geq t) \leq \frac{\mathbb{E}X}{t}, \quad \forall t > 0.$$

Proof. Using the non-negativity of X and the definition of expectation,

$$\begin{aligned} \mathbb{E}X &= \int_0^\infty x f(x) dx = \int_0^t x f(x) dx + \int_t^\infty x f(x) dx \\ &\geq \int_t^\infty x f(x) dx \geq t \int_a^\infty f(x) dx = tP(X \geq t). \end{aligned}$$

□

Thm. 4 (Chebyshev's inequality) For a random variable X with finite k -th order center momentum,

$$\begin{aligned} P(|X - \mu| \geq t) &= P(|X - \mu|^k \geq t^k) \\ &\leq \frac{\mathbb{E}|X - \mu|^k}{t^k}, \quad \forall t > 0. \end{aligned}$$

Proof. Replacing X with $|X - \mu|^k$ in Markov's inequality. □

Method. 1 (Chernoff bound) Apply Markov's inequality to random variable $Y = e^{\lambda(X - \mu)}$ ($0 \leq \lambda \leq b$), we get

$$P(X - \mu \geq t) = P(e^{\lambda(X - \mu)} \geq e^{\lambda t}) \leq \frac{\mathbb{E}e^{\lambda(X - \mu)}}{e^{\lambda t}}.$$

Then, optimizing λ to get a tighter bound,

$$\log P(X - \mu \geq t) \leq \inf_{\lambda} \left(\log \mathbb{E} e^{\lambda(X-\mu)} - \lambda t \right).$$

Chernoff bound will be used in achieving tail bounds of the sub-Gaussian and sub-exponential distribution.

2.2 Sub-Gaussian and Hoeffding bounds

Def. 3 (Sub-Gaussian) A random variable X with mean μ is sub-Gaussian with parameter σ if one of following holds:

- MGF condition:

$$\mathbb{E} e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2 \sigma^2}{2}}, \quad \forall \lambda \in \mathbb{R}.$$

- Tail bound condition ($Z \sim \mathcal{N}(0, \sigma^2)$):

$$P(|X - \mu| \geq t) \leq cP(|Z| \geq t), \quad \exists c > 0, \forall t \geq 0.$$

For simplicity, we denote the sub-Gaussian random variable X with mean μ and parameter σ^2 as $X \sim SG(\mu, \sigma^2)$.

Prop. 1 (Sub-Gaussian tail bound) For $X \sim SG(\mu, \sigma^2)$,

$$P(X - \mu \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}, \quad \forall t \in \mathbb{R}.$$

Proof. Applying Chernoff bound on $e^{\lambda(X-\mu)}$, we have

$$\begin{aligned} \log P(X - \mu \geq t) &\leq \inf_{\lambda > 0} \left(\log \mathbb{E} e^{\lambda(X-\mu)} - \lambda t \right) \\ &\leq \inf_{\lambda > 0} \left(\frac{\sigma^2 \lambda^2}{2} - \lambda t \right) = -\frac{t^2}{2\sigma^2}. \end{aligned}$$

Taking exponential on both sides gives the desired form. \square

Prop. 2 (Sum of sub-Gaussian RVs) For $X_1 \sim SG(\mu_1, \sigma_1^2)$, $X_2 \sim SG(\mu_2, \sigma_2^2)$,

$$X_1 + X_2 \sim SG(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Prop. 3 (Sub-Gaussian for bounded RV) For a RV $X \in [a, b]$ almost surely, X is a sub-Gaussian with parameter at most $\sigma = \frac{b-a}{2}$.

Proof. Define function $\phi(\lambda) = \log \mathbb{E} e^{\lambda X}$. It is easy to show that $\phi(0) = 0$ and $\phi'(0) = \mathbb{E} X := \mu$. The second derivative is

$$\phi''(\lambda) = \mathbb{E}_{\lambda}[X^2] - \mathbb{E}_{\lambda}[X]^2, \quad \text{where } \mathbb{E}_{\lambda}[X] = \frac{\mathbb{E} f(X) e^{\lambda X}}{\mathbb{E} e^{\lambda X}}$$

Taking a Taylor expansion of $\phi(\lambda)$ at $\lambda = 0$,

$$\begin{aligned} \phi(\lambda) &= \phi(0) + \lambda \phi'(0) + \frac{\lambda^2}{2} \phi''(\epsilon) \\ &\leq \lambda \mu + \frac{\lambda^2}{2} - \frac{(b-a)^2}{4}. \end{aligned}$$

\square

Thm. 5 (Hoeffding bound) For n independent sub-Gaussian random variables $X_i \in SG(\mu_i, \sigma_i^2)$, $i = [n]$,

$$P\left(\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq e^{-\frac{t^2}{2 \sum_{i=1}^n \sigma_i^2}}$$

Proof. Using the sub-Gaussian tail bound (Prop. 1) and the property of sum of Sub-Gaussian RVs (Prop. 2). \square

Thm. 6 (Sub-Gaussian maxima) For a sequence of sub-Gaussian RVs $\{X_i\}_{i=1}^n$, $X_i \sim SG(0, \sigma^2)$, the following bounds hold,

$$\begin{aligned} \mathbb{E} \max_{i=1, \dots, n} X_i &\leq \sqrt{2\sigma^2 \log n} \\ \mathbb{E} \max_{i=1, \dots, n} |X_i| &\leq \sqrt{2\sigma^2 \log 2n} \end{aligned}$$

Remark: This bound is frequently used in deriving other bounds.

Proof. Let $f(x) = e^{\lambda x}$ ($\lambda > 0$), we have

$$\begin{aligned} e^{\lambda \mathbb{E}[\max_i X_i]} &\leq \mathbb{E} \left[e^{\lambda \max_i X_i} \right] \quad (f(x) \text{ is convex}) \\ &= \mathbb{E} \left[\max_i e^{\lambda X_i} \right] \quad (f(x) \text{ is mono-increasing}) \\ &\leq \sum_{i=1}^n \mathbb{E} \left[e^{\lambda X_i} \right] \leq n e^{\sigma^2 \lambda^2 / 2} \end{aligned}$$

Therefore,

$$\mathbb{E}[\max_i X_i] \leq \frac{\log n}{\lambda} + \frac{\lambda \sigma^2}{2}, \quad \forall \lambda > 0.$$

Using Chernoff bound, we get

$$\mathbb{E}[\max_i X_i] \leq \inf_{\lambda > 0} \frac{\log n}{\lambda} + \frac{\lambda \sigma^2}{2} = \sqrt{2\sigma^2 \log n},$$

The second inequality can be proved by using the fact that

$$\max_i |X_i| = \max \{X_1, \dots, X_n, (-X_1), \dots, (-X_n)\}.$$

\square

2.3 Sub-exponential and Bernstein bounds

Def. 4 (Sub-exponential) A random variable X with mean μ is sub-exponential with parameter (v, α) if one of following holds:

- MGF condition:

$$\mathbb{E} e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2 v^2}{2}}, \quad \forall |\lambda| < \frac{1}{\alpha}.$$

- Tail bound condition:

$$P(|X - \mu| \geq t) \leq c_1 e^{-c_2 t}, \quad \exists c_1, c_2 > 0, \forall t \geq 0.$$

Thm. 7 (Bernstein-type bound) For any random variable satisfying the Bernstein condition $\left| \mathbb{E} \left[(X - \mu)^k \right] \right| \leq \frac{1}{2} k! \sigma^2 b^{k-2}$, we have,

$$P(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2(\sigma^2 + bt)}}$$

Remark: For bounded RV, this bound is more tighter compared with Hoeffding bound for sub-Gaussian with σ when $\sigma \ll b$.

