

Uniform Laws of Large Numbers

Jiuzhou Miao

School of Statistics and Mathematics, Zhejiang Gongshang University

March 19, 2025

- 1 Motivation
- 2 A uniform law via Rademacher complexity
- 3 Upper bounds on the Rademacher complexity

- 1 Motivation
- 2 A uniform law via Rademacher complexity
- 3 Upper bounds on the Rademacher complexity

Empirical cumulative distribution function (ECDF)

- Let X be a random variable with cumulative distribution function (CDF) $F(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$ and $\{X_i\}_{i=1}^n$ be independent samples which have same distribution with X .

Empirical cumulative distribution function (ECDF)

- Let X be a random variable with cumulative distribution function (CDF) $F(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$ and $\{X_i\}_{i=1}^n$ be independent samples which have same distribution with X .
- A natural estimation of F is the ECDF based on $\{X_i\}_{i=1}^n$, given by

$$\hat{F}_n(x) = n^{-1} \sum_{i=1}^n \mathbb{I}_{(-\infty, x]}(X_i) = n^{-1} \sum_{i=1}^n \mathbb{I}(X_i \leq x).$$

Empirical cumulative distribution function (ECDF)

- Let X be a random variable with cumulative distribution function (CDF) $F(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$ and $\{X_i\}_{i=1}^n$ be independent samples which have same distribution with X .
- A natural estimation of F is the ECDF based on $\{X_i\}_{i=1}^n$, given by

$$\hat{F}_n(x) = n^{-1} \sum_{i=1}^n \mathbb{I}_{(-\infty, x]}(X_i) = n^{-1} \sum_{i=1}^n \mathbb{I}(X_i \leq x).$$

- For any fixed $x \in \mathbb{R}$, the strong law of large numbers implies that $\hat{F}_n(x) \rightarrow F(x)$ almost surely as $n \rightarrow \infty$.

Empirical cumulative distribution function (ECDF)

- Let X be a random variable with cumulative distribution function (CDF) $F(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$ and $\{X_i\}_{i=1}^n$ be independent samples which have same distribution with X .
- A natural estimation of F is the ECDF based on $\{X_i\}_{i=1}^n$, given by

$$\hat{F}_n(x) = n^{-1} \sum_{i=1}^n \mathbb{I}_{(-\infty, x]}(X_i) = n^{-1} \sum_{i=1}^n \mathbb{I}(X_i \leq x).$$

- For any fixed $x \in \mathbb{R}$, the strong law of large numbers implies that $\hat{F}_n(x) \rightarrow F(x)$ almost surely as $n \rightarrow \infty$.
- A natural goal is to strengthen this pointwise convergence to a form of uniform convergence.

The functionals of CDFs

- In statistical settings, a typical use of the ECDF is to construct estimators of various quantities associated with the (population) CDF.

The functionals of CDFs

- In statistical settings, a typical use of the ECDF is to construct estimators of various quantities associated with the (population) CDF.
- Many such estimation problems can be formulated in a terms of functional γ which maps any CDF F to a real number $\gamma(F)$.

The functionals of CDFs

- In statistical settings, a typical use of the ECDF is to construct estimators of various quantities associated with the (population) CDF.
- Many such estimation problems can be formulated in a terms of functional γ which maps any CDF F to a real number $\gamma(F)$.
- Given a set of samples distributed according to F , the plug-in principle suggests replacing the unknown F by \hat{F}_n , thereby obtaining $\gamma(\hat{F}_n)$ as an estimation of $\gamma(F)$.

Expectation functionals

- Given some integrable function g , define the expectation functional γ_g by

$$\gamma_g(F) = \int g(x) dF(x).$$

Expectation functionals

- Given some integrable function g , define the expectation functional γ_g by

$$\gamma_g(F) = \int g(x) dF(x).$$

- For any g , the plug-in estimator is given by

$$\gamma_g(\hat{F}_n) = \int g(x) d\hat{F}_n(x) = n^{-1} \sum_{i=1}^n g(X_i).$$

Quantile functionals

- For any $\alpha \in [0, 1]$, the quantile functional Q_α is given by

$$Q_\alpha(F) = \inf \{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

Quantile functionals

- For any $\alpha \in [0, 1]$, the quantile functional Q_α is given by

$$Q_\alpha(F) = \inf \{x \in \mathbb{R} : F(x) \geq \alpha\}.$$

- The plug-in estimate is given by

$$Q_\alpha(\hat{F}) = \inf \{x \in \mathbb{R} : \hat{F}_n(x) \geq \alpha\}.$$

Goodness-of-fit functionals

- It is frequently of interest to test the hypothesis of whether or not a given set of data has been drawn from a known distribution F_0 .

