Wainwright High-Dimensional Statistics Notes

Chapter 4: Uniform Laws of Large Numbers

Annette Jing

April 27, 2019

I. Motivation

A. Uniform Convergence of CDFs

Def 1. Population CDF of r.v. X, F

$$F(t) := P(X \le t), \quad \forall t \in \mathbb{R}$$

Def 2. Empirical CDF corr. to i.i.d. r.v.s $\{X_i\}_{i=1}^n \sim F$, $\widehat{F}_n$

$$\widehat{F}_n(t) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(-\infty,t]}(X_i), \quad \forall t \in \mathbb{R}$$

Rmk. $\widehat{F}_n$ converges to F pointwise

$$\widehat{F}_n(t) \xrightarrow{a.s.} F(t) \quad \forall t \in \mathbb{R}$$

Proof. Observe that $F(t) = \mathbb{E}[\mathbf{1}_{(-\infty,1]}(X)]$. For each $t \in \mathbb{R}$, $\{\mathbf{1}_{(-\infty,1]}(X_i)\}_{i=1}^n$ are i.i.d. and $\mathbb{E}|\mathbf{1}_{(-\infty,1]}(X_i)| \le 1 < \infty$, so the result follows immediately after Khintchine's SLLN. □

Def 3. The Plug-in Principle
Many estimation techniques involve using the empirical CDF to construct estimators of quantities associated with the population CDF, which can be formulated in terms of a functional γ. The Plug-in Principle suggests estimating $\gamma(F)$ with $\gamma(\widehat{F}_n)$.

e.g. Expectation Functional corr. to integrable g, γ_g

$$\gamma_g(F) := \int g(x)dF(x) \equiv \mathbb{E}[g(X)]$$

e.g. Quantile Functional at $\alpha \in [0,1]$, Q_α

$$Q_\alpha(F) := \inf\{t \in \mathbb{R} | F(t) \ge \alpha\}$$

e.g. Goodness-of-fit Functionals
Measures the distance between F and a target CDF F_0. Following are two examples:

$$\gamma(F) = \|F - F_0\|_\infty \equiv \sup_{t\in\mathbb{R}} |F(t) - F_0(t)|$$

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

Def 4. γ is continuous at F in the sup-norm if

$$\forall \epsilon > 0, \exists \delta > 0 \ s.t. \ \|G - F\|_\infty \le \delta \Rightarrow |\gamma(G) - \gamma(F)| \le \epsilon$$

Thm 1. Glivenko-Cantelli Thm (Thm 4.4)

$$\|\widehat{F}_n - F\|_\infty \xrightarrow{a.s.} 0 \quad \text{as } n \to \infty$$

Proof. This can be proven as a corollary of more general results to follow (Corollary 7.3 of lemma 7). □

Corollary 1.1. (Ex 4.1: Continuity of functionals)
γ cont. at F in the sup-norm $\Rightarrow \gamma(\widehat{F}_n) \xrightarrow{p} \gamma(F)$

B. Uniform Laws for General Function Classes

Let X be an $\mathcal{X}$-valued random element on probability space (Ω, Σ, P) and denote its distribution as $\mathbb{P} \equiv P \circ X^{-1} : \mathcal{A} \to [0,1]$. Consider a class $\mathcal{F}$ of real-valued integrable functions on the measurable space $(\mathcal{X}, \mathcal{A})$ and a sample $\{X_i\}_{i=1}^n \sim$ i.i.d. $\mathbb{P}$. For any measurable $f : (\mathcal{X}, \mathcal{A}) \to \mathbb{R}$ and signed measure $Q : \mathcal{A} \to \overline{\mathbb{R}}$, define $Qf := \int f dQ$ and $\|Q\|_\mathcal{F} := \sup_{f\in\mathcal{F}} |Qf|$.

Def 5. Empirical Measure of $\{X_i\}_{i=1}^n$, $\mathbb{P}_n : \mathcal{A} \to [0,1]$

$$\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$$

where $\delta_{X_1}, ..., \delta_{X_n}$ are the Dirac measures at the observations (i.e. $\delta_{X_i}(C) = \mathbf{1}_C(X_i)$ for each $C \in \mathcal{A}$).