Proof. Using Taylor expansion of the the MGF of sub-exponential RV, we have

$$\begin{aligned} \mathbb{E} \left[e^{\lambda(X-\mu)} \right] &\leq 1 + \frac{\lambda^2 \sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E} \left[(X - \mu)^k \right]}{k!} \\ &\leq 1 + \frac{\lambda^2 \sigma^2}{2} + \frac{\lambda^2 \sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda|b)^{k-2} \\ &\leq 1 + \frac{\lambda^2 \sigma^2 / 2}{1 - b|\lambda|} \leq e^{\frac{\lambda^2 \sigma^2 / 2}{1 - b|\lambda|}} \end{aligned}$$

Then, based on Chernoff bound, we get

$$\begin{aligned} \log P(X - \mu \geq t) &\leq \inf_{\lambda} \left(\log \mathbb{E} e^{\lambda(X-\mu)} - \lambda t \right) \\ &\leq \inf_{\lambda} \left(\frac{\lambda^2 \sigma^2 / 2}{1 - b|\lambda|} - \lambda t \right) = -\frac{t^2}{2(\sigma^2 + bt)}. \end{aligned}$$

This gives the one-side bound. The two-side version can be obtained with additional factor of 2. \square

2.4 Martingale-based Methods

Def. 5 (Martingale) A sequence of RVs Y_1, Y_2, \dots is said to be a martingale with respect to another sequence of RVs X_1, X_2, \dots if for all n ,

$$\mathbb{E}[Y_n] < \infty \quad \text{and} \quad \mathbb{E}[Y_{n+1} | X_1, \dots, X_n] = Y_n.$$

Def. 6 (Doob martingale difference) For a function $g_n : \mathcal{X} \rightarrow \mathbb{R}$ on independent RV $X_i \in \mathcal{X}$ and the σ -field $F_i = \sigma(X_1, \dots, X_i)$, the Doob martingale difference $\{(D_i, F_i)\}_{i=1}^n$ is defined as

$$D_i := S_i - S_{i-1}, \quad \text{where } S_i := \mathbb{E}[g_n(X) | X_1, \dots, X_i]$$

Here, we often define $S_0 = \mathbb{E}[g_n(X)]$, so that we have the *telescoping decomposing*

$$S_n - S_0 = \sum_{i=1}^n D_i.$$

Thm. 8 (Azuma-Hoeffding) For a martingale difference sequence $\{(D_i, F_i)\}_{i=1}^n$, $D_i | F_{i-1} \in [a_i, b_i]$ almost surely for $i = [n]$, then

$$P\left(\sum_{i=1}^n D_i \geq t\right) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

Proof. Since $D_i | F_{i-1} \in [a_i, b_i]$, we have

$$\mathbb{E} \left[e^{\lambda D_i} | F_{i-1} \right] \leq e^{\frac{\lambda^2 (b_i - a_i)^2}{8}}$$

To show the $\sum_{i=1}^n D_i$ is sub-Gaussian, we use the iterated expectation as,

$$\begin{aligned} \mathbb{E} \left[e^{\lambda \sum_{i=1}^n D_i} \right] &= \mathbb{E}_{D_1, \dots, D_n} \left[e^{\lambda \sum_{i=1}^{n-1} D_i} \cdot \mathbb{E}_{D_n} \left[e^{\lambda D_n} | F_{n-1} \right] \right] \\ &= \mathbb{E}_{D_1, \dots, D_{n-1}} \left[e^{\lambda \sum_{i=1}^{n-1} D_i} \right] \cdot \mathbb{E}_{D_n} \left[e^{\lambda D_n} | F_{n-1} \right] \\ &\leq \mathbb{E}_{D_1, \dots, D_{n-1}} \left[e^{\lambda \sum_{i=1}^{n-1} D_i} \right] e^{\frac{\lambda^2 (b_n - a_n)^2}{8}} \end{aligned}$$

Using it recursively, we get $\mathbb{E} \left[e^{\lambda \sum_{i=1}^n D_i} \right] \leq e^{\frac{\lambda^2 \sum_{i=1}^n (b_i - a_i)^2}{8}}$. Then we can achieve the desired bound via Chernoff bound (similar to Prop. 1, sub-Gaussian tail bound). \square

Thm. 9 (McDiarmid inequality) If $g_n(Z)$ satisfies the bounded difference condition, i.e. $|g_n(z) - g_n(z^{\setminus k})| \leq \sigma_k$, for a random vector Z with n independent entries, then

$$P(g_n(Z) - \mathbb{E}g_n(Z) \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n \sigma_i^2}}$$

where $z = (z_1, \dots, z_k, \dots, z_n)$, $z^{\setminus k} = (z_1, \dots, z'_k, \dots, z_n)$.

Proof. First, please note that

$$g_n(Z) - \mathbb{E}g_n(Z) = S_n - S_0 = \sum_{i=1}^n D_i,$$

where $D_i = \mathbb{E}[g_n(Z) | Z_1, \dots, Z_i] - \mathbb{E}[g_n(Z) | Z_1, \dots, Z_{i-1}]$. Therefore, we can use the Azuma-Hoeffding inequality if we show that D_i is bounded.

$$\begin{aligned} D_k &\leq \sup_z \mathbb{E}_{Z_{k+1:n}} [g_n(Z_{1:k-1}, z, Z_{k+1:n})] \\ &\quad - \inf_z \mathbb{E}_{Z_{k+1:n}} [g_n(Z_{1:k-1}, z, Z_{k+1:n})] \\ &\leq \sup_{z, z'} \left| \mathbb{E}_{Z_{k+1:n}} [g_n(Z_{1:k-1}, z, Z_{k+1:n})] \right. \\ &\quad \left. - \mathbb{E}_{Z_{k+1:n}} [g_n(Z_{1:k-1}, z', Z_{k+1:n})] \right| \\ &\leq \sigma_k \end{aligned}$$

Then, we can plug it into the Azuma-Hoeffding inequality and conclude the proof. \square

2.5 Functional bounds

Thm. 10 (Lipschitz functions) If $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is L -Lipschitz function for vector z with edclidean norm, $z_i \sim \mathcal{N}(0, \sigma^2)$.

$$P(g(z) - \mathbb{E}g(z) \geq t) \leq e^{-\frac{ct^2}{L^2 \sigma^2}}$$

3 Uniform Laws of Large Numbers (Ch. 4)

3.1 Motivation

Def. 7 (Glivenko-Cantelli class) We say \mathcal{F} is Glivenko-Cantelli class if

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X_i) \right| \rightarrow 0,$$

in probability as $n \rightarrow \infty$.