Goodness-of-fit functionals

- It is frequently of interest to test the hypothesis of whether or not a given set of data has been drawn from a known distribution F_0 .
- Such tests can be performed using functionals that measure the distance between F and F_0 , including sup-norm distance $\|F - F_0\|_\infty$ and Cramér–von Mises criterion based on the functional

$$\gamma(F) = \int_{-\infty}^{\infty} \{F(x) - F_0(x)\}^2 dF_0(x).$$

The continuity of a functional

- Let F and G be two CDF both defined on \mathbb{R} . Define the sup-norm between them by

$$\|G - F\|_{\infty} = \sup_{x \in \mathbb{R}} |G(x) - F(x)|.$$

Definition 1 (The continuity of a functional)

Let F and G are two CDFs. We say that the functional γ is continuous at F in the sup-norm if for all $\epsilon > 0$, there exists a $\delta > 0$ such that $\|G - F\|_{\infty} \leq \delta$ implies that $|\gamma(G) - \gamma(F)| \leq \epsilon$.

Glivenko-Cantelli's Theorem

- For any continuous functional γ , the consistency question for the plug-in estimator $\gamma(\hat{F}_n)$ can be reduced to the issue of whether or not $\|\hat{F}_n - F\|_\infty$ tends to zero.

Theorem 2 (Glivenko-Cantelli's Theorem)

For any CDF F , as $n \rightarrow \infty$, the ECDF \hat{F}_n is a strongly consistent estimator of F in the uniform norm, i.e.,

$$\|\hat{F}_n - F\|_\infty = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \rightarrow 0$$

almost surely.

Uniform laws for more general function classes

- Let \mathcal{F} be a class of integrable real-valued functions with domain \mathcal{X} and X be a random variable with distribution \mathbb{P} . Let $\{X_i\}_{i=1}^n$ be independent random variables which have same distribution with X .

Uniform laws for more general function classes

- Let \mathcal{F} be a class of integrable real-valued functions with domain \mathcal{X} and X be a random variable with distribution \mathbb{P} . Let $\{X_i\}_{i=1}^n$ be independent random variables which have same distribution with X .
- Define the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^n f(X_i) - \mathbb{E}\{f(X)\} \right|.$$

Glivenko-Cantelli class

Definition 3 (Glivenko-Cantelli class)

We say that \mathcal{F} is \mathbb{P} -Glivenko-Cantelli [or strong \mathbb{P} -Glivenko-Cantelli] if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ converges to zero in probability [or almost surely].

- When $\mathcal{F} = \{\mathbb{I}_{(-\infty, x]}(\cdot) : x \in \mathbb{R}\}$, one has that

$$\mathbb{E}\{\mathbb{I}_{(-\infty, x]}(X)\} = \mathbb{P}(X \leq x)$$

for fixed x , so that the classical Glivenko-Cantelli theorem is equivalent to a strong uniform law for the class \mathcal{F} .

Failure of uniform law

- Let \mathcal{S} be the class of all subsets S of $[0, 1]$ such that the subset S has a finite number of elements. Consider the function class

$$\mathcal{F}_S = \{\mathbb{I}_S(\cdot) : S \in \mathcal{S}\}.$$

Failure of uniform law

- Let \mathcal{S} be the class of all subsets S of $[0, 1]$ such that the subset S has a finite number of elements. Consider the function class

$$\mathcal{F}_S = \{\mathbb{I}_S(\cdot) : S \in \mathcal{S}\}.$$

- Suppose that samples $\{X_i\}_{i=1}^n$ are drawn from some distribution \mathbb{P} over $[0, 1]$ which satisfies that $\mathbb{P}(\{x\}) = 0$ for all $x \in [0, 1]$.

Failure of uniform law

- Let \mathcal{S} be the class of all subsets S of $[0, 1]$ such that the subset S has a finite number of elements. Consider the function class

$$\mathcal{F}_S = \{\mathbb{I}_S(\cdot) : S \in \mathcal{S}\}.$$

- Suppose that samples $\{X_i\}_{i=1}^n$ are drawn from some distribution \mathbb{P} over $[0, 1]$ which satisfies that $\mathbb{P}(\{x\}) = 0$ for all $x \in [0, 1]$.
- This class includes any distribution that has a density with respect to Lebesgue measure. Then $\mathbb{P}(S) = 0$ for all $S \in \mathcal{S}$.

Failure of uniform law

- However, for any positive integer n , the discrete set $\{X_1, \dots, X_n\}$ belongs to \mathcal{S} , which implies that

$$\mathbb{P}[\{X_1, \dots, X_n\}] = 1$$

and

$$\sup_{S \in \mathcal{S}} |\mathbb{P}_n(S) - \mathbb{P}(S)| = 1.$$

Failure of uniform law

- However, for any positive integer n , the discrete set $\{X_1, \dots, X_n\}$ belongs to \mathcal{S} , which implies that

$$\mathbb{P}[\{X_1, \dots, X_n\}] = 1$$

and

$$\sup_{S \in \mathcal{S}} |\mathbb{P}_n(S) - \mathbb{P}(S)| = 1.$$

- \mathcal{F}_S is not \mathbb{P} -Glivenko–Cantelli.