Rmk. Some useful identities

1) $\mathbb{P}_n(C) = (\#\text{ of } X_i \in C)/n$ for all $C \in \mathcal{A}$
2) $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(X_i)$
3) $\mathbb{P}f = \int f d\mathbb{P} = \int f(X)dP = \mathbb{E}[f(X)]$
4) $\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} = \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f|$
   $= \sup_{f\in\mathcal{F}} |n^{-1}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]|$

Rmk. Relationship between $\|\widehat{F}_n - F\|_\infty$ and $\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F}$
$\|\widehat{F}_n - F\|_\infty = \|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F}$ if $\mathcal{F} = \{\mathbf{1}_{(-\infty,t]}(\cdot) | t \in \mathbb{R}\}$.

Def 6. $\mathcal{F}$ is a **Glivenko-Cantelli (GC) class** if

$$\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} \xrightarrow{p} 0 \quad \text{as } n \to \infty$$

Def 7. $\mathcal{F}$ is a **strong Glivenko-Cantelli (GC) class** if

$$\|\mathbb{P}_n - \mathbb{P}\|_\mathcal{F} \xrightarrow{a.s.} 0 \quad \text{as } n \to \infty$$

e.g. Failure of the uniform law
$\mathcal{F}_\mathcal{S} = \{\mathbf{1}_S(\cdot) \mid S \in \mathcal{S}\}$ where $\mathcal{S} = \{S \subset [0,1] \mid card(S) <$

$\infty\}$ is not a GC class if $\mathbb{P}$ has no atoms (i.e. $\mathbb{P}(\{x\}) = 0$ for all $x \in [0,1]$).

Here we abuse the notation of distributions so that for any $0 \le a < b \le 1$, $E \subset [0,1]$, and disjoint $\{E_m\}_{m\in\mathbb{N}}$ in $[0,1]$,

- $\mathbb{P}((a,b]) := \mathbb{P}(b) - \mathbb{P}(a)$
- $\mathbb{P}(\{a\}) := \mathbb{P}(a) - \mathbb{P}(a-)$
- $\mathbb{P}(E^c) := 1 - \mathbb{P}(E)$
- $\mathbb{P}(\cup_{m\in\mathbb{N}} E_m) := \sum_{m\in\mathbb{N}} \mathbb{P}(E_m)$

PROOF. For all $S \in \mathcal{S}$, $\mathbb{P}(S) = 0$, but any realization of $X_1, ..., X_n$ belongs to $\mathcal{S}$, which means

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{S\in\mathcal{S}} |\mathbb{P}_n(S) - \mathbb{P}(S)| = |1 - 0| = 1,$$

for every $n \in \mathbb{N}$. □

Now we consider a family of probability distributions $\mathcal{P}_\Theta := \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ indexed by a parameter set Θ. Fix a "true" parameter $\theta^* \in \Theta$ and let the sample $\{X_i\}_{i=1}^n$ be drawn according to $\mathbb{P}_{\theta^*}$. Let $\mathbb{E}_\theta$ denote the expectation corresponding to $\mathbb{P}_\theta$ (i.e. $\mathbb{E}_\theta[f(X)] = \int f d\mathbb{P}_\theta = \mathbb{P}_\theta f$).

The empirical risk minimization (ERM) approach of estimating θ^* begins by forming a cost function $\mathcal{L}_{(\cdot)}(\cdot) : \Theta \times \mathcal{X} \to \mathbb{R}$, where $\mathcal{L}_\theta(X)$ measures the "fit" between a parameter θ and the observation X.

Def 8. Population Risk at θ, $R(\theta, \theta^*)$

$$R(\theta, \theta^*) := \mathbb{E}_{\theta^*}[\mathcal{L}_\theta(X)] = \mathbb{P}_{\theta^*}\mathcal{L}_\theta \quad \text{where } X \sim \mathbb{P}_{\theta^*}$$

Def 9. Empirical Risk at θ, $\widehat{R}_n(\theta, \theta^*)$

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

Ideally, we want to minimize the population risk, but that is impossible since $\mathbb{P}_\theta^*$ is unknown. Instead, the estimator is obtained by minimizing the empirical risk over some $\Theta_0 \subset \Theta$, namely,

$$\widehat{\theta} := \arg\min_{\theta\in\Theta_0} \widehat{R}_n(\theta, \theta^*).$$

As a result, we want to bound the difference between the actual minimal population risk and the population risk at $\widehat{\theta}$. We define this difference to be the excess risk.