3.2 A uniform law via Rademacher complexity

Thm. 11 (Uniform law of large number) For b -uniformly bounded \mathcal{F} , we have

$$P \left(\sup_{f \in \mathcal{F}} R(f) - R_n(f) \geq 2\mathcal{R}_n(\mathcal{F}) + t \right) \leq e^{-\frac{nt^2}{2b^2}}.$$

Here, population risk is $R(f) = \mathbb{E}f(x)$ and empirical risk is $R_n(f) = \frac{1}{n} \sum_{i=1}^n f(x_i)$. The Rademacher complexity is

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\epsilon} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(z_i), \quad \epsilon_i \sim \text{Red}(\{-1, 1\})$$

Proof. 1. **Concentration around mean.** Let's denote

$$g_n(x_{1:n}) := \sup_{f \in \mathcal{F}} R(f) - R_n(f) = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}f(x).$$

Let $z = (x_1, \dots, x_k, \dots, x_n)$ and $z^{\setminus k} = (x_1, \dots, x'_k, \dots, x_n)$,

$$\begin{aligned} & \left| g_n(z) - g_n(z^{\setminus k}) \right| \\ &= \left| \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n [h(z_i) - \mathbb{E}h] - \sup_{\tilde{h} \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n [\tilde{h}(z_i^{\setminus k}) - \mathbb{E}\tilde{h}] \right| \\ &\leq \left| \sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^n h(z_i) - h(z_i^{\setminus k})}{n} \right| \leq \left| \sup_{h \in \mathcal{H}} \frac{h(x_k) - h(x'_k)}{n} \right| \leq \frac{2b}{n}, \end{aligned}$$

which means g_n has Lipschitz property. Using McDiarmid inequality, $P(g_n(X) - \mathbb{E}g_n(X) \geq t) \leq e^{-\frac{nt^2}{2b^2}}$.

2. **Upper bound on mean.** Using symmetrization technique and the definition of Rademacher complexity, we have

$$\begin{aligned} \mathbb{E}g_n(X) &= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}f(x) \right] \\ &\leq \mathbb{E}_{X,Y} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(x_i) - f(y_i) \right] \quad (\text{symmetrize}) \\ &= \mathbb{E}_{X,Y,\epsilon} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon(f(x_i) - f(y_i)) \right] \quad (\text{plug in } \epsilon) \\ &\leq 2\mathbb{E}_{X,\epsilon} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon f(x_i) \right] = 2\mathcal{R}_n(\mathcal{F}). \end{aligned}$$

Combining these two parts, we have the desired theorem that $P(g_n(X) - 2\mathcal{R}_n(\mathcal{F}) \geq t) \leq e^{-\frac{nt^2}{2b^2}}$. \square

Thm. 12 (Uniform law - corollary) If $\mathcal{R}_n(\mathcal{F}) = o(1)$, then when $n \rightarrow \infty$, we have

$$\sup_{f \in \mathcal{F}} R(f) - R_n(f) \rightarrow 0, \quad (a.s.)$$

In other words, $\mathcal{R}_n(\mathcal{F}) = o(1)$ implies that \mathcal{F} is a Glivenko-Cantelli class.

Proof. Let $\mathcal{E}_n(\alpha)$ denote the event that $\sup_{f \in \mathcal{F}} R(f) - R_n(f) \geq \alpha$. The uniform law shows that

$$P(\mathcal{E}_n(2\mathcal{R}_n(\mathcal{H}) + t)) \leq e^{-\frac{nt^2}{2b^2}}.$$

Using union bound, $\forall t > 0$, we have,

$$\begin{aligned} \sum_{n=1}^{\infty} P(\mathcal{E}_n(2t)) &\leq \sum_{n=1}^k P(\mathcal{E}_n(2t)) + \sum_{n=k+1}^{\infty} P(\mathcal{E}_n(2\mathcal{R}_n(\mathcal{H}) + t)) \\ &\leq \sum_{n=1}^k P(\mathcal{E}_n(2t)) + \sum_{n=k+1}^{\infty} e^{-\frac{nt^2}{2b^2}} < \infty. \end{aligned}$$

The k is chosen such that when $n > k$, $\mathcal{R}_n(f) < \frac{t}{2}$. Such k must exist, since $\mathcal{R}_n(f) = o(1)$. Using the Borel-Cantelli lemma, we show that

$$P\left(\lim_{n \rightarrow \infty} \sup_{f \in \mathcal{F}} \mathcal{E}_n(2t)\right) = 0.$$

Considering that t can be arbitrarily small, we have

$$P\left(\lim_{n \rightarrow \infty} \sup_{f \in \mathcal{F}} R(f) - R_n(f) = 0\right) = 1.$$

\square

3.3 Upper bounds on the Rademacher complexity

Def. 8 (Growth function) The growth function $N_{\mathcal{H}}$ for a hypothesis space \mathcal{H} is defined as: $\forall m \in \mathbb{N}$,

$$N_{\mathcal{H}}(n) := \sup_{z_{1:n} \subseteq Z} |\{(h(z_1), \dots, h(z_n)) \mid h \in \mathcal{H}\}|$$

Remark: Growth function measures the maximum number of distinct ways in which m points can be classified using hypotheses in \mathcal{H} .

Thm. 13 (Massart lemma) For a data set $z_{1:n} = \{z_1, \dots, z_n\}$, the hypothesis $h : Z \rightarrow \{0, 1\}$ and the hypothesis space $\mathcal{H}(z_{1:n}) := \{(h(z_1), \dots, h(z_n)) \mid h \in \mathcal{H}\}$, we have

$$\mathcal{R}_n(\mathcal{H}) := \mathbb{E}_{\epsilon} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(z_i) \right] \leq 2\sqrt{\frac{\log |\mathcal{H}(z_{1:n})|}{n}}$$

Remark: Using this result, we can now bound the Rademacher complexity in terms of the growth function.