Empirical risk minimization

- Consider an indexed family of probability distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$.

Empirical risk minimization

- Consider an indexed family of probability distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$.
- Let $\{X_i\}_{i=1}^n$ be i.i.d. samples lying in some space \mathcal{X} which are drawn according to \mathbb{P}_{θ^*} , where $\theta^* \in \Theta$ is fixed and unknown.

Empirical risk minimization

- Consider an indexed family of probability distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$.
- Let $\{X_i\}_{i=1}^n$ be i.i.d. samples lying in some space \mathcal{X} which are drawn according to \mathbb{P}_{θ^*} , where $\theta^* \in \Theta$ is fixed and unknown.
- Let $L : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ be a loss function. The quantity

$$R(\theta) = \mathbb{E}_{X \sim \mathbb{P}_{\theta^*}} \{L(X, \theta)\}$$

is called as population risk.

Empirical risk minimization

- Correspondingly, the quantity

$$\hat{R}_n(\theta) = n^{-1} \sum_{i=1}^n L(X_i, \theta)$$

is called as empirical risk.

Empirical risk minimization

- Correspondingly, the quantity

$$\hat{R}_n(\theta) = n^{-1} \sum_{i=1}^n L(X_i, \theta)$$

is called as empirical risk.

- A standard decision-theoretic approach to estimating θ^* is based on minimizing the empirical risk $\hat{R}_n(\theta)$ over $\Theta_0 \subseteq \Theta$, thereby obtaining an estimator $\hat{\theta}$.

Maximum likelihood

- Consider a parameterized family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ that each \mathbb{P}_θ has a strictly positive density $p_\theta(\cdot)$ defined with respect to a common underlying measure.

Maximum likelihood

- Consider a parameterized family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ that each \mathbb{P}_θ has a strictly positive density $p_\theta(\cdot)$ defined with respect to a common underlying measure.
- Let $\{X_i\}_{i=1}^n$ be i.i.d. samples from an unknown distribution \mathbb{P}_{θ^*} and we would like to estimate the unknown θ^* .

Maximum likelihood

- Consider a parameterized family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ that each \mathbb{P}_θ has a strictly positive density $p_\theta(\cdot)$ defined with respect to a common underlying measure.
- Let $\{X_i\}_{i=1}^n$ be i.i.d. samples from an unknown distribution \mathbb{P}_{θ^*} and we would like to estimate the unknown θ^* .
- Consider the loss function

$$L(x, \theta) = \log \{p_{\theta^*}(x)/p_\theta(x)\}.$$

The term $p_{\theta^*}(x)$ has no effect on the minimization over θ .

Maximum likelihood

- The maximum likelihood estimation is obtained by minimizing

$$\begin{aligned}\hat{\theta} &= \arg \min_{\theta \in \Theta_0 \subseteq \Theta} \underbrace{\left[n^{-1} \sum_{i=1}^n \log \{ p_{\theta^*}(X_i) / p_{\theta}(X_i) \} \right]}_{\hat{R}_n(\theta)} \\ &= \arg \min_{\theta \in \Theta_0 \subseteq \Theta} \left[n^{-1} \sum_{i=1}^n \log \{ 1 / p_{\theta}(X_i) \} \right].\end{aligned}$$

Maximum likelihood

- The maximum likelihood estimation is obtained by minimizing

$$\begin{aligned}\hat{\theta} &= \arg \min_{\theta \in \Theta_0 \subseteq \Theta} \underbrace{\left[n^{-1} \sum_{i=1}^n \log \{ p_{\theta^*}(X_i) / p_{\theta}(X_i) \} \right]}_{\hat{R}_n(\theta)} \\ &= \arg \min_{\theta \in \Theta_0 \subseteq \Theta} \left[n^{-1} \sum_{i=1}^n \log \{ 1 / p_{\theta}(X_i) \} \right].\end{aligned}$$

- The population risk is given by

$$R(\theta) = \mathbb{E}_{X \sim \mathbb{P}_{\theta^*}} \left[\log \{ p_{\theta^*}(x) / p_{\theta}(x) \} \right],$$

known as the Kullback-Leibler divergence between p_{θ^*} and p_{θ} .

Bound the excess risk

- The statistical question is how to bound the excess risk

$$\text{ER}(\theta) = R(\hat{\theta}) - \inf_{\theta \in \Theta_0} R(\theta).$$

Bound the excess risk

- The statistical question is how to bound the excess risk

$$\text{ER}(\theta) = R(\hat{\theta}) - \inf_{\theta \in \Theta_0} R(\theta).$$

- For simplicity, assume that there exists some $\theta_0 \in \Theta_0$ such that $R(\theta_0) = \inf_{\theta \in \Theta_0} R(\theta)$.

