Uniform laws of large numbers

Zhenduo Li & Yan Chen

September 17, 2022

Table of Contents

Table of Contents

Uniform convergence of cumulative distribution functions

- Cumulative distribution function (CDF): $\forall t \in \mathbb{R}, F(t) := \mathbb{P}[X \leq t]$.
- Now suppose that we are given a collection $\{X_i\}_{i=1}^n$ of n i.i.d. samples, each drawn according to the law specified by F. A natural estimate of F is the empirical *CDF* given by

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(-\infty, t]}[X_i],$$

where $\mathbb{I}_{(-\infty,t]}[x]$ is a $\{0,1\}$-valued indicator function for the event $\{x \leq t\}$. Since the population CDF can be written as $F(t) = \mathbb{E}\left[\mathbb{I}_{(-\infty,t]}[X]\right]$ the empirical CDF is an unbiased estimate.

Population and empirical CDFs

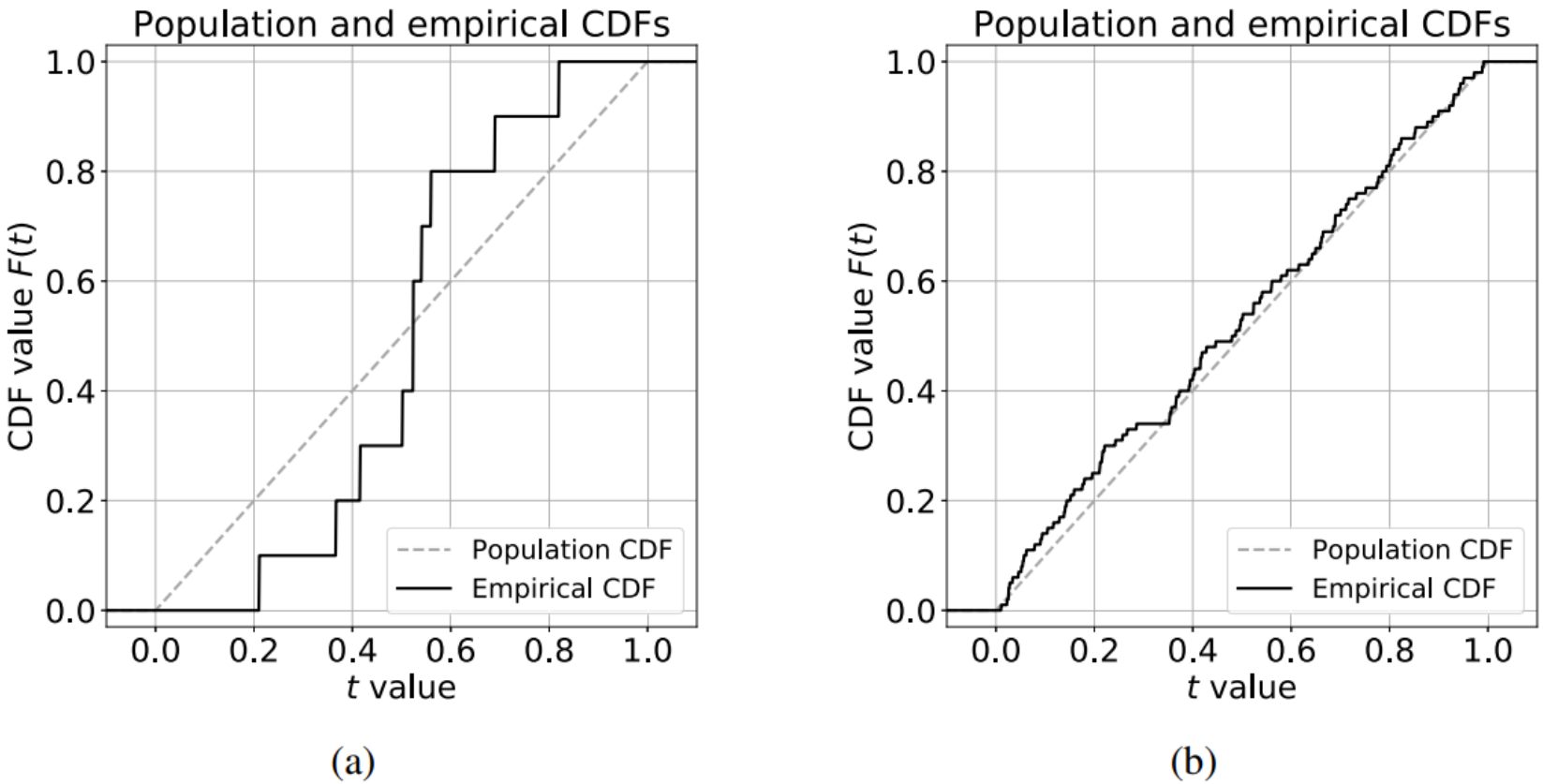

Figure: Plots of population and empirical CDF functions for the uniform distribution on $[0, 1]$.(a) Empirical CDF based on 10 samples. (b) Empirical CDF based on 100 samples

Why are uniform convergence results important?