Proof. First, we show that $\epsilon^T \theta$ is \sqrt{n} sub-Gaussian. Using the independence of z , $\forall \lambda \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}_\epsilon \left[\exp(\lambda \epsilon^T \theta) \right] &= \mathbb{E}_\epsilon \left[\exp \left(\lambda \sum_{i=1}^n \epsilon_i \theta_i \right) \right] \\ &= \mathbb{E}_{\epsilon_1} [\exp(\lambda \epsilon_1 \theta_1)] \cdots \mathbb{E}_{\epsilon_n} [\exp(\lambda \epsilon_n \theta_n)] \\ &\leq \exp \left(\frac{\lambda^2 \theta_1^2}{2} + \cdots + \frac{\lambda^2 \theta_n^2}{2} \right) \leq \exp \left(\frac{\lambda^2 n}{2} \right).\end{aligned}$$

Next, using the Gaussian maxima, we get

$$\mathbb{E}_{\theta \in \mathbb{T}} \epsilon^T \theta \leq 2\sqrt{n \log |\mathcal{H}(z_1^n)|}$$

Then, we can get the intended results. \square

Def. 9 (VC Dimension) The VC dimension of \mathcal{H} is the biggest $n \in \mathbb{N}$ such that there exists n samples in \mathcal{H} which can be arbitrarily scattered by a binary classifier, i.e.,

$$d_{VC} = \max_{n \in \mathbb{N}} n, \quad \text{s.t. } \exists z_{1:n} \in Z^n, \mathcal{H}(z_{1:n}) = \{0, 1\}^n.$$

Here, $\mathcal{H}(z_{1:n}) := \{(h(z_1), \dots, h(z_n)) \mid h \in \mathcal{H}\}$

Remark: Finite VC dimension can make \mathcal{H} a Glivenko-Cantelli class.

Thm. 14 (Sauer-Shelah) For a space \mathcal{H} with VC dimension d_{VC} , for any z_1, \dots, z_n , we have growth function

$$N_{\mathcal{H}}(n) := \sup_{z_{1:n} \in Z^n} |\mathcal{H}(z_{1:n})| \leq (n+1)^{d_{VC}}, \quad \forall n \geq d_{VC}$$

Proof. Proof by combination algebra. See the Chapter 4.3 for more details. \square

Thm. 15 (Rademacher contraction) For any $\mathbb{T} \subseteq \mathbb{R}^n$ and $\ell : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with univariate L -Lipschitz functions it holds that

$$\tilde{\mathcal{R}}_n(\ell \circ \mathbb{T}) \leq L \tilde{\mathcal{R}}_n(\mathbb{T})$$

Proof. See the slides of Lecture 5. \square

4 Non-uniform Learnability (Lec. 6-7)

4.1 Structural risk minimization (SRM)

Def. 10 (SRM) Say we have a nested family of function spaces $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \dots$, where $\mathcal{H} = \bigcup_k \mathcal{H}_k$. For each \mathcal{H}_k define the event

$$E_{k,h} = \left\{ R(h) - R_n(h) \leq c \sqrt{\frac{\log(1/\delta_k)}{n}} + 2\mathcal{R}_n(\mathcal{H}_k) \right\}.$$

Thm. 16 (Uniform law via the SRM) Define $k(h) = \min\{k \mid h \in \mathcal{H}_k\}$ which for each h finds the minimum set \mathcal{H}_k .

If $P(\bigcap_{h \in \mathcal{H}_k} E_{k,h}) \geq 1 - \delta_k$ for each k and if $\sum_k \delta_k \leq \delta$, with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}} R(h) - R_n(h) \leq c \sqrt{\frac{\log(1/\delta_{k(h)})}{n}} + 2\mathcal{R}_n(\mathcal{H}_{k(h)})$$

Proof. Observe that

$$\begin{aligned}A &:= \left\{ \sup_{h \in \mathcal{H}} R(h) - R_n(h) \leq c \sqrt{\frac{\log(1/\delta_{k(h)})}{n}} + 2\mathcal{R}_n(\mathcal{H}_{k(h)}) \right\} \\ &= \bigcap_{h \in \mathcal{H}} \bigcap_{k: h \in \mathcal{H}_k} E_{k,h} = \bigcap_{k \in \mathbb{N}} \bigcap_{h \in \mathcal{H}_k} E_{k,h}\end{aligned}$$

Then, we can use the union bound,

$$P(A) = 1 - P(\bigcup_{k \in \mathbb{N}} \bigcup_{h \in \mathcal{H}_k} E_{k,h}^c) \geq 1 - \sum_k \delta_k \geq 1 - \delta,$$

which concludes the proof. \square

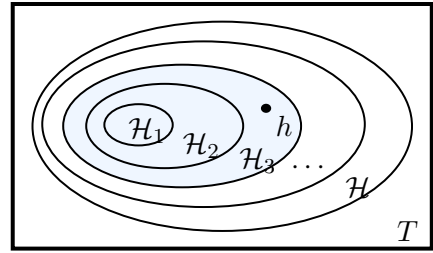


Figure: Demonstration of SRM when $k(h) = 3$.

4.2 Margin bound for linear classifiers

Def. 11 (Linear classifiers) Define the empirical risk

$$R_n^\gamma(f) = \frac{1}{n} \sum_{i=1}^n 1_{y_i f(x_i) \leq \gamma},$$

and the population risk

$$R^\gamma(f) := \mathbb{E}_{X,Y} 1_{y_i f(x_i) \leq \gamma}.$$

Thm. 17 (Rademacher complexity of L_2 -bounded linear class) For a class of linear functions $\mathcal{F}_{B,2} = \{f(x) = \langle w, x \rangle : \|w\|_2 \leq B\}$, we have

$$\mathcal{R}_n(\mathcal{F}_{B,2}) \leq \frac{B \max_i \|x_i\|_2}{\sqrt{n}}$$

Proof. Using the definition of Rademacher complexity and

Cauchy's inequality, we have

$$\begin{aligned}
n \cdot \mathcal{R}_n(\mathcal{F}_{B,2}) &= \mathbb{E}_\sigma \sup_{f \in \mathcal{F}_{B,2}} \sum_{i=1}^n \sigma_i f(x_i) \\
&= \mathbb{E}_\sigma \sup_{\|w\|_2 \leq B} \sum_{i=1}^n \sigma_i \langle w, x_i \rangle \\
&= \mathbb{E}_\sigma \sup_{\|w\|_2 \leq B} \left\langle w, \sum_{i=1}^n \sigma_i x_i \right\rangle \\
&\leq B \mathbb{E}_\sigma \sqrt{\left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2} \quad (\text{Cauchy's}) \\
&\leq B \sqrt{\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i x_i \right\|_2^2} \quad (\text{Jensen's})
\end{aligned}$$

