Contents

1 Fundamental Tools of Probability Theory

Law of Total Probability: if $B_1, \ldots, B_k$ partition the sample space Ω:

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Bayes Rule/theorem: if X and Y are random variables

$$\begin{aligned} p_{X|Y}(x|y) &= \frac{p_{XY}(x,y)}{p_Y(y)} \\ &= \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} \\ &= \begin{cases} \frac{p_{Y|X}(y|x)p_X(x)}{\int p_{Y|X}(y|x)p_X(x)} & \text{if continuous} \\ \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')} & \text{if discrete} \end{cases} \end{aligned}$$

Law of Total Expectation/Tower Law: $\mathbb{E}[g(X,Y)] = \mathbb{E}[\mathbb{E}[g(X,Y)|X]]$.

Law of Total Variance:

$$\begin{aligned} \text{Var}(Y) &= \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X]) \\ \text{Var}(Y) &= \mathbb{E}[\text{Var}(Y|X_1, X_2)] + \mathbb{E}[\text{Var}(\mathbb{E}[Y|X_1, X_2]|X_1)] + \text{Var}(\mathbb{E}[Y|X_1]) \end{aligned}$$

If $A_1, \ldots, A_n$ partition the whole outcome space then

$$\begin{aligned} \text{Var}(X) = &\sum_{i=1}^{n} \text{Var}(X|A_i)\text{Pr}(A_i) + \sum_{i=1}^{n} \mathbb{E}[X|A_i]^2(1 - \text{Pr}(A_i))\text{Pr}(A_i) \\ &- 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \mathbb{E}[X|A_i]\text{Pr}(A_i)\mathbb{E}[X|A_j]\text{Pr}(A_j) \end{aligned}$$

Law of Total Covariance:

$$\text{Cov}(X,Y) = \mathbb{E}(\text{Cov}(X,Y)|Z) + \text{Cov}(\mathbb{E}(X|Z), \mathbb{E}(Y|Z))\text{Cov}(X,Y) = \text{Cov}(X, \mathbb{E}(Y|X))$$

Moment Generating Functions: a powerful function that can be used to obtain polynomial moments of a random variable

$$M_X(t) = E(e^{tX}) = 1 + tE(X) + \frac{t^2E(X^2)}{2!} + \frac{t^3E(X^3)}{3!} + \ldots$$

The j-th moment of X is then:

$$E(X^j) = M^{(j)}(0) = \frac{d^j M_X(t)}{dt^j}\Big|_{t=0}$$

Properties

1. Location-scale: $M_{aX+b} = e^{bt}M_X(at)$
2. Multiplicity: $X \perp Y \implies M_{X+Y}(t) = M_X(t)M_Y(t)$

1.1 Useful facts of common parametric distributions

Normal Distribution:

1. Sums of Normals are Normal: For $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ independent, and $a_1, a_2 \in \mathbb{R}$,
$$a_1 X + a_2 Y \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$$
2. Connection to Chi-Square: if $X_1, \ldots, X_n \overset{iid}{\sim} N(0,1)$, then $Z_1 = X_1^2 \sim \chi_1^2$ and $Z_n = \sum_{i=1}^n X_i^2 \sim \chi_n^2$.
3. Stein's Lemma: useful for calculating higher order moments of normal distributions. For $X \sim N(\theta, \sigma^2)$:
$$\begin{aligned} \mathbb{E}(g(X)(X-\theta)) &= \sigma^2\mathbb{E}[g'(X)] \\ \text{In action: } \mathbb{E}[X^3] &= \mathbb{E}(X^2(X-\theta+\theta) \\ &= \mathbb{E}[X^2(X-\theta)] + \theta\mathbb{E}[X^2] \\ &= 2\sigma^2\mathbb{E}[X] + \theta(\sigma^2+\theta^2) \\ &= 3\sigma^2\theta + \theta^3 \end{aligned}$$

MVN Distribution:

1. Linear transformations are MVN:
$$Y = CX + b \sim N(C\mu + b, C\Sigma C^T)$$
2. Uncorrelated $\iff$ Independent: $\text{Cor}(X_i, X_j) \iff X_i \perp X_j$
3. Normal marginal: Suppose $X = (X_1, X_2)$, where $X_1 \in \mathbb{R}^{n_1}$ and $X_2 \in \mathbb{R}^{n_2}$. $\mu = (\mu_1, \mu_2)$, and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, then:
$$X_i \sim N(\mu_i, \Sigma_{ii})$$
4. Normal conditional:
$$X_1 | X_2 \sim N_{n_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11,2})$$
Where $\Sigma_{11,2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Chi-Square Distribution:

1. Centered and scaled MVN is Chi-Square: if $Y \sim N(\mu, \Sigma)$ of dimension n, then
$$(Y-\mu)^T\Sigma^{-1}(Y-\mu) \sim \chi_n^2$$
2. Projection property: if $Y \sim N(\mu, I)$ is n-dimensional MVN and P is a rank-p projection matrix
$$(Y-\mu)^T P(Y-\mu) \sim \chi_p^2$$
3. Special case of Gamma: $\chi_m^2 \equiv \text{Gamma}\left(\frac{m}{2}, 2\right)$.
4. Sample variance is Chi-square under normal model: when $X_i \sim N(\mu, \sigma)$, $(n-1)\frac{S_n^2}{\sigma^2} \sim \chi_{n-1}^2$

Uniform Distribution:

1. Scaled minimum of uniforms is exponential: If $X_1, \ldots, X_n \sim \text{Unif}(0,1)$, then $U = n \times \min(X_1, \ldots, X_n) \sim \text{Exp}(1)$

2. Probability integral transform: suppose X has continuous distribution with CDF F_X, then the random variable $Y := F_X(X) \sim \text{Unif}(0,1)$.

Poisson Distribution:

1. Sum of Poissons are Poisson: $X_1, \ldots, X_n \sim \text{Pois}(\lambda_i)$, then $\sum_{i=1}^n X_i \sim \text{Pois}\left(\sum_{i=1}^n \lambda_i\right)$
2. Connection to negative binomial: NB distribution can be considered a Poisson-Gamma mixture where $X|\lambda \sim \text{Pois}(\lambda)$ and $\lambda \sim \text{Gamma}(r, p/(1-p))$.

Exponential Distribution

1. Memoryless property: if $X \sim \text{Exp}(\lambda)$, then
$$P(X > x + y | X > x) = P(X > y)$$
2. Absolute Difference is Exponential: if $X, Y \sim \text{Exp}(1)$, then $|X - Y| \sim \text{Exp}(1)$
3. Minimum is Exponential: $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$, then $\min(X_1, \ldots, X_n) \sim \text{Exp}(n\lambda)$
4. Ratio is uniform: $X, Y \sim \text{Exp}(1)$, then $\frac{X}{X+Y} \sim \text{Unif}(0,1)$

Geometric Distribution

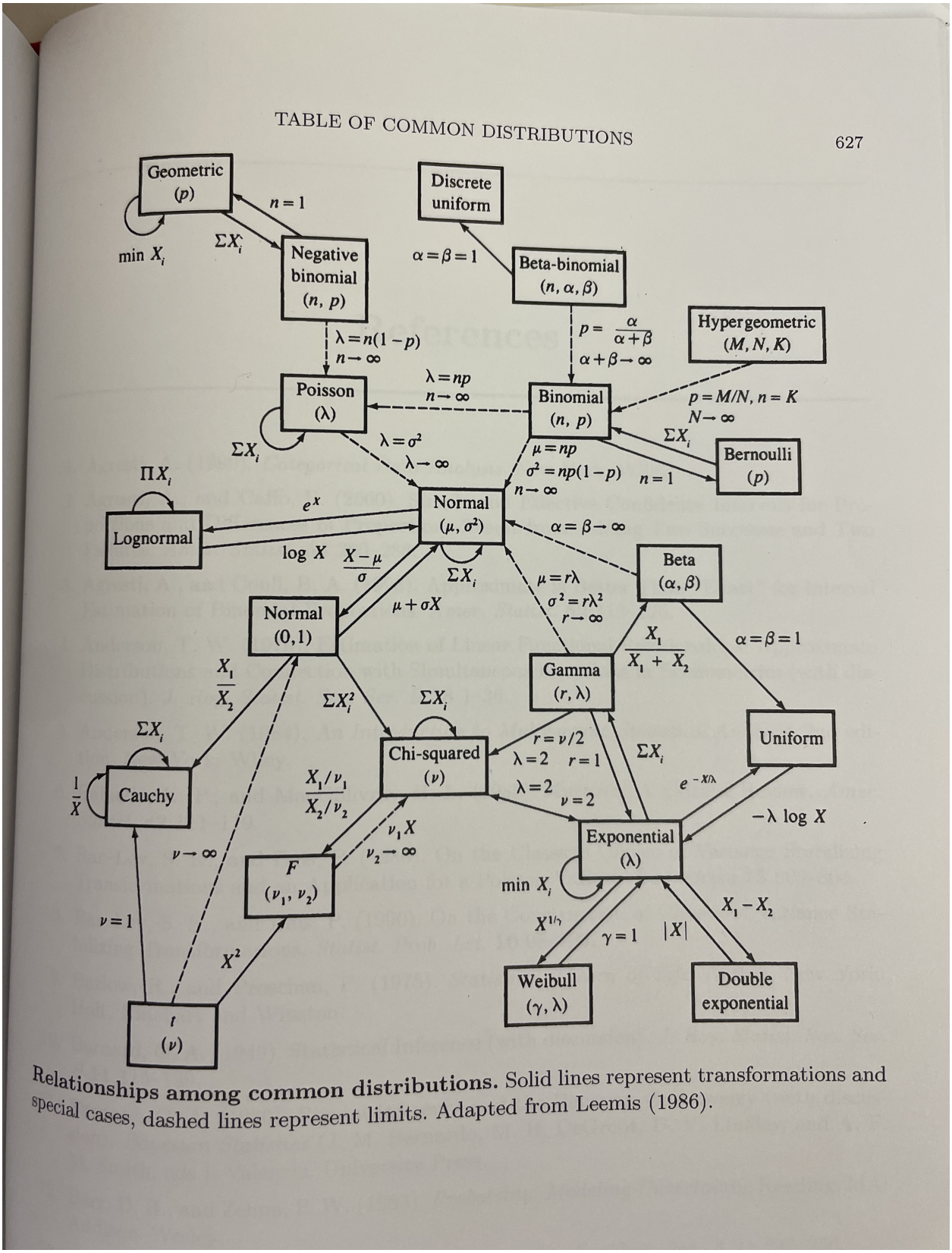
1. Memoryless property: if $X \sim \text{Geo}(p)$, then
$$P(X > x + y | X > x) = P(X > y)$$
2. Connection to negative binomial distribution: when parametrized a certain way, NB models either the number of failures before specified number of successes. Thus, the sum of r independent geometric random variables with parameter p is equivalent to $NB(r, p)$.

Gamma Distribution:

1. Sums of gammas are gamma: For $X_1, \ldots, X_n \sim \text{Gamma}(\alpha_i, \beta)$,
$$\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right)$$
2. Inverse Gamma Distribution: has pdf $f(x,\ \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp(-\beta/x)$. Has mean $\frac{\beta}{\alpha-1}$ and mode $\frac{\beta}{\alpha+1}$.

Multinomial Distribution: describes the frequency of observations over k categories:

1. Sum of independent multinomials (with same parameters) are multinomial: $X \sim M_k(n; p_1, \ldots, p_k)$, $Y \sim M_k(m; p_1, \ldots, p_k)$,
$$X + Y \sim M_k(n + m; p_1, \ldots, p_k)$$
2. Block decomposition: if we decompose $X_1, \ldots, X_k$ into r blocks B_j, they are conditionally independent given their block sum S_j:
$$B_j | S_j \sim M_{k_{j-1}+1, \ldots, k_j}\left(S_j; \frac{p_{k_{j-1}+1}}{\sum_{\ell=k_{j-1}+1}^{k_j} p_\ell}, \ldots, \frac{p_{k_j}}{\sum_{\ell=k_{j-1}+1}^{k_j} p_\ell}\right)$$
3. Negative correlation between entries: $\text{Cov}(X_i, X_j) = -np_ip_j$
4. MLE for p_j: if goal is to estimate probability of j-th category, requires optimization under constraint that $\sum_{i=1}^k p_i = 1$. Using Lagrange multipliers, we obtain $\hat{p}_j = \frac{X_j}{n}$
5. Dirichlet prior permits Bayesian inference on multinomial distribution.



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

1.2 Order Statistics

Marginal and Joint Distribution of Order Statistics:

1. Distribution of j-th order statistic:

$$p_{Y_{(j)}}(y) = \frac{n!}{(j-1)!(n-j)!} F_X(y)^{j-1} \left(1 - F_X(y)\right)^{n-j} p_X(y)$$

2. Joint distribution of ℓ and j-th order statistic:

$$p_{Y_{(j)},Y_{(\ell)}}(y,z) = \frac{n!}{(j-1)!(\ell-j-1)!(n-\ell)!} F_x(y)^{j-1} p_X(y) (F_X(z) - F_X(y))^{(\ell-j-1)} p_X(z) (1 - F_X(z))^{(n-\ell)}$$

3. Order statistics of $\text{Unif}(0,1)$ are Beta

$$Y_{(j)} \sim \text{Beta}(j, n-j+1)$$

4. Spacings between order statistics of $\text{Unif}(0,1)$ are Beta

$$W_i = Y_{(i)} - Y_{(i-1)} \sim \text{Beta}(1, n)$$

1.3 Identities

Calculus tricks:

1. Handy Derivative Rules

$$\frac{d}{dx} f(x)g(x) = f'(x)g(x) + f(x)g'(x)$$
$$\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}$$
$$\frac{d}{dx}[f^{-1}(x)] = \frac{1}{f'(f^{-1}(x))}$$

2. Integration by parts: $\int udv = uv - \int vdu$

Fundamental Theorem of Calculus/Leibniz Rule:

$$\frac{d}{dx} \int_{a(x)}^{b(x)} f(x,t)dt = f(x,b(x)) \cdot \frac{d}{dx} b(x) - f(x,a(x)) \cdot \frac{d}{dx} a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)dt$$

Binomial Theorem: For any real numbers, x, y and integer $n \geq 0$

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

Layer cake representation: for X a non-negative RV

$$\mathbb{E}[X] = \int_0^\infty P(X \geq t)dt$$

Woodbury identity: $(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$.

Blockwise matrix inversion: the inverse of a blocked matrix is

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

Taylor Expansion of a function: for f infinitely differentiable, Taylor expansion about x_0:

$$f(x) - f(x_0) = f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)(x - x_0)^n}{n!}$$

$L^r(P)$ space: set of all functions $f : \mathcal{X} \to \mathbb{R}$ such that $||f||_{L^r(P)} = \left[\int |f(x)|^r dP(x)\right]^{1/r} < \infty$.
Geometric series:

- Finite: for $r \neq 1$, $\sum_{i=0}^{n} ar^i = a\left(\frac{1-r^{n+1}}{1-r}\right)$
- Infinite: if $r < 1$, $\sum_{i=0}^{\infty} ar^i = \frac{a}{1-r}$

Exponential sums

- Finite: $\sum_{n=0}^{N-1} p^i e^{inx} = \frac{1-e^{iNx}}{1-e^{ix}}$
- Infinite: $\sum_{n=0}^{\infty} p^n e^{inx} = \frac{1}{pe^{ix}-1}$

Definitions of e

1. Limit definition: $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$
2. Power series: $\sum_{n\geq 0} \frac{x}{n!} = e^x$

1.4 Useful Inequalities

Markov Inequality: useful for probability tail bounds, if $X \geq 0, t > 0$

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Chebyshev Inequality: for k in natural numbers

$$P(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|^k)}{t^k}$$
$$\implies P(|X - \mathbb{E}(X)| \geq t) \leq \frac{\text{Var}(X)}{t^2}$$

Chernoff bound: if X has MGF M_X.

$$P\{X - \mathbb{E}(X) \geq t\} \leq \inf_{\lambda > 0} \frac{M_{X-\mu}(\lambda)}{e^{\lambda t}}$$
$$\log P\{X - \mathbb{E}(X) \geq t\} \leq -\sup_{\lambda > 0}\{\lambda t - \log M_{X-\mu}(\lambda)\}$$

Cauchy-Schwarz Inequality:

$$|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

Jensen Inequality: for RV X and f convex

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$$

Triangle/Reverse Triangle inequality:

$$||X + Y|| \leq ||X|| + ||Y||$$
$$\Big|||X|| - ||Y||\Big| \leq ||X - Y||$$

Integration by Parts: if g and h are cadlag functions with support $[a, b]$ (CDFs) it holds that

$$\int_{(a,b]} g(u)dh(u) + \int_{(a,b]} h(u)dg(u) = g(b)h(b) - g(a)h(a)$$

Max-Min Inequality: $\sup_z \inf_w f(z, w) \leq \inf_w \sup_z f(z, w)$

2 Convergence Theory

Convergence in probability: a sequence of random variables X_n converges in probability to X if

$$P(||X_n - \theta|| > \epsilon) \overset{n\to\infty}{\longrightarrow} 0$$

Weak Convergence: a sequence of random variables converges weakly/in distribution/in law iff for all bounded continuous functions $f : \mathbb{R}^d \to \mathbb{R}$:

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \text{ as } n \to \infty$$

Portmanteau Lemma: characterizes different definitions of weak convergence. For X_n as sequence of random variables and X a random variable, TFAE

1. $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ as $n \to \infty$ for all bounded continuous functions f.
2. CDF: for all continuity points $t \in \mathbb{R}^d$, $P(X_n \leq t) \to P(X \leq t)$ as $n \to \infty$
3. Levy Continuity (Characteristic Functions): for all $t \in \mathbb{R}^d$, $\mathbb{E}[\exp(it^T X_n)] \overset{n\to\infty}{\longrightarrow} \mathbb{E}[\exp(it^T X)]$.
4. Cramer-Wold: for all $t \in \mathbb{R}^d$, $t^T X_n \rightsquigarrow t^T X$

2.1 Convergence Theorems

Continuous Mapping Theorem: Let $X_n \in \mathbb{R}^d$ be a sequence of random variables and let $g : \mathbb{R}^d \to \mathbb{R}^m$ be a continuous function at every point such that $P(X \in C) = 1$. Then the following are valid

(i) if $X_n \rightsquigarrow X$, $g(X_n) \rightsquigarrow g(X)$.

(ii) if $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.

(iii) if $X_n \rightsquigarrow X$ and $||X_n - Y_n|| \xrightarrow{p} 0$, $Y_n \rightsquigarrow X$.

(iv) If $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{p} c$, then $(X_n, Y_n) \rightsquigarrow (X, c)$.

Slutsky's Lemma: Let X_n be a $\mathbb{R}^d$-valued sequence of random variables that converges weakly to X. If $\mathbb{R}^d$-valued sequence Y_n converges to c (in prob or a.s.) then the following are valid:

(i) $X_n + Y_n \rightsquigarrow X + c$.

(ii) $X_n \cdot Y_n \rightsquigarrow cX$.

(iii) if $c \neq 0$, $\frac{X_n}{Y_n} \rightsquigarrow \frac{X}{c}$.

Law of Large Numbers: describes the consistency of sample means. For $X_1, \ldots, X_n \overset{iid}{\sim} P$ s.t. $\mathbb{E}_P(|X|) < \infty$ and letting $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, we have

$$\bar{X}_n \xrightarrow{p} \mathbb{E}_P[X]$$

2.2 Central Limit Theorems

Levy's Central Limit Theorem: describes the limiting distribution of univariate sample means. For $X_1, \ldots, X_n \overset{iid}{\sim} P$ s.t. $\mathbb{E}_P(|X|) < \infty$ and $\mathbb{E}_P(X^2) < \infty$, then for $\sigma_P^2 := \text{Var}_P(X)$:

$$\sqrt{n}(\bar{X}_n - \mathbb{E}_P(X)) \rightsquigarrow N(0, \sigma_P^2)$$

Multivariate Central Limit Theorem: generalizes Levy's CLT to multivariate iid observations. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} P$ with support in $\mathbb{R}^d$ and $\mathbb{E}[||X||^2] < \infty$. Then

$$\sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow N(0, \Sigma)$$

Where $\mu := \mathbb{E}[X]$ and $\Sigma := \mathbb{E}_P[(X-\mu)(X-\mu)^T]$.

Lindeberg Feller Central Limit Theorem: generalizes Levy CLT to setting where observations are independent but not necessarily identically distributed.

Let $\{X_{ni}^n\}$ be an independent collection of $\mathbb{R}$-valued random variables for each n. Let $\mu_{ni} := \mathbb{E}[X_{ni}]$ and $\sigma_{ni}^2 := \text{Var}(X_{ni})$ exist and be finite. Define $\sigma_n^2 := \sum_{i=1}^n \sigma_{ni}^2 > 0$ and $Y_{ni} = \frac{(X_{ni}-\mu_{ni})}{\sigma_n}$. If the Lindeberg condition

$$\sum_{i=1}^n \mathbb{E}[Y_{ni}^2 1(|Y_{ni} \geq \epsilon|)] \overset{n\to\infty}{\longrightarrow} 0 \quad \text{for all } \epsilon > 0$$

or alternatively the Lyapunov condition holds

$$\sum_{i=1}^n \mathbb{E}[|Y_{ni}^{2+\delta}|] \overset{n\to\infty}{\longrightarrow} 0 \quad \text{for some } \delta > 0$$

then

$$\sum_{i=1}^n Y_{ni} \rightsquigarrow N(0,1)$$

We apply the LF-CLT to demonstrating ASN of OLS estimator in Example **??**.

2.3 Delta Methods

PDF of transformations of Random Variables (parametric):

1. Univariate: Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is monotone. Suppose $f_X(x)$ is continuous on $\mathcal{X}$ and $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then pdf of Y is:

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 \quad \text{else} \end{cases}$$

2. Multivariate: suppose (X,Y) jointly varies according to some $p_{X,Y}$ and $(U,V) := (T_1(X), T_2(Y))$ is unknown for T being 1-1 maps.

$$p_{U,V}(y) = p_{X,Y}(T^{-1}(x,y)) \left| \det \begin{pmatrix} \frac{\partial T_1^{-1}(u,w)}{\partial u} & \frac{\partial T_2^{-1}(u,w)}{\partial u} \\ \frac{\partial T_1^{-1}(u,w)}{\partial w} & \frac{\partial T_2^{-1}(u,w)}{\partial w} \end{pmatrix} \right|$$

Delta Method: functions of an estimator with known distribution:

1. Univariate ($\mathbb{R} \to \mathbb{R}$): suppose $f : \mathbb{R} \to \mathbb{R}$ is differentiable at ψ_0, and $r_n(\psi_n - \psi_0) \rightsquigarrow Z$ holds, then

$$r_n(f(\psi_n) - f(\psi_0)) \rightsquigarrow f'(\psi_0) \cdot Z$$

2. Multivariate ($\mathbb{R}^d \to \mathbb{R}$): if $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at $\psi_0 \in \mathbb{R}^d$, and $r_n(\psi_n - \psi_0) \rightsquigarrow Z$ holds, then

$$r_n(f(\psi_n) - f(\psi_0)) \rightsquigarrow \langle Z, \nabla f(\psi_0) \rangle$$

3. Multivariate ($\mathbb{R}^d \to \mathbb{R}^p$): suppose $f : \mathbb{R}^d \to \mathbb{R}^p$ is differentiable at ψ_0, meaning there exists an $\mathbb{R}^p \times \mathbb{R}^d$ *Jacobian matrix* $J_f = \begin{pmatrix} \nabla f_1 \\ \vdots \\ \nabla f_p \end{pmatrix}$. Suppose that $r_n(\psi_n - \psi_0) \rightsquigarrow Z$. Then it holds that:

$$r_n[f(\psi_n) - f(\psi_0)] \rightsquigarrow J_f \cdot Z$$

4. For influence functions: Suppose $\psi_n \in \mathbb{R}^d$ is an asymptotically linear estimator of $\psi_0 \in \mathbb{R}^d$, implying $\psi_{n,j}$ is ALE for $\psi_{0,j}$ for all $j \in \{1, \dots, d\}$. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable (at ψ_0). Then $f(\psi_n)$ is itself an asymptotically linear estimator for $f(\psi_0)$ with influence function equal to:

$$\tilde{\phi}_{P_0} : x \to \langle \nabla f(\psi_0), \phi_{P_0}(x) \rangle$$

Where $\phi_{P_0}(x)$ is the influence function of ψ_n. This implies

$$f(\psi_n) - f(\psi_0) = \frac{1}{n} \sum_{i=1}^{n} \langle \nabla f(\psi_0), \phi_{P_0}(X_i) \rangle + o_P(n^{-1/2})$$

A good Delta method example is shown in Example **??**.