In statistical settings, a typical use of the empirical CDF 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 that maps any CDF F to a real number $\gamma(F)$-that is, $F \mapsto \gamma(F)$.
- Given a set of samples distributed according to F, **the plug-in principle** suggests replacing the unknown F with the empirical CDF $\widehat{F}_n$, thereby obtaining $\gamma\left(\widehat{F}_n\right)$ as an estimate of $\gamma(F)$.

functionals

- **Example 4.1**(Expectation functionals) Given some integrable function g, we may define the expectation functional γ_g via

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

For any g, the plug-in estimate is given by $\gamma_g\left(\widehat{F}_n\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(X_i\right)$, corresponding to the sample mean of $g(X)$.

- **Example 4.2**(Quantile functionals)For any $\alpha \in [0,1]$, the quantile functional Q_α is given by

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

The plug-in estimate is given by

$$Q_\alpha\left(\widehat{F}_n\right) := \inf\left\{t \in \mathbb{R} \mid \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty,t]}\left[X_i\right] \geq \alpha\right\}.$$

and corresponds to estimating the α th quantile of the distribution by the α th sample quantile.

- **Example 4.3**(Goodness-of-fit functionals)
  For any plug-in estimator $\gamma\left(\widehat{F}_n\right)$, when does $\gamma\left(\widehat{F}_n\right)$ converge to $\gamma(F)$ in probability (or almost surely)? This question can be addressed in a unified manner for many functionals by defining a notion of continuity. Given a pair of CDFs F and G, define:

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

  Continuity of functionals(Exercise 4.1):
  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 $|\gamma(G) - \gamma(F)| \leq \epsilon$.

- **Theorem 4.4**(Glivenko-Cantelli) For any distribution, the empirical CDF $\widehat{F}_n$ is a strongly consistent estimator of the population CDF in the uniform norm, meaning that

$$\left\| \widehat{F}_n - F \right\|_\infty \xrightarrow{a.s.} 0.$$

More general function classes

We turn to more general consideration of uniform laws of large numbers. Let $\mathscr{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution $\mathbb{P}$ over X. Consider the random variable

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

More general function classes

- **Definition 4.5** We say that $\mathscr{F}$ is a Glivenko-Cantelli class for $\mathbb{P}$ if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ converges to zero in probability as $n \to \infty$.
  This notion can also be defined in a stronger sense, requiring almost sure convergence of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$, in which case we say that $\mathscr{F}$ satisfies a strong Glivenko-Cantelli law.

More general function classes

- **Example 4.6**(Empirical CDFs and indicator functions)Consider the function class

$$\mathscr{F} = \left\{ \mathbb{I}_{(-\infty, t]}(\cdot) \mid t \in \mathbb{R} \right\},$$

where $\mathbb{I}_{(-\infty,t]}$ is the $\{0,1\}$-valued indicator function of the interval $(-\infty, t]$. For each fixed $t \in \mathbb{R}$, we have the equality $\mathbb{E}\left[\mathbb{I}_{(-\infty,t]}(X)\right] = \mathbb{P}[X \leq t] = F(t)$, so that the classical Glivenko-Cantelli theorem is equivalent to a strong uniform law for the class $\mathscr{F}$.

More general function classes

- **Example 4.7**(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, and consider the function class $\mathscr{F}_S = \{\mathbb{I}_S(\cdot) \mid S \in \mathcal{S}\}$. Suppose that samples X_i are drawn from some distribution over $[0,1]$ that has no atoms (i.e., $\mathbb{P}(\{x\}) = 0$ for all $x \in [0,1]$). For any such distribution, we have $\mathbb{P}[S] = 0$ for all $S \in \mathcal{S}$. For any positive integer $n \in \mathbb{N}$, the discrete set $\{X_1, \ldots, X_n\}$ belongs to $\mathcal{S}$, by definition of the empirical distribution, we have $\mathbb{P}_n[X_1^n] = 1$.Thus,

$$\sup_{S \in \mathcal{S}} |\mathbb{P}_n[S] - \mathbb{P}[S]| = 1 - 0 = 1$$

Empirical risk minimization

- Given an indexed family of probability distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$, and suppose that we are given n samples $\{X_i\}_{i=1}^n$, each sample lying in some space X.
- Suppose that the samples are drawn i.i.d. according to a distribution $\mathbb{P}_{\theta^*}$, for some fixed but unknown $\theta^* \in \Omega$.
- The standard decision-theoretic approach to estimating θ^* is based on minimizing a cost function of the form $\theta \mapsto \mathcal{L}_\theta(X)$.