Finally, since the Rademacher variables σ are independent, we have

$$\begin{aligned}
\mathbb{E}_\sigma \left\| \sum_{i=1}^m \sigma_i x_i \right\|_2^2 &= \mathbb{E}_\sigma \sum_{i,j} \sigma_i \sigma_j \langle x_i, x_j \rangle \\
&= \sum_{i \neq j} \langle x_i, x_j \rangle \mathbb{E}_\sigma [\sigma_i \sigma_j] + \sum_{i=1}^m \langle x_i, x_i \rangle \mathbb{E}_\sigma [\sigma_i^2] \\
&= \sum_{i=1}^m \|x_i\|_2^2 \leq n \max_i \|x_i\|_2^2.
\end{aligned}$$

Plug into the previous inequality and we get the result. \square

Thm. 18 (Non-uniform margin bound) If the assumptions are valid for any fixed γ , with probability at least $1 - \delta$, for any $f \in \mathcal{F}_B$, we have

$$R^0(f) = P(y \neq \text{sign}(f(x))) \leq R_n^\gamma(f) + \frac{2DB}{\gamma\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)}{n}}$$

Proof. Please refer to Lec. 7 and Exercise class 1. \square

4.3 Margin bounds for SVM

Thm. 19 (Non-uniform margin bound for SVM) If the assumptions are valid for any fixed γ , with probability at least $1 - \delta$, for any $f \in \mathcal{F}_B$, we have

$$P(y \neq \text{sign}(f(x))) \leq R_n^\gamma(f) + \frac{2D\|w^*\|_2}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)}{n}}$$

Proof. Using the margin bound theorem (Thm. 18) with $\gamma = 1$ then yields the result. \square

Thm. 20 (Uniform margin bound for SVM)

$$\begin{aligned}
\mathbb{P}(yf_{SVM}(x) < 0) &\leq \frac{2eD\|w_{SVM}\|_2}{\sqrt{n}} \\
&\quad + c\sqrt{\frac{\log(1/\delta) + \log(4\log\|w_{SVM}\|_2)}{n}}
\end{aligned}$$

Proof. Choose $B_k = e^k$ and the nested function space is $\mathcal{F}_{B_k} := \{w \mid \|w\| \leq B_k\}$. According to the Non-uniform margin bound for SVM (Thm. 19), let $\delta_k = \frac{\delta}{2k^2}$. Then $k(w) = \log\|w\|$ and thus $B_{k(w)} = \|w\|e$ and

$$\frac{1}{\delta_{k(w)}} = \frac{2(k(w))^2}{\delta} \leq \frac{2(2\log\|w\|)^2}{\delta}.$$

Plugging in the quantities yields the results with probability at

$$1 - \sum_{i=1}^{\infty} \delta_k = 1 - \delta \sum_{i=1}^{\infty} \frac{1}{2k^2} \geq 1 - \delta,$$

which concludes the proof. \square

5 Metric Entropy (Ch. 5)

5.1 Covering and Packing

Def. 12 (Covering number) A δ -cover of a set \mathbb{T} with a metric ρ is a set $\{\theta^1, \dots, \theta^N\} \subset \mathbb{T}$, such that $\forall \theta \in \mathbb{T}, \exists i \in [N], \rho(\theta, \theta^i) \leq \delta$. The covering number $N(\delta; \mathbb{T})$ is the cardinality of the smallest δ -cover.

Def. 13 (Packing number) A δ -cover of a set \mathbb{T} with a metric ρ is a set $\{\theta^1, \dots, \theta^N\} \subset \mathbb{T}$, such that $\forall \theta \in \mathbb{T}, \exists i \in [N], \rho(\theta, \theta^i) \leq \delta$.

The covering number $N(\delta; \mathbb{T})$ is the cardinality of the smallest δ -cover.

Below are some common spaces' complexity.

Space	Rademacher C.	Gaussian C.
$B_1^d(1)$	1	$\sqrt{2\log d} \pm o(1)$
$B_2^d(1)$	\sqrt{d}	$\sqrt{d} - o(1)$
$B_q^d(1), q > 1$	-	$\sqrt{\frac{2}{\pi}} d^{1-1/q} \sim c_q d^{1-1/q}$

5.2 Metric entropy and sub-Gaussian processes

Def. 14 (sub-Gaussian processes) For zero-mean random variables $\{X_\theta, \theta \in \mathbb{T}\}$, we say it is a sub-Gaussian process with metric $\rho_X(\cdot, \cdot)$ on \mathbb{T} if

$$\mathbb{E}[e^{\lambda(X_\theta - X_{\theta'})}] \leq e^{\frac{\lambda^2 \rho_X^2(\theta, \theta')}{2}}, \quad \forall \lambda \in \mathbb{R} \text{ and } \theta, \theta' \in \mathbb{T}.$$

Remark: It can be shown that $X_\theta - X_{\theta'}$ is a sub-Gaussian RV with parameter $\sigma = \sup_{\theta, \theta'} \rho_X(\theta, \theta')$.

Thm. 21 (One-step discretization bound) For a zero-mean sub-Gaussian process $\{X_\theta, \theta \in \mathbb{T}\}$ with $\rho(\theta, \theta')$, we have an upper bound for any $\delta \in [0, \sigma]$,

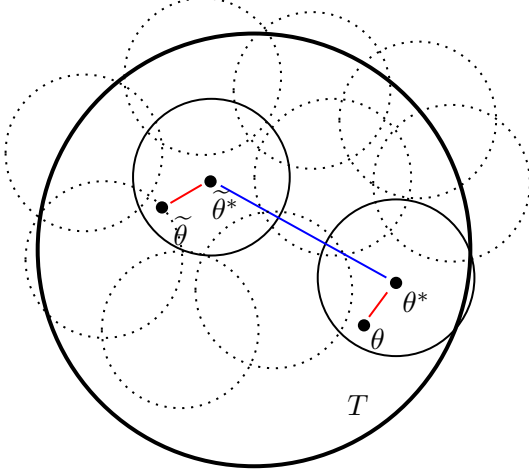
$$\mathbb{E} \sup_{\theta, \theta' \in \mathbb{T}} X_\theta - X_{\theta'} \leq 2\mathbb{E} \sup_{\substack{\theta, \theta' \in \mathbb{T} \\ \rho(\theta, \theta') \leq \delta}} (X_\theta - X_{\theta'}) + 4\sigma \sqrt{\log N(\delta)},$$

where $\sigma = \sup_{\theta, \theta' \in \mathbb{T}} \rho(\theta, \theta')$.

