Proof. $x^{(k)}(f) \multimap (2\pi i f)^k \hat{x}(f)$

$$|x^{(k)}(f)| \le C \int_{-\infty}^{\infty} |\hat{x}(f)| |f|^k \le C \int_{-\infty}^{\infty} |\hat{x}(f)| (1+|f|)^k df$$

$$\Rightarrow \int_{-\infty}^{\infty} |\hat{x}(f)| |f|^k df < \infty \quad , \text{for every } k \le p .$$

$$\boxed{|\hat{x}(f)| \le \frac{C'}{(1+|f|)^{p+1+\varepsilon}}}$$

11.1.3. Non-Linear Approximation

B - ONB

$$x = \sum_{j=0}^{\infty} \langle x, g_j \rangle g_j \quad , \quad x_M = \sum_{j \in I_M} \langle x, g_j \rangle g_j$$

$$\varepsilon[M] = \|x - x_M\|^2 = \| \sum_{j=0}^{\infty} \langle x, g_j \rangle g_j - \sum_{j \in I_M} \langle x, g_j \rangle g_j \|^2$$

$$= \| \sum_{j \notin I_M} \langle x, g_j \rangle g_j \|^2 = \sum_{j \notin I_M} |\langle x, g_j \rangle|^2$$

$$\|x\|^2 = \sum_j |\langle x, g_j \rangle|^2 = \sum_{j \in I_M} |\langle x, g_j \rangle|^2 + \sum_{j \notin I_M} |\langle x, g_j \rangle|^2$$

$$\varepsilon[M] = \|x\|^2 - \sum_{j \in I_M} |\langle x, g_j \rangle|^2$$

sort $\{|\langle x, g_j \rangle|\}_{j \in \mathbb{N}}$ in descending order and denote $x_B^r[k] = \langle x, g_{j_k} \rangle$

$$|x_B^r[k]| \ge |x_B^r[k+1]| \quad , k \ge 1$$

$$x_M = \sum_{k=1}^{M} x_B^r[k] g_{j_k}$$

$$\varepsilon_n[M] = \|x - x_M\|^2 = \sum_{k=M+1}^{\infty} |x_B^r[k]|^2$$

Thm. 11.4. Let $s>1/2$. If there exists $C>0$ s.t. $|x_B^r[k]| \le C k^{-s}$, then

$$\varepsilon_n[M] \le \frac{C^2}{2s-1} M^{1-s\cdot 2}. \qquad (*)$$

Conversely, if $\varepsilon_n[M]$ satisfies $(*)$, then

$$|x_B^r[k]| \le \left(1-\frac{1}{2s}\right)^{-s} C k^{-s}. \qquad (**)$$

$$\|x\|_{B,p} = \left(\sum_{j=0}^{\infty} |\langle x, g_j\rangle|^p \right)^{1/p}$$

Thm 11.5. Let $p<2$. If $\|x\|_{B,p} < \infty$, then

$$|x_B^r[k]| \le \|x\|_{B,p}\, k^{-1/p}$$

and

$$\varepsilon_n[M] = o(M^{1-2/p}), \text{ i.e.,}$$

i.e. $\lim_{M\to\infty} \varepsilon_n[M] M^{-1+2/p} = 0.$

$$\mathcal{B}_{B,p} = \{x \in \mathcal{H} : \|x\|_{B,p} < \infty\}$$

Summary. If $x \in \mathcal{B}_{B,p}$, then $\varepsilon_n[M] = o(M^{1-2/p})$. This is called a "Jackson type inequality". Conversely, if $\varepsilon_n[M] = o(M^{1-2/p})$, then the "Bernstein inequality" guarantees $|x_B^r[k]| \le C' k^{-1/p}$, which shows that $x \in \mathcal{B}_{B,q}$ for every $q>p$, we have

$$\sum_k |x_B^r[k]|^q \le \sum_k C' k^{-q/p} < \infty.$$

E.g. for wavelet bases, the space $\mathcal{B}_{B,p}$ is a Besov space.
WH bases, the space $\mathcal{B}_{B,p}$ is a Feichtinger–Gröchenig modulation space.