Empirical risk minimization

- Empirical risk:

$$\widehat{R}_n(\theta, \theta^*) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_\theta(X_i).$$

- Population risk:

$$R(\theta, \theta^*) := \mathbb{E}_{\theta^*}[\mathcal{L}_\theta(X)].$$

- Excess risk:

$$E\left(\widehat{\theta}, \theta^*\right) := R\left(\widehat{\theta}, \theta^*\right) - \inf_{\theta \in \Omega_0} R(\theta, \theta^*).$$

Empirical risk minimization

- **Example 4.8**(Maximum likelihood)In order to estimate the unknown parameter θ^*,we consider the cost function

$$\mathcal{L}_\theta(x) := \log\left[\frac{p_{\theta^*}(x)}{p_\theta(x)}\right].$$

The maximum likelihood estimate is obtained by minimizing the empirical risk defined by this cost function

$$\widehat{\theta} \in \arg\min_{\theta\in\Omega_0} \underbrace{\left\{\frac{1}{n}\sum_{i=1}^{n}\log\frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}\right\}}_{\widehat{R}_n(\theta,\theta^*)} = \arg\min_{\theta\in\Omega_0}\left\{\frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{p_\theta(X_i)}\right\}.$$

The population risk is given by $R(\theta,\theta^*) = \mathbb{E}_{\theta^*}\left[\log\frac{p_{\theta^*}(X)}{p_\theta(X)}\right]$, a quantity known as the Kullback-Leibler divergence between p_{θ^*} and p_θ.
In the special case that $\theta^* \in \Omega_0$, the excess risk is simply the Kullback-Leibler divergence between the true density p_{θ^*} and the fitted model $p_{\bar{\theta}}$.

Controlling the excess risk

Assume that there exists $\theta_0 \in \Omega_0$ such that $R(\theta_0, \theta^*) = \inf_{\theta \in \Omega_0} R(\theta, \theta^*)$.

$$E\left(\widehat{\theta}, \theta^*\right) = \underbrace{\left\{R\left(\widehat{\theta}, \theta^*\right) - \widehat{R}_n\left(\widehat{\theta}, \theta^*\right)\right\}}_{T_1} + \underbrace{\left\{\widehat{R}_n\left(\widehat{\theta}, \theta^*\right) - \widehat{R}_n\left(\theta_0, \theta^*\right)\right\}}_{T_2 \leq 0}$$

$$+ \underbrace{\left\{\widehat{R}_n\left(\theta_0, \theta^*\right) - R\left(\theta_0, \theta^*\right)\right\}}_{T_3}.$$

- T_2 is non-positive: $\widehat{\theta}$ minimizes the empirical risk over Ω_0.
- T_3 can be dealt with in a relatively straightforward manner, because θ_0 is an unknown but non-random quantity.

$$T_3 = \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\theta_0}\left(X_i\right)\right] - \mathbb{E}_X[\mathcal{L}_{\theta_0}(X)].$$

This can be controlled using the techniques introduced in Chap2.

Controlling the excess risk

- $$T_1 = \mathbb{E}_X\left[\mathcal{L}_{\hat{\theta}}(X)\right] - \left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\hat{\theta}}\left(X_i\right)\right].$$

  This quantity is more challenging to control, because the parameter $\widehat{\theta}$ is random, and moreover depends on the samples $\{X_i\}_{i=1}^n$, since it was obtained by minimizing the empirical risk. For this reason, controlling the first term requires a stronger result, such as a uniform law of large numbers over the cost function class $\mathfrak{L}\left(\Omega_0\right) := \{x \mapsto \mathcal{L}_\theta(x), \theta \in \Omega_0\}$. With this notation, we have

$$T_1 \leq \sup_{\theta\in\Omega_0}\left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_\theta\left(X_i\right) - \mathbb{E}_X\left[\mathcal{L}_\theta(X)\right]\right| = \left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathscr{Q}(\Omega_0)}.$$

Table of Contents

Introducing examples

- Given samples $S = ((x_1, y_1), (x_2, y_2) \ldots (x_m, y_m))$ where $y_i = \{-1, +1\}$
- A collection of training models H,for any $h \in H$,$h(X) \to \{-1 + 1\}$.
- Classification error:

$$\begin{aligned} \operatorname{err}(h) &= \frac{1}{m} \sum_{i=1}^{m} 1\{h(x_i) \neq y_i\} \\ &= \frac{1}{m} \sum_{i=1}^{m} \frac{1 - y_i h(x_i)}{2} \\ &= \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^{m} y_i h(x_i) \end{aligned}$$

- Replaced with Rademacher variables yields

$$\operatorname{Rad}(h) = \max_h \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i)$$

Rademacher complexity

- **Rademacher complexity of a set**:
  Let $\{\varepsilon_k\}_{k=1}^n$ be an i.i.d. sequence of Rademacher variables (i.e., taking the values $\{-1, +1\}$ equiprobably). Given a collection of vectors $\mathcal{A} \subset \mathbb{R}^n$, define the random variable

$$Z := \sup_{a \in \mathcal{A}} \left[\sum_{k=1}^{n} a_k \varepsilon_k \right] = \sup_{a \in \mathcal{A}} [\langle a, \varepsilon \rangle].$$

  Its expectation $\mathcal{R}(\mathcal{A}) := \mathbb{E}[Z(\mathcal{A})]$ is known as the Rademacher complexity of the set $\mathcal{A}$.