Proof. Let $\{\theta_1, \dots, \theta_N\}$ be a σ -cover of T . For any $\theta \in \mathbb{T}$, we can find at least one θ^* in the cover, s.t. $\rho(\theta, \theta^*) \leq \delta$, (the same for $\tilde{\theta}^*$), and hence,

$$\begin{aligned} X_\theta - X_{\tilde{\theta}} &= X_\theta - X_{\theta^*} + X_{\theta^*} - X_{\tilde{\theta}^*} + X_{\tilde{\theta}^*} - X_{\tilde{\theta}} \\ &\leq 2 \sup_{\rho(\theta, \theta') \leq \delta} (X_\theta - X_{\theta'}) + \max_{i,j=1, \dots, N} |X_{\theta_i} - X_{\theta_j}| \end{aligned}$$

Below is a figure to illustrate it.



Then, take the expectation on both sides. We use the sub-Gaussian maxima (Thm. 6) on $|X_{\theta_i} - X_{\theta_j}|$, which has parameter at most $\rho(\theta_i, \theta_j) \leq \sigma$. Therefore,

$$\mathbb{E} \sup_{\theta, \theta' \in \mathbb{T}} X_\theta - X_{\theta'} \leq 2 \mathbb{E} \sup_{\substack{\theta, \theta' \in \mathbb{T} \\ \rho(\theta, \theta') \leq \delta}} (X_\theta - X_{\theta'}) + 4\sigma \sqrt{\log N(\delta)}$$

□

Thm. 22 (One-step discretization bound - corollary) Let $X_\theta = \frac{1}{n} \sum_{i=1}^n \epsilon_i \theta_i$, where ϵ are Rademacher RVs. Then $\{X_\theta, \theta \in \mathbb{T}\}$ is a sub-Gaussian process with $\rho(\theta, \theta') = \frac{\|\theta - \theta'\|_2}{\sqrt{n}}$. We have an upper bound as

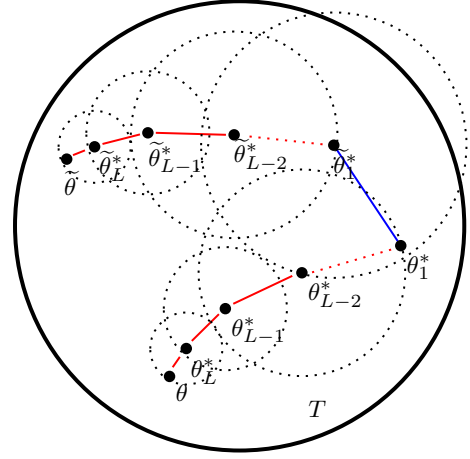
$$\mathcal{R}_n(T) \leq \frac{1}{\sqrt{n}} \mathbb{E} \sup_{\theta, \theta' \in \mathbb{T}} X_\theta - X_{\theta'} \leq \frac{2}{\sqrt{n}} [\delta \sqrt{n} + 2\sigma \sqrt{\log N(\delta)}],$$

Remark: This provides a usage of sub-Gaussian process, to bound the Rademacher complexity.

Thm. 23 (Dudley's integral) Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean sub-Gaussian process with a metric ρ . Define $D = \sup_{\theta, \theta'} \rho(\theta, \theta')$.

$$\mathbb{E} \sup_{\theta, \theta'} X_\theta - X_{\theta'} \leq 2 \mathbb{E} \sup_{\rho(\theta, \theta') \leq D} X_\theta - X_{\theta'} + 16 \int_{\delta/4}^D \sqrt{\log N(t)} dt.$$

Remark: The Dudley's integral (sometimes) achieves a tighter bound on the last term of $4\sigma \sqrt{\log N(\delta)}$ by chaining method.



Proof. Define $L = \log_2 \frac{D}{\delta}$ sets of δ_i -covers \mathcal{C}_i , where $\delta_i = \frac{1}{2^i} D$. Let $N(\delta_i)$ denote the covering number in \mathcal{C}_i .

$$\begin{aligned} X_{\theta_L} - X_{\tilde{\theta}_L} &\leq X_{\theta_L} - X_{\theta_{L-1}^*} + X_{\theta_{L-1}^*} - X_{\tilde{\theta}_{L-1}^*} + X_{\tilde{\theta}_{L-1}^*} - X_{\tilde{\theta}_L} \\ &= 2 \max_{\theta \in \mathcal{C}_L} X_\theta - X_{\theta_{L-1}^*} + \max_{\theta, \theta' \in \mathcal{C}_{L-1}} X_\theta - X_{\theta'} \\ &\dots \\ &\leq 2 \sum_{i=2}^L \max_{\theta \in \mathcal{C}_i} X_\theta - X_{\theta_{i-1}^*} + \max_{\theta, \theta' \in \mathcal{C}_1} X_\theta - X_{\theta'} \end{aligned}$$

Next, we use sub-Gaussian maxima,

$$\begin{aligned} \mathbb{E} \max_{\theta \in \mathcal{C}_i} X_\theta - X_{\theta_{i-1}^*} &\leq 2\delta_{i-1} \sqrt{\log |\mathcal{C}_i|} \leq 2 \cdot \frac{D}{2^{i-1}} \sqrt{\log N\left(\frac{D}{2^{i-1}}\right)} \\ &\leq 8 \int_{D/2^{i-1}}^{D/2^i} \sqrt{\log N(t)} dt \end{aligned}$$

Putting things together, we get

$$\begin{aligned} \mathbb{E} \max_{\theta, \tilde{\theta} \in \mathcal{C}_L} X_\theta - X_{\tilde{\theta}} &\leq 16 \sum_{i=2}^L \int_{\frac{D}{2^{i-1}}}^{\frac{D}{2^i}} \sqrt{\log N(t)} dt + 2D \sqrt{\log N\left(\frac{D}{2}\right)} \\ &\leq 16 \int_{\delta/4}^D \sqrt{\log N_{\mathbb{T}}(t)} dt \end{aligned}$$

This gives the desired form. □

6 Reproducing Kernel Hilbert Spaces (Ch. 12)

6.1 Basics of Hilbert space

Def. 15 (Hilbert spaces) A Hilbert space \mathcal{H} is a complete inner product space. In a Hilbert space

- There endows an inner product: $\langle \cdot, \cdot \rangle_{\mathcal{H}}$,
- Every Cauchy sequence $(f_n)_{n=1}^\infty$ in \mathcal{H} converges to some element $f^* \in \mathcal{H}$.

Thm. 24 (Riesz representation) Let L be a bounded linear functional on a Hilbert space \mathcal{H} . Then there exists a unique representer $g \in \mathcal{H}$ such that $L(f) = \langle f, g \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

Space	Kernel	Eigenfunctin
2-poly.	$K(x, z) = (1 + xz)^2$	$\phi_j(x) = a_0^j + a_1^j x + a_2^j x^2$, const via eigen-decomposition.
1-Sobolev $W_2^1([0, 1])$	$K(x, z) = \min\{x, z\}$	$\phi_j(t) = \sin \frac{(2j-1)\pi t}{2}$, $\mu_j = \left(\frac{2}{(2j-1)\pi}\right)^2$

Proof. Consider a null space $\text{Null}(L) := \{h \mid L(h) = 0\}$. If $\text{Null}(L) = H$, then $\text{Null}(L)^\perp = \{0\}$, we take $g = 0$.