2.4 Stochastic Order Notation and Prokohorv's Theorem

Stochastic Order Notation: helps us compare the magnitude of two sequences of random variables X_n and R_n:

1. Big O: $X_n = O_P(R_n)$ means X_n is stochastically bounded, or within a multiplicative constant of R_n:

$$P\left(\left|\frac{X_n}{R_n}\right| > \delta\right) < \epsilon \quad \forall\, n > N$$

2. Little o: $X_n = o_P(r_n)$ means X_n grows more slowly than r_n, and implies convergence of the ratio to 0.

$$\frac{X_n}{r_n} = o_P(1)$$

Operations on Stochastic Order Notation

1. Decomposition: $X_n = o_P(r_n)$ iff $X_n = Y_n \cdot r_n$ s.t. $Y_n = o_P(1)$
2. Decomposition: $X_n = O_P(R_n)$ iff $X_n = Y_n \cdot R_n$ s.t. $Y_n = O_P(1)$
3. Addition: $o_P(1) + o_P(1) = o_P(1)$
4. Addition: $o_P(1) + O_P(1) = O_P(1)$
5. Multiplication: $O_P(1) \cdot O_P(1) = O_P(1)$
6. Multiplication: $o_P(1) \cdot O_P(1) = o_P(1)$
7. Division: $[1 + o_P(1)]^{-1} = O_P(1)$
8. Convergence in probability implies uniformly tight: $X_n = o_P(1) \implies X_n = O_P(1)$

Prokhorov's Theorem: an analog of the Bolzano-Weierstrass theorem for sequences of random variables.

1. (Weak convergence implies uniformly tight): If $X_n \rightsquigarrow X$, then $X_n = O_P(1)$.
2. (Uniformly tight implies subsequence that converges weakly): If $X_n = O_P(1)$, there exists a subsequence $\{X_{n_j}\} \subset X_n$ s.t. $X_{n_j} \rightsquigarrow X$ for some X.

3 Decision Theory

Decision theory is a general framework that unites hypothesis testing and estimation.

A decision function is a probability of an $D(a, x) = d(a|X = x)$ is the probability of an action given data.

The loss $L(a, \theta)$ quantifies the quality of a decision at θ (could be squared error loss for estimation, 0-1 loss for hypothesis testing).

The risk is the expected loss, marginal over randomness in the data space and action space:

$$R(D, \theta) = \int_{\mathcal{X}} \int_{\mathcal{A}} L(a, \theta) D(a|x) P_\theta(x)$$

The Bayes risk is the risk marginal over the prior, Π, on the distribution of the parameter $\theta \in \Theta$.

$$r(D, \Pi) = \int R(D, \theta) d\Pi(\theta)$$

3.1 Bayes rules

A Bayes rule D_Π is optimal with regard to the Bayes risk

$$r(D_\Pi, \Pi) = \inf_{D \in \mathcal{D}} r(D, \Pi) = \inf_{D \in \mathcal{D}} \mathbb{E}_\Pi \left[\int_{\mathcal{A}} L(a, \theta) D(da|x) \middle| X = x \right]$$

To find Bayes rule under a convex loss, we know the action is deterministic. It is best to consider the action that minimizes the *Bayes risk function*:

$$f_x : a \to \mathbb{E}[\ell(a, \psi(\theta))|X = x]$$

Here are some Bayes rules for common loss functions:

1. Squared error loss: $L(a, \theta) = \{a - \psi(\theta)\}^2$, Posterior Mean: $T_\Pi = \mathbb{E}(\psi(\theta)|X = x]$
2. Absolute Deviation loss: $L(a, \theta) = |a - \psi(\theta)|$, Posterior Median: $T_\Pi = \text{median}(\psi(\theta)|X = x)$.
3. Weighted squared error loss: $L(a, \theta) = w(\theta)\{a - \psi(\theta)\}^2$, Weighted posterior mean: $T_\Pi = \frac{\mathbb{E}[w(\theta)\psi(\theta)|X=x]}{\mathbb{E}[w(\theta)|X=x]}$
4. 0-1 loss: $L(a, \theta) = \mathbb{I}(a \neq \psi(\theta))$, Maximum Posterior Probability: $T_\Pi = \underset{a}{\text{argmax}} \Pr(\psi(\theta) = a|X = x)$

3.2 Minimax rules

Minimaxity: the minimax framework seeks decision rules that minimize the maximal risk over $\theta \in \Theta$. We can find minimax rules by finding Bayes rules for priors that yield constant risk with respect to θ.

Admissibility: a good decision rule is one for which there does not exist a uniformly better one. A rule D is inadmissible if there exists another rule $\tilde{D}$ such that

$$\mathcal{R}(\tilde{D}, \theta) \leq \mathcal{R}(D, \theta) \text{ for all } \theta \in \Theta, \text{ and}$$
$$\mathcal{R}(\tilde{D}, \tilde{\theta}) < \mathcal{R}(D, \tilde{\theta}) \text{ for some } \tilde{\theta} \in \Theta$$

The rule is admissible otherwise.

For a given prior Π, a rule D_Π is unique Bayes if for all $\theta \in \Theta$, a rule is Bayes iff it equals D_Π almost everywhere.

A rule D^* is unique minimax if for all $\theta \in \Theta$, a rule is minimax iff it equals D^* almost everywhere.

Turns out unique Bayes/minimax rules are admissible!

James-Stein Estimator: when $d \geq 3$ and $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2 I_d)$, the sample mean is an inadmissible estimator under squared error loss. This is illustrated by the James-Stein estimator

$$T^{JS} : x \to \begin{cases} \left(1 - \frac{(d-2)}{n||\bar{x}_n||^2}\right) \bar{x}_n & \text{if } \bar{x}_n \neq (0, \ldots, 0) \\ 0 & \text{if } \bar{x}_n = (0, \ldots, 0) \end{cases}$$

The proof relies on Stein's Lemma, and is shown in Example 9.8.

3.3 Minimaxity

Note that the risk of a decision (estimator, hypothesis test) is the loss integrated over a measure. One quality that can judge the quality of a decision rule by is the maximal risk it attains over a statistical model. The minimax rule minimizes the maximal risk over a statistical model $\mathcal{P}$:

$$T^* = \underset{T\in\mathcal{T}}{\operatorname{argmin}} \sup_{P\in\mathcal{P}} R(T,P)$$

Minimax rate optimality: while we can't always derive the minimax estimator exactly, we can find sequences of estimators T_n that achieve the minimax optimal rate, i.e, don't dominate the minimax

$$\liminf_{n\to\infty} \frac{\inf_{T\in\mathcal{T}} \sup_{Q\in\mathcal{Q}} R(T,Q^n)}{\sup_{Q\in\mathcal{Q}} R(T_n,Q^n)} > 0$$

Our goal now is to find lower bounds on the minimax risk. We can do so with the following three strategies

1. Bayes risk under LFP: the minimax risk is bounded below by the Bayes risk, and the bound is tightest with the least favorable prior

$$\sup_{\Pi} \inf_{T\in\mathcal{T}} r(T,\Pi) \leq \inf_{T\in\mathcal{T}} \sup_{P\in\mathcal{P}} R(T,P)$$

2. Le Cam's Method: Define the following quantities

   (a) *Discrepancy*: measures difference in estimation procedures. $d(P_1,P_2) = \inf_{a\in\mathcal{A}}[L(a,P_1)+L(a,P_2)]$. Point estimation under squared error loss yields

   $$d(P_1,P_2) = \frac{1}{2}[\psi(P_1)-\psi(P_2)]^2$$

   Estimating a function with integrated squared error loss gives

   $$d(P_1,P_2) = \frac{1}{2}\int [f_{Q_1}(x)-f_{Q_2}(x)]^2 dx$$

   (b) *Testing affinity*: measures distributional overlap

   $$\begin{aligned} ||p_1 \wedge p_2||_1 &= \int \min\left(\frac{dP_1}{d\nu}(w), \frac{dP_2}{d\nu}(w)\right) d\nu(w) \\ &= \int \min(p_1,p_2) d\nu \\ &= 1 - \mathrm{TV}(P_1,P_2) \\ &= 1 - \sup_A |P_1(A) - P_2(A)| \end{aligned}$$

   (c) *KL divergence*: distance between distributions

   $$\mathrm{KL}(P_1,P_2) := \begin{cases} \int \log\left(\frac{dP_1}{dP_2}(w)\right) dP_1(w) & \text{if } P_1 << P_2 \\ +\infty & \text{else} \end{cases}$$

   Le Cam's Method yields the following lower bound on the minimax risk

   $$\begin{aligned} \inf_{T\in\mathcal{T}} \sup_{P\in\mathcal{P}} R(T,P) &\geq \frac{1}{2} d(P_1,P_2)||p_1 \wedge p_2||_1 \\ &\geq \frac{1}{4} d(P_1,P_2)\exp(-KL(P_1,P_2)) \end{aligned}$$

   See Example 9.4.

3. Fano's Method: letting $\eta := \min_{j \neq k} d(P_j, P_k)$ and $\bar{P} := \frac{1}{N}\sum_{j=1}^{N} P_j$, we obtain

$$\begin{aligned}\inf_{T \in \mathcal{T}} \sup_{P \in \mathcal{P}} R(T, P) &\geq \frac{\eta}{2}\left[1 - \frac{\log 2 + \frac{1}{N}\sum_{j=1}^{N} KL(P_j, \bar{P})}{\log(N)}\right] \\ &\geq \frac{\eta}{2}\left[1 - \frac{\log 2 + \max_{j \neq k} KL(P_j, \bar{P})}{\log(N)}\right]\end{aligned}$$

See Example 9.5.

4 Hypothesis Testing

4.1 Sufficiency, Minimal Sufficiency, Complete Sufficiency, Ancillarity

Sufficient statistic: given a distribution family $\mathcal{P}$ indexed by parameter θ, $T(X)$ is a sufficient statistic for θ if $T(X)$ is sufficient to generate new data X^* such that $X^* \overset{D}{=} X$.

Fisher Neyman Factorization Theorem: $T(X)$ is a sufficient statistic for θ in $\mathcal{P}$ iff the pdf of X factors as

$$f_\theta(X) = g_\theta(T(X)) \cdot h(X)$$

Minimal Sufficient Statistic: $T^*(X)$ is minimal sufficient statistic for $\mathcal{P}$ if for any sufficient statistic $T(X)$, there exists a function $h(\cdot)$, such that $T^*(X) = h(T(X))$,

Lehmann-Scheffé Theorem: Suppose $T(X)$ is sufficient. It is also minimal sufficient statistic if the following statement holds

$$T(X) = T(Y) \iff \frac{f_\theta(y)}{f_\theta(x)} \text{is } \theta\text{-free}$$

Ancillary Statistic: A statistic $V(X)$ is an *ancillary statistic* with respect to a distribution family $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ if the distribution of $V(X)$ is θ-free.

Location-scale invariance properties: if $\mathcal{P}$ is a location-scale family (e.g., $N(\mu, \sigma^2)$ because addition and multiplication by a constant reduce them to the same distribution $N(0, 1)$), then any location-scale invariant statistic is ancillary:

$$V(X_1, \ldots, X_n) = V(\sigma X_1 + \mu, \ldots, \sigma X_n + \mu)$$

Complete statistic: $T(X)$ is complete if

$$\forall \theta \in \Omega \quad \mathbb{E}_\theta(g(T)) = 0 \implies g(T) = 0$$

Basu's Theorem: if $T(X)$ is complete and sufficient statistic, it is independent of any ancillary statistic.

Here is a list of complete sufficient statistics in broad families

(a) When X follow a k-parameter exponential family

$$f_\theta(x) = a(\theta)e^{\theta_1 T_1(x) + \cdots + \theta_k T_k(x)} h(x), \qquad \theta \in \Omega \subset \mathbb{R}^k$$

With natural parameter space Ω containing an open rectangle, then $T(X) := (T_1(X), \ldots, T_k(X))$ is complete.

(b) When X is from a 2-parameter truncation family

$$f_{\theta_1,\theta_2}(x) = \frac{b(x)\mathbb{I}(\theta_1 \le x \le \theta_2)}{B(\theta_1,\theta_2)} \qquad \infty < x < \infty$$

Where $-\infty < \theta_1 < \theta_2 < \infty$, $b(x) > 0$, $B(\theta_1,\theta_2) = \int_{\theta_1}^{\theta_2} b(x)dx < \infty$. Then $T(X) = (X_{(1)}, X_{(n)})$ is complete sufficient.

4.2 UMVUE

UMVUE: an estimator is UMVUE if it has minimal variance over all unbiased estimators.

Rao-Blackwell theorem: provides a way to compute the UMVUE and guarantees its uniqueness. Given a sufficient statistic for θ, $T(X)$, and an unbiased estimator of $\tau(\theta)$, $\tau'(X)$, then the following admits the unique UMVUE

$$\mathbb{E}[\tau'(X)|T(X)]$$

UMVUE supermarket: the easiest way to find a UMVUE. Suppose $\hat{\tau}(X)$ is an unbiased estimator for our target of interest $\tau(\theta)$. If $\hat{\tau}(X)$ depends on X through a complete sufficient statistic $T(X)$, then $\hat{\tau}(X)$ is UMVUE for its own expectation, UMVUE for $\tau(\theta)$.
Thus, if we can find a complete sufficient statistic, and find a function of the CSS $\phi : \mathcal{T} \to \Omega$ such that $\mathbb{E}[\phi(T)] = \tau(\theta)$.

4.3 Common Tests

Suppose our objective is to test

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \notin \Theta_0$$

Test function: $\phi_n : \mathcal{X}^n \to [0,1]$ where 0 denotes deterministic fail to reject H_0 and 1 denotes deterministic reject H_0.
Power function: probability of rejecting H_0:

$$\pi_n(\theta) = \mathbb{E}_\theta[\phi_n(X_1,\ldots,X_n)]$$

Neyman-Pearson Paradigm: choose the highest-power test under alternatives that controls the T1ER under null:

(a) T1ER control at level α: $\sup_{\theta_0 \in \Theta_0} \pi_n(\theta_0) \le \alpha$

(b) Higher power under H_1: make $\pi_n(\theta)$ are large as possible under H_1.

Asymptotically level-α test:

$$\limsup_{n\to\infty} \pi_n(\theta_0) \le \alpha \text{ for all } \theta \in \Theta_0$$

4.3.1 Two-point and One-Sided Alternative Tests

Neyman-Pearson Lemma: under a two-point hypothesis of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, the likelihood ratio test is the *uniform most powerful* test, meaning it has the highest power among all α-level tests:

$$\phi_c(x) = \begin{cases} 0 & \text{if } \frac{f_1(x)}{f_0(x)} < c \\ 1 & \text{if } \frac{f_1(x)}{f_0(x)} > c \\ \gamma(x) & \text{if } \frac{f_1(x)}{f_0(x)} = c \end{cases}$$

Monotone likelihood ratio: suppose we are interested in testing one of the two sets of hypotheses (a) $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ or (b) $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Se say that $f_\theta(x)$ has strict monotone likelihood ratio in sufficient statistic $T(X)$ if for each pair $\theta_1 < \theta_2 \in \Omega$, the LR is strictly increasing as a function of $T(X)$:

$$\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = g_{\theta_1,\theta_2}(T(x)) \qquad \forall\, \theta_1 < \theta_2 \in \Omega$$

Under monotone likelihood ratio, the test based on the sufficient statistic T is the UMP test for the hypotheses above

$$\phi(t) = \begin{cases} 0 & \text{if } T(X) < c_\alpha \\ 1 & \text{if } T(X) > c_\alpha \\ \gamma_\alpha & \text{if } T(X) = c_\alpha \end{cases}$$

Where c_α and γ_α chosen to satisfy:

$$P_{\theta_0}[T > c_\alpha] + \gamma_\alpha P_{\theta_0}[T = c_\alpha] = \alpha$$

.

4.3.2 General Testing Strategies

Common tests: suppose $\theta = (\psi, \eta)$ where $\psi \in \mathbb{R}^m$ is the parameter of interest and $\eta \in \mathbb{R}^{d-m}$ is the nuisance. WLOG assume $\Theta_0 = \{\theta = (\psi, \eta) : \psi = 0\}$

(a) Wald test: reject H_0 when ψ_n is far from its value under the null:

$$\begin{aligned} &\sqrt{n}(\theta_n - \theta) \rightsquigarrow N(0, I_\theta^{-1}) \\ \implies\ & n^{1/2}[\hat{\psi} - \psi] \rightsquigarrow N(0, A_\theta^{-1}) \qquad \text{by Woodbury } A_\theta = I_{\theta,11} - I_{\theta,12} I_{\theta,22}^{-1} I_{\theta,12}^T \\ \implies\ & n^{1/2} A_\theta^{1/2}[\hat{\psi} - \psi] \rightsquigarrow N(0, I_m) \implies nA_\theta[\hat{\psi} - \psi] \rightsquigarrow \chi^2(m) \end{aligned}$$

So an asymptotic α-level test is we reject H_0 when $nA_\theta[\hat{\psi} - \psi]$ is outside the $(1-\alpha)$ quantiles of $\chi^2(m)$.

(b) Likelihood Ratio test: the likelihood ratio test rejects H_0 when $\inf_{\theta_0 \in \Theta_0} D_{KL}(P_\theta, P_{\theta_0})$ is large.

$$\inf_{\theta_0 \in \Theta_0} D_{KL}(P_\theta || P_{\theta_0}) = \inf_{\theta_0 \in \Theta_0} P_\theta[\ell_\theta - \ell_{\theta_0}]$$

We base our test statistic on the empirical risk minimizer scaled by $2n$:

$$L_n := 2n \cdot \left(P_n[\ell_{\hat{\theta}}] - \sup_{\theta_0 \in \Theta_0} P_n[\ell_{\theta_0}] \right)$$

Thus, the LRT compares the log likelihood of the unrestricted MLE to the log likelihood ratio of the null restricted MLE. Under QMD model, the log-likelihood ratio affords the expansion

$$\begin{aligned} L_n &= \underbrace{-2(\hat{\theta}_0 - \hat{\theta})^T \sum_{i=1}^n \dot{\ell}_{\hat{\theta}}(X_i)}_{=0} - (\hat{\theta}_0 - \hat{\theta})^T \sum_{i=1}^n \ddot{\ell}_{\tilde{\theta}}(X_i)(\hat{\theta}_0 - \hat{\theta}) \\ &= -\sqrt{n}(\hat{\theta}_0 - \hat{\theta})^T P_n \ddot{\ell}_{\tilde{\theta}}(X_i) \sqrt{n}(\hat{\theta}_0 - \hat{\theta}) \\ &= (\sqrt{n} I_\theta (\hat{\theta}_0 - \hat{\theta})^T) I_\theta^{-1} \sqrt{n} I_\theta (\hat{\theta}_0 - \hat{\theta}) + o_P(1) \end{aligned}$$

where $\tilde{\theta}_n$ is between $\hat{\theta}_0$ and $\hat{\theta}$.
Also under H_0, by asymptotic results of MLEs we have:

$$\begin{aligned}\sqrt{n}I_\theta(\hat{\theta}_0 - \hat{\theta}) &= \sqrt{n}I_\theta(P_n - P_0)\left(\begin{pmatrix}0\\ I_{\theta,22}^{-1}\dot{\ell}_{\theta,2}\end{pmatrix} - I_\theta^{-1}\dot{\ell}_\theta\right) + o_P(1)\\ &= \sqrt{n}(P_n - P_0)\left(\begin{pmatrix}I_{\theta,11} & I_{\theta,12}\\ I_{\theta,21} & I_{\theta,22}\end{pmatrix}\begin{pmatrix}0\\ I_{\theta,22}^{-1}\dot{\ell}_{\theta,2}\end{pmatrix} - \begin{pmatrix}\dot{\ell}_{\theta,1}\\ \dot{\ell}_{\theta,2}\end{pmatrix}\right) + o_P(1)\\ &= \sqrt{n}(P_n - P_0)\left(\begin{pmatrix}-[\dot{\ell}_{\theta,1} - I_{\theta,12}I_{\theta,22}^{-1}\dot{\ell}_{\theta,2}]\\ 0\end{pmatrix}\right) + o_P(1)\\ &\rightsquigarrow \begin{pmatrix}N(0, A_\theta)\\ 0\end{pmatrix} \qquad A_\theta = I_{\theta,11} - I_{\theta,12}I_{\theta,22}^{-1}I_{\theta,12}^T\end{aligned}$$

Thus, under H_0, we have Wilk's Theorem:

$$\begin{aligned}L_n &= (\sqrt{n}I_\theta(\hat{\theta}_0 - \hat{\theta})^T)I_\theta^{-1}\sqrt{n}I_\theta(\hat{\theta}_0 - \hat{\theta}) + o_P(1)\\ &\rightsquigarrow \begin{pmatrix}N(0, A_\theta)\\ 0\end{pmatrix}\begin{pmatrix}A_\theta^{-1} & \dots\\ \dots & \dots\end{pmatrix}\begin{pmatrix}N(0, A_\theta)\\ 0\end{pmatrix} \rightsquigarrow \chi^2(m)\end{aligned}$$

Thus, an α-level test rejects when L_n exceeds the $1-\alpha$ quantile of a $\chi^2(m)$ distribution.

(c) Score test: heuristically, scores have mean zero $P_\theta\dot{\ell}_{\theta_0}$. Idea is to reject if estimate of this expectation: $P_n\dot{\ell}_{\psi_0,\eta}$ is far from 0. Define

$$Z_n(\theta) := \sqrt{n}P_n\dot{\ell}_\theta$$

Note that based on the restricted MLE $\hat{\theta}_0 \in \Theta_0$

$$\begin{aligned}Z_n(\hat{\theta}_0) &= Z_n(\hat{\theta}_0) + \sqrt{n}P_\theta\dot{\ell}_{\hat{\theta}_0} - \sqrt{n}P_\theta\dot{\ell}_{\hat{\theta}_0}\\ &= \sqrt{n}(P_n - P_\theta)\dot{\ell}_{\hat{\theta}_0} + \sqrt{n}(P_\theta\dot{\ell}_{\hat{\theta}_0} - P_\theta\dot{\ell}_\theta)\\ &= \underbrace{\sqrt{n}(P_n - P_\theta)\dot{\ell}_\theta}_{\text{CLT}} + \underbrace{\sqrt{n}(P_\theta\dot{\ell}_{\hat{\theta}_0} - P_\theta\dot{\ell}_\theta)}_{\text{Delta method}} + \underbrace{\sqrt{n}(P_n - P_\theta)(\dot{\ell}_{\hat{\theta}_0} - \dot{\ell}_{\hat{\theta}})}_{\text{Donsker}}\end{aligned}$$

Under conditions on the score function, the third term is $o_P(1)$. By the multivariate delta method, we have

$$P_\theta\dot{\ell}(\psi, \hat{\eta}_0) - P_\theta\dot{\ell}(\psi, \eta) = -\begin{pmatrix}I_{\theta,12}\\ I_{\theta,22}\end{pmatrix}(\hat{\eta}_0 - \eta) + o_P(n^{-1/2})$$

Recalling that

$$\hat{\eta}_0 - \eta = I_{\theta,22}^{-1}(P_n - P_\theta)\dot{\ell}_{\theta,2} + o_P(n^{-1/2})$$

the above becomes

$$P_\theta\dot{\ell}(\psi, \hat{\eta}_0) - P_\theta\dot{\ell}(\psi, \eta) = -(P_n - P_\theta)\begin{pmatrix}I_{\theta,12}I_{\theta,22}^{-1}\dot{\ell}_{\theta,2}\\ \dot{\ell}_{\theta,2}\end{pmatrix} + o_P(n^{-1/2})$$

Plugging into the earlier expansion we have

$$\begin{aligned}Z_n(\hat{\theta}_0) &= \sqrt{n}(P_n - P_\theta)\begin{pmatrix}\dot{\ell}_{\theta,1}\\ \dot{\ell}_{\theta,2}\end{pmatrix} - \sqrt{n}(P_n - P_\theta)\begin{pmatrix}I_{\theta,12}I_{\theta,22}^{-1}\dot{\ell}_{\theta,2}\\ \dot{\ell}_{\theta,2}\end{pmatrix} + o_P(1)\\ &\rightsquigarrow \begin{pmatrix}N(0, A_\theta)\\ 0\end{pmatrix}\end{aligned}$$

By the continuous mapping theorem and Slutsky's Lemma, under H_0 we have:

$$Z_n(\hat{\theta}_0)I_{\hat{\theta}_0}^{-1}Z_n(\hat{\theta}_0) \rightsquigarrow \chi^2(m)$$

Thus, an asymptotic α-level test compares the test statistic above to the $1-\alpha$ quantile of a $\chi^2(m)$ distribution.