Rademacher complexity

- **Rademacher complexity of a function class**:
  For any fixed collection $x_1^n := (x_1, \ldots, x_n)$ of points, consider the subset of $\mathbb{R}^n$ given by

$$\mathscr{F}(x_1^n) := \{(f(x_1), \ldots, f(x_n)) \mid f \in \mathscr{F}\}.$$

  and the empirical Rademacher complexity is given by

$$\mathcal{R}(\mathscr{F}(x_1^n)/n) := \mathbb{E}_\varepsilon \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right].$$

  Given a collection $X_1^n := \{X_i\}_{i=1}^n$ of random samples, then the empirical Rademacher complexity $\mathcal{R}(\mathscr{F}(X_1^n)/n)$ is a random variable. Taking its expectation yields the Rademacher complexity of the function class $\mathscr{F}$:

$$\mathcal{R}_n(\mathscr{F}) := \mathbb{E}_X[\mathcal{R}(\mathscr{F}(X_1^n)/n)] = \mathbb{E}_{X,\varepsilon} \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right].$$

Rademacher complexity

- **Exercise 4.7**(Basic properties of Rademacher complexity)
  (a) $\mathcal{R}_n(\mathscr{F}) = \mathcal{R}_n(\mathrm{conv}(\mathscr{F}))$.
  (b) $\mathcal{R}_n(\mathscr{F} + \mathscr{G}) \leq \mathcal{R}_n(\mathscr{F}) + \mathcal{R}_n(\mathscr{G})$.
  (c) Given a fixed and uniformly bounded function g, show that

$$\mathcal{R}_n(\mathscr{F} + g) \leq \mathcal{R}_n(\mathscr{F}) + \frac{\|g\|_\infty}{\sqrt{n}}.$$

A uniform law via Rademacher complexity

- **Theorem 4.10** For any b-uniformly bounded class of functions $\mathscr{F}$, any positive integer $n \geq 1$ and any scalar $\delta \geq 0$, we have

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

with $\mathbb{P}$-probability at least $1 - \exp\left(-\frac{n\delta^2}{2b^2}\right)$. Consequently, as long as $\mathcal{R}_n(\mathscr{F}) = o(1)$, we have $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \xrightarrow{\text{a.s.}} 0$.

Proof

- **Concentration around mean**:
  Define the recentered functions: $\bar{f}(x) := f(x) - \mathbb{E}[f(X)]$, thus $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(X_i)\right|$.Consider the function

$$G(x_1, \ldots, x_n) := \sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right|.$$

  Define the vector $y \in \mathbb{R}^n$ with $y_i = x_i$ for all $i \neq 1$. For any function $\bar{f} = f - \mathbb{E}[f]$,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \sup_{h\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n \bar{h}(y_i)\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(y_i)\right| \\ &\leq \frac{1}{n}\left|\bar{f}(x_1) - \bar{f}(y_1)\right| \\ &\leq \frac{2b}{n}.\end{aligned}$$

Proof

Take the supremum over $f \in \mathscr{F}$ on both sides yields

$$G(x) - G(y) \leq \frac{2b}{n}.$$

Reverse x and y yields

$$|G(x) - G(y)| \leq \frac{2b}{n}.$$

Therefore, by the bounded differences method (Corollary 2.21),

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} - \mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] \leq t \quad \text{with } \mathbb{P}\text{-prob. at least } 1 - \exp\left(-\frac{nt^2}{2b^2}\right),$$

valid for all $t \geq 0$.

- **Upper bound on mean**:
  $(Y_1, \ldots, Y_n)$,a second i.i.d. sequence, independent of $(X_1, \ldots, X_n)$.

$$\mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] = \mathbb{E}_X\left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \{f(X_i) - \mathbb{E}_{Y_i}\left[f(Y_i)\right]\} \right|\right]$$

Proof

$$= \mathbb{E}_X \left[\sup_{f \in \mathscr{F}} \left| \mathbb{E}_Y \left[\frac{1}{n} \sum_{i=1}^{n} \{ f(X_i) - f(Y_i) \} \right] \right| \right]$$

$$\overset{(Ex4.4)}{\leq} \mathbb{E}_{X,Y} \left[\sup_{f \in \mathscr{F}} | \frac{1}{n} \sum_{i=1}^{n} \{ f(X_i) - f(Y_i) \} \right].$$