In a non-trivial case (i.e., $\text{Null}(L)^\perp \neq \{0\}$), there exist a non-zero element $g \in \text{Null}(L)^\perp$ such that $\|g\|_{\mathcal{H}} = L(g)$. Define $h := L(f)g - L(g)f$, then we note

$$L(h) = L(f)L(g) - L(g)L(f) = 0,$$

which means $h \in \text{Null}(L)$. Therefore, we have $h \perp g$, i.e.,

$$0 = \langle h, g \rangle_{\mathcal{H}} = L(f) \|g\|_{\mathcal{H}}^2 - L(g) \langle f, g \rangle_{\mathcal{H}}.$$

This implies $L(f) = \langle f, g \rangle_{\mathcal{H}}$. \square

6.2 Reproducing kernel Hilbert space

Def. 16 (Kernel function) Given a feature map $\phi : \mathcal{X} \rightarrow \mathcal{H}$ and the Hilbert space \mathcal{H} that ϕ maps to, the kernel function is defined as

$$K(x, y) := \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}.$$

Def. 17 (RKHS - defined by kernel) A reproducing kernel Hilbert space is a Hilbert space \mathcal{H} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ with a kernel function $K(\cdot, \cdot)$ that

- For any $x \in \mathcal{X}$, $K(\cdot, x) \in \mathcal{H}$,
- Satisfies the “reproducing property”:

$$\langle f(\cdot), K(\cdot, x) \rangle_{\mathcal{H}} = f(x), \quad \forall f \in \mathcal{H}.$$

Thm. 25 (RKHS from kernel function) Given any positive semi-definite kernel function $K(\cdot, \cdot)$, there is a **unique** Hilbert space \mathcal{H} in which the kernel $K(\cdot, \cdot)$ satisfies the reproducing property.

Given some data $\{x_i\}_{i=1}^n$, such Hilbert space \mathcal{H} is

$$\mathcal{H} := \left\{ f(\cdot) = \sum_{i=1}^n \alpha_i K(\cdot, x_i) \mid x_i \in \mathcal{X} \right\}$$

with the norm

$$\langle f, f' \rangle_{\mathcal{H}} := \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha'_j K(x_i, x_j)$$

Remark: For a fixed kernel function, there will be (infinitely) many feature maps, and thus many Hilbert spaces of the feature. But the Hilbert space that **satisfies the reproducing property** is unique!

Def. 18 (RKHS - defined by evaluation functional) A reproducing kernel Hilbert space is a Hilbert space \mathcal{H} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that for each $x \in \mathcal{X}$, the evaluation functional $L_x : \mathcal{H} \rightarrow \mathbb{R}$ that performs the operation $L_x(f) = f(x)$ is bounded, i.e.,

$$f(x) = |L_x(f)| \leq M \|f\|_{\mathcal{H}}, \quad \exists M < \infty, \forall f \in \mathcal{H}.$$

Proof. Let's prove the equivalence to the first definition. When L_x is a bounded linear functional, the Riesz theorem shows that there exists a unique $R_x \in \mathcal{H}$ such that

$$L_x(f) = \langle f, R_x \rangle_{\mathcal{H}}.$$

Similarly, we can get a unique R_y based on y . The kernel is defined as $K(x, y) = \langle R_x, R_y \rangle_{\mathcal{H}}$. Next, we can verify that K is positive semidefinite.

$$\begin{aligned} \alpha^T K \alpha &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) = \left\langle \sum_{i=1}^n \alpha_i R_{x_i}, \sum_{j=1}^n \alpha_j R_{x_j} \right\rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^n \alpha_i R_{x_i} \right\|_{\mathcal{H}}^2 \geq 0. \end{aligned}$$

\square

6.3 Mercer's theorem and its consequences

Thm. 26 (Mercer's) For a continuous and PSD kernel function K that satisfies the Hilbert-Schmidt condition. Then there exist a sequence of eigenfunctions $(\phi_i)_{i=1}^\infty$ that form an orthonormal basis of $L^2(\mathcal{X}; P)$ and non-negative eigenvalues $(\mu_i)_{i=1}^\infty$ such that

$$T_K(\phi_i) = \mu_i \phi_i, \quad \forall i = 1, 2, \dots \quad (6.1)$$

Moreover, the kernel function has the expansion

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\ell^2(\mathbb{N})} := \sum_{i=1}^{\infty} \mu_i \phi_i(x) \phi_i(y)$$

7 Non-parametric Least Squares (Ch. 13)

7.1 Fixed design

Def. 19 (Least square regression) For a function f^* and data collection $\{(x_i, y_i)\}_{i=1}^n$, where $y_i = f^*(x_i) + \sigma w_i$, $w_i \sim \mathcal{N}(0, 1)$, the least-squares estimator or estimator is given by

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2.$$

Prop. 4 (Basic inequality) In the non-parametric least squares, we have the optimality of \hat{f} , which means

$$\frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - y_i)^2 \leq \frac{1}{n} \sum_{i=1}^n (f^*(x_i) - y_i)^2$$

Or, equivalently,

$$\|f - f^*\|_n^2 \leq \frac{2\sigma}{n} \sum_{i=1}^n w_i (f(x_i) - f^*(x_i)).$$

Proof. Let's prove the equivalence. Note that $y_i = f^*(x_i) + \sigma w_i$, we have

$$\begin{aligned} 0 &\geq \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - y_i)^2 - \frac{1}{n} \sum_{i=1}^n (f^*(x_i) - y_i)^2 \\ &\geq \frac{1}{n} \sum_{i=1}^n ((\hat{f}(x_i)^2 - f^*(x_i)^2) - 2y_i(\hat{f}(x_i) - f^*(x_i))) \\ &\geq \frac{1}{n} \sum_{i=1}^n ((\hat{f}(x_i) - f^*(x_i))^2 - 2\sigma w_i(\hat{f}(x_i) - f^*(x_i))), \end{aligned}$$

which means

$$\frac{1}{n} \sum_{i=1}^n ((\hat{f}(x_i) - f^*(x_i))^2) \leq \frac{2\sigma}{n} \sum_{i=1}^n (w_i(\hat{f}(x_i) - f^*(x_i))).$$

□

Def. 20 (Localized Gaussian complexity) If we restrict the radius of function space \mathcal{F} , i.e., $\|f\|_n \leq \delta$, we have the localized Gaussian complexity for \mathcal{F} ,

$$\mathcal{G}_n(\delta; \mathcal{F}) := \mathbb{E}_{w \sim \mathcal{N}(0,1)} \left[\sup_{f \in \mathcal{F}, \|f\|_n \leq \delta} \frac{1}{n} \sum_{i=1}^n w_i f(x_i) \right]$$

Def. 21 (Critical radius) For a local Gaussian complexity \mathcal{G}_n around f^* with radius δ , the radius δ is said to be *valid* if the following *critical inequality* satisfies,

$$\frac{\mathcal{G}_n(\delta; \mathcal{F}^*)}{\delta} \leq \frac{\delta}{2\sigma}$$

The smallest δ satisfies it is the *critical radius*, which must exist for any star-shaped function class \mathcal{F} .