Def 10. Excess Risk, $E(\widehat{\theta}, \theta^*)$

$$E(\widehat{\theta}, \theta^*) := R(\widehat{\theta}, \theta^*) - \inf_{\theta\in\Theta_0} R(\theta, \theta^*) \ge 0$$

Rmk. Excess risk decomposition
Assume there exists some $\theta_0 \in \Theta_0$ such that $R(\theta_0, \theta^*) = \inf_{\theta\in\Theta_0} R(\theta, \theta^*)$ (if not, one can choose some θ_0 for which the equality holds up to an arbitrarily small tolerance $\epsilon > 0$), then $E(\widehat{\theta}, \theta^*) = T_1 + T_2 + T_3$ where

$$T_1 = R(\widehat{\theta}, \theta^*) - \widehat{R}_n(\widehat{\theta}, \theta^*) = \mathbb{P}\mathcal{L}_{\widehat{\theta}} - \mathbb{P}_n\mathcal{L}_{\widehat{\theta}}$$
$$T_2 = \widehat{R}_n(\widehat{\theta}, \theta^*) - \widehat{R}_n(\theta_0, \theta^*) \le 0$$
$$T_3 = \widehat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*) = \mathbb{P}\mathcal{L}_{\theta_0} - \mathbb{P}_n\mathcal{L}_{\theta_0}$$

This illustrates the importance of GC classes - while T_3 can be bounded via the Hoeffding bound given that $\theta \mapsto \mathcal{L}_\theta$ is bounded, T_1 cannot be easily controlled due to the randomness of $\widehat{\theta}$.

Let $\mathcal{F} = \{\mathcal{L}_\theta : \mathcal{X} \to \mathbb{R} \mid \theta \in \Theta_0\}$, then

$$T_1, T_3 \le \sup_{\theta\in\Theta_0} |\mathbb{P}\mathcal{L}_\theta - \mathbb{P}_n\mathcal{L}_\theta| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}},$$

so if $\mathcal{F}$ is a GC class we have

$$0 \le E(\widehat{\theta}, \theta^*) \le 2\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{p} 0.$$

e.g. Maximum Likelihood Estimation
Let $\mathcal{P}_\Theta$ be a set of distributions with strictly positive densities $\{p_\theta \mid \theta \in \Theta\}$, then for

$$\mathcal{L}_\theta(x) := log\Big(\frac{p_{\theta^*}(x)}{p_\theta(x)}\Big)$$

we have

$$R(\theta, \theta^*) = \mathbb{E}\Big[log\Big(\frac{p_{\theta^*}(X)}{p_\theta(X)}\Big)\Big] = D_{KL}(p_{\theta^*} \| \, p_\theta),$$

and if $\theta^* \in \Theta_0$,

$$E(\widehat{\theta}, \theta^*) = R(\widehat{\theta}, \theta^*) = D_{KL}(p_{\widehat{\theta}} \, \| \, p_{\theta^*}),$$

II. A UNIFORM LAW VIA RADEMACHER COMPLEXITY

For any $x_1^n := (x_1, ..., x_n) \subset \mathcal{X}$, define $\mathcal{F}(x_1^n) := \{(f(x_1), ..., f(x_n)) \mid f \in \mathcal{F}\} \subset \mathbb{R}^n$. Let $\{\varepsilon_i\}_{i=1}^n$ denote i.i.d. Rademacher variables (i.e. $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 0.5$). Moreover, for clarity let subscripts of expectations denote the random element(s) being integrated with respect to $\mathbb{P}$.

Def 11. Empirical Rademacher Complexity of $\mathcal{F}$ corr. to x_1^n, $\mathcal{R}(\mathcal{F}(x_1^n)/n)$

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

Def 12. Rademacher Complexity of $\mathcal{F}$, $\mathcal{R}_n(\mathcal{F})$

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

Lemma 2. (Ex 4.4: Details of symmetrization)
Let $\mathcal{G}$ be a class of real-valued, integrable functions on $(\mathcal{X}, \mathcal{A})$, then

1) $\sup_{g\in\mathcal{G}} \mathbb{E}[g(X)] \le \mathbb{E}[\sup_{g\in\mathcal{G}} |g(X)|]$
2) $\phi : \mathbb{R} \to \mathbb{R}$ convex, non-decreasing $\Rightarrow$ $\sup_{g\in\mathcal{G}} \phi(\mathbb{E}[g(X)]) \le \mathbb{E}[\phi(\sup_{g\in\mathcal{G}} |g(X)|)]$