Let $(\varepsilon_1, \ldots, \varepsilon_n)$ be an i.i.d. sequence of Rademacher variables, independent of X and Y. For any function $f \in \mathscr{F}$, the random vector with components $\varepsilon_i (f(X_i) - f(Y_i))$ has **the same joint distribution** as the random vector with components $f(X_i) - f(Y_i)$, whence

$$\mathbb{E}_{X,Y} \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \{ f(X_i) - f(Y_i) \} \right| \right] = \mathbb{E}_{X,Y,\varepsilon} \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f(X_i) - f(Y_i)) \right| \right]$$

$$\leq 2\mathbb{E}_{X,\varepsilon} \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right] = 2\mathcal{R}_n(\mathscr{F})$$

Necessary conditions with Rademacher complexity

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

$$\|\mathbb{S}_n\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i) \right|.$$

- **Proposition 4.11** For any convex non-decreasing function $\Phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\|\mathbb{S}_n\|_{\overline{\mathscr{F}}}\right)\right] \overset{(a)}{\leq} \mathbb{E}_X\left[\Phi\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right)\right] \overset{(b)}{\leq} \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\|\mathbb{S}_n\|_{\mathscr{F}}\right)\right],$$

where $\overline{\mathscr{F}} = \{f - \mathbb{E}[f], f \in \mathscr{F}\}$ is the recentered function class. When $\Phi(t) = t$, proposition 4.11 yields

$$\frac{1}{2}\mathbb{E}_{X,\varepsilon}\|\mathbb{S}_n\|_{\overline{\mathscr{F}}} \leq \mathbb{E}_X\left[\|\mathbb{P}_n - \mathbb{P}\|\,\mathscr{F}\right] \leq 2\mathbb{E}_{X,\varepsilon}\|\mathbb{S}_n\|_{\mathscr{F}}.$$

Proof

$$\mathbb{E}_X\left[\Phi\left(\left\|\mathbb{P}_n-\mathbb{P}\right\|_{\mathscr{F}}\right)\right]=\mathbb{E}_X\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-\mathbb{E}_Y\left[f(Y_i)\right]\right|\right)\right]$$

$$\overset{\text{(i)}}{\leq}\ \mathbb{E}_{X,Y}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(X_i)-f(Y_i)\right|\right)\right]$$

$$\overset{\text{(ii)}}{=}\ \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left\{f(X_i)-f(Y_i)\right\}\right|\right)\right].$$

$T_1=\mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left\{f(X_i)-f(Y_i)\right\}\right|\right)\right]$, By the triangle inequality, we have

Proof

$$
\begin{aligned}
T_1 &\leq \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i)\right|\right)\right] \\
&\overset{\text{(iii)}}{\leq} \frac{1}{2}\mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right)\right] \\
&\quad + \frac{1}{2}\mathbb{E}_{Y,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i)\right|\right)\right] \\
&\overset{\text{(iv)}}{=} \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right)\right].
\end{aligned}
$$

Proof

$$\mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\left\|\mathbb{S}_n\right\|_{\mathscr{F}}\right)\right] = \mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left\{f(X_i)-\mathbb{E}_{Y_i}\left[f(Y_i)\right]\right\}\right|\right)\right]$$

$$\overset{\text{(i)}}{\leq} \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left\{f(X_i)-f(Y_i)\right\}\right|\right)\right]$$

$$\overset{\text{(ii)}}{=} \mathbb{E}_{X,Y}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i)-f(Y_i)\right\}\right|\right)\right]$$

Let $T_2 := \frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i)-f(Y_i)\right\}\right|$, we have

Proof

$$T_2 \leq \frac{1}{2} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \{f(X_i) - \mathbb{E}[f]\} \right| + \frac{1}{2} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \{f(Y_i) - \mathbb{E}[f]\} \right|.$$

Since Φ is convex and non-decreasing, we have

$$\Phi(T_2) \leq \frac{1}{2}\Phi\left(\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \{f(X_i) - \mathbb{E}[f]\} \right| \right) + \frac{1}{2}\Phi\left(\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \{f(Y_i) - \mathbb{E}[f]\} \right| \right)$$

The claim follows by taking expectations and using the fact that X and Y are identically distributed.

Necessary conditions with Rademacher complexity

- **Proposition 4.12** For any b-uniformly bounded function class $\mathscr{F}$, any integer $n \geq 1$ and any scalar $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \geq \frac{1}{2}\mathcal{R}_n(\mathscr{F}) - \frac{\sup_{f\in\mathscr{F}} |\mathbb{E}[f]|}{2\sqrt{n}} - \delta$$

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

- **Sketch to the proof**:
  Using exercise 4.5

$$\mathbb{E}_{X,\varepsilon}\left[\|\mathbb{S}_n\|_{\mathscr{F}}\right] \geq \mathbb{E}_{X,\varepsilon}\left[\|\mathbb{S}_n\|_{\mathscr{F}}\right] - \frac{\sup_{f\in\mathscr{F}} |\mathbb{E}[f]|}{\sqrt{n}}.$$

  Using(4.16)

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} - \mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] \leq t \quad \text{with } \mathbb{P}\text{-prob. at least } 1 - \exp\left(-\frac{nt^2}{2b^2}\right)$$

  valid for all $t \geq 0$.
  Using proposition 4.11

Table of Contents

Upper bounds on the Rademacher complexity

- Obtaining concrete results using Theorem 4.10 requires methods for upper bounding the Rademacher complexity.
- There are a variety of such methods, ranging from simple union bound methods (suitable for finite function classes) to more advanced techniques involving the notion of metric entropy and chaining arguments. We explore the latter techniques in Chapter 5 to follow.
- Now we just focus on the former techniques.