Bound the excess risk

- The statistical question is how to bound the excess risk

$$\text{ER}(\theta) = R(\hat{\theta}) - \inf_{\theta \in \Theta_0} R(\theta).$$

- For simplicity, assume that there exists some $\theta_0 \in \Theta_0$ such that $R(\theta_0) = \inf_{\theta \in \Theta_0} R(\theta)$.
- With this notation, the excess risk can be decomposed as

$$\text{ER}(\theta) = \underbrace{R(\hat{\theta}) - \hat{R}_n(\hat{\theta})}_{T_{n1}} + \underbrace{\hat{R}_n(\hat{\theta}) - \hat{R}_n(\theta_0)}_{T_{n2}} + \underbrace{\hat{R}_n(\theta_0) - R(\theta_0)}_{T_{n3}}.$$

Bound the excess risk

- The statistical question is how to bound the excess risk

$$\text{ER}(\theta) = R(\hat{\theta}) - \inf_{\theta \in \Theta_0} R(\theta).$$

- For simplicity, assume that there exists some $\theta_0 \in \Theta_0$ such that $R(\theta_0) = \inf_{\theta \in \Theta_0} R(\theta)$.
- With this notation, the excess risk can be decomposed as

$$\text{ER}(\theta) = \underbrace{R(\hat{\theta}) - \hat{R}_n(\hat{\theta})}_{T_{n1}} + \underbrace{\hat{R}_n(\hat{\theta}) - \hat{R}_n(\theta_0)}_{T_{n2}} + \underbrace{\hat{R}_n(\theta_0) - R(\theta_0)}_{T_{n3}}.$$

- Obviously, $T_{n2} \leq 0$.

Bound the excess risk

- Recall that

$$T_{n1} = \mathbb{E}_{X \sim \mathbb{P}_{\theta^*}} \{L(X, \hat{\theta})\} - n^{-1} \sum_{i=1}^n L(X_i, \hat{\theta}),$$

$$T_{n3} = n^{-1} \sum_{i=1}^n L(X_i, \theta_0) - \mathbb{E}_{X \sim \mathbb{P}_{\theta^*}} \{L(X, \theta_0)\}.$$

Bound the excess risk

- Recall that

$$T_{n1} = \mathbb{E}_{X \sim \mathbb{P}_{\theta^*}} \{L(X, \hat{\theta})\} - n^{-1} \sum_{i=1}^n L(X_i, \hat{\theta}),$$

$$T_{n3} = n^{-1} \sum_{i=1}^n L(X_i, \theta_0) - \mathbb{E}_{X \sim \mathbb{P}_{\theta^*}} \{L(X, \theta_0)\}.$$

- Define the function class

$$\mathcal{L}(\Theta_0) = \{x \mapsto L(x, \theta) : \theta \in \Theta_0\},$$

then $T_{n1} + T_{n3}$ is bounded by $2\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{L}(\Theta_0)}$.

- 1 Motivation
- 2 A uniform law via Rademacher complexity
- 3 Upper bounds on the Rademacher complexity

Rademacher complexity of the function class

- Let \mathcal{F} be a function class. For any collection $x_1^n = \{x_1, \dots, x_n\}$, consider the subset of \mathbb{R}^n given by

$$\mathcal{F}(x_1^n) = \left\{ (f(x_1), \dots, f(x_n))^T : f \in \mathcal{F} \right\}.$$

Rademacher complexity of the function class

- Let \mathcal{F} be a function class. For any collection $x_1^n = \{x_1, \dots, x_n\}$, consider the subset of \mathbb{R}^n given by

$$\mathcal{F}(x_1^n) = \left\{ (f(x_1), \dots, f(x_n))^T : f \in \mathcal{F} \right\}.$$

- Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables. The Rademacher complexity of $\mathcal{F}(x_1^n)/n$ is given by

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) = \mathbb{E}_{\varepsilon} \left\{ \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right\},$$

where $\mathcal{F}(x_1^n)/n$ denotes the set with elements $(f(x_1)/n, \dots, f(x_n)/n)^T$ for $f \in \mathcal{F}$.

Rademacher complexity of the function class

- Let $\mathbf{X} = (X_1, \dots, X_n)^\top$ where $\{X_i\}_{i=1}^n$ are i.i.d. random variables. The quantity $\mathcal{R}(\mathcal{F}(X_1^n)/n)$ is still a random variable.

Rademacher complexity of the function class

- Let $\mathbf{X} = (X_1, \dots, X_n)^\top$ where $\{X_i\}_{i=1}^n$ are i.i.d. random variables. The quantity $\mathcal{R}(\mathcal{F}(X_1^n)/n)$ is still a random variable.
- The quantity

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\mathbf{X}} \{ \mathcal{R}(\mathcal{F}(X_1^n)/n) \} = \mathbb{E}_{\mathbf{X}, \epsilon} \left\{ \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^n \epsilon_i f(X_i) \right| \right\}$$

is called as Rademacher complexity of the function class \mathcal{F} .