Lemma 3. Inequality from symmetrization
Let $Y_1^n \equiv (Y_1, ..., Y_n)$ be an i.i.d. sequence equal in distri-

bution to and independent of X_1^n, $\phi : \mathbb{R} \to \mathbb{R}$ is convex and non-decreasing, then

$$\mathbb{E}_X[\phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})] \leq$$
$$\mathbb{E}_{X,Y,\varepsilon}\Big[\phi\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i)-f(Y_i))\Big)\Big|\Big]$$

PROOF.

$$\begin{aligned}
&\mathbb{E}_X[\phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})]\\
&= \mathbb{E}_X\Big[\phi\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)]\Big|\Big)\Big]\\
&= \mathbb{E}_X\Big[\phi\Big(\sup_{f\in\mathcal{F}}\Big|\mathbb{E}_Y\Big[\frac{1}{n}\sum_{i=1}^{n} f(X_i) - f(Y_i)\Big]\Big|\Big)\Big]\\
&\leq \mathbb{E}_X\Big[\phi\Big(\mathbb{E}_Y\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - f(Y_i)\Big|\Big]\Big)\Big]
\end{aligned}$$

by Lemma 2 and the fact that ϕ is non-decreasing

$$\leq \mathbb{E}_{X,Y}\Big[\phi\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - f(Y_i)\Big|\Big)\Big] \text{ by Jensen's ineq.}$$
$$= \mathbb{E}_{X,Y,\varepsilon}\Big[\phi\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(Y_i))\Big|\Big)\Big]$$

since each $f(X_i) - f(Y_i)$ is symmetric. □

Thm 4. (Thm 4.10)
$\mathcal{F}$ is b-uniformly bounded ($\|f\|_\infty \leq b \ \forall f \in \mathcal{F}$), $n \in \mathbb{N}$, $\delta > 0$, then

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

with P-probability $\geq 1 - exp(-n\delta^2/(2b^2))$

PROOF. *(Step 1: Concentration around mean)*
Recall the bounded differences inequality (Corollary 2.21): Suppose $X_1^n \equiv (X_1, ..., X_n)$ has independent components and $G : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$|G(x) - G(x^{\backslash k})| \leq L_k \quad \forall x, x' \in \mathbb{R}^n, k = 1, ..., n,$$

where

$$x_j^{\backslash k} := \begin{cases} x_j & \text{if } j \neq k \\ x'_k & \text{if } j = k, \end{cases}$$

then for all $\delta \geq 0$,

$$P(G(X_1^n) - \mathbb{E}_X[G(X_1^n)] \geq \delta) \leq exp(-\frac{2\delta^2}{\sum_{k=1}^n L_k^2}).$$

Here we define $\bar{f}(x) := f(x) - \mathbb{E}_X[f(X)] = f(x) - \mathbb{P}f$ and consider $G(x_1^n) := \sup_{f\in\mathcal{F}}|n^{-1}\sum_{i=1}^n \bar{f}(x_i)|$ and note that $G(X_1^n) = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$.

Observe that G is invariant to permutations of its coordinates, so it suffices to bound the difference when the first coordinate is perturbed. For simplicity, let $y_j = x_j^{\backslash 1}$.

For any $f \in \mathcal{F}$, we have

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

Taking the supremum over $f \in \mathcal{F}$ gives

$$G(x_1^n) - G(y_1^n) = \sup_{f\in\mathcal{F}}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\bar{f}(x_i)\Big| - G(y_1^n)\Big) \leq \frac{2b}{n},$$

and by applying the same argument with x_1^n, y_1^n reversed we get $|G(x) - G(y)| \leq 2b/n$. Thus, for all $\delta \geq 0$,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}_X[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \leq \delta$$

with P-probability $\geq 1 - exp(-2\delta^2/(n(2b/n)^2)) = 1 - exp(-n\delta^2/(2b^2))$.