4.4 Distribution under Alternatives and Local Power Analysis

Contiguity: we can only describe the behavior of statistics $T(X_1^n)$ with respect to sequences of measures. Contiguity generalizes the concept of absolute continuity to sequences of measures. We say Q_n is contiguous wrt P_n if for all sequences $\{A_n\}_{n=1}^{\infty}$, $P_n(A_n) \to 0 \implies Q_n(A_n) \to 0$. It is denotes by $Q_n \lhd P_n$.

Local Asymptotic Normality: the power of local asymptotic normality is it shows that sampling under a local alternative is equivalent to sampling from a shifted null model.

Let h_n be the local parameter for fixed θ. Suppose the model $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean (QMD) at θ. Then for every $h_n \to h$, the likelihood ratio affords the expansion:

$$\log \prod_{i=1}^{n} \frac{p_{\theta+h_n/\sqrt{n}}}{p_\theta}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^T \dot{\ell}_\theta(X_i) - \frac{1}{2} h^T I_\theta h + o_P(1)$$

This result is important in deriving the distribution of the MLE under local alternatives.

Le Cam's Third Lemma: enables us to obtain the limiting distribution of a statistic under an alterative law Q_n based on laws P_n.

(a) *General Version*: letting $T_n : \Omega_n \to \mathbb{R}^k$ be a sequence of test statistics, assuming $Q_n \lhd P_n$ and

$$\left(T_n, \log \frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} (T, V)$$

Defining for all measurable A, $R(A) := \mathbb{E}[\mathbb{I}(T \in A)V]$. Then:

$$T_n \overset{Q_n}{\rightsquigarrow} R$$

(b) *User-friendly Version*:

$$\left(T_n, \log \frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} N_{k+1}\left(\begin{pmatrix} \mu \\ -\frac{\sigma^2}{2} \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right)$$

Then

$$T_n \overset{Q_n}{\rightsquigarrow} N_k(\mu + \tau, \Sigma)$$

Distribution of MLE under local alternatives: asymptotic analysis of the MLE, multivariate CLT, and QMD $\implies$ local shows that for $L_n = \log \frac{d\mathbb{Q}_n}{d\mathbb{P}_n}$ with $\mathbb{Q}_n \lhd \mathbb{P}_n$:

$$\begin{pmatrix} \sqrt{n}(\hat{\theta} - \theta) \\ \log L_n \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} I_\theta^{-1} \dot{\ell}_{\theta_0}(X_i) \\ h^T \dot{\ell}_{\theta_0}(X_i) \end{pmatrix} + \begin{pmatrix} 0 \\ -\frac{1}{2} h^T I_\theta h \end{pmatrix} + o_P(1) \overset{P_\theta}{\rightsquigarrow} N\left(\begin{pmatrix} 0 \\ -\frac{1}{2} h^T I_\theta h \end{pmatrix}, \begin{pmatrix} I_\theta^{-1} & \tau \\ \tau & h^T I_\theta h \end{pmatrix}\right)$$

where $\tau = I_\theta^{-1} \mathbb{E}[(\dot{\ell}\dot{\ell}^T)h] = h$. By *Le Cam's third lemma*, this implies that under the local alternative model, $\mathbb{Q}_n := P_{\theta+h/\sqrt{n}}$ for $h \in \mathbb{R}$:

$$\sqrt{n}(\hat{\theta} - \theta) \overset{\mathbb{Q}_n}{\rightsquigarrow} N(h, I_\theta^{-1}) \implies \sqrt{n}(\hat{\theta} - (\theta + h/\sqrt{n})) \overset{\mathbb{Q}_n}{\rightsquigarrow} N(0, I_\theta^{-1})$$

Demonstrating that the MLE is a regular estimator.

Distribution under local alternatives for ALEs: consider a general asymptotic linear estimator, μ_n, with mean-zero influence function, ϕ_θ:

$$\sqrt{n}[\mu_n - \mu(\theta_0)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_\theta(X_i) + o_P(1) \overset{P_n}{\rightsquigarrow} N(0, P_\theta \phi_\theta^2)$$

If μ_n is ALE and $\mathcal{M} := \{P_\theta : \theta \in \Theta\}$ is QMD, for any h, including random $h_n \to h$:

$$\begin{pmatrix} \sqrt{n}(\mu_n - \mu(\theta)) \\ \log \frac{d\mathbb{Q}_n}{d\mathbb{P}_n}(X_1^n) \end{pmatrix} \overset{\mathbb{P}_n}{\rightsquigarrow} N\left(\begin{pmatrix} 0 \\ -\frac{1}{2}h^T I_\theta h \end{pmatrix}, \begin{pmatrix} P_\theta \phi_\theta^2 & P_\theta \phi_\theta(\dot{\ell} h) \\ P_\theta \phi_\theta(\dot{\ell} h) & h^T I_\theta h \end{pmatrix} \right)$$

Implying by Le Cam's third lemma

$$\sqrt{n}(\mu_n - \mu(\theta)) \overset{\mathbb{Q}_n}{\rightsquigarrow} N(P_\theta \phi_\theta(\dot{\ell} h), P_\theta \phi_\theta^2)$$

By a Taylor expansion,

$$\begin{aligned} \mu(\theta + h/\sqrt{n}) - \mu(\theta) &= \dot{\mu}(\theta)^T h/\sqrt{n} + o_P(n^{-1/2}) \\ \implies \sqrt{n}(\mu_n - \mu(\theta + h/\sqrt{n})) &\overset{\mathbb{Q}_n}{\rightsquigarrow} N([P_\theta(\phi_\theta \dot{\ell}_\theta) - \dot{\mu}(\theta)]^T h, P_\theta \phi_\theta^2) \end{aligned}$$

Thus, μ_n is regular iff

$$P_\theta(\phi_\theta \dot{\ell}_\theta) = \langle \phi_\theta, \dot{\ell}_\theta \rangle = \dot{\mu}(\theta)$$

where $\dot{\mu}(\theta)$ is the *pathwise derivative* any ϕ_θ satisfying the above is the *gradient* of μ at θ in $\mathcal{M}$.

Power under local alternatives for ALEs: suppose μ_n is a regular asymptotic linear estimator and $\mathcal{M}$ is a collection of QMD distributions. By regularity, μ is a pathwise differentiable parameter at 0 with influence function g_0 that is also a gradient with $\sigma(0) = P_0 g_0^2$. Under these conditions, an asymptotic level-α test rejects $H_0 : \theta = 0$ satisfies for all $h \in \mathbb{R}^d$:

$$\pi_n\left(\frac{h}{\sqrt{n}}\right) \overset{n\to\infty}{\rightsquigarrow} 1 - \Phi\left(z_{1-\alpha} - h^T \frac{\dot{\mu}(0)}{\sigma(0)}\right)$$

Thus, the power of a test under local alternatives is determined by the slope: the pathwise derivative $\dot{\mu}(0)$ and the variance of the influence function $\sigma(0) := P_0 g_0^2$. See Example 9.9.

5 Empirical Process Theory

5.1 Concentration Inequalities

See Wainwright "High Dimensional Statistics; A Nonasymptotic Point of View" Ch 2 for more details.

Concentration inequalities give finite sample guarantees on tail bounds of form $P(f(X_1,\dots,X_n) \geq t)$. We can loosely group them into moment-based inequalities, MGF-based bounds, and Martingale-based bounds.

5.1.1 Moment-based bounds

Markov Inequality: if $X \geq 0, t > 0$

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Chebyshev Inequality: for k in natural numbers

$$P(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|^k)}{t^k}$$
$$\implies P(|X - \mathbb{E}(X)| \geq t) \leq \frac{\text{Var}(X)}{t^2}$$

Chernoff bound: the tail bound depends on the growth rate of the MGF. If X has MGF M_X.

$$P\{X - \mathbb{E}(X) \geq t\} \leq \inf_{\lambda > 0} \frac{M_{X-\mu}(\lambda)}{e^{\lambda t}}$$
$$\log P\{X - \mathbb{E}(X) \geq t\} \leq -\sup_{\lambda > 0}\{\lambda t - \log M_{X-\mu}(\lambda)\}$$

Sub-Gaussian bound: based on the Chernoff bound for a Gaussian random variable. A random variable X is sub-Gaussian with parameter σ if it has cumulant generating function that satisfies

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$$

And also satisfies the following tail bounds, where tail probabilities are less than a normal random variable

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

A few examples of sub-Gaussian random variables

(a) Any bounded random variable on $[a, b]$ is sub-Gaussian with parameter $\sigma = (b-a)/2$.

(b) If two zero-mean independent random variables X_1, X_2 are sub-Gaussian with parameters σ_1, σ_2 then $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

(c) Two (non-independent) RVs X_1, X_2 sub-G with parameters σ_1, σ_2, then $X_1 + X_2$ is sub-G with parameter $\sigma_1 + \sigma_2$.

Hoeffding equality:

(a) General Case for sub-G random variables: suppose $X_1, \dots, X_n$ are independent with X_i mean μ_i and sub-G parameter σ_i. Then the following bound on the sum holds

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

(b) Special Case for Bounded Random variables: If P has bounded support on $[a, b]$, then X is sub-Gaussian with parameter $\sigma^2 = (b-a)^2/4$. Therefore

$$\log P(X - \mu \geq t) \leq -\frac{2t^2}{(b-a)^2}$$

The concentration of the mean is obtained as

$$\log P(\bar{X}_m - \mu \geq t) \leq -\frac{2nt^2}{(b-a)^2}$$

Subexponential random variable: a random variable is sub-exponential with parameters (σ^2, b) if for all $|\lambda| < \frac{1}{b}$:

$$\log M_{x-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$$

By Chernoff, a sub-exponential random variable also satisfies the following tail probability

$$\log P(X \geq \mu + t) \leq \begin{cases} -\frac{t^2}{2\sigma^2}, & \text{if } 0 \leq t \leq \sigma^2/b \\ -\frac{t}{2b}, & \text{if } t > \sigma^2/b \end{cases}$$

Meaning that the concentration in Gaussian within a certain proximity to 0, but has thicker tails as t increases.
Some examples of sub-exponential random variables

(a) All sub-Gaussian RVs are sub-Exponential.

(b) If $X_1, \ldots, X_n$ are sub-Exponential random variables with parameters $(\sigma_1^2, b_1), \ldots, (\sigma_n^2, b_n)$, then their mean-centered sum is sub-Exponential with parameters $(\sum_{i=1}^n \sigma_i^2, \max_{1 \leq i \leq n} b_i)$.

(c) If $X = Z^2$ for $Z \sim N(0,1)$, then for $|\lambda| < 1/4$, the $\log M_{X-\mu}(\lambda) = \frac{4\lambda^2}{2}$ implying X is sub-Exponential with parameters $(\sigma^2, b) = (2, 4)$.

Bernstein's Inequality:

(a) General form: if a random variable with mean μ and variance σ^2 satisfies the Bernstein condition with parameter b:

$$|\mathbb{E}[(X-\mu)^k]| \leq \frac{1}{2}k!\sigma^2 b^{k-2}$$

Then the following tail bound holds:

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

(b) Bounded random variables: Suppose X is bounded so that $|X - \mu| \leq b$ and $\text{Var}(X) = \sigma^2$. Whenever, $|\lambda| < \frac{1}{2b}$, then $M_{X-\mu}(\lambda) \leq \exp(\lambda^2\sigma^2)$, meaning X is sub-E with parameters $(2\sigma^2, 2b)$. This implies that

$$P\{X - \mu \geq t\} \leq \exp\left(-\frac{t^2}{2[\sigma^2 + bt]}\right)$$

For $X_1, \ldots, X_n$ independent bounded random variables $|X_i - \mu_i| \leq b$.
For sample means: Let $X_1, \ldots, X_n$ be independent random variables such that $|X_i - \mu_i| \leq b$ and let $\bar{\sigma}_n = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$:

$$P\{\bar{X}_n - \mathbb{E}[\bar{X}_n] \geq t\} \leq \exp\left(-\frac{t^2}{2[\bar{\sigma}^2 + bt]}\right)$$

5.1.2 Martingale-based Bounds

Useful for dealing with functions that are non-sums of random variables.

Given a martingale $\{(D_k, F_k)\}_{k=1}^{\infty}$, define a *martingale difference sequence* as $D_k = Y_k - Y_{k-1}$ such that such that $\mathbb{E}[|D_k|] < \infty$ and $\mathbb{E}[D_{k+1}|F_k] = 0$. Any Martingale difference sequence has the following telescoping decomposition $Y_n - Y_0 = \sum_{k=1}^{n} D_k$.

Bounded differences property: ensures that a function does not depend too heavily on one input. A function satisfies the bounded differences property if for all i there exists a finite c_i such that the following holds for all $x_1, \dots, x_n, x_i^{'} \in \mathcal{X}$:

$$|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i$$

Bounded differences/McDiarmind's Inequality: let f be an arbitrary function that satisifies the bounded differences property with $c_1, \dots, c_n$. Then

$$P(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

See Example 9.10 for an example with U statistics.

5.1.3 Lipschitz Functions of Gaussian variables

Lipschitz transform of Gaussian vector: suppose $X_1, \dots, X_n$ is a vector of standard normal random variables and $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz wrt the Euclidean norm: $|f(x) - f(y)| \leq L||x - y||_2$. Then $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L implying

$$P(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2\exp\left(-\frac{t^2}{2L^2}\right)$$

Thus, the concentration of any Lipschitz function of a standard Gaussian random vector, regardless of the dimension, exhibits concentration like a scalar Gaussian random variable with variance L^2.

See Example 9.11 for the concentration of the Gaussian maximum.

5.2 Empirical Process Theory

The main goal of empirical process theory is to study the behavior of $\mathbb{P}_n f$, the empirical measure indexed by $\mathcal{F}$ *uniformly* over the function class $\mathcal{F}$. This differs from traditional LLNs and CLT because the sequences of random variables (functions) are not fixed but vary within a class. This means understanding under what conditions

(a) Uniform law of large numbers/consistency: $||\mathbb{P}_n - P||_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| = o_P(1)$

(b) Uniform convergence to Gaussian process: $\{\mathbb{G}_n(f) : f \in \mathcal{F}\} := \{\sqrt{n}(\mathbb{P}_n - P)f : f \in \mathcal{F}\} \rightsquigarrow \mathbb{G}$

Why are we interested in these uniform convergence results? Many statistical estimands can be written in the form of a *functional* of the distribution function $\Psi(F)$. Uniform consistency and uniform convergence conditions can guarantee consistency and asymptotic normality of *plug-in estimators* $\Psi(\hat{F}_n)$ when the functional is continuous in the supnorm and hadamard differentiable respectively.

There is also great interest in understanding the concentration of $||\mathbb{P}_n - P||_{\mathcal{F}}$ about its mean, which can help establish performance guarantees for many algorithms. A key result is that many problems in nonparametric statistics involves estimating a finite-dimensional/infinite-dimensional parameter θ^* through minimizing the empirical risk.

$$\hat{\mathcal{R}}_n(\theta, \theta^*) := \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, X_i)$$

In practice we minimize the risk over a subset of our parameter space $\Omega_0 \subset \Omega$. We assess performance via the regret, which describes the additional risk incurred by using the empirical risk minizer compared to the true minimizer

$$\text{Reg}(\hat{\theta}) = P\ell(\hat{\theta}) - \inf_{\theta \in \Omega_0} P\ell(\theta)$$

The statistical question becomes how to bound the regret on the empirical risk minimizer? Turns out when $\mathcal{F}$ is the class of loss functions:

$$\text{Reg}(\hat{\theta}) \leq ||\mathbb{P}_n - P||_{\mathcal{F}}$$

(For example, see Example **??** for an example dealing with the regret of an empirical risk minimizer.

5.2.1 Uniform Consistency for function classes on $[0,1]$

The following subsection is useful for classification problems with respect to the 0-1 loss.

Suppose $\mathcal{F}$ consists of $[0,1]$-valued functions with boolean valued functions as a special case. Then $||P_n - P||_{\mathcal{F}}$ satisfies the bounded differences property with $c_i = \frac{1}{n}$, implying:

$$P(||P_n - P||_{\mathcal{F}} - \mathbb{E}||P_n - P||_{\mathcal{F}} \geq t) \leq 2\exp(-2nt^2)$$

So with high probability (asymptotically), $||P_n - P||_{\mathcal{F}}$ is close to its mean so it suffices to study the mean.

Rademacher Complexity: Rademacher complexity characterizes the complexity of a function class by characterizing the maximum correlation achievable between a function $f \in \mathcal{F}$ and a noise vector of Rademacher RVs that takes $\{-1, +1\}$ with probability $1/2$

$$\mathbb{E}||R_n||_{\mathcal{F}} := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right]$$

Bounding $||P_n - P||_{\mathcal{F}}$ via Rademacher Complexity: Suppose $\mathcal{F}$ is a collection of $[0,1]$-valued functions. Then with probability at least $1 - 2\exp(-2nt^2)$, it holds that

$$\begin{aligned} \frac{1}{2}\mathbb{E}||R_n||_{\mathcal{F}} - \sqrt{\frac{\log 2}{2n}} - t \leq \mathbb{E}||P_n - P||_{\mathcal{F}} - t \\ \leq ||P_n - P||_{\mathcal{F}} \\ \leq \mathbb{E}||P_n - P||_{\mathcal{F}} + t \\ \leq 2\mathbb{E}||R_n||_{\mathcal{F}} + t \end{aligned}$$

Thus, bounding the Rademacher complexity allows us to bound the empirical process term. How do we go about bounding the Rademacher complexity? On approach is studying the VC dimension.

VC dimension: let $\mathcal{F} : \mathcal{X} \to \{0,1\}$ be a class of binary functions. The *projection* of $\mathcal{F}$ onto $x_1^n := (x_1, \ldots, x_n)$ is given by $\mathcal{F}_{x_1^n} := \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$. We say $\mathcal{F}$ *shatters* x_1^n if every possible 0-1 labelling of the n points can be accommodated by $\mathcal{F}$, i.e., $|\mathcal{F}_{x_1^n}| = 2^n$. The *growth function* is the maximal cardinality that the projection can take under n inputs:

$$\Pi_{\mathcal{F}}(n) := \sup_{x_1^n \in \mathcal{X}^n} |\mathcal{F}_{x_1^n}|$$

with the following properties. Letting $\mathcal{A}$ and $\mathcal{B}$ denote two families of sets, then the growth function satisfies

- $\Pi_{\mathcal{A}}(n+m) \leq \Pi_{\mathcal{A}}(n)\Pi_{\mathcal{A}}(m)$
- $\Pi_{\mathcal{A}\cup\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n) + \Pi_{\mathcal{B}}(m)$
- $\Pi_{A\cup B:A\in\mathcal{A},B\in\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n)\Pi_{\mathcal{B}}(m)$
- $\Pi_{A\cap B:A\in\mathcal{A},B\in\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n)\Pi_{\mathcal{B}}(m)$

The VC dimension is the largest natural number n such that there exists *some* collection x_1^n shattered by $\mathcal{F}$. The VC index is the smallest natural number n such that x_1^n is NOT shattered by $\mathcal{F}$ (VC index = VC dim + 1).

$$VC_{dim}(\mathcal{F}) := \sup\{n \in \mathcal{N} : \Pi_{\mathcal{F}}(n) = 2^n\}$$
$$VC_{ind}(\mathcal{F}) := \inf\{n \in \mathcal{N} : \Pi_{\mathcal{F}}(n) < 2^n\}$$

Sauer's Lemma describes that when $n > VC_{dim}(\mathcal{F})$, the growth function attains polynomial order. The *Finite class lemma* upper bounds the Rademacher complexity by the log of the growth function, so polynomial growth leads to control of the Rademacher complexity and uniform norm. Let $d \geq VC_{dim}(\mathcal{F})$

$$\Pi_{\mathcal{F}}(n) \leq \begin{cases} 2^n & n \leq d \\ \left(\frac{e}{d}\right)^d \cdot n^d & n > d \end{cases}$$

Implying by the corollary of the Finite Class Lemma

$$\begin{aligned}\mathbb{E}||P_n - P||_{\mathcal{F}} &\leq 2\sqrt{\frac{2\log 2 + 2d\log(\frac{e}{d}n)}{n}} \\ &= \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)\end{aligned}$$

Upper bound VC dim by number of operations: consider a family of binary functions $\mathcal{F}$ which can be computed using no more than t arithmetic or comparison operations $(+, -, \div, \times, >, \geq, <, \leq, =, \neq)$. Then

$$VC(\mathcal{F}) \leq 4p(t+2)$$

VC dimension of $\mathbb{R}$-valued functions: if $\mathcal{F}$ consists of $\mathbb{R}$-valued functions, it defines a class of sets by the operation of subgraphs $\mathcal{A} := \{\{(x,t) \in \mathcal{X} \times \mathbb{R} : t < f(x)\} : f \in \mathcal{F}\}$. If $\mathcal{F}$ forms a vector space of finite dimension (e.g., polynomials of degree at most n), the VC dimension of the function class is equal to the dimension of the vector space.

5.2.2 Uniform Consistency for richer function classes

How can we ensure uniform consistency over richer classes of functions? Consider $\mathbb{R}$-value functions of interest like regression problems, ML, density estimation, etc? We have two approaches in general: Dudley's entropy integral and bracketing integrals.

Bracketing numbers: $[\ell, u]$ is an ϵ-bracket is $||u-\ell||_{L^r(P)} \leq \epsilon$. The bracketing number, $N_{[]}(\epsilon, \mathcal{F}, L^r(P))$, is the minimal number of ϵ-brackets needed to cover $\mathcal{F}$.
Glivenko-Cantelli Theorem: Finite bracketing numbers imply $||P_n - P||_{\mathcal{F}} = o_P(1)$.

Covering numbers: for a set T equipped with pseudometric d, T_1 is an ϵ-cover if for each $\theta \in T$, there exists $\theta_1 \in T_1$ such that $d(\theta, \theta_1) \leq \epsilon$. The ϵ covering number, $N(\epsilon, T, d)$ is the size of the minimal ϵ-cover. The following relationship holds between covering and bracketing numbers:

$$N_{[]}(2\epsilon, \mathcal{F}, L^r(P)) \leq N(\epsilon, \mathcal{F}, ||\cdot||_\infty)$$

Canonical Rademacher Process: a stochastic process is a collection of random variables. Let $\{X_\theta : \theta \in T\}$ for an index set $T \subset \mathbb{R}^n$ and let ϵ be a vector of Rademacher RVs. Then

$$X_\theta := \sum_{i=1}^{n} \theta_i \epsilon_i \equiv \langle \theta, \epsilon \rangle$$

is the Canonical Rademacher Process, that is mean-zero, and sub-Gaussian wrt the Euclidean metric. A stochastic process is sub-Gaussian if for all $\theta, \theta^{'} \in T$ and $(X_\theta - X_{\theta'})$ is a sub-Gaussian random variable with parameter $\sigma^2 = d(\theta, \theta^{'})^2$.

Dudley's Entropy Integral: for any mean-zero sub-G process wrt pseudometric d with D the diameter of index set T:

$$\mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] \leq \mathbb{E}\left[\sup_{\theta, \theta' : d(\theta, \theta') \leq \epsilon} (X_\theta - X'_\theta)\right] + 8 \int_{\epsilon/2}^{D} \sqrt{\log(N(\tilde{\epsilon}, T, d))} d\tilde{\epsilon}$$

And if $\{X_\theta : \theta \in T\}$ is a separable process (such as Rademacher process), then

$$\mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] \leq 8 \int_{0}^{D} \sqrt{\log(N(\tilde{\epsilon}, T, d))} d\tilde{\epsilon}$$

This bound is not vacuous when the integral is finite, i.e., $\log N(\epsilon) = C\epsilon^{-r}$ for $r < 2$.

Applying to the Rademacher Complexity: we obtain if $\mathcal{F}$ is a real-valued function such that $\mathcal{F} = -\mathcal{F}$:

$$\begin{aligned} \mathbb{E}||R_n||_{\mathcal{F}} &\leq \frac{8}{\sqrt{n}} \mathbb{E}_{P_n}\left[\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}, L^2(P_n))} d\epsilon\right] \\ &\leq \frac{8}{\sqrt{n}} \sup_Q \left[\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}, L^2(Q))} d\epsilon\right] \end{aligned}$$

Thus, if the entropy integral is finite $\mathbb{E}||P_n - P||_{\mathcal{F}} = \mathcal{O}(n^{-1/2})$, we control $||P_n - P||_{\mathcal{F}}$ and establish a uniform law of large numbers. Note: we can replace ∞ by D if $\mathcal{F}$ maps to $[-D, D]$. Also, the empirical process $\sqrt{n}(P_n - P)f = \mathcal{O}(1)$ converges to a tight limit process uniformly in $\mathcal{F}$.