Classes with polynomial discrimination

Definition 4.13(Polynomial discrimination) A class $\mathscr{F}$ of functions with domain $\mathcal{X}$ has polynomial discrimination of order $\nu \geq 1$ if, for each positive integer n and collection $x_1^n = \{x_1, \ldots, x_n\}$ of n points in $\mathcal{X}$, the set $\mathscr{F}(x_1^n)$ has cardinality upper bounded as

$$card(\mathscr{F}(x_1^n)) \leq (n+1)^\nu$$

Classes with polynomial discrimination

- In the case of polynomial discrimination, we have the lemma below.
  Lemma 4.14 Suppose that $\mathscr{F}$ has polynomial discrimination of order ν. Then for all positive integers n and any collection of points $x_1^n = (x_1, \ldots, x_n)$,

$$\underbrace{\mathbb{E}_\epsilon \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right]}_{\mathcal{R}(\mathscr{F}(x_1^n)/n)} \leq 4 D(x_1^n) \sqrt{\frac{\nu \log(n+1)}{n}}$$

  where $D(x_1^n) := \sup_{f \in \mathscr{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$
- We leave the proof of this claim for the reader (see Exercise 4.9).

Classes with polynomial discrimination

Now, we suppose that the function class is b uniformly bounded, combined with **Lemma 4.14**, we reach the conclusion easily:

$$\mathcal{R}_n(\mathscr{F}) \leq 4b\sqrt{\frac{\nu log(n+1)}{n}} \quad for\ all\ n \geq 1$$

Then combined with **Theorem 4.10**, we finally conclude that, any bounded function class with polynomial discrimination is Glivenko–Cantelli.

Classes with polynomial discrimination

Now we will return to **Theorem 4.4**, we will observe that it's a direct corollary of **Theorem 4.10** and **Lemma 4.14**. We express the theorem in another way:

Corollary 4.15(Classical Glivenko–Cantelli) Let $F(t) = \mathbb{P}[X \leq t]$ be the CDF of a random variable X, and let $\hat{F}_n$ be the empirical CDF based on n i.i.d. samples $X_i \sim P$. Then

$$\mathbb{P}\left[\|\hat{F}_n - F\|_\infty \geq 8\sqrt{\frac{log(n+1)}{n}} + \delta\right] \leq e^{-\frac{n\delta^2}{2}} \quad for\ all\ \delta \geq 0$$

and hence $\|\hat{F}_n - F\|_\infty \xrightarrow{a.s.} 0$

Hint: For Classical Glivenko–Cantelli, $b = 1$, directly apply **Lemma 4.14** on the basis of **Theorem 4.10**.

Vapnik–Chervonenkis dimension

- More broadly, it is of interest to develop techniques for certifying this property in a less laborious manner.
- The theory of Vapnik–Chervonenkis (VC) dimension provides one such class of techniques.
- Accordingly, we now turn to defining the notions of shattering and VC dimension.
- Let us consider a function class $\mathscr{F}$ in which each function f is binary-valued, taking the values $\{0, 1\}$ for concreteness. In this case, the set $\mathscr{F}(x_1^n)$ can have at most 2^n elements.

Vapnik–Chervonenkis dimension

Definition 4.16 Given a class $\mathscr{F}$ of binary-valued functions, we say that the set $x_1^n = (x_1, \ldots, x_n)$ **is shattered by** $\mathscr{F}$ if $card(\mathscr{F}(x_1^n)) = 2^n$. The VC dimension $\nu(\mathscr{F})$ is the largest integer n for which there is some collection $x_1^n = (x_1, \ldots, x_n)$ of n points that is shattered by $\mathscr{F}$.

- When the quantity $\nu(\mathscr{F})$ is finite, then the function class $\mathscr{F}$ is said to be a VC class.
- Let us illustrate the notions of shattering and VC dimension with some examples:

Vapnik–Chervonenkis dimension

Other examples: In fact, the concept of VC dimension derived from, given n points in the sample space, each with a binary label, whether there exists a function in the set class or not which can tell the difference between them.