(Step 2: Upper bound on mean)
We shall show $\mathbb{E}_X[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \leq 2\mathcal{R}_n(\mathcal{F})$ using Lemma 3. Let Y_1^n be an i.i.d. sequence equal in distribution to but independent of X_1^n, then

$$\begin{aligned}
&\mathbb{E}_X[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \leq \mathbb{E}_{X,Y,\varepsilon}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(Y_i))\Big|\Big]\\
&\leq 2\mathbb{E}_{X,\varepsilon}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\Big|\Big] = 2\mathcal{R}_n(\mathcal{F}),
\end{aligned}$$

since $h(t) := t$ is a convex and non-decreasing function. □

Corollary 4.1. $\mathcal{F}$ is b-uniformly bounded, then

$$\mathcal{R}_n(\mathcal{F}) = o(1) \Rightarrow \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{a.s.} 0$$

PROOF. We make use of the following lemma:

$$Y_n \xrightarrow{a.s.} Y \Leftrightarrow P(\|Y_n - Y\| > \eta \ i.o.) = 0 \ \forall \eta > 0.$$

Fix any $\eta > 0$ and let $\delta = \eta/2$. Given $\mathcal{R}_n(\mathcal{F}) = o(1)$, there exists $N \in \mathbb{N}$ such that $\mathcal{R}_n(\mathcal{F}) < \eta/4$ for all $n \geq N$. Then, for each $n \geq N$,

$$\begin{aligned}
P(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > \eta) &\leq P\Big(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 2\mathcal{R}_n(\mathcal{F}) + \frac{\eta}{2}\Big)\\
&= exp(-n\delta^2/(2b^2)),
\end{aligned}$$

and hence

$$\begin{aligned}
&\sum_{n=1}^{\infty} P(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > \eta) \leq \sum_{n=1}^{N-1} 1 + \sum_{n=N}^{\infty} exp(-n\delta^2/(2b^2))\\
&= (N-1) + \frac{exp(-N\delta^2/(2b^2))}{1 - exp(-\delta^2/(2b^2))} < \infty.
\end{aligned}$$

By the first Borel-Cantelli lemma ($\sum_{n=1}^{\infty} P(A_n) < \infty \Rightarrow$

$P(A_n\ i.o.) = 0)$, we have

$$P(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > \eta\ i.o.) = 0.$$

Since $\eta > 0$ is chosen arbitrarily, the desired result follows immediately. $\square$

A. Necessary conditions with Rademacher complexity

Here we define $\|\mathbb{S}_n\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |n^{-1}\sum_{i=1}^n \varepsilon_i f(X_i)|$ (so $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\varepsilon,X}[\|\mathbb{S}_n\|_{\mathcal{F}}]$) and $\bar{\mathcal{F}} := \{f - \mathbb{P}f \mid f \in \mathcal{F}\}$.

Thm 5. (Proposition 4.11)
$\phi : \mathbb{R} \to \mathbb{R}$ is convex, non-decreasing, then
$\mathbb{E}_{\varepsilon,X}[\phi(\frac{1}{2}\|\mathbb{S}_n\|_{\bar{\mathcal{F}}})] \le \mathbb{E}_X[\phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})] \le \mathbb{E}_{\varepsilon,X}[\phi(2\|\mathbb{S}_n\|_{\mathcal{F}})]$

PROOF. Again, let Y_1^n be an i.i.d. sequence equal in distribution but independent of X_1^n.
(Step 1: $\mathbb{E}_{\varepsilon,X}[\phi(1/2\|\mathbb{S}_n\|_{\bar{\mathcal{F}}})] \le \mathbb{E}_X[\phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})]$)
Let $T_1 := n^{-1}\sum_{i=1}^n \varepsilon_i(f(X_i) - f(Y_i))$, then by the fact that $\mathbb{E}_X[X_i] = \mathbb{E}_Y[Y_i]$ and the linearity of expectations we have

$$\mathbb{E}_{\varepsilon,X}[\phi(\frac{1}{2}\|\mathbb{S}_n\|_{\bar{\mathcal{F}}})] = \mathbb{E}_{X,\varepsilon}\Big[\phi\Big(\frac{1}{2}\sup_{f\in\mathcal{F}}|\mathbb{E}_Y[T_1]|\Big)\Big]$$
$$\le \mathbb{E}_{X,\varepsilon}\Big[\phi\Big(\mathbb{E}_Y\Big[\frac{1}{2}\sup_{f\in\mathcal{F}}|T_1|\Big]\Big)\Big] \quad \text{by Lemma 2}$$
$$\le \mathbb{E}_{X,Y,\varepsilon}\Big[\phi\Big(\frac{1}{2}\sup_{f\in\mathcal{F}}|T_1|\Big)\Big] \quad \text{by Jensen's inequality}$$
$$= \mathbb{E}_{X,Y}\Big[\phi\Big(\frac{1}{2}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\Big|\Big)\Big] \quad \text{by symmetry.}$$