Upper bound of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$

Theorem 4 (Upper bound of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$)

Let \mathcal{F} be a function class which satisfies that $\|f\|_{\infty} \leq b$ for each $f \in \mathcal{F}$. For any positive integer n and $t > 0$, one has that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + t$$

with \mathbb{P} -probability at least $1 - e^{-\frac{nt^2}{2b^2}}$. Consequently, as long as $\mathcal{R}_n(\mathcal{F}) = o(1)$, one has that \mathcal{F} is \mathbb{P} -Glivenko-Cantelli.

Outline of proof

- The proof of Theorem 4 involves two steps.

Outline of proof

- The proof of Theorem 4 involves two steps.
- Concentration around mean: Show that for any $t > 0$,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \mathbb{E}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}) + t$$

with \mathbb{P} -probability at least $1 - e^{-\frac{nt^2}{2b^2}}$.

Outline of proof

- The proof of Theorem 4 involves two steps.
- Concentration around mean: Show that for any $t > 0$,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \mathbb{E}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}) + t$$

with \mathbb{P} -probability at least $1 - e^{-\frac{nt^2}{2b^2}}$.

- Upper bound on mean: Show that

$$\mathbb{E}(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}) \leq 2\mathcal{R}_n(\mathcal{F}).$$

Necessary conditions with Rademacher complexity

- The proof of Theorem 4 illustrates an important technique known as symmetrization, which relates the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ to its symmetrized version

$$\|S_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i) \right|.$$

Necessary conditions with Rademacher complexity

- The proof of Theorem 4 illustrates an important technique known as symmetrization, which relates the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ to its symmetrized version

$$\|S_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i) \right|.$$

- It is natural to wonder whether much was lost in moving from the variable to its symmetrized version.

Necessary conditions with Rademacher complexity

- The proof of Theorem 4 illustrates an important technique known as symmetrization, which relates the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ to its symmetrized version

$$\|S_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i) \right|.$$

- It is natural to wonder whether much was lost in moving from the variable to its symmetrized version.
- Denote

$$\check{\mathcal{F}} = \left\{ f - \mathbb{E}_X \{f(X)\} : f \in \mathcal{F} \right\},$$

where X is a random variable from \mathbb{P} .

Necessary conditions with Rademacher complexity

Proposition 5 (Sandwich results on $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ and $\|S_n\|_{\mathcal{F}}$)

For any convex and non-decreasing function $\phi : \mathbb{R} \rightarrow \mathbb{R}$, one has that

$$\mathbb{E}_{\mathbf{X}, \epsilon} \left\{ \phi(\|S_n\|_{\tilde{\mathcal{F}}}/2) \right\} \leq \mathbb{E}_{\mathbf{X}} \left\{ \phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}) \right\} \leq \mathbb{E}_{\mathbf{X}, \epsilon} \left\{ \phi(2\|S_n\|_{\mathcal{F}}) \right\}.$$

- When $\phi(t) = t$, Proposition 5 implies that

$$\mathbb{E}_{\mathbf{X}, \epsilon} \{ \|S_n\|_{\tilde{\mathcal{F}}} \} / 2 \leq \mathbb{E}_{\mathbf{X}} \{ \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \} \leq 2 \mathbb{E}_{\mathbf{X}, \epsilon} \{ \|S_n\|_{\mathcal{F}} \}.$$

Lower bound of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$

Theorem 6 (Lower bound of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$)

Under the assumption of Theorem 4, for any positive integer n and $t > 0$, one has that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq 2^{-1} \mathcal{R}_n(\mathcal{F}) - 2^{-1} n^{-1/2} \sup_{f \in \mathcal{F}} \left| \mathbb{E}\{f(X)\} \right| - t$$

with \mathbb{P} -probability at least $1 - e^{-\frac{nt^2}{2b^2}}$.

- 1 Motivation
- 2 A uniform law via Rademacher complexity
- 3 Upper bounds on the Rademacher complexity

Classes with polynomial discrimination

- Recall the notation

$$\mathcal{F}(x_1^n) = \left\{ (f(x_1), \dots, f(x_n))^T : f \in \mathcal{F} \right\}.$$

For a given collection of points $x_1^n = \{x_1, \dots, x_n\}$, the “size” of $\mathcal{F}(x_1^n)$ provides a sample-dependent measure of the complexity of \mathcal{F} .

Classes with polynomial discrimination

- Recall the notation

$$\mathcal{F}(x_1^n) = \left\{ (f(x_1), \dots, f(x_n))^T : f \in \mathcal{F} \right\}.$$

For a given collection of points $x_1^n = \{x_1, \dots, x_n\}$, the “size” of $\mathcal{F}(x_1^n)$ provides a sample-dependent measure of the complexity of \mathcal{F} .

- Consider that $\mathcal{F}(x_1^n)$ contains only a finite number of vectors for all sample sizes, so that its “size” can be measured via its cardinality.