VC classes: for example, the covering number for VC classes satisfies the entropy integral, allowing us to obtain tighter bounds on empirical process term than those obtained via Sauer's Lemma. For $\mathcal{F}$ a VC class of functions that map to $[-1, 1]$ with $V_i(\mathcal{F})$ denoting the VC index, then

$$\sup_Q N(\epsilon, \mathcal{F}, L^2(Q)) \leq kV_i(\mathcal{F})(16e)^{V_i(\mathcal{F})} \cdot \left(\frac{1}{\epsilon}\right)^{2(V_i(\mathcal{F})-1)}$$

Since the $\sup_Q \log N(\epsilon) = C\epsilon^{-1}$, the entropy integral is finite implying

$$\mathbb{E}||P_n - P||_{\mathcal{F}} \lesssim \mathbb{E}||R_n||_{\mathcal{F}} \lesssim \mathcal{O}(n^{-1/2})$$

See Examples 9.12 and 9.13 for applications to functions that are Lipschitz in Indexing Parameters.

Bracketing Integral Bound: suppose $\mathcal{F}$ is a class of functions that maps to $[-1, 1]$. Then it holds that

$$\mathbb{E}||P_n - P||_{\mathcal{F}} \leq \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P))} d\epsilon$$

In the case of envelope function F

$$\mathbb{E}||P_n - P||_{\mathcal{F}} \leq \frac{c}{\sqrt{n}} ||F||_{L_2(P)} \int_0^1 \sqrt{\log N_{[]}(\epsilon ||F||_{L_2(P)}, \mathcal{F}, L^2(P))} d\epsilon$$

Thus, if the bracketing integral is finite $\mathbb{E}||P_n - P||_{\mathcal{F}} = \mathcal{O}(n^{-1/2})$ and then the empirical process $\sqrt{n}(P_n - P)f = \mathcal{O}(1)$ converges to a tight limit process uniformly in $\mathcal{F}$. See Example 9.14 for an example using Sobolev classes.

5.3 Uniform Convergence of Empirical Process

Suppose we are interested in inference about the *random function* $t \to F_n(t)$ and hope to get inference uniform over its domain. We may also be interested in whether *plug-in estimators* $\Psi(F_n)$, *continuous functionals* of the empirical distribution are consistent and asymptotically normal estimators for population parameters $\Psi(F_0)$.

Glivenko-Cantelli Theorem (CDF): Suppose $X_1, \dots, X_n \overset{iid}{\sim} F_0$, then $||F_n - F_0||_\infty = \sup_t |F_n(t) - F_0(t)| \overset{a.s.}{\to} 0$. Two corollaries of this theorem

(a) Concentration of supnorm

$$P\left[||F_n - F||_\infty \geq \frac{c}{\sqrt{n}} + \delta\right] \leq \exp(-n\delta^2/28)$$

(b) Plug-in Functionals: Any plug-in estimator $\Psi(\hat{F}_n)$ for a functional $\Psi(F_0)$ is almost surely consistent provides Ψ is continuous wrt the supnorm metric.

Donsker Theorem (CDF): Suppose $X_1, \dots \overset{iid}{\sim} F$, the sequence of empirical processes $\sqrt{n}(F_n - F_0) \rightsquigarrow \mathbb{G}$, a mean zero Gaussian process with covariance function $F_0(\min(t_i, t_j)) - F_0(t_i)F_0(t_j)$. See Example 9.21 for confidence bands on the CDF.

$\ell^\infty(\mathcal{F})$: in order to study convergence of a stochastic process (the empirical process) in a metric space, we need a metric space in which to describe the convergence.

$$\ell^\infty(\mathcal{F}) := \{H : \mathcal{F} \to \mathbb{R} \text{ such that } ||H(f)||_\mathcal{F} < \infty\}$$

A set of maps equipped with the sup norm $||\cdot||_\mathcal{F}$.
Glivenko-Cantelli Class: a class of functions $\mathcal{F}$ is P_0-G-C if

$$||P_n f - P_0 f||_\mathcal{F} = \sup_{f \in \mathcal{F}} |P_n f - P_0 f| \overset{a.s.}{\to} 0$$

Donsker Class: a class of functions $\mathcal{F}$ is P_0-Donsker if $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F}) \iff ||\mathbb{G}_n||_\mathcal{F} \to ||\mathbb{G}||_\mathcal{F}$ where $\mathbb{G}$ is a mean-0 Gaussian process with covariance function:

$$(f_1, f_2) \to \mathbb{P}_0(f_1 f_2) - \mathbb{P}_0(f_1)\mathbb{P}_0(f_2)$$

Donsker permanance properties: if $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker classes, then the following are also P-Donsker

(a) $\mathcal{F} + \mathcal{G} = \{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$
(b) $-\mathcal{F} = \{-f : f \in \mathcal{F}\}$
(c) $\mathcal{F} \cup \mathcal{G}$
(d) Suppose that only $\mathcal{F}$ is P-Donsker, then if $\mathcal{G} \subset \mathcal{F}$, $\mathcal{G}$ is P-Donsker.
(e) If $\mathcal{F}$ is Donsker, $\bar{\mathcal{F}}$ (i.e., the closure, the set of all elements of $\mathcal{F}$ and its $L^2(P)$ limit points) is also Donsker.

Sufficient conditions to prove a Donsker Class:

(a) Satisfy finite bracketing integral: for $\delta > 0$, $\mathcal{F}$ is P_0-Donsker if:

$$J_{[]}(\delta = 1, \mathcal{F}, L^2(P)) := \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P))} d\epsilon < \infty$$

(b) Satisfy uniform entropy integral: $\mathcal{F}$ is P_0-Donsker if it has an envelope function $\bar{F}$ satisfying $P\bar{F}^2 < \infty$ and

$$J(\delta = 1, \mathcal{F}, L^2(P)) = \int_0^\infty \sqrt{\log \sup_Q N(\epsilon ||\bar{F}||_{Q,2}, \mathcal{F}, L^2(Q))} < \infty$$

Where $||\bar{F}||_{Q,2} = Q\bar{F}^2$

6 Estimation Paradigms

6.1 M and Z estimation

6.1.1 M estimation

M-estimation: a collection of functions $\{m_\theta\}$ indexed by parameter/functional θ identifies θ_0 if the following holds uniquely;

$$\theta_0 = \underset{\theta}{\text{argmax}}\ P_0[m_\theta(X)]$$

M-estimator replaces the expectation over P_0 with the empirical expectation

$$\theta_n = \underset{\theta}{\text{argmax}}\ P_n[m_\theta(X)]$$

Consistency of M-estimators (vdv 5.21): since M-estimator θ_n is a near maximizer of the random criterion function $M_n := P_n m_\theta(X)$. In order for $\theta_n \xrightarrow{p} \theta_0$.

(a) Near solution is available:

$$M_n(\theta_n) \geq \sup_\theta M_n(\theta) - o_P(1)$$

(b) Identification: for all $\epsilon > 0$:

$$M_0(\theta_0) > M_0(\theta) \quad \forall\ \theta\ : ||\theta - \theta_0|| > \epsilon$$

(c) *Uniform consistency* of criterion/estimating function: since criterion/estimating function is random, we require that it uniformly converge to true criterion/estimating function.

$$\sup_\theta |M_n(\theta) - M_0(\theta)| \xrightarrow{p} 0$$

More generally, we require $\{m_\theta : \theta \in \Theta\}$ lies in a Glivenko-Cantelli Class: meaning $\sup_\theta |(P_n - P_0)m_\theta| = o_P(1)$. A sufficient condition for this is the function maximized is continuous in x, Θ has compact support, and is dominated by integrable function.

Asymptotic Linearity and Normality of M-estimators (vdv 5.23): suppose the M-estimators is a near-maximizer of $M_n := P_n m_\theta(X)$. Under the following conditions

(a) Suppose $\hat{\theta}_n \xrightarrow{p} \theta$ (condition (a-c) above).

(b) Interior and differentiability: suppose θ, θ_0 are on the interior (not on boundary) of the parameter space, and that m_θ is differentiable at θ_0 with derivative $\dot{m}_{\theta_0}(x)$.

(c) m_θ is sufficiently smooth: uniform convergence under local alternatives: there exists a a nonsingular symmetric matrix V_{θ_0}

$$\lim_{\epsilon \to 0} \sup_{||h||=1} \frac{|P_0 m_{\theta_0 + \epsilon h} - P_0 m_{\theta_0} - \frac{1}{2}\epsilon^2 h^T V_{\theta_0} h|}{\epsilon^2} \xrightarrow{\epsilon \to 0} 0$$

Note this assumption is equivalent to assuming that $\left\{ \frac{m_\theta - m_{\theta_0} - (\theta - \theta_0)^T \dot{m}_{\theta_0}}{||\theta - \theta_0||} : ||\theta - \theta_0 < \epsilon|| \right\}$ forms a Donsker class. We can replace this hard-to-verify assumption with two conditions.

i. Condition 1: assuming that $P_0 m_\phi$ admits a second order Taylor expansion at θ_0

$$P_0 m_\theta = P_0 m_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^T V_{\theta_0}(\theta - \theta_0) + o(||\theta - \theta_0||^2)$$

Where V_{θ_0} is the matrix of second derivatives of m_θ at θ_0.

ii. Condition 2: Lipschitz $\forall\ x \in \mathcal{X}$ and every $\theta_1, \theta_2 \in U(\phi_0)$ (neighborhood of ϕ_0):

$$||m_{\theta_1}(x) - m_{\theta_2}(x)|| \leq \dot{m}(x)||\theta_1 - \theta_2||$$

Under these conditions, recalling V_{θ_0} is the matrix of **second** derivatives of m_θ at θ_0:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{m}_{\theta_0}(X_i) + o_P(1)$$
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\left(0, V_{\theta_0}^{-1}P(\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T)V_{\theta_0}^{-1}\right)$$

Convergence rate of M-estimator (vdv 5.52) Suppose the following conditions hold for constants $C > 0$ for sufficiently small $\delta > 0$

$$\sup_{\delta/2 < d(\theta,\theta_0) < \delta} P(m_\theta - m_{\theta_0}) \leq -C\delta^\alpha$$
$$\mathbb{E}\sup_{d(\theta,\theta_0)|<\delta} |\sqrt{n}(P_n - P)(m_\theta - m_{\theta_0})| \leq C\delta^\beta$$

If $P_n m_{\hat{\theta}_n} \geq P_n m_{\theta_0} - O_p(n^{\alpha/(2\beta-2\alpha)})$, then $n^{1/(2\alpha-2\beta)}d(\hat{\theta}_n - \theta_0) = O_p(1)$, implying the optimal balance rate for weak convergence is $n^{1/(2\alpha-2\beta)}$.
Typically, $\alpha = 2$ when Pm_θ is twice differentiable at θ_0 affording a 2nd order Taylor expansion when the second derivatives exist. The second condition (the maximal inequality) is harder to verify but can be determined based on the entropy of the function classes (see Lemmas 19.34-19.38).

6.1.2 Z estimation

Z-estimation: for a collection of *estimating functions* z_θ indexed by parameter or functional θ, the *population estimating equation* identifies θ_0 if the following holds

$$\theta_0 \text{ is solution in } \theta \text{ to } P_0[z_\theta(X)] = (0, \ldots, 0)$$

The Z-estimator is the solution in ψ to the sample estimating equation which replaces the expectation over P_0 with the empirical expectation:

$$\hat{\theta}_n \text{ is solution in } \theta \text{ to } P_n[z_\theta(X)] = (0, \ldots, 0)$$

Consistency of Z-estimators in 1-D case (vdv 5.10)

(a) Let $\Theta \subset \mathbb{R}$ and $Z_n := P_n z_\theta$ be a random function and $Z_0 := P_0 z_\theta$ such that

$$Z_n(\theta) \xrightarrow{p} Z_0(\theta) \quad \forall\theta$$

(b) Let $Z_n(\theta)$ is continuous in θ with exactly one zero $\hat{\theta}_n$ and let θ_0 be a point where

$$Z_0(\theta_0 - \epsilon) < 0 < Z_0(\theta_0 + \epsilon)$$

Then $\hat{\theta}_n \xrightarrow{p} \theta_0$.
Consistency of Z-estimators (vdv 5.9)

(a) Near solution is available: θ_n is near solution to estimating equation.

$$P_n z_{\theta_n} = Z_n(\theta_n) = o_P(1)$$

(b) Identification: well separated minimizer. For all $\epsilon > 0$:

$$0 = -||Z_0(\theta_0)|| > -||Z_0(\theta)|| \quad \forall\ \theta\ : ||\theta - \theta_0|| > \epsilon$$

(c) *Uniform consistency* of estimating equation across all $\theta \in \Theta$

$$\sup_{\theta \in \Theta} ||Z_n(\phi) - Z_0(\phi)|| \xrightarrow{p} 0$$

Equivalent to requiring class of estimating functions $\{z_{\theta,j} : \theta \in \Theta, j = 1, \ldots, k\}$ lies in a Glivenko-Cantelli Class. A sufficient condition for this is the estimating function is continuous in x, Θ has compact support, and is dominated by integrable function.

Asymptotic Linearity and Normality of Z estimators: suppose $\hat{\theta}_n$ and θ_0 are the (near) solutions in θ to $P_n z_\theta = 0$ and $P_0 z_\theta = 0$.

(a) $\hat{\theta}_n \xrightarrow{p} \theta_0$ (conditions (a-c) above.

(b) Conditions on estimating function: suppose the estimating function is squared integrable, $P||z_\theta||^2 < \infty$, and Pz_θ is differentiable at θ_0 with **first** derivative matrix V_{θ_0}.

(c) Suppose the class of estimating functions $\{z_\theta : \theta \in \Theta\}$ is a Donsker class. A sufficient condition is that the estimating functions be lipschitz in their indexing parameters

$$||z_{\theta_1} - z_{\theta_2}|| \leq \dot{z}(x)||\theta_1 - \theta_2||$$

Then under these conditions

$$\sqrt{n}(\theta_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_{\theta_0}(X_i) + o_P(1) \rightsquigarrow N(0, V_{\theta_0}^{-1} P_0[z_{\theta_0} z_{\theta_0}^T](V_{\theta_0}^{-1})^T)$$

6.2 Kernel Density Estimation

Kernel density estimators are useful for functionals that depend on local features (density, regression function), like estimating the average density or the value of a density at a point.

A kernel is a function satisfying $\int K(u)du = 1$. An s-order kernel satsifies $\int u^r K(u)du = 0$ for all $r \in \{1, \ldots, s-1\}$ and $|\int u^s K(u)du| < \infty$.

A Kernel density estimator takes the form

$$\hat{f}_{n,h} : x \to \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

Common choices of K include a Uniform Kernel $K(U) = \frac{1}{2}\mathbb{I}(|u| \leq 1)$ or a Gaussian Kernel $K(U) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$.

Kernel density estimators are often judged by their *mean-integrated squared error* (MISE)

$$\text{MISE}(\hat{f}) = \int E_f[(\hat{f}(x) - f(x)^2]dx$$

Second order kernel: suppose the density f is twice continuously differentiable, and that K is second-order ($\int uK(u)du = 0$, $\int u^2 K(u)du < \infty$). There exists a constant C such that

$$\text{MISE}(\hat{f}) := \int E_f[(\hat{f}(x) - f(x)^2]dx \leq C\left(\frac{1}{nh} + h^4\right)$$

consequently for $h_n = \mathcal{O}(n^{-1/5})$, $\text{MISE}(\hat{f}) = \mathcal{O}(n^{-4/5})$ which is slower than parametric rate ($\mathcal{O}(n^{-1})$).

Higher order kernels: suppose the density f is m-times differentiable with $\int |f^{(m)}(x)|^2 dx < \infty$. Then there exists a constant C such that

$$\text{MISE}(\hat{f}) := \int E_f[(\hat{f}(x) - f(x)^2]dx \leq C\left(\frac{1}{nh} + h^{2m}\right)$$

consequently for $h_n = \mathcal{O}(n^{-1/(2m+1)})$, we have $\text{MISE}(\hat{f}) = \mathcal{O}(n^{-2m/(2m+1)})$ which approaches a parametric rate.

Undersmoothing: if we're interested in a statistical functional Ψ that depends on a local feature of a parameter (such as an average density) we require smoothing and it is reasonable to base estimation on the KDE $\Psi(\hat{P}_n)$. However, using the above results optimizes the bias-variance tradeoff for the density itself (a nuisance) and the estimate of $\Psi(\hat{P}_n)$ will inherit bias. One can show the order of the bias in h_n and n (order of variance remains unchanged), and setting the rates equal ensures optimal rate on MISE.

6.3 Asymptotic Linearity

An estimator ψ_n of ψ_0 is asymptotically linear with influence function ϕ_{P_0} if it can be written as:

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n}\phi_{P_0}(X_i) + o_P(1/\sqrt{n}) \tag{1}$$

Where ϕ_{P_0} satisfies:

(a) P_0-Mean 0: $P_0\phi_{P_0} = 0$

(b) P_0-squared integrable: $P_0\phi_{P_0}^2 < \infty$

Asymptotically linear estimators are both consistent and ASN with limiting distribution $N(0, \text{Var}[\phi_{P_0}])$. Below are some examples of asymptotically linear estimators:

(a) Sample mean: $\psi = \frac{1}{n}\sum_{i=1}^{n} X_i$ is linear estimator for $\psi_0 := \mathbb{E}_{P_0}[X]$ with $\phi_{P_0}(x) = x - \psi_0$:

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} X_i - \psi_0$$

(b) Sample variance: $\psi_n = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}[X_i - \bar{X}_n]^2$ is ALE for σ_0^2 with influence function $\phi_{P_0}(x) = [x-\mu_0]^2 - \sigma_0^2$.

$$\sigma_n^2 - \sigma_0^2 = \frac{1}{n}\sum_{i=1}^{n}\left([X_i - \mu_0]^2 - \sigma_0^2\right) + o_P(n^{-1/2})$$

(c) Sample median: ψ_n is an ALE estimator for the ψ_0 population median with influence function $\phi_{P_0}(x) = \frac{\mathbb{I}(x>\psi_0)-\frac{1}{2}}{f_0(\psi_0)}$:

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}\left(X_i > \psi_0\right) - \frac{1}{2}}{f_0(\psi_0)} + o_p(n^{-1/2})$$

(d) p-th sample quantile: let $Q_0(p)$ be the p-th quantile. Let P_0 have distribution function F_0 and density f_0. Let $Q_n(p)$ denote the p-th sample quantile. Then:

$$Q_n(p) - Q_0(p) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{F_0(Q_0(p)) - 1(X_i \le Q_0(p))}{f_0(Q_0(p))}\right] + o_P(n^{-1/2})$$

(e) Z-estimators (no nuisance): if ψ_0 is the unique solution to $P_0U(\psi) = 0$ and ψ_n is the (near) solution to the estimating equation $P_nU(\psi) = 0$, then ψ_n sastifies

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\partial}{\partial\psi}P_0U(\psi)\Big|_{\psi=\psi_0}\right)^{-1}U(\psi_0)(X_i) + o_P(n^{-1/2})$$

To confirm asymptotic linearity, we require that the influence function $\phi_{P_0}(x) = \left(-\frac{\partial}{\partial\psi}P_0U(\psi)\Big|_{\psi=\psi_0}\right)^{-1}U(\psi_0)(x)$ be finite squared integrable.

(f) Z-estimator (with nuisance): suppose that the estimating function now depends on a nuisance parameter η: $U(\psi, \eta)$. Suppose that ψ_0 is the solution in ψ to the equation $P_0 U(\psi, \eta_0) = 0$. Suppose η_n is an ALE for η_0 with IF φ_{P_0}. Define ψ_n to be a solution or near solution in ψ to:

$$\frac{1}{n}\sum_{i=1}^{n} U(\psi, \eta_n) = 0$$

Assuming ψ_n is consistent for ψ_0, ψ_n is ALE for ψ_0 with influence function:

$$\phi_{P_0}(x) := -\left(\frac{\partial}{\partial\psi} P_0 U(\psi, \eta_0)\Big|_{\psi=\psi_0}\right)^{-1}\left[U(\psi_0, \psi_n)(x) + \left(\frac{\partial}{\partial\eta} P_0 U(\psi_0, \eta)\Big|_{\eta=\eta_0} \varphi_{P_0}(x)\right)\right]$$

6.4 V/U statistics

Many parameters of interest can be written in the following form:

$$V(P) = \int\int\cdots\int H(x_1, \ldots, x_m) dP(x_1)\ldots dP(x_m$$

with H a function known as a kernel.
V-statistics: Natural estimators of these quantities are plug-in estimators $V(P_n)$ called V-statistics:

$$V_n := V(P_n) = \frac{1}{n^m}\sum_{i_1=1}^{n}\ldots\sum_{i_m=1}^{n} H(X_{i_1}, \ldots, X_{i_m})$$

Some examples include:

(a) General moment: $V(P) = \int g(x) dP(x)$ with V-statistic

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

(b) Variance: $V(P) = \int\int \frac{1}{2}(x_1 - x_2)^2 dP(x_1) dP(x_2)$ with V-statistic

$$V_n = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2$$

(c) Kendall's Tau: $V(P) = 4P(X_1 < X_2, Y_1 < Y_2) - 1$ or

$$V(P) = \int\int [2\mathbb{I}(x_1 < x_2, y_1 < y_2) + 2\mathbb{I}(x_2 < x_1, y_2 < y_1) - 1] P(dx_1, dy_1) P(dx_2, dy_2)$$

with V-statistic

$$V_n = 2 \times \left(1 - \frac{1}{n}\right) \times \text{fraction of pairs with positive slopes} - 1$$

(d) Cramer-von Mises GOF criterion: $V(P) = \int [F_P(x) - F_P^*(x)]^2 F^*(dx)$ for given F^*:

$$V(P) = \int\int \left[\int \{\mathbb{I}(x_1 \le u) - F^*(u)\}\{\mathbb{I}(x_2 \le u) - F^*(u)\} F^*(du)\right] dP(x_1) dP(x_2)$$

with V statistic

$$V_n = \int [F_n(x) - F^*(x)]^2 F^*(dx)$$

Linearization (V-statistic): the key tool we use to determine the asymptotic distribution of V-statistics. Suppose V_n, V_0 have m inputs and is symmetric in its arguments (we can symmetrize if we compute average under permutations):

$$V_n - V_0 = (P_n^m - P_0^m) = \sum_{k=1}^{m} \binom{m}{k} (P_n - P_0)^k H_k$$

Where $H_k := (P_n - P_0)^{m-k} H$ is where we've integrated all terms excluding the first k in the kernel. If we let $\tau_k^2 = \text{Var}(H_k(X_1, \ldots, X_k))$ denote the variance of the k-th variate, and let a be the minimum index such that $\tau_a^2 > 0$, then the dominant term in the above expansion is

$$\binom{m}{a} (P_n - P_0)^a H_a$$

When $a = 1$, we can write the expansion above as

$$V_n - V_0 = (P_n^m - P_0^m) = m(P_n - P)H_1 + \sum_{k=2}^{m} \binom{m}{k} (P_n - P_0)^k H_k$$

And the asymptotic behavior is determined by the first order term. The V-statistic is non-degenerate, and provided the kernel $H \in \mathcal{H}$ a Donsker class, then V_n is asymptotically linear with influence function $m(H_1(X_i) - V_0)$:

$$V_n - V_0 = \frac{1}{n} \sum_{i=1}^{n} m(H_1(X_i) - V_0) + o_P(n^{-1/2})$$

U-statistics: V-statistics have finite sample bias because they use matching pairs of indices. A U-statistic averages out H over unique indices

$$U_n := \binom{n}{m}^{-1} \sum_{\underline{i}_m \in \mathcal{D}_{m,n}} H(X_{i_1}, \ldots, X_{i_m})$$

where $\mathcal{D}_{m,n} := \{\underline{i}_m \subset \{1, \ldots, n\} := (i_1, \ldots, i_2, \ldots, i_m) : 1 \leq i_1 < \ldots < i_m \leq n\}$ denotes the set of unique indices.

Linearization (U-statistic): consider the case where number of arguments $m = 2$. Define the following quantities:

$$V_n := \frac{1}{n^2} \sum_{i,j} H(X_i, X_j)$$

$$U_n := \frac{1}{n(n-1)} \sum_{i \neq j} H(X_i, X_j)$$

$$D_n := \frac{1}{n} \sum_{i=1}^{n} H(X_i, X_i)$$

Then

$$V_n = \left(1 - \frac{1}{n}\right) U_n + \frac{1}{n} D_n$$

$$\implies U_n - V_n = \frac{1}{n}(U_n - D_n)$$

$$\implies n^{1/2}(U_n - V_n) = n^{-1/2}(U_n - D_n) = O_P(n^{-1/2}) \quad \text{(WLLN)}$$

Hence, $U_n = V_n + O_P(n^{-1})$, implying:

$$\begin{aligned} U_n - V_0 &= (V_n - V_0) + (U_n - V_n) \\ &= m(P_n - P_0)H_1 + o_P(n^{-1/2}) \end{aligned}$$

Thus, U_n is ALE for V_0 with IF: $\phi : x \to m[H_1(x) - V_0]$.
The same result holds for general m. V and U statistics are hence asymptotically equivalent.