Now let $T_2 := 1/2\sup_{f\in\mathcal{F}}|n^{-1}\sum_{i=1}^n f(X_i) - f(Y_i)|$ and $T_Z := \sup_{f\in\mathcal{F}}|n^{-1}\sum_{i=1}^n f(Z_i) - \mathbb{P}f|$ for any Z_1^n, then

$$T_2 = \frac{1}{2}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n(f(X_i) - \mathbb{P}f) - (f(Y_i) - \mathbb{P}f)\Big|$$
$$\le \frac{1}{2}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{P}f\Big| + \frac{1}{2}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{P}f\Big|$$
$$= \frac{1}{2}T_X + \frac{1}{2}T_Y.$$

Given ϕ is non-decreasing and convex,

$$\phi(T_2) \le \phi(\frac{1}{2}T_X + \frac{1}{2}T_Y) \le \frac{1}{2}\phi(T_X) + \frac{1}{2}\phi(T_Y),$$

and it follows after the fact $X_i \stackrel{d}{=} Y_i$ that

$$\mathbb{E}_{\varepsilon,X}[\phi(\frac{1}{2}\|\mathbb{S}_n\|_{\bar{\mathcal{F}}})] \le \mathbb{E}_{X,Y}[\phi(T_2)]$$
$$\le \frac{1}{2}\mathbb{E}_X[\phi(T_X)] + \frac{1}{2}\mathbb{E}_Y[\phi(T_Y)]$$
$$= \mathbb{E}_X[\phi(T_X)] = \mathbb{E}_X[\phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})].$$

(Step 2: $\mathbb{E}_X[\phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})] \le \mathbb{E}_{\varepsilon,X}[\phi(2\|\mathbb{S}_n\|_{\mathcal{F}})]$)
Define for any Z_1^n, $S_Z := \sup_{f\in\mathcal{F}}|n^{-1}\sum_{i=1}^n \varepsilon_i f(Z_i)|$, then

$$\mathbb{E}_X[\phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})]$$
$$\le \mathbb{E}_{X,Y,\varepsilon}\Big[\phi\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - f(Y_i))\Big)\Big|\Big] \quad \text{by Lemma 3}$$
$$\le \mathbb{E}_{X,Y,\varepsilon}[\phi(S_X + S_Y)] \quad \text{since } \phi \text{ is non-decreasing}$$
$$\le \frac{1}{2}\mathbb{E}_{X,\varepsilon}[\phi(2S_X)] + \frac{1}{2}\mathbb{E}_{Y,\varepsilon}[\phi(2S_Y)] \quad \text{by Jensen's ineq.}$$
$$= \mathbb{E}_{X,\varepsilon}[\phi(2S_X)] = \mathbb{E}_{X,\varepsilon}[\phi(2\|\mathbb{S}_n\|_{\mathcal{F}})]$$

since $X_i \stackrel{d}{=} Y_i$ and $S_X = \|\mathbb{S}_n\|_{\mathcal{F}}$. $\square$

Thm 6. (Proposition 4.12, Ex. 4.5)
$\mathcal{F}$ is b-uniformly bounded, $n \in \mathbb{N}$, $\delta > 0$, then

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

with P-probability $\ge 1 - exp(-n\delta^2/(2b^2))$

Corollary 6.1. $\mathcal{F}$ is b-uniformly bounded, then

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{p} 0 \Rightarrow \mathcal{R}_n(\mathcal{F}) = o(1)$$

III. Upper bounds on the Rademacher complexity

A. Classes with polynomial discrimination

Def 13. $\mathcal{F}$ has **polynomial discrimination of order $v \in \mathbb{N}$ (PD_v)** if

$$card(\mathcal{F}(x_1^n)) \le (n+1)^v \quad \forall n \in \mathbb{N},\ x_1^n \subset \mathcal{X}$$

(i.e. $f \in \mathcal{F}$ maps x_1^n to at most $(n+1)^v$ points in $\mathbb{R}^n$)

Lemma 7. (Lemma 4.14, Ex. 4.9)
$\mathcal{F}$ has PD_v, then for all $n \in \mathbb{N}$, $x_1^n \subset \mathbb{R}$,

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) \le 4D(x_1^n)\sqrt{\frac{vlog(n+1)}{n}}$$

where $D(x_1^n) := \sup_{p\in\mathcal{F}(x_1^n)/\sqrt{n}}\|p\|_2 = \sup_{f\in\mathcal{F}}\sqrt{n^{-1}\sum_{i=1}^n f^2(x_i)}$