Chapter 12. Uniform Laws of large numbers

12.1. Uniform convergence of CDFs

X ... RV

$F(t) = P[X \le t]$, $t \in \mathbb{R}$.

Problem. We want to estimate $F(t)$ from a given collection $\{X_i\}_{i=1}^n$ of i.i.d. samples each drawn according to the law specified by F.

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, t]}[X_i], \quad \text{empirical CDF}$$

where $1_{(-\infty,t]}[x]$ is the indicator function of the event $\{x \le t\}$.

$$F(t) = E[1_{(-\infty,t]}[X]] \quad \leftarrow \text{population CDF}$$

$$= \int_{-\infty}^{\infty} 1_{(-\infty,t]}(x) f(x)\, dx$$

$$= \int_{-\infty}^{t} f(x)\, dx \quad (= F(t))$$

want to understand the properties of $\hat{F}_n(t)$.

$$E[\hat{F}_n(t)] = \frac{1}{n} \sum_{i=1}^{n} E[1_{(-\infty,t]}[X_i]]$$

$$= \frac{1}{n} \sum_{i=1}^{n} F(t) = F(t)$$

One can show that, indeed, $\hat{F}_n(t)$ converges almost surely to $F(t)$, i.e.,

$$P\left[\lim_{n \to \infty} \hat{F}_n(t) = F(t)\right] = 1, \quad \forall t \in \mathbb{R}.$$

Functionals ρ that map a CDF F to a real number, namely $\rho(F)$, i.e., $F \mapsto \rho(F)$

Problem: We want to estimate $\rho(F)$ from $\{X_i\}_{i=1}^n$

Q: Can we get good estimates of $\rho(F)$ by replacing F by $\hat{F}_n$, i.e.,

what can we say about $\rho(\hat{F}_n)$? In particular, what can we say about the convergence of $\rho(\hat{F}_n)$ to $\rho(F)$ as $n \to \infty$.

Example 12.1. (Expectation functional). Given an integrable g, we define

$$\rho_g(F) = E[g(X)] = \int g(x) f(x) dx = \int g(x) dF(x).$$

e.g. $g(x) = x \Rightarrow \rho_g(F) = \int x f(x) dx = E[X]$.

$\rho_g(\hat{F}_n)$ relative to $\rho_g(F)$

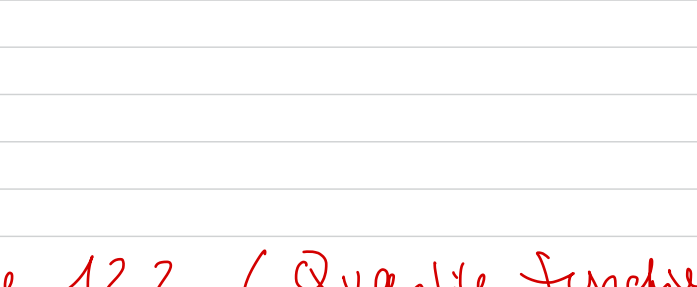
↑ plug-in estimator

$$\rho_g(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

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

$$\frac{dF(x)}{dx} = \frac{1}{n} \sum_{i=1}^{n} \delta(x - X_i)$$

↑ Dirac delta



$$1_{(-\infty, x]}[X_i] = \begin{cases} 1, & X_i \le x \\ 0, & \text{else} \\ & (x < X_i) \end{cases}$$

plug-in $F \Rightarrow \hat{F}_n$

$$\int g(x) dF(x) = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

↑ to be shown

$$\int_{-\infty}^{\infty} g(x) \frac{1}{n} \sum_{i=1}^{n} \delta(x - X_i) dx = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} g(x) \delta(x - X_i) dx$$

$$= \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Example 12.2. (Quantile functionals). For any $\alpha \in (0,1)$, the quantile functional Q_α is given by

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

(median corr. to $\alpha = 0.5$). The plug-in est. is

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

Want to understand convergence of $\rho_g(\hat{F}_n)$ to $\rho_g(F)$ as $n \to \infty$.