6.5 Functional Delta Method

The delta method allows us to study distribution of a fixed, differentiable *function* f of an estimator θ_n of θ_0. What if we want to study a fixed *functional* Ψ of an estimator F_n of the true distribution function F_0?

The functional delta method is a general method for determining the asymptotic distribution of plug-in estimators of the form $\Psi(P_n)$ where Ψ is appropriately differentiable (hadamard differentiable). If the functional is hadamard differentiable and the derivative $\dot{\psi}_{P_0}$ is finite-squared integrable, then the functional delta method ensures the plug-in estimator is asymptotic linear.

Heuristically, if the functional is "differentiable" in an appropriate sense, and we consider a perturbation from P in the direction of $H = \sqrt{n}(P_n - P)$ of length $t = \frac{1}{\sqrt{n}}$, we obtain the von Mises Expansion.

$$\psi(P + tH) - \psi(P) = t\delta'_P(H) + \ldots + \frac{1}{m!}t^m\psi_p^{(m)}(H) + o(t^m)$$
$$\psi(P_n) - \psi(P) = \frac{1}{\sqrt{n}}\psi'_P(\mathbb{G}_n) + \ldots + \frac{1}{m!}\frac{1}{n^{m/2}}\psi_P^{(m)}(\mathbb{G}_n)$$

Where the asymptotic distribution is determined by the first order term. If the function ψ'_P is linear, then

$$\psi(P_n) - \psi(P) \approx \frac{1}{n}\sum_{i=1}^{n}\psi'_P(\delta_{X_i} - P)$$

where δ_{X_i} are the dirac delta measures on the observations. This function $\psi_P(\delta_x - P)$ is known as the influence function of the function ψ.

In order for a delta-method to be appropriate for functionals, we require appropriately defined notions of continuity and differentiability for functionals.

(a) ρ-continuity: a functional is ρ-continuous at $\tilde{F} \in \mathcal{P}$ if for all deterministic sequences $\{\tilde{F}_k\}_{k=1}^{\infty} \subset \mathcal{P}$ s.t.

$$\rho(\tilde{F}_k - \tilde{F}) \longrightarrow 0 \implies \psi(\tilde{F}_k) \xrightarrow{p} \psi(F_0)$$

(b) Hadamard differentiability: we first define the Gâteaux derivative of Ψ at $F \in \mathcal{P}$ in the direction of $h \in \mathcal{Q}(F) := \{c(F_1 - F) : c \in \mathbb{R}, F_1 \in \mathcal{P}\}$ is given by:

$$\dot{\Psi}(F; h) = \lim_{\epsilon \to 0}\left[\frac{\Psi(F + \epsilon h) - \Psi(F)}{\epsilon}\right] = \frac{d}{d\epsilon}\Psi(F + \epsilon h)\Big|_{\epsilon=0}$$

However, Gâteaux differentiability only implies that a Taylor expansion holds in a fixed direction h. We want the expansion to hold uniformly for all directions. Let $\epsilon_n := n^{-1/2}$ and $h_n := \sqrt{n}(F_n - F_0)$. A functional Ψ is *Hadamard differentiable* with respect to the supnorm $||\cdot||_\infty$ if there exists a *continuous linear map* $\dot{\psi}$ between normed spaces such that

$$\sup_{h \in H}\left|\left(\frac{\Psi(F_0 + \epsilon_n h_n) - \Psi(F_0)}{\epsilon_n} - \dot{\psi}(F_0; h_n)\right)\right| \xrightarrow{\epsilon_n \to 0} 0$$

Hadamard differentiability implies the validity of the following expansion, which comprises the functional delta method when $\epsilon_n = 1/\sqrt{n}$ and $h_n = \sqrt{n}(F_n - F_0)$.

$$\begin{aligned}\Psi(F_n) - \Psi(F_0) &= \epsilon_n\,\dot{\psi}(F_0; h_n) + \epsilon_n\left(\frac{\Psi(F_0 + \epsilon_n h_n) - \Psi(F_0)}{\epsilon_n} - \dot{\psi}(F_0; h_n)\right) \\ &= \dot{\psi}(F_0; F_n - F_0) + o_P(n^{-1/2})\end{aligned}$$

Functional Delta Method: if Ψ is a Hadamard differentiable functional relative to the supnorm metric, letting $\epsilon_n = n^{-1/2}$ and $h_n := \sqrt{n}(F_n - F_0)$, it holds that $R_{F_0,\epsilon_n}(h_n) = o_P(1)$ yielding

$$\begin{aligned}\Psi(F_n) - \Psi(F_0) &= \dot{\psi}(F_0; F_n - F_0) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \dot{\psi}(F_0; \delta(X_i) - F_0) + o_P(n^{-1/2})\end{aligned}$$

By linearity of $\dot{\psi}$, the term in the sum is mean 0. If the term is also finite squared integrable, then $\Psi(F_n)$ is an ALE for $\Psi(F_0)$ with influence function $\phi_{P_0}(x) = \dot{\psi}(F_0; \delta(x) - F_0)$.

Functional Chain rule: suppose there are two functionals Ψ and Φ which are Hadamard differentiable at F_0 and $\Psi(F_0)$ respectively. Then the composed map $\Phi \circ \Psi$ is also hadamard differentiable at F_0 with derivative (influence function) given by

$$\phi'_{\Psi(F_0)} \circ \psi'(F_0) = \phi'(\Psi(F_0); \psi'(F_0; h_n))$$

7 Efficiency Theory

7.1 Parametric Efficiency

Score function: the score function is the gradient of the log-likelihood, $g_n(\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta}\ell(\theta|X_i)$.

Maximum Likelihood Estimator: is both an M and Z estimator. Maximizes the (log)-likelihood, minimizer of the KL divergence, also can be the solution in θ to the score equation $g_n(\theta) = 0$.

Quadratic Mean Differentiability: ensures the likelihood is sufficiently smooth, affording us a Taylor expansion of the likelihood ratio process. A root density $\sqrt{p_\theta}$ is called QMD (or differentiable in quadratic mean) at θ if there exists a function $\dot{\ell}_\theta$ s.t.:

$$\sup_{h\in\mathbb{R}^d:||h||=1}\int\left[\frac{\sqrt{p_{\theta+\epsilon h}(x)}-\sqrt{p_\theta(x)}}{\epsilon}-\frac{1}{2}h^T\dot{\ell}_\theta(x)\sqrt{p_\theta(x)}\right]^2 d\mu(x)\overset{\epsilon\to 0}{\longrightarrow}0$$

When a model is QMD, it has mean zero score and FIM exists.

Properties of MLE: Fisher-Cramer Theorem states under a QMD model

(a) Consistent: $\hat{\theta}\xrightarrow{p}\theta_0$

(b) Asymptotically normal and efficient:

$$\sqrt{n}(\hat{\theta}-\theta)\rightsquigarrow N(0, I_X^{-1}(\theta_0))$$

(c) Invariance Property: if $\hat{\theta}$ is the MLE for θ_0, then $f(\hat{\theta})$ is the MLE for $f(\theta_0)$ and has limiting distribution

$$\sqrt{n}(f(\hat{\theta})-f(\theta_0))\rightsquigarrow N(0,\nabla f(\theta_0)^T I^{-1}(\theta_0)\nabla f(\theta_0))$$

Fisher information: describes the curvature of the log-likelihood surface

$$I_n(\theta) = -\mathbb{E}[\nabla_\theta\nabla_\theta\ell_n(\theta|X_1)] = nI_1(\theta) = n\cdot -\mathbb{E}[\nabla_\theta\nabla_\theta p(X_1;\theta)]$$

Regular Estimator: an estimator T_n is regular for a parameter $\psi(\theta)$ if for every h

$$\sqrt{n}(T_n-\psi(\theta+h/\sqrt{n}))\overset{\theta+h/\sqrt{n}}{\rightsquigarrow}L_\theta$$

where L_θ is a probability measure that does not depend on h.

Why are we interested in regular estimators?

(a) *Hodge's Estimator* mimics the sample mean when the value of the $\bar{X}_n > n^{-1/4}$ elects 0 when $|\bar{X}_n| < n^{-1/4}$. It's "improvement" over the sample mean is deceptive, and illustrates the importance of evaluating estimator performance in neighborhoods that shrink with n.

(b) When we restrict attention to regular estimators $\psi(\theta)$, they have best possible limiting distribution $N(0,\dot{\psi}_\theta I_\theta^{-1}\dot{\psi}_\theta)$. The AE convolution theorem says that this limiting distribution can only be improved over a Lebesgue null set of parameters.

(c) Regular estimators are locally asymptotic minimax, meaning the risk over $N(0,\dot{\psi}_\theta I_\theta^{-1}\dot{\psi}_\theta)$ lower bounds the asymptotic minimax risk of any estimator in a neighborhood about θ_0.

Cramer-Rao Bound: provides the lower bound on the variance of a regular estimator.

$$\text{Var}(T(X)) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)]\right)^2}{I_X(\theta)}$$

Nuisances: Suppose $\theta = (\theta_1, \ldots, \theta_k)$ and our sole interest is estimating $\tau(\theta_1)$, while $(\theta_2, \ldots, \theta_k)$ are nuisances.

$$\begin{aligned}
\nabla_\theta(\tau) &= \left(\frac{d\tau}{d\theta_1}, 0, \ldots, 0\right) \\
\text{Var}_\theta[T(X)] &\geq \left(\frac{d\tau}{d\theta_1}, 0, \ldots, 0\right) \begin{pmatrix} \mathrm{I}_{X,11} & \mathbf{I}_{X,12} \\ \mathbf{I}_{X,12} & \mathbf{I}_{X,22} \end{pmatrix}^{-1} \left(\frac{d\tau}{d\theta_1}, 0, \ldots, 0\right)^T \\
&= \frac{\left(\frac{d\tau}{d\theta_1}\right)^2}{I_{11\cdot 2}(\theta)}
\end{aligned}$$

Where $I_{11\cdot 2}(\theta) = I_{11}(\theta) - \mathbf{I}_{X,12}(\theta)[\mathbf{I}_{X,22}(\theta)]^{-1}\mathbf{I}_{X,21}(\theta)$.

7.2 General Efficiency Theory

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} P_0 \in M$ with functional $\psi : M \to \mathbb{R}$. If our goal is to estimate $\psi_0 := \psi(P_0)$ from the observed data, efficiency theory asks what is the best variance we can achieve in a model M indexed by a possible infinite-dimensional parameter.

The **guiding idea** of efficiency theory is that estimation of ψ_0 in the M should be *at least as hard as in any parametric submodel* $M_1 \subset M$ containing P_0. Let $\mathcal{H}(P_0)$ index all smooth (QMD) one-dimensional parametric submodels of M centered at P_0. In other words, for each $h \in \mathcal{H}(P_0)$ there exists a $\delta > 0$ s.t.

(a) P_0 origin: $P_{\theta,h} = P_0$ when $\theta = 0$

(b) Submodel is in broader model M: $P_{\theta,h} \in M$ for all $\theta \in [0, \delta)$

(c) QMD at origin: $M_h = \{P_{\theta,h} : \theta \in [0, \delta)\}$ is QMD at $\theta = 0$.

Our objective is to find the lower bound on the variance of a regular estimator's asymptotic distribution in our model M, where a regular estimator is regular wrt all parametric submodels.

Generalized Cramer Rao Lower Bound: The variance of any regular estimator in the infinite dimensional model, $v_0^*(M)$, can be lower bounded by the variance in any parametric submodel. To achieve the tightest lower bound, we appeal to the *least favorable parametric submodel*:

$$\begin{aligned}
v_0^*(M) &\geq \sup_{h \in \mathcal{H}(P_0)} v_0(M_h) \\
&= \sup_{h \in \mathcal{H}(P_0)} \frac{\left(\frac{\partial}{\partial\theta}\psi(P_{\theta,h})\Big|_{\theta=0}\right)^2}{\mathcal{I}_{M_h}(0)}
\end{aligned}$$

Where

$$\mathcal{I}_{M_h}(0) := P_{\theta,h}\left(\frac{\partial}{\partial\theta}\log p_{\theta,h}\right)^2\Big|_{\theta=0} \equiv P_0 g_h^2$$

Where g_h is the score function. Thus, the Fisher information in the least favorable submodel depends on h completely through the score.

Pathwise differentiability & Gradient: when the functional ψ is pathwise differentiable, there exists a P_0-mean-zero square integrable function $D(P_0) : \mathcal{X} \to \mathbb{R}$, the gradient, such that for all $h \in \mathcal{H}(P_0)$ (all QMD submodels)

$$\frac{\partial}{\partial\theta}\psi(P_{\theta,h})\Big|_{\theta=0} = P_0[D(P_0)g_h]$$

Riez Representation Theorem: guarantees the existence of a gradient for pathwise differentiable parameters.

Tangent set/Space: the tangent set of a statistical model M at P_0, denoted $G(P_0)$ is the collection of scores of all QMD submodels of M centered at $\theta = 0$. The tangent space denoted $T_M(P_0)$ is the $L^2(P)$-closure of the linear span of the tangent set.

(a) In parametric models: the tangent space is just the linear span of the score vector for the parameter $\beta \in \mathbb{R}^q$.

$$T_M = \{u^T s_0(x) : u \in \mathbb{R}^q\}$$

Where $s_0(x) = \frac{\partial}{\partial\theta}\log p_\theta(x)\big|_{\theta=0}$ for the parametric model p.

(b) In semiparametric models: the more restrictive the semiparametric model, the smaller the tangent space. Will be proper subspace of $L_0^2(P)$.

(c) In nonparametric models: the tangent space is $L_0^2(P)$

Canonical Gradient: let $D^*(P_0) := \Pi_{T_M}(P_0)\,(D(P_0))$ denote the projection of the gradient onto the tangent space of the model T_M. This implies that $D*(P_0) \in T_M$ and $D(P_0) - D^*(P_0) \in T_M^\perp$ by the orthogonality of Hilbert space projections.

Then, the Generalized Cramer-Rao lower bound writes as

$$\begin{aligned} v_0^*(M) &\geq \sup_{g\in T_M(P_0)} \frac{[\langle D(P_0), g\rangle]^2}{P_0 g^2} \\ &= \sup_{g\in T_M(P_0)} \frac{[\langle \cancelto{0}{D(P_0) - D^*(P_0), g\rangle} + \langle D^*(P_0), g\rangle]^2}{P_0 g^2} \qquad (D(P_0) - D^*(P_0) \perp T_M(P_0)) \\ &= P_0(D^*(P_0)^2) \end{aligned}$$

We have that $D^*(P_0)$ is the unique gradient in $T_M(P_0)$, the *canonical gradient*.

Gradients in nested models: let $M_1 \subseteq M_2$ be two models. Suppose $P \in M_1$ and $\psi : M_2 \to \mathbb{R}$ is pathwise differentiable at P relative to M_2. Then ψ is pathwise differentiable at P relative to M_1a nd

$$\text{Grad}_{M_2}(P) \subseteq \text{Grad}_{M_1}(P)$$

This means we can pick bigger models, find gradients in bigger models, and apply them to smaller models.

Efficient Influence Functions: there exists a connection between influence functions of RAL estimators and gradients of pathwise differentiable parameters.

(a) If ψ_n is RALE for $\psi(P_0)$ with IF $\phi_{P_0} \implies \psi$ is pathwise differentiable at P_0 with gradient ϕ_{P_0}

(b) If ψ is pathwise differentiable with gradient $D(P_0) \implies$ there exists an ALE with influence function $D(P_0)$

Therefore, if $D^*(P_0)$ is the canonical gradient of ψ at P_0 in M, this ensures the existence of an ALE with influence function equal $D^*(P_0)$ to the efficient influence function. Thus, the variance lower bound given in the GCRLB is achievable using an ALE, and this estimator is termed efficient.

A regular asymptotic linear estimator is efficient iff

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^n D^*(P_0)(X_i) + o_P(n^{-1/2})$$

7.3 Constructing Efficient Estimators

When parameters depend on "local" information, such as densities or regression functions, plug-in estimators may not be well-defined. We must rely on smoothing, which can introduce first order bias to simplistic plug-in estimators. We require more sophisticated approaches to obatin asymptotic linear estimators.

Undersmoothing: if our parameter of interest depends on a density, there does not exist a plug-in estimator for the density because the density is "too local" of a parameter. Instead, we can use a plug-in estimator where a KDE is used to estimate the density. By examining the bias of our estimator, supplying a bandwidth $h = \mathcal{O}(n^{-1/5})$, while optimal for estimating the density, leads to first order bias in estimation of the average density for example. Thus, we inspect the bias tune the bandwidth of the KDE to ensure bias decays at rate $\mathcal{O}(n^{-1/2})$. Not a general approach.

Estimating Equations Approach: we saw if ψ_n is consistent for ψ_0 where ψ_n is a near solution in ψ to

$$\mathbb{P}_n U(\psi, \eta_n) = 0$$

Then ψ_n is asymptotically linear with the form

$$\psi_n - \psi_0 = -a_0^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} U(\psi_0, \eta_0)(X_i) + b_0 \underbrace{(\eta_n - \eta_0)}_{\text{or IF } \eta_n} \right] + o_P(n^{-1/2})$$

Where $a_0^{-1} := \left(\frac{\partial}{\partial \psi} P_0 U(\psi, \eta_n) \Big|_{\psi=\psi_0} \right)^{-1}$ and $b_0 := \frac{\partial}{\partial \eta} P_0 U(\psi_0, \eta) \Big|_{\eta=\eta_0}$. Motivating the possibility of using the efficient influence function of ψ_n as an estimating function to obtain an efficient estimator when $a_0 = -1$, $b_0 = 0$!

If we assume that (a) ψ and η are variationally independent and (b) the estimating function is continuous in $L_2(P_0)$, and (c) η_n tends to η_0 at $o_P(n^{-1/4})$ rates then **Neyman-Orthogonality** holds ($b_0 = 0$). Under other conditions, $a_0 = -1$. And using the EIF as the estimating function U will yield an efficient solution.

7.4 One-step estimation

Based on a first-order (linear) expansion of the functional with $R(P, P_0) := \psi(P) - \psi(P_0) + P_0 D(P)$ often second order in the nuisances:

$$\begin{aligned} \psi(P) - \psi(P_0) &= -P_0 D(P_n) + R(P_n, P_0) \\ &= (P_n - P_0) D(P_n) - P_n D(P_n) + R(P_n, P_0) \\ &= (P_n - P_0) D(P_0) - P_n D(P_n) + (P_n - P_0)\left[D(P_n) - D(P_0) \right] + R(P_n, P_0) \end{aligned}$$

Where D is the EIF and

(a) Term 1: is a linear term

(b) Term 2: is the source of excess bias of $\psi(P_n)$

(c) Term 3: is an empirical process term that is negligible under certain conditions

i. $P_0[D(P_n) - D(P_0)]^2 = o_P(1)$

ii. There exists a P_0-Donsker class s.t. $D(P_n) \in \mathcal{F}$ w.p. tending to 1. Note: this condition can be removed if we use cross-fitting.

(d) Term 4: is a second order remainder term.

Motivating the definition of a one-step estimator, which removes the source of excess bias:

$$\psi_{os,n} := \psi(P_n) + P_n D(P_n)$$

Under the conditions above, Terms 3 and 4 are $o_P(n^{-1/2})$, implying

$$\psi_{os,n} - \psi_0 = P_n D(P_0) + o_P(n^{-1/2})$$

Implying that the one-step estimator is asymptotically linear with influence function equal to the EIF, meaning it is efficient.

8 Strategies

8.1 Identify a Bayes Rule

Strategy 1 (Squared Error Loss): if we are using squared error loss, the Bayes rule for estimating $\Psi(\theta)$ is the posterior mean:

$$D_\Pi := \mathbb{E}(\Psi(\theta)|X = x)$$

Strategy 2 (Abs deviation loss): if using absolute deviation loss, the Bayes rule for estimating $\Psi(\theta)$ is the posterior median:

$$D_\Pi := \text{median}(\Psi(\theta)|X = x)$$

Strategy 3 (Minimize Bayes Risk Function): if the loss is convex, we can minimize the bayes risk function

$$f : x \to \mathbb{E}[\{L(a, \psi(\theta))|X = x\}]$$

8.2 Prove Admissible Rule

Strategy 1 (Contradiction): assume another rule uniformly dominates it and achieve contradiction. A good strategy is to show a rule is *unique bayes* or unique minimax, which guarantees admissibility.

Strategy 2 (Connect to squared error loss): if you have an admissible estimator under a certain loss (e.g., squared error loss) and you want to assess admissibility under a related (e.g., weighted) loss, assume not admissible and manipulate the inequalities to be in terms of the loss for which you have admissibility.

8.3 Prove Minimax Rule

Strategy 1 (Submodel approach): Well suited for demonstrating minimaxity over semi/nonparametric models. if D_1 is minimax over $\mathcal{P}_1$, $\mathcal{P}_1 \subset \mathcal{P}_2$, and

$$\sup_{P \in \mathcal{P}_1} \mathcal{R}(D_1, P) = \sup_{P \in \mathcal{P}_1} \mathcal{R}(D_1, P)$$

Then D_1 is minimax over $\mathcal{P}_2$. For instance, the sample mean is minimax under squared error loss in models with bounded variance because its risk is σ^2/n, which is independent of the model family.

Strategy 2 (Connect to Bayes Rule): demonstrate a prior Π or prior sequence Π_k such that

$$r(D_\Pi, \Pi) = \sup_\theta \mathcal{R}(D_\Pi, \theta)$$
$$\liminf_{k \to \infty} r(D_{\Pi_k}, \Pi_k) = \sup_\theta \mathcal{R}(D_\Pi, \theta)$$

Or alternatively, if the Bayes rule has a risk function that does not depend on θ, naturally the first condition is satisfied.

8.4 Prove consistency of an estimator

Strategy 0 (WLLN): sample means are guaranteed consistent by the WLLN.

Strategy 0.5 (CMT): is the estimator a continuous function of a consistent estimator?

Strategy 1 (Concentration results)

(a) Hoeffding for bounded random variables.

(b) Chebyshev inequality for variables with finite expectations and shrinking variance.

(c) Chernoff bound if we have access to the MGF/Cumulant Generating function. Can leverage chernoff is the variable is sub-Gaussian/sub-Exponential.

(d) For non-sample means, we rely on the bounded differences/McDiarmind Inequality.

[Strategy 2 (Plug-in estimator)]

(a) A plug-in estimator of the form $\Psi(F_n)$ is consistent almost surely for $\Psi(F)$ when Ψ is a fixed functional that is continuous wrt the supnorm metric.

Strategy 3 (Random Function / ERM – Prove $\mathcal{F}$ is Glivenko-Cantelli Class)

(a) If the target of inference is a *random function*, uniform consistency $||P_n - P||_{\mathcal{F}} = o_P(1)$ over a function class $\mathcal{F}$ holds for Glivenko-Cantelli classes.

(b) If the loss function identifies the true (possibly infinite dimensional) parameter θ_0 and $\hat{\theta}$ is an ERM, controlling the regret ensures that $\hat{\theta} \xrightarrow{p} \theta_0$.

(c) Here is an incomplete list of GC classes.

i. Any VC class.

ii. VC permanence properties: unions, intersections, positive/negative restrictions, add, multiply, compositions with fixed functions.

iii. If we are dealing with a real-valued function class that forms a vector space (e.g., polynomials of degree at most n), the VC dimension is the dimension of the vector space.

(d) These results can also be useful for characterizing convergence rates.

8.5 Find asymptotic distribution of an random variable/estimator

Strategy 0: for random variables

(a) Work from first principles by writing and manipulating CDF

(b) Is it a sum of random variables with known CDFs?

(c) Are they transformations of random variables with known distributions?

Strategy 1: Central Limit Theorems for sample means.

(a) Levy CLT for univariate sample means.

(b) Multivariate CLT for multivariate iid data.

(c) LF CLT for independent but not IID data.

Strategy 1.5: Convergence Theory Results

(a) Slutsky's Theorem guarantees convergence of sums, differences, products, quotients of random variables where one converges weakly and another converges in probability.

(b) Portmanteau Lemma for other versions of weak convergence. Can we work out the CDF or show convergence of CDFs?

Strategy 2: Delta methods if a differentiable function of statistics with known distributions

(a) Univariate delta method for real-valued functions

(b) Multivariate delta method for $(\mathbb{R}^d \to \mathbb{R})$ and $(\mathbb{R}^d \to \mathbb{R}^p)$ functions.

(c) Delta method for ALEs: if your estimator can be written as a differentiable function of other ALEs.