Corollary 7.1. $\mathcal{F}$ b-uniformly bounded and has PD_v, then for all $n \in \mathbb{N}$,

$$\mathcal{R}_n(\mathcal{F}) \le 4b\sqrt{\frac{vlog(n+1)}{n}}$$

PROOF. $D(x_1^n) = \sup_{f\in\mathcal{F}}\sqrt{n^{-1}\sum_{i=1}^n|f(x_i)|^2} \le b$ since $|f(x_i)| \le b$ for each x_i. $\square$

Corollary 7.2. $\mathcal{F}$ b-uniformly bounded and has $PD_v \Rightarrow \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \xrightarrow{a.s.} 0$

PROOF. Follows immediately after Corollary 4.1 of Thm. 4 since $\sqrt{n^{-1}vlog(n+1)} \to 0$ as $n \to \infty$. $\square$

Corollary 7.3. Classical Glivenko-Cantelli (Cor. 4.15)

$$P\Big(\|\widehat{F}_n - F\|_\infty \ge 8\sqrt{\frac{log(n+1)}{n}} + \delta\Big) \le exp(-\frac{n\delta^2}{2}) \quad \forall\delta \ge 0$$

PROOF. Consider $\mathcal{F} := \{\mathbf{1}_{(-\infty,t]} \mid t \in \mathbb{R}\}$. For any $x_1^n \subset \mathcal{X}$, we can reorder it as $x_{(1)}^{(n)}$ such that $x_{(1)} \le \cdots \le x_{(n)}$, and

$$card(\mathcal{F}(x_1^n)) = card(\mathcal{F}(x_{(1)}^{(n)}))$$
$$\le card(\{(0,\ldots,0),(1,0,\ldots,0),\ldots,(1,\ldots,1)\})$$
$$= n+1.$$

Hence, $\mathcal{F}$ has PD_1. $\mathcal{F}$ is 1-uniformly bounded, then by Cor. 7.1 for all $n \in \mathbb{N}$,

$$\mathcal{R}_n(\mathcal{F}) \leq 4\sqrt{\frac{log(n+1)}{n}}.$$

Apply Thm. 4 and the result follows. □

This result can be utilized to prove Thm. 1 using a similar argument to that in the proof of Corollary 4.1.

B. Vapnik-Chervonekis dimension

Following on, suppose the functions in $\mathcal{F}$ are binary-valued and define $\mathcal{F}_\mathcal{S} := \{\mathbf{1}_S(\cdot) \mid S \in \mathcal{S}\}$ for an arbitrary class $\mathcal{S}$ of subsets in $\mathcal{X}$. For simplicity, let $\mathcal{S}(x_1^n) := \mathcal{F}_\mathcal{S}(x_1^n)$.

Def 14. $\boldsymbol{\mathcal{F}}$ **shatters** $\boldsymbol{x_1^n \subset \mathcal{X}}$ if $card(\mathcal{F}(x_1^n)) = 2^n$

Def 15. VC Dimension of $\boldsymbol{\mathcal{F}}$**,** $\boldsymbol{v(\mathcal{F})}$

$$v(\mathcal{F}) := max\{n \in \mathbb{N} \mid \exists x_1^n \subset \mathcal{X} \text{ shattered by } \mathcal{F}\}$$

Def 16. $\boldsymbol{\mathcal{S}}$ **picks out** $\boldsymbol{C \subset x_1^n}$ if $\exists S \in \mathcal{S}\ s.t.\ C = S \cap x_1^n$

Def 17. $\boldsymbol{\mathcal{S}}$ **shatters** $\boldsymbol{x_1^n \subset \mathcal{X}}$ if $\mathcal{S}$ picks out all $C \in 2^\mathcal{S}$

Def 18. VC Dimension of $\boldsymbol{\mathcal{S}}$**,** $\boldsymbol{v(\mathcal{S})}$

$$v(\mathcal{S}) := max\{n \in \mathbb{N} \mid \exists x_1^n \subset \mathcal{X} \text{ shattered by } \mathcal{S}\}$$

Rmk. Relationship between $\boldsymbol{\mathcal{S}}$ **and** $\boldsymbol{\mathcal{F}_\mathcal{S}}$