Classes with polynomial discrimination

- Recall the notation

$$\mathcal{F}(x_1^n) = \left\{ (f(x_1), \dots, f(x_n))^T : f \in \mathcal{F} \right\}.$$

For a given collection of points $x_1^n = \{x_1, \dots, x_n\}$, the “size” of $\mathcal{F}(x_1^n)$ provides a sample-dependent measure of the complexity of \mathcal{F} .

- Consider that $\mathcal{F}(x_1^n)$ contains only a finite number of vectors for all sample sizes, so that its “size” can be measured via its cardinality.
- If \mathcal{F} consists of a family of binary-valued functions, then $\mathcal{F}(x_1^n)$ can contain at most 2^n elements. Of interest to us are function classes for which this cardinality grows only as a polynomial function of n .

Classes with polynomial discrimination

Definition 7 (Polynomial discrimination)

Let \mathcal{F} be a class consisting a family of binary-valued functions on \mathcal{X} . We say that \mathcal{F} has polynomial discrimination of order $\nu \geq 1$ if for each positive integer n and collection $x_1^n = \{x_1, \dots, x_n\}$ of n points in \mathcal{X} , $\mathcal{F}(x_1^n)$ has cardinality upper bounded as

$$\text{Card}(\mathcal{F}(x_1^n)) \leq (n + 1)^\nu.$$

- The significance of this property is that it provides a straightforward approach to controlling the Rademacher complexity.

Upper bound of the Rademacher complexity

Proposition 8 (Upper bound of $\mathcal{R}(\mathcal{F}(x_1^n)/n)$)

Suppose that \mathcal{F} has polynomial discrimination of order ν . Then for all positive integers n and any collection of points $x_1^n = \{x_1, \dots, x_n\}$, one has that

$$\underbrace{\mathbb{E}_\varepsilon \left\{ \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right\}}_{\mathcal{R}(\mathcal{F}(x_1^n)/n)} \leq 4D(x_1^n) \sqrt{\nu \log(n+1)/n},$$

where $D(x_1^n) = n^{-1/2} \sup_{f \in \mathcal{F}} \sqrt{\sum_{i=1}^n f^2(x_i)}$.

Upper bound of the Rademacher complexity

- When the function class is b uniformly bounded, then one has that $D(x_1^n)$ is bounded by b uniformly for all points x_1^n , which implies that

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\mathbf{X}} \{ \mathcal{R}(\mathcal{F}(X_1^n)/n) \} \leq 4b \sqrt{\nu \log(n+1)/n}.$$

Upper bound of the Rademacher complexity

- When the function class is b uniformly bounded, then one has that $D(x_1^n)$ is bounded by b uniformly for all points x_1^n , which implies that

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\mathbf{X}} \{ \mathcal{R}(\mathcal{F}(X_1^n)/n) \} \leq 4b \sqrt{\nu \log(n+1)/n}.$$

- As discussed previously, the classical Glivenko-Cantelli law is based on indicator functions of $(-\infty, t]$, which are uniformly bounded by $b = 1$.

Upper bound of the Rademacher complexity

- When the function class is b uniformly bounded, then one has that $D(x_1^n)$ is bounded by b uniformly for all points x_1^n , which implies that

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\mathbf{X}} \{ \mathcal{R}(\mathcal{F}(X_1^n)/n) \} \leq 4b\sqrt{\nu \log(n+1)/n}.$$

- As discussed previously, the classical Glivenko-Cantelli law is based on indicator functions of $(-\infty, t]$, which are uniformly bounded by $b = 1$.
- We will apply Proposition 8 and Theorem 4 to give a version proof of Theorem 2.

Classical Glivenko-Cantelli's Theorem

Theorem 9 (Classical Glivenko-Cantelli)

Let $F(x) = \mathbb{P}(X \leq x)$ be the CDF of a random variable X and $\hat{F}_n(x)$ be the ECDF based on n i.i.d. samples $\{X_i\}_{i=1}^n$ from \mathbb{P} . Then one has that for all $t > 0$

$$\|\hat{F}_n - F\|_\infty \leq 8\sqrt{\log(1+n)/n} + t$$

with \mathbb{P} -probability at least $1 - e^{-nt^2/2}$, which implies that as $n \rightarrow \infty$, $\|\hat{F}_n - F\|_\infty \rightarrow 0$ almost surely.

Vapnik–Chervonenkis (VC) dimension

Definition 10 (Shattering and VC dimension)

Given a class \mathcal{F} of binary-valued functions, we say that the set $x_1^n = \{x_1, \dots, x_n\}$ is shattered by \mathcal{F} if $\text{Card}(\mathcal{F}(x_1^n)) = 2^n$. The VC dimension $\nu(\mathcal{F})$ is the largest integer n for which there is some collection x_1^n of n points that is shattered by \mathcal{F} .

- When $\nu(\mathcal{F}) < \infty$, \mathcal{F} is said to be a VC class.