Strategy 2.5 (M/Z Estimators): M and Z estimators are consistent and asymptotically normal under conditions (Glivenko-Cantelli and Donsker respectively) on the loss functions (for M-estimators) or estimating functions (Z-estimators).

Strategy 3: Donsker if an target of inference is a function-valued parameter.

(a) Common Examples
   i. VC classes of functions: half line indicators, functions with bounded support, boolean-valued functions with finite number of arithmetic operations, polynomials of degree less than some number, and union/intersection/positive or negative restriction of other VC classes.
   ii. Lipschitz Conditions: lipschitz functions or functions lipschitz in indexing parameters,
   iii. Monotonicity and Bounded Variation (which are differences of bounded monotone functions).

(b) Permanence Properties
   i. Negations, subsets, and closures of Donsker classes are Donsker.
   ii. Sums, products, and unions of two Donsker classes are Donsker.

(c) Bracketing/entropy integral. In order to have a finite bracketing/entropy integral, one of the following conditions must hold
   i. Bracketing number:
$$\log N_{[]}(\epsilon, \mathcal{F}, L^2(P) = \mathcal{O}\left(\frac{1}{\epsilon^d}\right) \qquad d < 2$$
   ii. Covering number with envelope function $\bar{F}$ satisfying $P\bar{F}^2 < \infty$ and
$$\log \sup_Q N(\epsilon ||\bar{F}||_{Q,2}, \mathcal{F}, L^2(Q)) = \mathcal{O}\left(\frac{1}{\epsilon^d}\right) \qquad d < 2$$
   Where $||\bar{F}||_{Q,2} = Q\bar{F}^2$

8.6 Establishing asymptotic linearity (ALE)

Strategy 1 (Expansion): Notice that $\psi_n = P_n f_n$ and $\psi_0 = P_0 f_0$, then
$$\psi_n - \psi_0 = P_n f_n - P_0 f_0 = (P_n - P_0)f_0 + P_0(f_n - f_0) + (P_n - P_0)(f_n - f_0)$$

Red term: linear, the piece we want to isolate.

Blue term: To establish $(P_n - P_0)(h_n - h_0) = o_P(1/\sqrt{n})$ we require vdV 19.24:

(a) $\{h_k\}_{k=1}^{\infty}$ is a sequence of random functions in $L^2(P)$ s.t., $P(h_n \in \mathcal{F}) \to 1$ for Donsker class $\mathcal{F} \subset L^2(P)$

(b) $P(h_n - h_0)^2 = o_P(1)$ for some $h_0 \in \mathcal{F}$ (shrinking variance of random function)

Strategy 2 (ALE Delta Method): if $\psi_n \in \mathbb{R}^d$ is a multivariate asymptotic linear estimator, and we want an asymptotic linear estimator for a real-valued, differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, $f(\psi_n)$ is ALE for $f(\psi_0)$ with influence w/ IF
$$\tilde{\phi}_{P_0} : x \to \langle \nabla f(\psi_0), \phi_{P_0}(x) \rangle$$

Strategy 3 (Functional Delta Method): best used via chain rule. Suppose we have a fixed, Hadamard differentiable functional (at F_0 relative to $||\cdot||_\infty$) of the empirical distribution $\psi_n = \Psi(F_n)$. Then
$$\Psi(F_n) - \Psi(F_0) = \frac{1}{n}\sum_{i=1}^{n} \dot{\Psi}(F_0; 1(X_i \leq \cdot) - F_0) + o_p(n^{-1/2})$$

Where $\dot{\Psi}$ is the linear Gateaux derivative of Ψ perturbed in the direction $F_n - F_0$.

Strategy 4 (U/V Statistic): does the functional depend on two independent draws from the same distribution $X_1, X_2 \sim P$? If so, linearization of the U/V statistic and identifying the dominant nondegenerate term will establish asymptotic linearity.

8.7 Computing a gradient of a pathwise differentiable parameter

Strategy 1: Computing a gradient of a pathwise differentiable functional $\psi(P)$ at P_0 relative to M

(a) **Before you begin**, recalling if $M_1 \subset M_2$ are nested models, then $\text{Grad}(M_2) \subset \text{Grad}(M_1)$. Thus, we can always find a gradient in a large nonparametric model and apply it to the submodel.

(b) If unfamiliar, choose a simple parametric submodel centered at P_0. Often the linear submodel will suffice

$$p_{\theta,h}(x) = [1 + \theta h(x)]p_0(x)$$

If you're working with a model with logs, you can use the exponential submodel.

$$p_{\theta,h}(x) := \frac{\exp(\theta h(x))p_0(x)}{\int \exp(\theta h(x))dP_0(x)}$$

(c) Compute the pathwise derivative of your parameter along the path

$$\frac{\partial}{\partial\theta}\psi(P_\theta)\Big|_{\theta=0}$$

(d) Write the pathwise derivative from step 2 as an inner product of the score g and some function $\tilde{D}(P_0)$

$$\frac{\partial}{\partial\theta}\psi(P_\theta)\Big|_{\theta=0} = \langle\tilde{D}(P_0), g\rangle$$

Note: $\tilde{D}(P_0)$ cannot depend on choice of g and must lie in $L^2(P_0)$

(e) To ensure the gradient lies in the (nonparametric) tangent space, recenter to mean 0

$$D(P_0) := \tilde{D}(P_0) - P_0[\tilde{D}(P_0)]$$

Strategy 2: do you have access to a RAL estimator? If so, the influence function of the RAL estimator is also a gradient.

8.8 Deriving the tangent space

Strategy 1: does tangent space take the following forms?

(a) If model is parametric: the tangent space is just the linear span of the score vector for the parameter $\beta \in \mathbb{R}^q$. Thus, the tangent space is a linear span in $\mathbb{R}^q$, implying it is a finite-dimensional subset of $L_0^2(P)$.

(b) If model is nonparametric: the tangent space is $L_0^2(P)$

(c) If model is semiparametric: the more restrictive the semiparametric model, the smaller the tangent space. Will be proper subspace of $L_0^2(P)$.

Strategy 2: if we're deriving the tangent space to a model M subject to a moment restriction, such as $P(g_0) = 0$, we use the linear submodel to obtain the restriction on what scores allow us to remain in our tangent space

$$\begin{aligned}\int g_0[1 + \theta h]f_0 dx = 0 \\ \implies P_0(g_0 h) = 0\end{aligned}$$

Projections of elements $h \in L_0^2(P)$ onto T_M is given by

$$h(x) - \frac{P(g_0 h)}{P(g_0^2)}g_0(x)$$

So the EIF of $\psi(P) = P(f)$ in this model is

$$D^*(P) = f(x) - P(f) - \frac{P(g_0 f_0)}{P(g_0^2)} g_0(x)$$

Strategy 3: write the model in terms of fluctuating each component of the model separately. Break model into **variationally independent** components.

For instance, suppose that $X := (Y, Z) \sim P \in M$. We can write

$$M \equiv M_Z \otimes M_{Y|Z}$$

Where M_Z and $M_{Y|Z}$ are models for P_Z and $P_{Y|Z}$ separately such that the two are variationally independent of each other. Then the tangent space can be written as the sum of orthogonal subspaces

$$T_M = T_{M_Z} + T_{M_{Y|Z}}$$

And the projection onto M can be obtained by projection onto each of the subspaces.

Strategy 4: suppose $P = AB$ and the parameter ψ depends on P only through A, B is an orthogonal nuisance. We can write the total tangent space as

$$T_M(P) = T_{M_A}(P) + \cancelto{0}{T_{M_B}(P)}$$

Projection of the gradient onto T_M is tantamount to projection onto T_{M_A}, since the pathwise derivative is 0 along paths that fluctuate P_B. Thus, the EIF is entirely contained in T_{M_A} and restricting the form of the model P_B, even assuming it is known, does not impact the efficiency of your estimator.

8.9 Computing projections onto the tangent space

Strategy 0: can you guess at the form of the projection? The guess must (a) lie in $L^2(P_0)$, (b) lie in the tangent space, and (c) the residual must lie in the orthogonal complement.

Strategy 1: consider $X := (Y, Z) \sim P_0 \in M$. The tangent space is $T_M = T_{M_Z} + T_{M_{Y|Z}}$. If M_Z and $M_{Y|Z}$ are nonparametric:

$$T_{M_Z}(P) := \{s \in L_0^2(P) : s(y_1, z) = s(y_2, z) \; \forall \; z, y_1, y_2\}$$
$$T_{M_{Y|Z}}(P) := \{s \in L_0^2(P) : \mathbb{E}_P[s(Y, Z)|Z = z] = 0 \; \forall \; z\}$$

Then the projections onto the components of the tangent space are

$$\Pi[s|T_{M_Z}(P)] = \mathbb{E}_P(s(Y, Z)|Z = z)$$
$$\Pi[s|T_{M_{Y|Z}}(P)] = s(y, z) - \mathbb{E}_P(s(Y, Z)|Z = z)$$

Strategy 2: if the model is composed of independent components, modelled nonparametrically, $X := (Y, Z) \sim P_0 \in M$ s.t. $T_M = T_{M_Z}(P) + T_{M_{Y|Z}}(P)$, then the projection onto each of the subspaces is given by

$$\Pi[s|T_{M_Y}(P)] = \mathbb{E}_P(s(Y, Z)|Y = y)$$
$$\Pi[s|T_{M_Z}(P)] = \mathbb{E}_P(s(Y, Z)|Z = z)$$

Strategy 3: if M is a parametric model $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$, then the tangent space is a finite-dimensional subspace corresponding to the linear span spanned of the score

$$T_M = \left\{ u^T \frac{\partial}{\partial \theta} \log p_\theta(x) : u \in \mathbb{R}^q \right\}$$

Letting $g_\theta(x) := \frac{\partial}{\partial \theta} \log p_\theta(x)$ denote the score with respect to θ, the projection of $s \in L_0^2(P_0)$ onto the space is obtained by

$$\Pi[s|T_M] = \frac{\mathbb{E}_0[s(X) g_\theta(X)]}{\mathbb{E}[g_\theta(x)^2]} g_\theta(x)$$

9 Examples

9.1 Decision Theory

9.1.1 Bayes Rules

Example 9.1 (Bayes rule, admissible, minimax rule under modified squared error loss (P4 Theory Exam 2021))**.** *Suppose Z is a random variable with PMF*

$$p_\theta(z) = (1-\theta)\theta^z \qquad z \in \{0, 1, \ldots\}$$

For $\theta[0,1)$. We sih to study the performance of estimators of θ which will be judged by the risk function

$$R(T,\theta) = \frac{E_{P_\theta}(\{\theta - T(Z)\}^2)}{1-\theta}$$

(a) ***Calculate the Bayes rule****. Suppose we have a prior Π with nondegenerate support on $[0,1)$. To find the Bayes rule, we minimize the Bayes risk function wrt the action a.*

$$T_\Pi = \underset{a}{argmin}\ \mathbb{E}_\theta\left[\frac{\{\theta - a\}^2}{1-\theta}|Z=z\right]$$
$$\implies a = \frac{\mathbb{E}_\theta\left[\frac{\theta}{1-\theta}|Z=z\right]}{\mathbb{E}_\theta\left[\frac{1}{1-\theta}|Z=z\right]}$$

We can write these posterior expectations conditional on $Z=z$ by integrating the value against the PMF.

$$T_\Pi = \frac{\int \frac{\theta}{1-\theta}(1-\theta)\theta^z d\Pi}{\int \frac{1}{1-\theta}(1-\theta)\theta^z d\Pi} = \frac{\mathbb{E}_\Pi(\theta^{z+1})}{\mathbb{E}_\Pi(\theta^z)}$$

Which is a ratio of posterior expectations.

(b) ***Prove T_Π is admissible*** *if Π is a fixed prior with nondegenerate support. We know that all unique Bayes rules are admissible. As shown in part (a), the Bayes rule must satisfy:*

$$T_\Pi = \frac{\mathbb{E}_\Pi(\theta^{z+1})}{\mathbb{E}_\Pi(\theta^z)} \implies [\mathbb{E}_\Pi(\theta^z)]\, T_\Pi - \left[\mathbb{E}_\Pi(\theta^{z+1})\right] = 0$$

Thus, when Π is fixed, T_Π is a solution in X to the problem $aX - b = 0$ for fixed $a, b \in R$. This is a linear system of equations with only one solution. Thus, T_Π is the unique Bayes rule and therefore is admissible.

(c) ***Show constant risk****: consider the estimator $T(z) = 0.5\mathbb{I}(z=0) + \mathbb{I}(z \geq 1)$. Show the risk function is constant over all $\theta \in \Theta$. Notice that $P(Z=0) = (1-\theta)\theta^0 = (1-\theta)$ therefore $P(Z \geq 1) = \theta$.*

$$\begin{aligned}
R(T,\theta) &= \frac{E_{P_\theta}(\{\theta - T(Z)\}^2)}{1-\theta} \\
&= \frac{E_{P_\theta}(\{\theta - (0.5\mathbb{I}(z=0) + \mathbb{I}(z\geq 1))\}^2)}{1-\theta} \\
&= \frac{\theta^2 - 2\theta\mathbb{E}_{P_\theta}(0.5\mathbb{I}(z=0) + \mathbb{I}(z\geq 1)) + \mathbb{E}_{P_\theta}((0.5\mathbb{I}(z=0) + \mathbb{I}(z\geq 1))^2)}{1-\theta}]] \\
&= \frac{\theta^2 - \theta(1-\theta) - 2\theta^2 + 0.25(1-\theta) + \theta}{1-\theta} \\
&= 0.25
\end{aligned}$$

Therefore, this particular form of $T(z)$ ensures that the risk function R is constant over θ.

(d) ***Exhibit a minimax estimator**: the idea is to find a minimax estimator by finding a prior such that the Bayes rule developed in part (a) equals the estimator developed in part (c) which has constant risk. Bayes rule + constant risk implies minimax! Setting our bayes estimator equal to our estimator with constant risk, we see*

$$\frac{\mathbb{E}_\Pi(\theta^{z+1})}{\mathbb{E}_\Pi(\theta^z)} = \frac{1}{2}\mathbb{I}(z=0) + \mathbb{I}(z \geq 1)$$

$$(z=0 \text{ case}) \quad \mathbb{E}_\Pi(\theta^1) = \mathbb{E}(\theta^0) \cdot \frac{1}{2}\mathbb{I}(z=0) = \frac{1}{2}$$

$$(z=1 \text{ case}) \quad \mathbb{E}_\Pi(\theta^2) = \mathbb{E}(\theta^1) \cdot \mathbb{I}(z \geq 1) = \frac{1}{2}$$

$$\vdots$$

This implies that all the moments of $\mathbb{E}_\Pi[\theta] = \frac{1}{2}$. The only distribution with constant raw moments is a Bernoulli distribution with $p = 1/2$. Thus, T_Π is minimax!

Example 9.2 (Bayes, Admisible, Minimax Rules in Poisson-Gamma Model (581 Midterm P3))**.** *Suppose $X \sim Pois(\lambda)$. Consider the weighted squared error loss for λ:*

$$L(T(X), \lambda) := \frac{(T(X) - \lambda)^2}{\lambda}$$

(a) ***Compute Bayes Estimator** when $\Pi \equiv Gamma(\lambda|a, b)$ with density $b^a\lambda^{a-1}\exp(-b\lambda)/\Gamma(a)$. First calculate the form of the posterior:*

$$\begin{aligned}\lambda|X &\propto X|\lambda \times \Pi \\ &\propto \frac{\lambda^x \exp(-\lambda)}{x!} \times b^a\lambda^{a-1}\exp(-b\lambda)/\Gamma(a) \\ &\propto \lambda^{x+a-1}\exp(-(b+1)\lambda) \equiv Gamma(x+a, b+1)\end{aligned}$$

Next, we find the form of the Bayes Estimator by minimizing the Bayes risk function with respect to the action

$$\frac{\partial f}{\partial a} = \frac{\partial}{\partial a}\mathbb{E}\left[\frac{(a-\lambda)^2}{\lambda}\Big|X = x\right] = 0$$

$$a = \frac{1}{\mathbb{E}\left[\frac{1}{\lambda}|X = x\right]}$$

Now, since $\lambda \sim Gamma(x+1, b+1)$, $1/\lambda \sim Inv\ Gamma(x+a, b+1)$ which has mean $(b+1)/(x+a-1)$. Therefore, the bayes rule takes value

$$T_\Pi(x) = \frac{1}{(b+1)/(x+a-1)} = \frac{x+a-1}{b+1}$$

(b) ***Prove** $T(X) = X$ **is Minimax** under loss. Note that under the specified loss*

$$\mathcal{R}(X, \lambda) = \mathbb{E}\left(\left(\frac{(X-\lambda)}{\lambda^{1/2}}\right)^2\right) = \mathbb{E}(\chi_1^2) = 1$$

Thus, the risk function is constant over $\lambda \in (0, \infty)$. Our new goal is to construct a sequence of priors Π_k such that

$$\lim_{k\to\infty} r(D_{\Pi_k}, \Pi_k) = \sup_\lambda \mathcal{R}(X, \theta)$$

We derived the Bayes estimator for $\Pi \sim \Gamma(a, b)$ prior to be $\frac{x+a-1}{b+1}$. To prove $T(x) = x$ is minimax, we can choose the prior sequence $\Pi_k \sim \Gamma(a = 1+1/k, b = 1/k)$ such that asymptotically, the Bayes rule $D_{\Pi_k} \to X$ which will attain the constant risk value demonstrated above. Thus, X is minimax.

9.1.2 Minimax Rules

Example 9.3 (Sample mean is minimax)**.** *Consider $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ with σ^2 known. We claim $\bar{X}_n$ is minimax. Under squared error loss, letting $T : X_1, \dots, X_n \to \bar{X}_n$*

$$R(\bar{X}_n, \theta) = \mathbb{E}[(\bar{X}_n - \theta)^2] = \frac{\sigma^2}{n}$$

Consider the prior sequence $\Pi_k := N(0, k)$. Under this model, the posterior takes the form

$$\theta | X \sim N\left(\frac{\bar{x}_n n/\sigma^2}{1/k + n/\sigma^2}, \frac{1}{1/k + n/\sigma^2}\right)$$

Under squared error loss, the Bayes rule is the posterior mean

$$r(T_{\Pi_k}, \Pi_k) - \mathbb{E}[(\bar{x}_n - \theta)^2] = \mathbb{E}\left[\left(\frac{\bar{x}_n n/\sigma^2}{1/k + n/\sigma^2} - \theta\right)^2\right] = \mathbb{E}[(\bar{x}_n - \theta)^2] \to 0$$

Thus, $\sup_{\theta \in \Theta} R(D, \theta) = \frac{\sigma^2}{n} = \lim_{k \to \infty} r(T_{\Pi_k}, \Pi_k)$. This implies $\bar{X}_n$ is minimax in $P_1 := \{N(\theta, \sigma^2) : \theta \in \mathbb{R}, \sigma^2$ known$\}$ by Strategy 3 under Finding Minimax Rules.

We can go further and show that $\bar{X}_n$ is minimax with respect to distributions with bounded variance. Consider $P_2 := \{P \in Q^n; support(Q) \subset \mathbb{R}, Var_Q(X) \leq \sigma^2)\}$. Note that for any distribution in P_2, by CLT

$$R(\bar{X}_n, \theta) = \frac{Var_Q(X)}{n} \leq \frac{\sigma^2}{n}$$

Thus, $\sup_{P \in \mathcal{P}_1} \mathcal{R}(D_1, P) = \sup_{P \in \mathcal{P}_2} \mathcal{R}(D_1, P)$ implying by Strategy 4 in Finding Minimax Rules that $\bar{X}_n$ is minimax over P_2.

Example 9.4 (Lower Bounding the Minimax Risk of a density at a point – Le Cam's Method)**.** *Let $\mathcal{P}(\beta, L)$ be the collection of densities ($q \geq 0, \int q(x)dx = 1$) that belong in a Holder class $\Sigma(\beta, L)$ meaning the density is $(\beta - 1)$-times differentiable with derivative $q^{(\beta-1)}$ that satisfies for all x_1, x_2*

$$\left|q^{(\beta-1)}(x_1) - q^{(\beta-1)}(x_2)\right| \leq L|x_1 - x_2|$$

If our goal is to estimate the density at a point, $p(x_0)$, we can pursue Le Cam's Method.

(a) Propose two candidate distributions with large discrepancy and small KL diveregence. Let ϕ denote the density of a standard normal RV;

$$p_1 : x \to \sigma^{-1}\phi\left(\frac{x - x_0}{\sigma}\right)$$
$$p_2 : x \to p_1 + Lh_n^\beta\left[K\left(\frac{x - x_0}{h_n}\right) - K\left(\frac{x - 1 - x_0}{h_n}\right)\right]$$

Where for sufficiently small $a > 0$, $K : x \to a \exp\left(-\frac{1}{1-4x^2}\right)\mathbb{I}(|x| \leq 1/2)$.

(b) Verify $p_1, p_2 \in \mathcal{P}$.

i. p_2: Let $H_\beta(x)$ is the β-the Hermite polynomial.

$$\frac{d^\beta}{dx^\beta} p_1(x) = (-1)^\beta H_\beta(x)\phi(x)$$

Since $\lim_{|x| \to \infty} \frac{1}{\sqrt{2\pi}} H_\beta(x) e^{-x^2/2} = 0$ and the derivative is continuous, $\left|\frac{d^\beta}{dx^\beta} p_1(x)\right|$ is bounded uniformly by a constant. We can make this constant $\leq L$ by choosing σ large enough.

ii. p_2*: clearly integrates to 1 because of bump term cancellations. In order for* p_2 *to be positive, we need to choose* a^* *such that*

$$\begin{aligned}
&0 < p_1(x) - Lh_n^\beta K\left(\frac{x-h_n-x_0}{h_n}\right) \\
\implies &0 < p_1(x) - Lh_n^\beta a \exp\left(-\frac{1}{1-4\left(\frac{x-h_n-x_0}{h_n}\right)^2}\right)\mathbb{I}\left(\left|\frac{x-1-x_0}{h_n}\right| \le 1/2\right) \\
\implies &0 < p_1(x) - Lh_n^\beta a \exp\left(-\frac{1}{1-4\left(\frac{x-h_n-x_0}{h_n}\right)^2}\right)\mathbb{I}\left(x_0+1-\frac{h_n}{2} \le x \le x_0+1+\frac{h_n}{2}\right) \\
\implies &a^* < \inf_{x\in\mathbb{I}(\dots)} \frac{p_1(x)}{Lh_n^\beta \exp\left(-\frac{1}{1-4\left(\frac{x-h_n-x_0}{h_n}\right)^2}\right)}
\end{aligned}$$

To ensure p_2 *is in the Holder class, it is sufficient to show that* q *is* β*-times differentiable with bounded derivatives. We note that the Bump functions* K *and its* β *derivatives are continuous functions defined on a compact interval, therefore they obtain their maxima and minima. This means that the* β*-th derivative is upper bounded by a constant, and we can force this constant to be less than* L *by choosing* $\sigma, a > 0$ *small enough.*

(c) Study KL divergence, using the Taylor expansion $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \dots$

$$\begin{aligned}
-KL(P_1,P_2) &= \int \log\left(\frac{p_2}{p_1}\right)p_1 d\nu \\
&= \int \log\left(\frac{p_1+bump}{p_1}\right)p_1 d\nu \\
&= \int \left(\sum_{i=1}^{\infty}(-1)^{i+1}\frac{\left(\frac{bump}{p_1}\right)^i}{i}\right)p_1 d\nu \\
\textit{1st order term} &= \int bump\, d\nu = 0 \\
\textit{2nd order term} &= \frac{1}{2}\int \frac{bump^2}{p_1(x)}d\nu \\
&= \frac{1}{2}\int L^2h^{2\beta}p_1(x)^{-1}\left[K\left(\frac{x-x_0}{h_n}\right) - K\left(\frac{x-1-x_0}{h_n}\right)\right]^2 d\nu \\
&\overset{h_n small}{=} c_1h_n^{2\beta}\int p^{-1}(x)\left[K\left(\frac{x-x_0}{h_n}\right)^2 - K\left(\frac{x-1-x_0}{h_n}\right)^2\right] \quad (h_n \textit{ small bumps don't overlap}) \\
&= c_1h_n^{2\beta+1}\int p^{-1}(h_nU+x_0)\left[K(U)^2 - K\left(U-\frac{1}{h_n}\right)^2\right]dU \\
&= c_2h_n^{2\beta+1} \\
\textit{3rd order term} &= o(h^{3\beta})
\end{aligned}$$

Now the KL-divergence under n-iid draws yields

$$-KL(P_1^n, P_2^n) \ge cnh_n^{2\beta+1}$$

To get a stable lower bound on the KLD, we require $h_n = \mathcal{O}(n^{-\frac{1}{2\beta+1}})$.