1) $\mathcal{S}$ shatters x_1^n iff $\mathcal{F}_\mathcal{S}$ shatters x_1^n
2) $v(\mathcal{S}) = v(\mathcal{F}_\mathcal{S})$

Def 19. $\boldsymbol{\mathcal{F}}$ or $\boldsymbol{\mathcal{S}}$ is a **VC class** if $v(\mathcal{F})$ or $v(\mathcal{S}) < \infty$

Lemma 8. (Ex. 4.10)

$$\binom{n-1}{k} + \binom{n-1}{k-1} = \binom{n}{k}$$

Thm 9. (Prop. 4.18, Ex. 4.11: VC, Sauer and Shelah)
$\mathcal{S}$ is a VC class, then for all $n \geq v(\mathcal{S})$, $x_1^n \subset \mathcal{X}$,

$$card(\mathcal{S}(x_1^n)) \leq \sum_{i=0}^{v(\mathcal{S})} \binom{n}{i} \leq (n+1)^{v(\mathcal{S})}$$

PROOF. This proof is left as an exercise for the reader XD. □

C. Controlling the VC dimension

Thm 10. (Prop. 4.19, Ex. 4.8)
$\mathcal{S}$, $\mathcal{T}$ are VC classes, then the following are also VC classes:

- $\mathcal{S}^c := \{S^c \mid S \in \mathcal{S}\}$
- $\mathcal{S} \sqcup \mathcal{T} := \{S \cup T \mid S \in \mathcal{S}, T \in \mathcal{T}\}$
- $\mathcal{S} \sqcap \mathcal{T} := \{S \cap T \mid S \in \mathcal{S}, T \in \mathcal{T}\}$

Def 20. Subgraph of $\boldsymbol{g : \mathcal{X} \to \mathbb{R}}$ **at level 0,** $\boldsymbol{S_g}$

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

Def 21. Subgraph Class of $\boldsymbol{\mathcal{G} \subset \mathbb{F}(\mathcal{X}, \mathbb{R})}$**,** $\boldsymbol{\mathcal{S}(\mathcal{G})}$

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

Thm 11. (Prop. 4.20: Finite-dim vector spaces)
$\mathcal{G} \subset \mathbb{F}(\mathbb{R}^d, \mathbb{R})$ is a vector space, $dim(\mathcal{G}) < \infty$, then $v(\mathcal{S}(\mathcal{G})) \leq dim(\mathcal{G})$

PROOF. We show that no collection of $n := dim(\mathcal{G}) + 1$ points can be shattered by $\mathcal{S}(\mathcal{G})$. Fix any $x_1^n \subset \mathbb{R}^d$ and define $L : \mathcal{G} \to \mathbb{R}^n$ by $L(g) = (g(x_1), ..., g(x_n))$. It can be easily verified that L is linear, so $L(\mathcal{G})$ is a linear subspace of $\mathbb{R}^n$. Let $g_1, \ldots, g_{n-1}$ be a basis of $\mathcal{G}$, then $L(\mathcal{G}) = L(span(\{g_1, \ldots, g_{n-1}\})) = span(L(\{g_1, \ldots, g_{n-1}\}))$. So, $dim(L(\mathcal{G})) \leq card(L(\{g_1, \ldots, g_{n-1}\}) \leq n - 1$. It follows $L(\mathcal{G})^\perp \neq \{\mathbf{0}_n\}$ and hence we can find $\gamma \in \mathbb{R}^n \setminus \{\mathbf{0}_n\}$ such that $\sum_{i=1}^n \gamma_i g(x_i) = \langle \gamma, L(g) \rangle = 0$ for all $g \in \mathcal{G}$. Then,

$$LHS := \sum_{\{i|\gamma_i \leq 0\}} (-\gamma_i) g(x_i) = \sum_{\{i|\gamma_i > 0\}} \gamma_i g(x_i) =: RHS.$$

It suffices to prove that $\mathcal{S}(\mathcal{G})$ cannot pick out $\{x_i | \gamma_i \leq 0\}$. Suppose the contrary, and let $S_g \in \mathcal{S}(\mathcal{G})$ be the set that satisfies $S_g \cap x_1^n = \{x_i | \gamma_i \leq 0\}$. In this case,

$$g(x_i) \begin{cases} \leq 0 & \text{if } \gamma_i \leq 0 \\ > 0 & \text{if } \gamma_i > 0. \end{cases}$$

Then, $LHS \leq 0$ while $RHS > 0$, which contradicts the previous equation. □