Vapnik–Chervonenkis (VC) dimension

Definition 10 (Shattering and VC dimension)

Given a class \mathcal{F} of binary-valued functions, we say that the set $x_1^n = \{x_1, \dots, x_n\}$ is shattered by \mathcal{F} if $\text{Card}(\mathcal{F}(x_1^n)) = 2^n$. The VC dimension $\nu(\mathcal{F})$ is the largest integer n for which there is some collection x_1^n of n points that is shattered by \mathcal{F} .

- When $\nu(\mathcal{F}) < \infty$, \mathcal{F} is said to be a VC class.
- When \mathcal{F} is consisted by indicator functions $\mathbb{I}_S(\cdot)$ for $S \in \mathcal{S}$, we use $\mathcal{S}(x_1^n)$ and $\nu(\mathcal{S})$ to denote $\mathcal{F}(x_1^n)$ and $\nu(\mathcal{F})$ respectively.

Intervals in \mathbb{R}

- Consider the class of all indicator functions for left-sided half-intervals on the real line, i.e., the class

$$\mathcal{S}_1 = \{(-\infty, a] : a \in \mathbb{R}\}.$$

We have shown that for any collection $x_1^n = \{x_1, \dots, x_n\}$, $\text{Card}(\mathcal{S}_1(x_1^n)) \leq n + 1$.

Intervals in \mathbb{R}

- Consider the class of all indicator functions for left-sided half-intervals on the real line, i.e., the class

$$\mathcal{S}_1 = \{(-\infty, a] : a \in \mathbb{R}\}.$$

We have shown that for any collection $x_1^n = \{x_1, \dots, x_n\}$, $\text{Card}(\mathcal{S}_1(x_1^n)) \leq n + 1$.

- For any single point x_1 , the collection $\{x_1\}$ can be picked out by the class \mathcal{S}_1 . But given two distinct points $x_1 < x_2$, it is impossible to find a left-sided interval that contains x_2 but not x_1 . Therefore, we conclude that $\nu(\mathcal{S}_1) = 1$.

Intervals in \mathbb{R}

- Consider the class of all indicator functions for two-sided intervals on the real line, i.e., the class

$$\mathcal{S}_2 = \{(a, b] : a, b \in \mathbb{R}, a < b\}.$$

Intervals in \mathbb{R}

- Consider the class of all indicator functions for two-sided intervals on the real line, i.e., the class

$$\mathcal{S}_2 = \{(a, b] : a, b \in \mathbb{R}, a < b\}.$$

- The class \mathcal{S}_2 can shatter any two-point set. But given three distinct points $x_1 < x_2 < x_3$, it cannot pick out the subset $\{x_1, x_3\}$, which implies that $\nu(\mathcal{S}_2) = 2$.

Intervals in \mathbb{R}

- Consider the class of all indicator functions for two-sided intervals on the real line, i.e., the class

$$\mathcal{S}_2 = \{(a, b] : a, b \in \mathbb{R}, a < b\}.$$

- The class \mathcal{S}_2 can shatter any two-point set. But given three distinct points $x_1 < x_2 < x_3$, it cannot pick out the subset $\{x_1, x_3\}$, which implies that $\nu(\mathcal{S}_2) = 2$.
- Note that any collection of n distinct points $x_1 < \cdots < x_n$ divides up the real line into $n + 1$ intervals. Thus, any set of the form $(a, b]$ can be specified by choosing one of $n + 1$ intervals for a and a second interval for b , which implies that this class has polynomial discrimination with $\nu = 2$.

Connection between VC dimension and polynomial discrimination

Theorem 11 (Vapnik-Chervonenkis, Sauer and Shelah)

Consider a set class \mathcal{S} with $\nu(\mathcal{S}) < \infty$. Then for any collection of points $x_1^n = \{x_1, \dots, x_n\}$ with $n \geq \nu(\mathcal{S})$, one has that

$$\text{Card}(\mathcal{S}(x_1^n)) \leq \sum_{i=0}^{\nu(\mathcal{S})} \binom{n}{i} \leq (n+1)^{\nu(\mathcal{S})}.$$

Operations on VC classes

Proposition 12 (Operations on VC classes)

Let \mathcal{S} and \mathcal{T} be set classes, each with finite VC dimensions $\nu(\mathcal{S})$ and $\nu(\mathcal{T})$ respectively. Then each of the following set classes also have finite VC dimension:

- (1) $\mathcal{S}^c = \{S^c : S \in \mathcal{S}\}.$
- (2) $\mathcal{S} \sqcup \mathcal{T} = \{S \cup T : S \in \mathcal{S}, T \in \mathcal{T}\}.$
- (3) $\mathcal{S} \sqcap \mathcal{T} = \{S \cap T : S \in \mathcal{S}, T \in \mathcal{T}\}.$

Vector space structure

Definition 13 (Subgraph)

Let $g : \mathcal{X} \rightarrow \mathbb{R}$ be a function, the subset of \mathcal{X}

$$S_g = \{x \in \mathcal{X} : g(x) \leq 0\}$$

is called as the subgraph of g at level zero. Let \mathcal{G} be a function class, the collection of subsets

$$\mathcal{S}(\mathcal{G}) = \{S_g : g \in \mathcal{G}\}$$

is called as the subgraph class of \mathcal{G} .