(d) Study Discrepancy:

$$\begin{aligned} d(P_1, P_2) &= \frac{1}{2}(p_1(x_0) - p_2(x_0))^2 \\ &= \frac{1}{2}\left(p_1(x_0) - p_1(x_0) - Lh_n^\beta\left[K(0) - K\left(-\frac{1}{h_n}\right)\right]\right)^2 \\ &= Ch_n^{2\beta} \end{aligned}$$

We have all the pieces we need now. Applying Le Cam's method, we obtain

$$\inf_{T\in\mathcal{T}}\sup_{P\in\mathcal{P}} R(T,P) \geq \frac{1}{4}d(P_1,P_2)\exp(-KL(P_1,P_2)) \geq c\cdot h_n^{2\beta}$$

In order for the KL divergence to have a stable lower bound, we required $h_n = \mathcal{O}(n^{-\frac{1}{2\beta+1}})$.

$$\inf_{T\in\mathcal{T}}\sup_{P\in\mathcal{P}} R(T,P) \leq c^* n^{-\frac{2\beta}{2\beta+1}}$$

Thus a lower bound on the minimax rate is $\mathcal{O}(n^{-\frac{2\beta}{2\beta+1}})$ *which is a slightly slower than parametric rate.*

Example 9.5 (Lower bounding minimax risk of smooth regression function – Fano's Method)**.** *Suppose we observe* $(X_1,Y_1),\ldots,(X_n,Y_n) \overset{iid}{\sim} Q \in \mathcal{Q}$*, where* $X \sim U[0,1]$ *and* $Y|X = x \sim N(f_Q(x),1)$ *where* $f_Q(x) \in \mathcal{F}(\beta, L)$ *a Holder class. Suppose our objective is to estimate* $f_Q(x)$*, with performance quantified by the mean integrated squared error:*

$$L(a, Q^n) = \int_0^1 [a(x) - f_Q(x)]^2 dx$$

We take the following few steps

(a) Define candidate function class. Let $\mathcal{F}_1$ *denote a convex combination of orthonormal basis functions where the elements of the basis are scaled bump functions:*

$$\mathcal{F}_1 := \left\{ x \to \sum_{j=1}^m w_j\phi_j(x) : w \in \{0,1\}^m, \phi_j(x) = Lh^\beta K\left(\frac{x - \frac{j}{m+1}}{h}\right), m \in \left[8, \frac{1}{h-1}\right] \right\}$$

Where for sufficiently small a*,*

$$K : x \to a\exp\left(-\frac{1}{1-4x^2}\right)\mathbb{I}(|x| < 1/2)$$

So $\mathcal{F}_1$ *is a collection of functions that are sums of* m *bump functions centered at* $\frac{j}{m+1}$ *for* $j = 1,\ldots,m$*, that are multiplied by 0 or 1, and that do not overlap since* $m \leq \frac{1}{h} = 1 \implies h \leq \frac{1}{m+1}$*. Recall that* $\Omega := \{0,1\}^m$ *indexes the collection of functions in* $\mathcal{F}_1$*. Thus,* $|\mathcal{F}_1| = |\Omega| = 2^m$*.*

(b) Study the discrepancy:

$$\begin{aligned} d(P_w, P_\nu) &= \frac{1}{2}\int [f_w(x) - f_\nu(x)]^2 dx \\ &= \frac{1}{2}\sum_{j=1}^{m}[w_j - \nu_j]^2 \int \phi_j(x)^2 dx \quad \textit{(Bases orthogonal so cross terms cancel)} \\ &= \frac{1}{2}\sum_{j=1}^{m}[w_j - \nu_j]^2 L^2 h^{2\beta+1} \underbrace{\int K(u)^2 du}_{c_2} \quad \textit{(U-sub)} \\ &= \frac{1}{2} c_2 L^2 h^{2\beta+1} \underbrace{\sum_{j=1}^{m}[w_j - \nu_j]^2}_{\textit{Hamming dist}} \\ &= c_3 h^{2\beta+1} H(w, \nu) \quad \left(c_3 := \frac{c_2 L^2}{2}\right) \end{aligned}$$

The minimal Hamming distance for two functions in that differ is exactly 1, yielding;

$$\min_{j \neq k} d(P_j, P_k) = c_3 h^{2\beta+1}$$

(c) Study the KL divergence. Turns out KL divergence takes the form:

$$\begin{aligned} KL(P_w, P_\nu) &= \frac{n}{2}\int_0^1 [f_w(x) - f_\nu(x)]^2 dx \\ &= c_3 n h^{2\beta+1} H(w, \nu) \quad \textit{(By same logic)} \\ &\leq c_3 n h^{2\beta+1} m \quad \textit{(since } H(w,\nu) \leq m\textit{)} \end{aligned}$$

(d) Plug into Fano's Bound: recall that $\Omega := \{0,1\}^m$ indexes the collection of functions in $\mathcal{F}_1$.

$$\begin{aligned} \inf_{T \in \mathcal{T}} \sup_{P \in \mathcal{P}} R(T, P) &\geq \frac{\min_{j \neq k} d(P_j, P_k)}{2}\left[1 - \frac{\log 2 + \max_{j \neq k} KL(P_j, \bar{P})}{\log(|\Omega|)}\right] \\ &\geq \frac{c_3 h^{2\beta+1}}{2}\left(1 - \frac{\log 2 + c_3 n h^{2\beta+1} m}{\log|\Omega|}\right) \\ &= \frac{c_3 h^{2\beta+1}}{2}\left(1 - \frac{\log 2 + c_3 n h^{2\beta+1} m}{m \log 2}\right) \end{aligned}$$

For this bound to be informative, $h = \mathcal{O}(n^{-1/(2\beta+1)})$. But this produces a lower bound on the minimax risk of $\mathcal{O}(n^{-1})$, meaning the problem is as least as difficult as a parametric problem. This suggests that the bound may not be tight.

(e) Tighten the bound using the Varshamov-Gilbert Lemma. For $m \geq 8$, there exista an $\Omega \subset \Omega$ s.t. $|\Omega| \geq 2^{m/8}$ and $\min_{w \neq v} H(w, v) \geq \frac{m}{8}$. If we choose this subset

$$\begin{aligned} \inf_{T \in \mathcal{T}} \sup_{P \in \mathcal{P}} R(T, P) &\geq \frac{c_3 h^{2\beta+1} m}{16}\left(1 - \frac{\log 2 + c_3 n h^{2\beta+1} m}{\frac{m}{8}\log 2}\right) \\ &= \frac{c_3 h^{2\beta+1} m}{16}\left(1 - \frac{8}{m} - \frac{8 c_3 n h^{2\beta+1}}{\log 2}\right) \end{aligned}$$

Goal is to choose m as large as possible to provide the tightest bound. If we choose $m = \lfloor \frac{1}{h} - 1 \rfloor$.

We know that $\frac{1}{2h} < m < \frac{1}{h}$. Plugging in the lower bound, we have

$$\begin{aligned}\inf_{T\in\mathcal{T}}\sup_{P\in\mathcal{P}} R(T,P) &\geq \frac{c_3 h^{2\beta+1} m}{16}\left(1 - \frac{8}{m} - \frac{8c_3 n h^{2\beta+1}}{\log 2}\right) \\ &\geq \frac{c_3 h^{2\beta}}{32}\left(1 - 16h - \frac{8c_3 n h^{2\beta+1}}{\log 2}\right)\end{aligned}$$

To ensure that the negative term above is bounded, we require $n = h^{2\beta+1} \implies h = \mathcal{O}(n^{-1/(2\beta+1)})$. Since the bandwidth is $h = \mathcal{O}(n^{-1/(2\beta+1)})$, the lower bound on the minimax risk is $\mathcal{O}(n^{-2\beta/(2\beta+1)})$.

9.1.3 Admissible Rules

Example 9.6 (Posterior Mean is Admissible in Normal Model)**.** *Let $X_1,\dots,X_n \overset{iid}{\sim} N(\theta,\sigma^2)$ and $\theta \sim N(\mu,\tau^2)$. We will show that the following estimator is admissible*

$$T_\Pi : (X_1,\dots,X_n) \to \left(1 - \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}\right)\bar{X}_n + \left(\frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}\right)\mu$$

By Strategy 1 in finding admissible estimators, we are using squared error loss and the Bayes risk is finite because all the random quantities are finite. Also, the normal distribution is absolutely continuous wrt the Lesbegue measure and vice versa. Therefore, T_Π is unique Bayes and therefore admissible for $\left(\frac{1/\tau^2}{1/\tau^2+n/\sigma^2}\right) \in (0,1)$.
When $\left(\frac{1/\tau^2}{1/\tau^2+n/\sigma^2}\right) = 0$, $T : x \to \mu$ is admissible because it is a constant estimator that achieves risk 0 when $\theta = \mu$. When $\left(\frac{1/\tau^2}{1/\tau^2+n/\sigma^2}\right) = 1$, turns out the sample mean is admissible, but this requires further proof.

Example 9.7 (Sample mean is Admissible in Normal Model)**.** *Let $X_1,\dots,X_n \overset{iid}{\sim} N(\theta,\sigma^2)$ with σ^2 known. We claim that $\bar{X}_n$ is admissible in the model.*
We will show by demonstrating either

(a) $R(T,\theta) \geq R(\bar{X}_n,\theta)\ \forall\theta \in R$
(b) There exists some θ for which $R(T,\theta) > R(\bar{X}_n,\theta)$

Consider WLOG $\sigma^2 = 1$. Suppose (a) does not hold. We will show that (b) holds. If (a) does not hold, there exists a θ_1 s.t. $R(T,\theta_1) < R(\bar{X}_n,\theta_1)$. By continuity of R, there exists $\epsilon,\delta > 0$ s.t. for all $\theta \in (\theta_1-\delta,\theta_1+\delta)$,

$$R(T,\theta) < R(\bar{X}_n,\theta) - \epsilon = \frac{1}{n} - \epsilon$$

Specifying the prior $\Pi = N(0,\tau^2)$ and the Bayes rule T_Π as the posterior mean, we obtain

$$\begin{aligned}r(T_\Pi,\Pi) - R(\bar{X}_n,\theta) &= \int R\left(\frac{n}{1/\tau^2+n}\bar{X}_n,\theta\right)d\Pi(\theta) - \frac{1}{n} \\ &= \int\left(\frac{n}{1/\tau^2+n}\bar{X}_n - \theta\right)^2 \cdot \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{1}{2\tau^2}\theta^2\right)d\theta - \frac{1}{n} \\ &= \frac{\tau^2}{1+n\tau^2} - \frac{1}{n} = -\frac{1}{n(1+n\tau^2)}\end{aligned}$$

By optimality of the Bayes rule

$$\begin{aligned}&r(T_\Pi,\Pi) - R(\bar{X}_n,\theta) \leq r(T_1,\Pi) - R(\bar{X}_n,\theta) \\ \implies &\frac{\tau^2}{1+n\tau^2} - \frac{1}{n} = -\frac{1}{n(1+n\tau^2)} \leq \int\left[\mathcal{R}(T_1,\theta) - \frac{1}{n}\right]^+\Pi(d\theta) - \int\left[\mathcal{R}(T_1,\theta) - \frac{1}{n}\right]^-\Pi(d\theta)\end{aligned}$$

Recall that for $\theta \in (\theta_1 - \delta, \theta_1 + \delta)$ and $R(T_1, \theta) < \frac{1}{n} - \epsilon$ implying $[R(T_1, \theta) - 1/n]^- > \epsilon$. Then simple bounding yields

$$\begin{aligned}\int \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^- \Pi(d\theta) &\leq \int_{\theta_1 - \delta}^{\theta_1 + \delta} \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^- \Pi(d\theta) \\ &\leq \epsilon \int_{\theta_1 - \delta}^{\theta_1 + \delta} d\Pi(\theta) \\ &= \epsilon \Pi(\theta_1 - \delta \leq \theta \leq \theta_1 + \delta)\end{aligned}$$

Implying

$$\int \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^+ d\Pi(\theta) \geq -\frac{1}{n(1 + n\tau^2)} + \epsilon \Pi(\theta_1 - \delta \leq \theta \leq \theta_1 + \delta)$$

Noting that

$$\sqrt{2\pi}\tau \left(-\frac{1}{n(1 + n\tau^2)} + \epsilon \Pi(\theta_1 - \delta \leq \theta \leq \theta_1 + \delta)\right) \overset{\tau \to \infty}{\longrightarrow} 2\epsilon\delta$$

Thus, choosing τ_0 s.t. $\sqrt{2\pi}\tau_0 \left(-\frac{1}{n(1+n\tau_0^2)} + \epsilon \Pi(\theta_1 - \delta \leq \theta \leq \theta_1 + \delta)\right) > \epsilon\delta$ we obtain

$$\int \left[\mathcal{R}(T_1, \theta) - \frac{1}{n}\right]^+ d\Pi(\theta) \leq \left(-\frac{1}{n(1 + n\tau_0^2)} + \epsilon \Pi(\theta_1 - \delta \leq \theta \leq \theta_1 + \delta)\right) > \frac{\epsilon\delta}{\sqrt{2\pi}\tau_0} > 0$$

Thus, ther exists θ for which $R(T, \theta) > R(\bar{X}_n, \theta)$ implying condition (b) holds. Thus, the sample mean is admissible.

Example 9.8 (Sample mean is inadmissible in $d \geq 3$)**.** *Suppose $X_1, \ldots, X_n \sim N(\theta, I_d)$ for $d \geq 3$. Let T^{JS} be the James-Stein estimator:*

$$T^{JS} : x \to \begin{cases} \left(1 - \frac{(d-2)}{n||\bar{x}_n||^2}\right) \bar{x}_n & \text{if } \bar{x}_n \neq (0, \ldots, 0) \\ 0 & \text{if } \bar{x}_n = (0, \ldots, 0) \end{cases}$$

Under MSE loss, letting T denote the sample mean

$$\begin{aligned}R(T^{JS}, \theta) &= \mathbb{E}[||T^{JS}(||X||)X - \theta||^2] \quad \text{(T is spherically symmetric est)} \\ &= \mathbb{E}\left[||[T^{JS}(||X||) - 1]X + [X - \theta]||^2\right] \\ &= \mathbb{E}\left[||[T^{JS}(||X||) - 1]X||^2\right] + \mathbb{E}[||X - \theta||^2] - 2\mathbb{E}\left[\langle [1 - T^{JS}(||X||)]X,\ X - \theta\rangle\right] \\ &= \mathbb{E}\left[\frac{(d-2)^2}{||X||^2}\right] + R(T, \theta) - 2(d-2)\mathbb{E}\left[\left\langle \frac{X}{||X||^2}, X - \theta \right\rangle\right]\end{aligned}$$

To show the third term in the above display is $-2\mathbb{E}\left[||[T^{JS}(||X||) - 1]X||^2\right]$, we appeal to Stein's Lemma.

Stein's Lemma: Letting $Y \sim N(\mu, \sigma^2 I_d)$ and $g_1, \ldots, g_d$ be functions from $\mathbb{R}^d \to \mathbb{R}$ s.t. for all $j = 1, \ldots, d$, $\mathbb{E}\left|\frac{\partial}{\partial y_j} g_j(y)|_{y=Y}\right| < \infty$. Letting $g : y \to (g_1(y), \ldots, g_d(y))$, we have

$$\mathbb{E}[\langle g(Y), Y - \mu\rangle] = \sigma^2 \mathbb{E}\left[\sum_{j=1}^{d} \frac{\partial}{\partial y_j} g_j(y)\right]$$

Define $g_j : z \to \frac{z_j}{||z||^2}$. Then we see that

$$\begin{aligned}
\mathbb{E}\left[\left\langle \frac{X}{||X||^2}, X-\theta \right\rangle\right] &= \mathbb{E}\left[\langle g(X), X-\theta\rangle\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{d} \frac{\partial}{\partial y_j} g_j(y)\right] \qquad \textit{(Stein's Lemma)} \\
&= \mathbb{E}\left[\sum_{j=1}^{d} \frac{1}{||X||^2} - \frac{2X_j}{||X||^4}\right] \qquad \textit{(Quotient Rule)} \\
&= \mathbb{E}\left[\frac{d}{||X||^2} - \frac{2||X||^2}{||X||^4}\right] \\
&= \mathbb{E}\left[\frac{d-2}{||X||^2}\right]
\end{aligned}$$

Therefore,

$$\begin{aligned}
R(T^{JS},\theta) &= \mathbb{E}\left[\frac{(d-2)^2}{||X||^2}\right] + R(T,\theta) - 2(d-2)\mathbb{E}\left[\left\langle \frac{X}{||X||^2}, X-\theta \right\rangle\right] \\
&= \mathbb{E}\left[\frac{(d-2)^2}{||X||^2}\right] + R(T,\theta) - 2\mathbb{E}\left[\frac{(d-2)^2}{||X||^2}\right] \\
&= R(T,\theta) - \mathbb{E}\left[\frac{(d-2)^2}{||X||^2}\right]
\end{aligned}$$

Therefore, $R(T^{JS},\theta) < R(T,\theta)$ for all θ.

9.2 Hypothesis Testing

Example 9.9 (Power under local alternatives for location family)**.** *Suppose $X_1^n \sim P_\theta$ for location family where (i) P_θ has density $f(x-\theta)$, (ii) f is symmetric about 0, (iii) f is positive and continuously differentiable with finite second moment.*
Suppose we wish to test $H_0 : \theta = 0$ and $H_1 : \theta > 0$ with the following sign and t-statistics:

(a) Sign: $S_n = \frac{1}{n}\sum I(X_i > 0)$

(b) t-statistic: $T_n = \frac{1}{n}\sum \frac{X_i}{\hat{\sigma}_n}$ where $\hat{\sigma}_n$ is the empirical standard deviation.

The estimators are both asymptotically linear:

$$\begin{aligned}
\sqrt{n}\left(S_n - \frac{1}{2}\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I(X_i > 0) - \frac{1}{2}\right) \rightsquigarrow N(0, 1/4) \\
\sqrt{n}T_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i}{\sigma} + o_p(1)
\end{aligned}$$

Let's show that both estimators are regular. For the sign statistic, $\mu(\theta) := P_\theta(X > 0)$

$$\begin{aligned}
\dot{\mu}(0) &= \frac{\partial}{\partial\theta} P_\theta(X>0)\Big|_{\theta=0} = \frac{\partial}{\partial\theta}\int_0^\infty f(x-\theta)dx\Big|_{\theta=0} \\
&= \int_0^\infty \frac{\partial}{\partial\theta} f(x-\theta)\Big|_{\theta=0} dx \\
&= -\int_0^\infty \dot{f}(x)dx = \int \left[-\frac{\dot{f}(x)}{f(x)}\right] I(x>0)dP_0(x) \\
&= \int \dot{\ell}_0(x) I(x>0) dP_0(x) \\
&= P_0(\dot{\ell}_0 s_0)
\end{aligned}$$

Recall regularity is equivalent to $P_0(g_\theta \dot{\ell}_\theta) = \dot{\mu}(\theta)$. *Note that* $\theta = 0$ *under* H_0 *implies* S_n *is regular. Now for the t-statistic, let* $\mu(\theta) := \theta/\sigma$. *We have*

$$\begin{aligned}
\int \dot{\ell}_0 t_0(x) dP_0(x) &= \int \dot{\ell}_0(x)\frac{x}{\sigma} dP_0(x) \\
&= -\sigma^{-1}\int \frac{\dot{f}(x)}{f(x)} x dP_0(x) \\
&= -\sigma^{-1}\int \dot{f}(x)\cdot x dx \\
&= \sigma^{-1}\int \left(f(x) - \frac{d}{dx}[xf(x)]\right)dx \quad \textit{(Product rule and add subtract)} \\
&= \sigma^{-1} - \lim_{a\to\infty}\int_{-a}^{a}\left(\frac{d}{dx}[xf(x)]\right)dx \quad \textit{(Pdf integrates to 1)} \\
&= \sigma^{-1} - \lim_{a\to\infty} a[f(a) - f(-a)] \\
&= \sigma^{-1} = \dot{\mu}(0)
\end{aligned}$$

Thus, T_n *is also regular.*

Knowing S_n *and* T_n *are regular ALEs, we know that their corresponding tests*

$$\mathbb{I}(\sqrt{n}(2S_n - 1) > z_{1-\alpha})$$
$$\mathbb{I}(\sqrt{n}T_n > z_{1-\alpha})$$

have power functions under local alternatives take the form for all h*:*

$$\begin{aligned}
\pi_n\left(\frac{h}{\sqrt{n}}\right) &\overset{n\to\infty}{\rightsquigarrow} 1 - \Phi\left(z_{1-\alpha} - h^T\frac{\dot{\mu}(0)}{\sigma(0)}\right) \\
\implies P_{\theta+h/\sqrt{n}}(\sqrt{n}(2S_n-1) > z_{1-\alpha}) &= 1 - \Phi\left(z_{1-\alpha} - 2hf(0)\right) \\
\implies P_{\theta+h/\sqrt{n}}(\sqrt{n}(T_n) > z_{1-\alpha}) &= 1 - \Phi\left(z_{1-\alpha} - h\sigma^{-1}\right)
\end{aligned}$$

Thus, we can compare the relative power of the sign test to the t-test under local alternatives by the ratio of their two slopes:

(a) If $2f(0)\sigma > 1$, *the sign test has greater local power.*

(b) If $2f(0)\sigma < 1$, *the t test has greater local power.*

This indicates when $f(0)$ *is very large relative to the variance* σ, *the sign test is more powerful under local alternatives asymptotically. For instance, if we consider* f *to be a density that for small* $\epsilon > 0$, *places mass* $(1-\epsilon)$ *on Unif*$(-1,1)$ *and* ϵ *mass at* $N(0,\epsilon^4)$, *then the sign test will have much greater power.*

9.3 Empirical Process Theory

9.3.1 Concentration Inequalities

Example 9.10 (Bivariate U statistic (McDiarmind's Inequality)). *A good use case of McDiarmind's inequality is in the study of the concentration of U-statistics, where $g : \mathbb{R}^2 \to \mathbb{R}$ and*

$$U := \binom{n}{2}^{-1} \sum_{j \leq k} g(X_j, X_k)$$

If g is bounded, say $||g||_\infty \leq b$, then McDiarmind's inequality yields for a given coordinate k:

$$\begin{aligned} |f(x) - f(x^{\setminus k})| &\leq \binom{n}{2}^{-1} \sum_{j \neq k} |g(x_j, x_k) - g(x_j, x_k')| \\ &\leq \frac{(n-1)(2b)(2)}{(n)(n-1)} = \frac{4b}{n} \end{aligned}$$

So the bounded differences property holds with parameter $c_i = \frac{4b}{n}$ in each coordinate. By McDiarmind's Inequality

$$P(|U - \mathbb{E}(U)| \geq t) \leq 2 \exp\left(-\frac{nt^2}{8b^2}\right)$$

Example 9.11 (Gaussian Order Statistics (Lipschitz Transformation of Gaussian)). *Let $X_{(k)}$ denote the k-th order statistic of a Gaussian random vector. Let $Y_{(k)}$ denote the k-th order statistic from an iid ghost sample from the same Gaussian distribution. Turns out*

$$|X_{(k)} - Y_{(k)}| \leq ||X - Y||_2$$

so each order statistic is 1-Lipschitz. Based on the concentration result for lipschitz transformations of Gaussian random vectors

$$P[|X_{(k)} - \mathbb{E}[X_{(k)}]| \geq \delta] \leq 2 \exp\left(-\frac{\delta^2}{2}\right)$$

9.3.2 Establish Uniform LLN and Upper Bounding Empirical Process Terms

Example 9.12 (Establish Uniform LLN in Lipschitz Function Class – Dudley). *If $\mathcal{F}$ denotes a class of $[0,1] \to \mathbb{R}$-valued Lipschitz functions s.t., $|f(x) - f(y)| \leq L|x - y|$.*
Let's first derive the metric entropy (log covering number) of the function class $\mathcal{F}$. Create $M = \lfloor \frac{1}{\epsilon} \rfloor$ grid points $x_i = (i-1)\epsilon$ for $i = 1, \ldots, M$ on $[0,1]$. Defining ϕ as

$$\phi(u) := \begin{cases} 0 & \text{if } u < 0 \\ u & \text{if } 0 \leq u \leq 1 \\ 1 & \text{else} \end{cases}$$