$VC : 3$ ∞ ∞

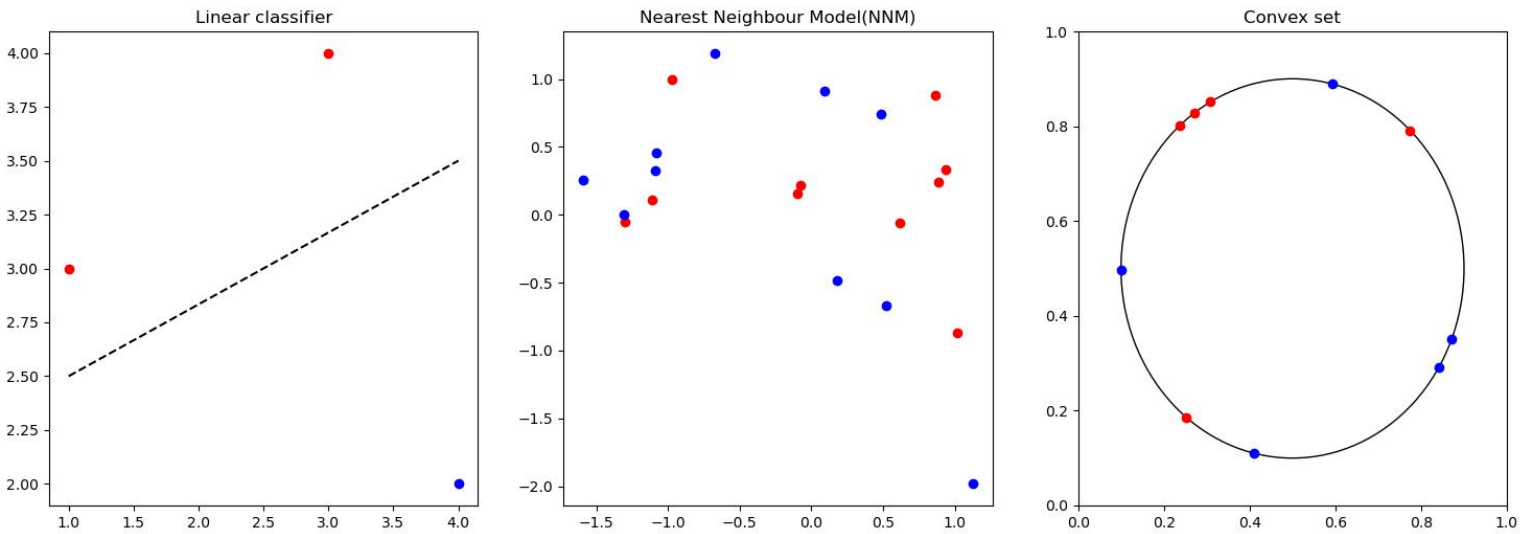

Figure: Other examples

Vapnik–Chervonenkis dimension

Proposition 4.18 (Vapnik–Chervonenkis, Sauer and Shelah) Consider a set class $\mathcal{S}$ with $\nu(\mathcal{S}) < \infty$. Then for any collection of points $P = (x_1, \ldots, x_n)$ with $n \geq \nu(\mathcal{S})$, we have

$$card(\mathcal{S}(P)) \overset{(i)}{\leq} \sum_{i=0}^{\nu(\mathcal{S})} \binom{n}{i} \overset{(ii)}{\leq} (n+1)^{\nu(\mathcal{S})}$$

- (ii) is trivial through some combinatorics methods.
- (i) is comparatively complex, we just sketch the proof.

Vapnik–Chervonenkis dimension

Proof: Given a subset of points Q and a set class T, we let $\nu(T; Q)$ denote the VC dimension of T when considering only whether or not subsets of Q can be shattered. Note that $\nu(T) \leq k$ implies that $\nu(T; Q) \leq k$ for all point sets Q. For positive integers (n, k), define the functions

$$\Phi_k(n) := \sup_{Q, card(Q) \leq n} \sup_{T, \nu(T;Q) \leq k} card(T(Q))$$

$$\Psi_k(n) := \sum_{i=0}^{k} \binom{n}{i}$$

In terms of this notation, we claim that it suffices to prove that $\Phi_k(n) \leq \Psi_k(n)$

Vapnik–Chervonenkis dimension

- We consider Mathematical Induction for $n + k$.
- *Base case:* $n + k = 2$, trivial.
- *Induction step:* Now assume that, for some integer $l > 2$, the inequality holds for all pairs with $n + k < l$. We claim that it then holds for all pairs with $n + k = l$. Fix an arbitrary pair (n, k) such that $n + k = l$, a point set $P = \{x_1, \ldots, x_n\}$ and a set class S such that $\nu(S; P) = k$.Define the point set $P' = P \backslash \{x_1\}$, and let $S_0 \subseteq S$ be the smallest collection of subsets that labels the point set P in the maximal number of different ways. Let S_1 be the smallest collection of subsets inside $S \backslash S_0$ that produce binary labelings of the point set P that are not in $S_0(P)$. (The choices of S_0 and S_1 need not be unique.) With this decomposition we have

$$card(S(P)) = card(S_0(P)) + card(S_1(P'))$$

Vapnik–Chervonenkis dimension

- The decomposition above may sound complex, but below we will give an example to help understand it.
- Given a set class $S = \{s_1, s_2, s_3, s_4\}$ and a point set $P = \{x_1, x_2, x_3\}$, suppose that the sets generated the binary labelings

$$s_1 \leftrightarrow (0, 1, 1), s_2 \leftrightarrow (1, 1, 1), s_3 \leftrightarrow (0, 1, 0), s_4 \leftrightarrow (0, 1, 1)$$