Vector space structure

Proposition 14 (Finite-dimensional vector spaces)

Let \mathcal{G} be a vector space of functions $g : \mathbb{R}^d \rightarrow \mathbb{R}$ with dimension $\dim(\mathcal{G}) < \infty$. Then the subgraph class $S(\mathcal{G})$ has VC dimension at most $\dim(\mathcal{G})$.

Linear functions in \mathbb{R}^d

- For a pair $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, define $f_{a,b}(x) = \langle a, x \rangle + b$ and consider the family

$$\mathcal{L}^d = \{f_{a,b} : (a, b) \in \mathbb{R}^d \times \mathbb{R}\}.$$

Linear functions in \mathbb{R}^d

- For a pair $(\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}$, define $f_{\mathbf{a},b}(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b$ and consider the family

$$\mathcal{L}^d = \{f_{\mathbf{a},b} : (\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}\}.$$

- The associated subgraph class $\mathcal{S}(\mathcal{L}^d)$ corresponds to the collection of all half-spaces of the form

$$H_{\mathbf{a},b} = \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{a}, \mathbf{x} \rangle + b \leq 0\}.$$

Linear functions in \mathbb{R}^d

- For a pair $(\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}$, define $f_{\mathbf{a},b}(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b$ and consider the family

$$\mathcal{L}^d = \{f_{\mathbf{a},b} : (\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}\}.$$

- The associated subgraph class $\mathcal{S}(\mathcal{L}^d)$ corresponds to the collection of all half-spaces of the form

$$H_{\mathbf{a},b} = \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{a}, \mathbf{x} \rangle + b \leq 0\}.$$

- \mathcal{L}^d forms a vector space of dimension $d + 1$, one has that $\mathcal{S}(\mathcal{L}^d)$ has VC dimension at most $d + 1$.

Spheres in \mathbb{R}^d

- Consider the sphere

$$S_{a,b} = \{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\|_2 \leq b, (\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}_+ \}$$

and let \mathcal{S}^d be the collection of all such spheres.

Spheres in \mathbb{R}^d

- Consider the sphere

$$S_{\mathbf{a},b} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{a}\|_2 \leq b, (\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}_+\}$$

and let \mathcal{S}^d be the collection of all such spheres.

- Define

$$f_{\mathbf{a},b}(\mathbf{x}) = \|\mathbf{x}\|_2^2 - 2\langle \mathbf{a}, \mathbf{x} \rangle + \|\mathbf{a}\|_2^2 - b^2,$$

then one has that

$$S_{\mathbf{a},b} = \{\mathbf{x} \in \mathbb{R}^d : f_{\mathbf{a},b}(\mathbf{x}) \leq 0\},$$

so that the sphere $S_{\mathbf{a},b}$ is a subgraph of the function $f_{\mathbf{a},b}$.

Spheres in \mathbb{R}^d

- Define a feature map $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d+2}$ via

$$\phi(\mathbf{x}) = (1, x_1, \dots, x_d, \|x\|_2^2)^\top$$

and then consider functions of the form $g_{\mathbf{c}}(\mathbf{x}) = \langle \mathbf{c}, \mathbf{x} \rangle$,
 $\mathbf{c} \in \mathbb{R}^{d+2}$.

Spheres in \mathbb{R}^d

- Define a feature map $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d+2}$ via

$$\phi(\mathbf{x}) = (1, x_1, \dots, x_d, \|x\|_2^2)^\top$$

and then consider functions of the form $g_{\mathbf{c}}(\mathbf{x}) = \langle \mathbf{c}, \mathbf{x} \rangle$, $\mathbf{c} \in \mathbb{R}^{d+2}$.

- The family of functions $\{g_{\mathbf{c}} : \mathbf{c} \in \mathbb{R}^{d+2}\}$ is a vector space of dimension $d+2$ and contains the function class $\{f_{\mathbf{a},b} : (\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}_+\}$.

Spheres in \mathbb{R}^d

- Define a feature map $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d+2}$ via

$$\phi(\mathbf{x}) = (1, x_1, \dots, x_d, \|x\|_2^2)^\top$$

and then consider functions of the form $g_{\mathbf{c}}(\mathbf{x}) = \langle \mathbf{c}, \mathbf{x} \rangle$, $\mathbf{c} \in \mathbb{R}^{d+2}$.

- The family of functions $\{g_{\mathbf{c}} : \mathbf{c} \in \mathbb{R}^{d+2}\}$ is a vector space of dimension $d+2$ and contains the function class $\{f_{\mathbf{a},b} : (\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}_+\}$.
- By applying Proposition 14 to this larger vector space, one has that $\nu(\mathcal{S}^d) \leq d+2$.

Thank You