Def. 12.4. We say that a sequence of random variables X_n converges in probability to the random variable X if

$$\lim_{n\to\infty} P\{|X - X_n| > \varepsilon\} = 0, \quad \forall \varepsilon > 0.$$

a.s. $\Rightarrow$ p. (Fatou's lemma)

$\|G - F\|_\infty = \sup_{t\in\mathbb{R}} |G(t) - F(t)|$, distance between F and G

$\left(\|x\|_\infty = \sup_{t\in\mathbb{R}} |x(t)|\right)$

Def. 12.5. The functional ρ is continuous at F in the sup-norm, if, for all $\varepsilon > 0$, there exists a $\delta > 0$ s.t. $\|G - F\|_\infty \leq \delta \Rightarrow |\rho(G) - \rho(F)| < \varepsilon$.

We establish that if ρ_g is continuous in the sup-norm (at F), then $\rho_g(\hat{F}_n) \xrightarrow{p} \rho_g(F)$.

Proof follows by combining continuity of the functional with

Thm. 12.6. (Glivenko-Cantelli). For any distribution, the empirical CDF $\hat{F}_n$ satisfies

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

We need to show that

$$\lim_{n\to\infty} P[|\rho(\hat{F}_n) - \rho(F)| > \varepsilon] = 0, \quad \forall \varepsilon > 0.$$

(conv. in prob.)

$$\|\hat{F}_n - F\|_\infty \le \delta \Rightarrow |\rho(\hat{F}_n) - \rho(F)| \le \varepsilon \qquad \text{(continuity)}$$

Hence, it follows that

$$|\rho(\hat{F}_n) - \rho(F)| > \varepsilon \Rightarrow \|\hat{F}_n - F\|_\infty > \delta$$

$$\Rightarrow$$

$$P\left[|\rho(\hat{F}_n) - \rho(F)| > \varepsilon\right] \le P\left[\|\hat{F}_n - F\|_\infty > \delta\right]$$

now, Glivenko-Cantelli $\|\hat{F}_n - F\|_\infty \xrightarrow{a.s.} 0 \Rightarrow \|\hat{F}_n - F\|_\infty \xrightarrow{p.} 0$, i.e., ($a.s. \Rightarrow p.$)

$$\lim_{n\to\infty} P\left[\|\hat{F}_n - F\|_\infty > \varepsilon'\right] = 0, \forall \varepsilon' > 0$$

This establishes that

$$\lim_{n\to\infty} P\left[|\rho(\hat{F}_n) - \rho(F)| > \varepsilon\right] \le \lim_{n\to\infty} P\left[\|\hat{F}_n - F\|_\infty > \delta\right] = 0$$

$$\Rightarrow \lim_{n\to\infty} P\left[|\rho(\hat{F}_n) - \rho(F)| > \varepsilon\right] = 0, \forall \varepsilon > 0$$

12.2. Uniform laws for more general function classes

Let $\mathcal{F}$ be a class of integrable functions with domain $\mathcal{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution P over $\mathcal{X}$.

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

Def. 12.7. We say that $\mathcal{F}$ is a Glivenko-Cantelli class for P if $\|P_n - P\|_{\mathcal{F}}$ converges to zero in prob. as $n\to\infty$.

Ex. 12.8. (Empirical CDF and indicator functions).

$$\mathcal{F} = \{1_{(-\infty,t]}(\cdot) \mid t\in\mathbb{R}\}.$$

For fixed $t\in\mathbb{R}$, we have

$$E[\mathbb{1}_{(-\infty,t]}(X)] = P[X \le t] = F(t)$$

GC. $\|\hat{F}_n - F\|_\infty \xrightarrow{a.s.} 0$, i.e., (a.s. $P(\lim_{n\to\infty} X_n = x) = 1$

$$P\left[\lim_{n\to\infty} \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| = 0\right] = 1$$

$$\Downarrow$$

$$P\left[\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i) - E[\mathbb{1}_{(-\infty,t]}(X)]\right| = 0\right] = 1$$

$$= P\left[\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E[f(X)]\right| = 0\right] = 1.$$

$\|P_n - P\|$

Where does this occur?