- In this particular case, we have $S(P) = \{(0, 1, 1), (1, 1, 1), (0, 1, 0)\}$, and one valid choice of the pair (S_0, S_1) would be $S_0 = \{s_1, s_3\}$ and $S_1 = \{s_2\}$, generating the labelings $S_0(P) = \{(0, 1, 1), (0, 1, 0)\}$ and $S_1(P) = \{(1, 1, 1)\}$

Vapnik–Chervonenkis dimension

- $\nu(S_0; P') \leq \nu(S_0; P) \leq k$, hence $card(S_0(P')) \leq \Psi_k(n-1)$
- On the other hand, we claim that $\nu(S_1; P') \leq k-1$, then $card(S_1(P')) \leq \Psi_{k-1}(n-1)$, combined, we finally conclude that $card(S(P)) \leq \Psi_k(n-1) + \Psi_{k-1}(n-1) = \Psi_k(n)$
- In fact, suppose that S_1 shatters some subset $Q \subseteq P$ of cardinality m; it suffices to show that $m \leq k-1$. If S_1 shatters such a set Q , then S would shatter the set $Q = Q \cup \{x_1\} \subseteq P$, since $\nu(S; P) \leq k$, it must be the case that $card(Q) = m + 1 \leq k$, which implies that $\nu(S_1; P) \leq k-1$.□

Controlling the VC dimension

- *Basic operations:* The property of having finite VC dimension is preserved under a number of basic operations, as summarized in the following.
- **Proposition 4.19:** 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:
  (a) The set class $\mathcal{S}^c := \{S^c | S \in \mathcal{S}\}$, where S^c denotes the complement of S
  (b) The set class $\mathcal{S} \sqcup \mathcal{T} := \{S \cup T | S \in \mathcal{S}, T \in \mathcal{T}\}$
  (c) The set class $\mathcal{S} \sqcap \mathcal{T} := \{S \cap T | S \in \mathcal{S}, T \in \mathcal{T}\}$
- We leave the proof of this result as an exercise for the reader (Exercise 4.8).

Controlling the VC dimension

- *Vector space structure:* Any class $\mathscr{G}$ of real-valued functions defines a class of sets by the operation of taking subgraphs.We define the subgraph at level zero $S_g := \{x \in \mathcal{X} | g(x) \leq 0\}$ and the collection of subgraphs $\mathcal{S}(\mathscr{G}) := \{S_g | g \in \mathscr{G}\}$, when $\mathscr{G}$ is a vector space, we have the following result
- **Proposition 4.20**(Finite-dimensional vector spaces) Let $\mathscr{G}$ be a vector space of functions $g : \mathbb{R}^d \to \mathbb{R}$ with dimension $dim(\mathscr{G}) < \infty$. Then the subgraph class $\mathcal{S}(\mathscr{G})$ has VC dimension at most $dim(\mathscr{G})$.

Controlling the VC dimension

- **Proof:** By the definition of VC dimension, we need to show that no collection of $n = dim(\mathscr{G}) + 1$ points in $\mathbb{R}_d$ can be shattered by $\mathcal{S}(\mathscr{G})$. Fix an arbitrary collection $x_1^n = \{x_1, \ldots, x_n\}$ of n points in $\mathbb{R}^d$, and consider the linear map $L : \mathscr{G} \to \mathbb{R}^n$ given by $L(g) = (g(x_1), \ldots, g(x_n))$. By construction, the range of the mapping L is a linear subspace of $\mathbb{R}^n$ with dimension at most $dim(\mathscr{G}) = n-1 < n$. Therefore, there must exist a non-zero vector $\gamma \in \mathbb{R}^n$ such that $< \gamma, L(g) >= 0$ for all $g \in \mathscr{G}$. We may assume without loss of generality that at least one coordinate is positive, and then write

$$\sum_{\{i|\gamma_i \leq 0\}} (-\gamma_i) g(x_i) = \sum_{\{i|\gamma_i \geq 0\}} \gamma_i g(x_i)$$

Controlling the VC dimension

- Proceeding via proof by contradiction, suppose that there were to exist some $g \in \mathscr{G}$ such that the associated subgraph set $S_g = \{x \in \mathbb{R}^d | g(x) \leq 0\}$ included only the subset $\{x_i | \gamma_i \leq 0\}$. For such a function g, the right-hand side of equation would be strictly positive while the left-hand side would be non-positive, which is a contradiction. We conclude that $\mathcal{S}(\mathscr{G})$ fails to shatter the set $\{x_1, \ldots, x_n\}$, as claimed.$\square$