Remark: One intuition is that the satisfaction of critical inequality means the optimality of \hat{f} . One example is that if we have the optimality of \hat{f} , i.e.,

$$\frac{1}{2n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \leq \frac{1}{2n} \sum_{i=1}^n (y_i - f^*(x_i))^2$$

And let $\delta = \mathbb{E} \|\hat{f} - f^*\|_n$, then we can show the δ satisfies $\frac{1}{2}\delta^2 \leq \sigma \mathcal{G}_n(\delta; \mathcal{F}^*)$, which is equivalent to the critical inequality.

Thm. 27 (Risk decomposition) In the least-squares regression, where $y_i = f^*(x_i) + \sigma w_i, w_i \sim \mathcal{N}(0, 1)$, we have risk decomposition

$$R(f) = (\mathbb{E}f(x) - \mathbb{E}f^*(x))^2 + \text{var}(f(x)) + \sigma^2$$

Proof. Using the fact $\text{var}(x) = \mathbb{E}x^2 - (\mathbb{E}x)^2$ and the independence of y and $f(x)$, we have

$$\begin{aligned} R(f) &= \mathbb{E}[(y - f(x))^2] \\ &= \mathbb{E}y^2 + \mathbb{E}f(x)^2 - 2\mathbb{E}yf(x) \quad (y \text{ and } f(x) \text{ are indep.}) \\ &= \text{var}(y) + (\mathbb{E}y)^2 + \text{var}(f(x)) - (\mathbb{E}f(x))^2 + 2(\mathbb{E}y)(\mathbb{E}f(x)) \\ &= \text{var}(y) + \text{var}(f(x)) + (\mathbb{E}f(x) - \mathbb{E}y)^2 \\ &= \underbrace{(\mathbb{E}f(x) - \mathbb{E}f^*(x))^2}_{\text{bias}^2} + \underbrace{\text{var}(f(x))}_{\text{variance}} + \underbrace{\sigma^2}_{\text{irreducible noise}}. \end{aligned}$$

If there is no variance in model, we can also get the decomposition as,

$$\begin{aligned} R(f) &= \mathbb{E} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \\ &= \mathbb{E} \frac{1}{n} \sum_{i=1}^n (f^*(x_i) + \sigma w_i - f(x_i))^2 \\ &= \mathbb{E} \frac{1}{n} \sum_{i=1}^n (f^*(x_i) - f(x_i))^2 + \mathbb{E} \frac{1}{n} \sum_{i=1}^n \sigma^2 w_i^2 \\ &\quad + 2\mathbb{E} \frac{1}{n} \sum_{i=1}^n w_i (f^*(x_i) - f(x_i)) \\ &= \|f - f^*\|_n^2 + \sigma^2 \end{aligned}$$

□

Thm. 28 (Prediction error bound) If \mathcal{F}^* is star-shaped, then for any δ satisfies the critical inequality and $t \geq \delta$, the non-parametric least square estimate \hat{f}_n satisfies

$$P\left(\|\hat{f}_n - f^*\|_n^2 \geq 16t\delta_n\right) \leq \exp\left(-\frac{nt\delta_n}{2\sigma^2}\right),$$

which also means for some constant c ,

$$\mathbb{E} \|\hat{f}_n - f^*\|_n^2 \leq c \left(\delta_n^2 + \frac{\sigma^2}{n} \right).$$

Remark: This theorem links the local Gaussian complexity to the error bounds.

Proof. Let $\hat{\Delta} = f - f^*$, and thus $\|\hat{\Delta}\| = \|f - f^*\|_n^2$. Our goal is to show for any $\|\hat{\Delta}\|_n \geq \sqrt{t}\delta_n$,

$$\|\hat{\Delta}\|_n^2 \leq \frac{2\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) \stackrel{(i)}{\leq} 4\|\hat{\Delta}\|_n \sqrt{t}\delta_n$$

with probability $1 - e^{-\frac{nt\delta_n^2}{2\sigma^2}}$. The theorem follows from rearranging terms. We organize the proof in multiple steps.

For a given scalar $u \geq \delta_n$, define the event

$$\mathcal{A}(u) := \left\{ \exists g \in \mathcal{F}^* \cap \{\|g\|_n \geq u\} \mid \left| \frac{\sigma}{n} \sum_{i=1}^n \tilde{w}_i g(x_i) \right| \geq 2u \|g\|_n \right\}$$

□

Thm. 29 (Bounds via metric entropy) Any $\delta \in (0, \sigma]$ such that

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n(t; \mathbb{B}_n(\delta; F^*))} dt \leq \frac{\delta^2}{4\sigma}$$

satisfies the critical inequality.

7.2 Error bounds for RKHS

Def. 22 (R -restrained space) The function space inside the radius R is

$$\mathcal{F}_R := \{f \in \mathcal{F} \mid \|f\|_{\mathcal{F}} \leq R\}$$

Thm. 30 (Localized Gaussian complexity for RKHS) Let an RKHS \mathcal{F} with kernel function $K(\cdot, \cdot)$. Defining $\hat{\mu}_j$ as eigenvalues of the kernel matrix K , we have the local Gaussian complexity bounded by

$$\mathcal{G}_n(\delta; \mathcal{F}_R) \leq \sqrt{\frac{R+1}{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}}$$

Note that the localized Gaussian complexity is computed under $\|f\|_{\mathcal{F}} \leq R, \|f\|_n \leq \delta$.

Thm. 31 (Prediction error for norm-bounded RKHS)

When $\lambda_n \geq 2\delta_{n;R}^2$ there is universal constants c_0, c_1, c_3 such that

$$P\left(\left\|\hat{f}_{\lambda_n} - f^*\right\|_n^2 \geq cR^2\left(\delta_{n;R}^2 + \lambda_n\right)\right) \leq c_0 e^{-c_1 \frac{nR^2\delta_{n;R}^2}{\sigma^2}}$$

7.3 Random design

WIP...