family of prob. distributions $\{P_\theta \mid \theta \in \Omega\}$

given $\{X_i\}_{i=1}^n$ drawn i.i.d. according to P_{θ^*}, for some fixed, but unknown $\theta^* \in \Omega$

$\Omega = \mathbb{R}^d$ vector estimation (finite-dimensional) parametric

$\Omega = \mathcal{G}$ function class non-parametric

A standard approach for estimating θ^* from $\{X_i\}_{i=1}^n$ is based on minimizing a cost function $L_\theta(x)$, which quantifies the fit between $\theta \in \Omega$ and the sample $\{X_i\}_{i=1}^n$.

$$\hat{R}_n(\theta, \theta^*) = \frac{1}{n}\sum_{i=1}^{n} L_\theta(X_i)$$ empirical risk

$$R(\theta, \theta^*) = E_{\theta^*}[L_\theta(X)]$$ population risk

in practice, typically minimize over a subset Ω_0 of Ω, $\Rightarrow$ est. $\hat{\theta}$

$$E(\hat{\theta}, \theta) = R(\hat{\theta}, \theta^*) - \inf_{\theta\in\Omega_0} R(\theta, \theta^*)$$

Ex. 12.9. (Maximum likelihood). $\{P_\theta | \theta \in \Omega\}$, θ^* unknown parameter

$\{X_i\}_{i=1}^n$

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

ML estimator

$$\hat{\theta} = \underset{\theta \in \Omega_0}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n \log\left[\frac{p_{\theta^*}(x_i)}{p_\theta(x_i)}\right] \right\}$$

$$= \underset{\theta \in \Omega_0}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n \log\left[\frac{1}{p_\theta(x_i)}\right] \right\}$$

population risk: $R(\theta, \theta^*) = E_{\theta^*}\left[\log \frac{p_{\theta^*}(x)}{p_\theta(x)}\right] = \underbrace{D(p_{\theta^*} \| p_\theta)}_{\text{Kullback-Leibler divergence}}$

$$\inf_{\theta \in \Omega_0} R(\theta, \theta^*) = \inf_{\theta \in \Omega_0} E_{\theta^*}\left[\log\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right)\right].$$

$D(p\|q) \geq 0$ with $=$ iff $p(x) = q(x), \forall x$.

$$E(\hat{\theta}, \theta^*) = R(\hat{\theta}, \theta^*) - \underbrace{\inf_{\theta \in \Omega_0} R(\theta, \theta^*)}_{=0 \text{ if } \theta^* \in \Omega_0} = R(\hat{\theta}, \theta^*) = D(p_{\theta^*} \| p_{\hat{\theta}})$$

□

How does ERM relate to $\|P_n - P\|_F$

assume that $\vec{\theta}_0 \in \Omega_0$ s.t. $R(\theta_0, \theta^*) = \inf_{\theta \in \Omega_0} R(\theta, \theta^*)$

$$E(\hat{\theta}, \theta^*) = \underbrace{R(\hat{\theta}, \theta^*) - \hat{R}_n(\hat{\theta}, \theta^*)}_{T_1} + \underbrace{\hat{R}_n(\hat{\theta}, \theta^*) - \hat{R}_n(\theta_0, \theta^*)}_{T_2} + \underbrace{\hat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*)}_{T_3}$$

$T_2 \leq 0$ because $\hat{\theta}$ minimizes ER over Ω_0.

$$T_3 = \frac{1}{n}\sum_{i=1}^{n} \ell_{\theta_0}(X_i) - E_X[\ell_{\theta_0}(X)]$$

concentration of measure inequ.

$$T_1 = E_X[\ell_{\hat\theta}(X)] - \frac{1}{n}\sum_{i=1}^{n} \ell_{\hat\theta}(X_i)$$

challenging because $\hat\theta$ is a random quantity

$$\mathcal{F} = \{\ell_\theta(\cdot) \mid \theta \in \Theta_0\}$$

$$T_1 \le \sup_{\theta\in\Theta_0} \left| \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i) - E_X[\ell_\theta(X)] \right| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$$

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