For any binary sequence $\beta = \{-1, +1\}^M$, define a function f_β such that

$$f_\beta(y) = \sum_{i=1}^{M} \beta_i L \epsilon \phi\left(\frac{y - x_i}{\epsilon}\right)$$

consider the interval from $y \in (0, x_1)$. ϕ increases linearly in $y - x_i$ the interval with slope $\pm L$. Thus, $f_\beta(y)$ is piecewise linear with slope $\pm L$ over each pair of gridpoints. For any two functions $f_\beta, f_{\beta'}$, there is at least one interval where the two functions start at the same point and have opposite slopes,

implying that $||f_\beta - f_{\beta'}||_\infty \geq 2L\epsilon$. *Thus,* $\{f_\beta, \beta \in \{-1,+1\}^M\}$ *forms a* $2L\epsilon$*-packing in the supnorm. By relationships between covering and packing numbers*

$$2^M = |f_\beta| \leq M(2L\epsilon, \mathcal{F}, ||\cdot||_\infty) \leq N(L\epsilon, \mathcal{F}, ||\cdot||_\infty)$$

Defining $\delta = \epsilon L$, *and recalling that* $M = \lfloor \frac{1}{\epsilon} \rfloor$, *we have*

$$C \cdot \frac{L}{\delta} \leq \log N(\delta, \mathcal{F}, ||\cdot||_\infty)$$

Also by plotting $\{f_\beta, \beta \in \{-1,+1\}^M\}$, *one can see that the farthest an element of* $\mathcal{F}$ *can be from a given* f_β *pointwise is* $L\epsilon$. *Thus,* $\{f_\beta, \beta \in \{-1,+1\}^M\}$ *is a* δ*-cover for* $\mathcal{F}$. *The covering number (size of smallest cover), then:*

$$N(\delta, \mathcal{F}, ||\cdot||_\infty) \leq |f_\beta| = C^* \cdot \frac{L}{\delta}$$

Therefore, going back to $\epsilon > 0$ *notation*

$$\sup_Q \log(N(\epsilon, \mathcal{F}, L^2(Q))) = \log(N(\epsilon, \mathcal{F}, ||\cdot||_\infty) = \mathcal{O}\left(\frac{L}{\epsilon}\right)$$

Recognizing that $D = 2L < \infty$, *Dudley's entropy integral gives:*

$$\mathbb{E}||R_n||_\mathcal{F} \leq \frac{8}{\sqrt{n}} \sup_Q \left[\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}, L^2(P_n))} d\epsilon\right] \equiv \frac{8}{\sqrt{n}} \left[\int_0^D \mathcal{O}\left(\frac{L}{\epsilon}\right) d\epsilon\right] = \mathcal{O}(n^{-1/2})$$

Therefore, the entropy integral is satisfied, and the empirical process term is controlled. Also, $\mathcal{F}$ *is Donsker since is satisfies the entropy integral.*

Example 9.13 (Establish Uniform LLN in Class of Functions Lipschitz in Indexing Parameters – Dudley)**.** *Let* $\mathcal{F} := \{g_\beta : \beta \in \mathbb{R}^p; ||\beta||_2 \leq 1\}$ *be a collection of functions indexed by parameter* β *where* $|g_{\beta_1}(x) - g_{\beta_2}(x)| \leq L||\beta_1 - \beta_2||$.
Step 1: Note that the indexing parameter set $B = \{\beta \in \mathbb{R}^p : ||\beta||_2 = 1\}$ *is a sphere of radius 1. We previously proved that the* ϵ*-covering number of a ball of radius* r *has the upper bound*

$$N(\epsilon, B(0,r), ||\cdot||_{L^p(P)}) \leq \left(\frac{2r}{\epsilon} + 1\right)^p$$

Step 2: we also know that functions Lipschitz in their indexing parameters also satisfy the following covering number bound on their function space $\mathcal{F}$

$$N(\epsilon, \mathcal{F}, ||\cdot||_\mathcal{F}) \leq N(\epsilon/L, B, ||\cdot||_B)$$

Step 3: bringing these two together

$$N(\epsilon, \mathcal{F}, ||\cdot||_\mathcal{F}) \leq N(\epsilon/L, B(0,1), ||\cdot||_2) \leq \left(\frac{2 \cdot 1}{\epsilon/L} + 1\right)^p$$
$$\implies \log N(\epsilon, \mathcal{F}, ||\cdot||_\mathcal{F}) \leq p \log\left(\frac{2L}{\epsilon} + 1\right) \approx p \log\left(\frac{L}{\epsilon}\right)$$

And the Dudley integral is:

$$\begin{aligned}\frac{8}{\sqrt{n}}\left[\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}, L^2(P_n))} d\epsilon\right] &\leq \frac{8}{\sqrt{n}} \int_0^{2L} \sqrt{p \log\left(\frac{L}{\epsilon}\right)} d\epsilon \\ &\lessapprox \frac{8}{\sqrt{n}} L\sqrt{p} \int_0^1 \log(1/\delta) d\delta \\ &\lessapprox \frac{8}{\sqrt{n}} L\sqrt{p}\end{aligned}$$
$$\implies \mathbb{E}||P_n - P||_\mathcal{F} \lessapprox \mathbb{E}||R_n||_\mathcal{F} = \mathcal{O}\left(\frac{L\sqrt{p}}{\sqrt{n}}\right)$$

Thus, a function that is Lipschitz in its 1-dimensional indexing parameter controls the empirical process term at a $\mathcal{O}(n^{-1/2})$ rate! However, as the dimension of the indexing parameter increases, we get slower convergence.

Example 9.14 (Establish Uniform LLN in Sobolev Class)**.** *Let $\mathcal{F}$ be a collection of functions $f : [0,1] \to \mathbb{R}$ such that*

(a) Uniformly bounded: $||f||_\infty \leq 1$

(b) Absolutely continuity of $(k-1)$-th derivative

(c) $\int f^{(k)}(x)^2 dx \leq 1$ for some $k \in \mathbb{N}$

There exists a constant C such that the log bracketing number wrt the supnorm metric takes form for all $\epsilon \in [0,1]$:

$$\log N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_\infty) \leq C\left(\frac{1}{\epsilon}\right)^{1/k}$$

Suppose $k \geq 1$, then by the bracketing integral bound is finite;

$$\begin{aligned}\mathbb{E}||P_n - P||_{\mathcal{F}} &\leq \frac{C}{\sqrt{n}}\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_\infty)}d\epsilon \\ &\leq \frac{C^*}{\sqrt{n}}\int_0^1 \sqrt{\epsilon^{-1/k}}d\epsilon \\ &= \mathcal{O}(n^{-1/2})\end{aligned}$$

Thus, we can control the empirical process term at $\mathcal{O}(n^{-1/2})$ rates. Also since the function class satisfies the uniform entropy integral, it is Donsker.

9.4 M/Z Estimation

Example 9.15 (Limiting distributions and Regularity of Mean and Median (581 HW 8 P3))**.** *Let μ_n and m_n be the sample mean and median respectively. Find the limit distributions of each under P_{θ_0} and $P_{\theta_0 + h/\sqrt{n}}$ given $P_{\theta'} \equiv N(\theta', 1)$.*

We start with the sample mean, μ_n. By WLLN, the sample mean is consistent and by the central limit theorem,

$$\sqrt{n}(\mu_n - \theta_0) \overset{P_{\theta_0}}{\rightsquigarrow} N(0,1)$$

Recalling that the normal distribution is QMD and both distributions are mutually contiguous, local asymptotic normality holds, the log likelihood ratio affords a taylor expansion, and is asymptotically normal. The joint distribution between the sample mean and log likelihood ratio is normal with covariance h. Le Cam's third lemma says that the distribution of the MLE under sampling from the local alternative is

$$\sqrt{n}(\mu_n - \theta_0) \overset{P_{\theta_0+h/\sqrt{n}}}{\rightsquigarrow} N(h,1) \implies \sqrt{n}\left(\mu_n - \left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) \overset{P_{\theta_0+h/\sqrt{n}}}{\rightsquigarrow} N(0,1)$$

Thus, the MLE is invariance to local perturbations in the parameter, implying that it is a regular estimator.

We now turn our attention to the sample median. The sample median can be defined as a z-estimator that solves the estimating equation $P_n z_\theta = 0$ where $z_\theta(x) = \mathbb{I}(x \leq \theta) - \frac{1}{2}$.

We start by proving consistency. This is a 1-dimensional Z-estimator, where the estimating function is decreasing in the parameter θ and has exactly one root. The sample estimating equation converges point wise to the population estimating equation by WLLN. We also know that for the population median equal to θ_0 and small $\epsilon > 0$,

$$P_0(\mathbb{I}(x \leq \theta_0 + \epsilon) - 0.5) < 0 < P_0(\mathbb{I}(x \leq \theta_0 - \epsilon) - 0.5)$$

Thus, $m_n \xrightarrow{p} \theta_0$.

Now we characterize the asymptotic distribution of the sample median under P_{θ_0}. Checking the conditions for asymptotic normality, we know that

(a) The estimating function z_θ is squared differentiable because it is bounded.

(b) Pz_θ is differentiable at θ_0:

$$\frac{\partial}{\partial\theta} Pz_\theta\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta} P\left(\mathbb{I}(x \leq \theta) - \frac{1}{2}\right)\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta} F_x(\theta) - \frac{1}{2}\Big|_{\theta=\theta_0} = f_x(\theta_0)$$

Recalling the form of the normal density, $f_x(\theta_0) = \frac{1}{\sqrt{2\pi}}$.

(c) $\{z_\theta(x) : \theta \in \mathbb{R}\}$ forms a Donsker class because it is a shifted indicator function, and indicator functions are VC class.

Under these conditions

$$\sqrt{n}(m_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_{\theta_0}(X_i) + o_P(1) \overset{P_{\theta_0}}{\rightsquigarrow} N\left(0, V_{\theta_0}^{-1} P_0[z_{\theta_0} z_{\theta_0}^T](V_{\theta_0}^{-1})^T\right)$$

$$\equiv N\left(0, \frac{\mathbb{E}((\mathbb{I}(X \leq \theta_0) - 0.5)^2}{\left(\frac{1}{\sqrt{2\pi}}\right)^2}\right) \equiv N(0, \pi/2)$$

Lastly, we investigate the distribution of m_n under the local alternative. This relies on the applying Le Cam's third lemma for asymptotic linear estimators. For an asymptotic linear estimator with influence function ϕ_θ, the asymptotic distribution under the local alternative is

$$\sqrt{n}(m_n - \theta_0) \overset{P_{\theta_0 + h/\sqrt{n}}}{\rightsquigarrow} N(P_0(\phi_{\theta_0} \cdot \dot{\ell})h, P_0\phi_{\theta_0}^2)$$

Thus, we must evaluate the $P_0(\phi_{\theta_0} \cdot \dot{\ell})$ to learn the limiting distribution.

(a) Recall the influence function of m_n is given by $\phi_0(x) = \frac{\mathbb{I}(x \leq \theta_0) - \frac{1}{2}}{f(\theta_0)} = \sqrt{2\pi}\left(\mathbb{I}(x \leq \theta_0) - \frac{1}{2}\right) = \sqrt{2\pi}\left(\frac{1}{2} - \mathbb{I}(x \geq \theta_0)\right)$.

(b) Recall the score is given by $\dot{\ell}(x) = \frac{\partial}{\partial\theta} - \frac{1}{2}\sum_{i=1}^{n}(X_i - \theta_0)^2 = \sum_{i=1}^{n}(X_i - \theta_0)$.

Writing the inner product of these quantities we obtain by using the mean of a positive-restricted normal.

$$\begin{aligned} P_0(\phi_{\theta_0} \cdot \dot{\ell}) &= \int \sqrt{2\pi}\left(\frac{1}{2} - \mathbb{I}(x \leq \theta_0)\right)(x - \theta_0) dP_0(x) \\ &= \int \sqrt{2\pi}\left(\mathbb{I}(x \geq \theta_0) - \frac{1}{2}\right)(x - \theta_0) dP_0(x) \\ &= \sqrt{2\pi}\int_0^\infty (x - \theta_0) dP_0(x) = \sqrt{2\pi}\frac{1}{\sqrt{2\pi}} = 1 \end{aligned}$$

Therefore, the sample median m_n is also a regular estimator

$$\sqrt{n}(m_n - \theta_0) \overset{P_{\theta_0 + h/\sqrt{n}}}{\rightsquigarrow} N(h, \pi/2) \implies \sqrt{n}\left(m_n - \left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) \overset{P_{\theta_0 + h/\sqrt{n}}}{\rightsquigarrow} N(0, \pi/2)$$

9.5 Calculating Influence Functions

Each of these examples are taken from Chapter 20 in van der Vaart.

Example 9.16 (Mean functional)**.** *Suppose the sample mean $\psi(P_n)$ is the plug-in estimator of the mean functional $\psi(P) = \int x dP(x)$. By the von-Mises expansion, the influence function is*

$$\begin{aligned}\psi'_P(\delta_x - P) &= \frac{d}{d\epsilon}\int x\, d[(1-\epsilon)P + \epsilon\delta_x](x)\Big|_{\epsilon=0} \\ &= x - \int x dP(x)\end{aligned}$$

Example 9.17 (Wilcoxon Mann-Whitney Statistic)**.** *Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are random sample from a bivariate distribution with empirical distributions F_n and G_n for each margin. The Mann-Whitney Statistic is a plug-in estimator of the functional $\psi_0 = \psi(P, G) = \int F dG$:*

$$\psi(P_n, G_n) = \int F_n dG_n = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{I}(X_i \le Y_j)$$

The influence function of the Mann-Whitney statistic can also be calculated from the von-Mises expansion

$$\begin{aligned}\psi'_P(\delta_x - F, \delta_y - G) &= \frac{d}{d\epsilon}\int ((1-\epsilon)F + \epsilon\delta_x) d\left[(1-\epsilon)G + \epsilon\delta_y\right]\Big|_{\epsilon=0} \\ &= \frac{d}{d\epsilon}\int (1-\epsilon)^2 F dG + \int (1-\epsilon)\epsilon F d\delta_y + \int \epsilon(1-\epsilon)\delta_x dG + \int \epsilon^2 \delta_x d\delta_y\Big|_{\epsilon=0} \\ &= F(y) + 1 - G_-(x) - 2\int F dG\end{aligned}$$

Example 9.18 (Z estimators)**.** *The Z-estimator $\psi(P_0)$ is the solution to the population-based estimating equation $P_0 z_{\psi(P_0)} = 0$. Differentiating with respect to ϵ across the identity*

$$0 = (P + \epsilon\delta_x) z_{\psi(P+\epsilon\delta_x)} = P z_{\psi(P+\epsilon\delta_x)} + \epsilon z_{\psi(P+\epsilon\delta_x)}(x)$$

Assumin the derivatives exist and z_ψ is continuous, we have that

$$0 = \left(\frac{\partial}{\partial\theta} P z_\theta\right)_{\theta=\psi(P)} \left[\frac{d}{dt}\psi(P + t\delta_x)\right]_{t=0} + z_{\psi(P)}(x)$$

Where the expression in parentheses is the influence function and is given by

$$-\left(\frac{\partial}{\partial\theta} P z_\theta\right)^{-1}_{\theta=\psi(P)} z_{\psi(P)}(x)$$

Example 9.19 (Quantiles)**.** *The pth qauntile of distribution function F is $\psi(F) = F^{-1}(p)$. We set $F_\epsilon = (1-\epsilon)F + \epsilon\delta_x$ and differentiate wrt ϵ the identity*

$$p = F_\epsilon F_\epsilon^{-1}(p) = (1-t)F(F_\epsilon^{-1}(p)) + \epsilon\delta_x(F_t^{-1}(p))$$

We find that

$$0 = -F(F^{-1}(p)) + f(F^{-1}(p))\left[\frac{d}{d\epsilon}F_\epsilon^{-1}(p)\right]_{t=0} + \delta_x(F^{-1}(p))$$

Where the influence function is given by

$$\left[\frac{d}{d\epsilon}F_\epsilon^{-1}(p)\right]_{t=0} = \psi'_F(\delta_x - F) = -\frac{\mathbb{I}(x \le F^{-1}(p)) - p}{f(F^{-1}(p))}$$

This implies that the sequence of empirical quantiles is asymptotically normal

$$\sqrt{n}(F_n^{-1}(t) - F^{-1}(t)) \rightsquigarrow N\left(0, P_0\left[-\frac{\mathbb{I}(x \le F^{-1}(p)) - p}{f(F^{-1}(p))}\right]^2\right) \equiv N\left(0, \frac{p(1-p)}{f(F^{-1}(p))^2}\right)$$

Example 9.20 (Cramer-von Mises statistic: higher order expansion)**.** *The Cramer-von Mises statistic $\psi(F_n)$ estimates the following parameter $\psi(F) = \int (F - F_0)^2 dF_0$ for some fixed F_0. The von-Mises expansion yields*

$$\psi(F + \epsilon H) = \int (F + \epsilon H - F_0)^2 dF_0 = \int (F - F_0)^2 dF_0 + 2\epsilon \int (F - F_0) H dF_0 + \epsilon^2 \int H^2 dF_0$$

The first derivative is

$$\frac{\partial}{\partial \epsilon} \psi(F + \epsilon H)\Big|_{\epsilon=0} = 2 \int (F - F_0) H dF_0$$

Plugging in $\epsilon = 1/\sqrt{n}$ and $H = \mathbb{G}_n = \sqrt{n}(F - F_0)$, we have

$$\psi'_{F_0}(H) = \frac{\partial}{\partial \epsilon} \psi(F_0)\Big|_{\epsilon=0} = 2 \int (F_0 - F_0) H dF_0 = 0$$

Therefore, first order expansion is degenerate. To determine the asymptotic distribution, we must go to the second order derivative

$$\psi''_{F_0}(H) = \frac{\partial^2}{\partial \epsilon^2} \psi(F + \epsilon H)\Big|_{\epsilon=0} = 2 \int H^2 dF_0$$

Which for $\epsilon = 1/\sqrt{n}$ and $H = \mathbb{G}_n = \sqrt{n}(F - F_0)$ produces

$$\psi''_{F_0}(\mathbb{G}_n) = 2 \int \mathbb{G}_n dF_0$$

The von Mises expansion suggests the following approximation

$$\begin{aligned} \psi(F_n) - \psi(F) &= \frac{1}{\sqrt{n}} \psi'_{F_0}(\mathbb{G}_n) + \frac{1}{2!} \frac{1}{n^{2/2}} \psi''_{F_0}(\mathbb{G}_n) + \dots \quad (\psi'_{F_0} \to 0) \\ &\approx \frac{1}{n} \int \mathbb{G}_n dF_0 \end{aligned}$$

9.6 Semiparametric/Nonparametric Inference

9.6.1 Function-valued parameters

Example 9.21 (Uniform confidence bands for CDF)**.** *Suppose our goal is to construct confidence bands for the CDF $F_0(t)$ uniformly over all $t \in \mathbb{R}$. We estimate $F_0(t)$ with the class of functions $\mathcal{F} := \{x \to \mathbb{I}(x \le t) : t \in \mathbb{R}\}$. By Donsker's Theorem and the continuous mapping theorem,*

$$\begin{aligned} \mathbb{G}_n &\rightsquigarrow \mathbb{G} \text{ in } \ell^\infty(\mathcal{F}) \\ ||\mathbb{G}_n||_{\mathcal{F}} &\rightsquigarrow ||\mathbb{G}||_{\mathcal{F}} \end{aligned}$$

Our goal of constructing valid confidence bands is equivalent to finding $\{L_n(t), U_n(t)\}$ such that

$$\lim_{n \to \infty} P(L_n(t) \le F_0(t) \le U_n(t)) \ge 1 - \alpha \quad \forall\, t \in \mathbb{R}$$

We propose the following bounds where c is the $(1-\alpha)$-quantile of $||\mathbb{G}||_{\mathcal{F}}$

$$L_n(t) := F_n(t) - \frac{c}{\sqrt{n}} \qquad U_n(t) := F_n(t) - \frac{c}{\sqrt{n}}$$

These bounds are asymptotically valid because

$$\begin{aligned}
&\lim_{n\to\infty} P_0(L_n(t) \le F_0(t) \le U_n(t)) \qquad \forall t \in \mathbb{R} \\
&= \lim_{n\to\infty} P_0\left(F_n(t) - \frac{c}{\sqrt{n}} \le F_0(t) \le F_n(t) - \frac{c}{\sqrt{n}}\right) \qquad \forall t \in \mathbb{R} \\
&= \lim_{n\to\infty} P_0\left(-c \le \sqrt{n}(F_0(t) - F_n(t)) \le c\right) \qquad \forall t \in \mathbb{R} \\
&= \lim_{n\to\infty} P_0\left(\sqrt{n}|F_n(t) - F_0(t)| \le c\right) \qquad \forall t \in \mathbb{R} \\
&= \lim_{n\to\infty} P_0\left(\sup_t \sqrt{n}|F_n(t) - F_0(t)| \le c\right) \\
&= \lim_{n\to\infty} P_0\left(\sup_{f\in\mathcal{F}} \sqrt{n}|(P_n - P_0)h| \le c\right) \\
&= \lim_{n\to\infty} P_0\left(||\mathbb{G}_n||_{\mathcal{F}} \le c\right) \\
&= P_0\left(||\mathbb{G}||_{\mathcal{F}} \le c\right) \\
&= (1-\alpha)
\end{aligned}$$

9.6.2 Proving Asymptotic Linearity and Delta Method for ALEs

Example 9.22 (Coefficient of Variation is ALE)**.** *Let $C_n := \frac{\sigma_n}{\mu_n}$ be the plug-in estimator for $C_0 := \frac{\sigma_0}{\mu_0}$. Let $h(u,v) = u^{1/2}v^{-1}$ and let $C_0 := h(\sigma_0^2, \mu_0), C_n := h(\sigma_n^2, \mu_n)$. We know that:*

$$\begin{pmatrix}\sigma_n^2 \\ \mu_n\end{pmatrix} - \begin{pmatrix}\sigma_0^2 \\ \mu_0\end{pmatrix} = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix}(X_i-\mu_0)^2 - \sigma_0^2 \\ (X_i - \mu_0)\end{pmatrix} + o_P(n^{-1/2})$$

By the Delta Method for ALEs/Influence Functions, we have

$$\begin{aligned}
C_n - C_0 = h(\sigma_n^2, \mu_n) - h(\sigma_0^2, \mu_0) &= \frac{1}{n}\sum_{i=1}^n \left\langle \nabla h(\sigma_0^2, \mu_0)^T, \begin{pmatrix}(X_i-\mu_0)^2 - \sigma_0^2 \\ (X_i - \mu_0)\end{pmatrix} \right\rangle + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^n \left\langle \left(\frac{1}{2\sigma_0\mu_0}, -\frac{\sigma_0}{\mu_0^2}\right)^T \begin{pmatrix}(X_i-\mu_0)^2 - \sigma_0^2 \\ (X_i - \mu_0)\end{pmatrix} \right\rangle + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{(X_i-\mu_0)^2 - \sigma_0^2}{2\mu_0\sigma_0} - \frac{\sigma_0(X_i-\mu_0)}{\mu_0^2} + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\mu_0(X_i-\mu_0)^2 - \mu_0\sigma_0^2 - 2\sigma_0^2(X_i-\mu_0)}{2\mu_0^2\sigma_0} + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^n C_0\left[\frac{\mu_0^2(X_i-\mu_0)^2 - \mu_0^2\sigma_0^2 - 2\mu_0\sigma_0^2(X_i-\mu_0)}{2\mu_0^2\sigma_0^2}\right] + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^n C_0\left[\frac{1}{2}\left(\frac{X_i-\mu_0}{\sigma_0}\right)^2 - \frac{X_i}{\mu_0} + \frac{1}{2}\right] + o_P(n^{-1/2})
\end{aligned}$$

So C_n is asymptotically linear with influence function $\phi_{P_0}(x) := C_0\left[\frac{1}{2}\left(\frac{x-\mu_0}{\sigma_0}\right)^2 - \frac{x}{\mu_0} + \frac{1}{2}\right]$

Example 9.23 (Average Absolute Deviation from Mean is ALE)**.** *Suppose we wish to infer about $\psi_0 := \int P_0|x-\mu_0|$ for $\mu_0 = \int x dP_0(x)$. Consider the plug-in estimator:*

$$\psi_n := \frac{1}{n}\sum_{i=1}^n |X_i - \bar{X}_n|$$

Noting that $\psi_n = P_n f_n$ and $\psi_0 = P_0 f_0$ for $f_n(x) = |x - \bar{X}_n|$ and $f_0(x) = |x - \mu_0|$, we write the following expansion

$$\psi_n - \psi_0 = (P_n - P_0)f_0 + P_0(f_n - f_0) + (P_n - P_0)(f_n - f_0)$$

Where term 1 is linear. The other two terms require further inspection. Let's study term 2. Letting $h(u) : u \to \int |x - u| dP_0(x)$, we have that

$$P_0(f_n - f_0) = h(\bar{X}_n) - h(\mu_0) = h'(\mu_0)(\bar{X}_n - \mu_0) + o_P(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n} h'(\mu_0)(X_i - \mu_0) + o_P(n^{-1/2})$$

By the delta method. Let $F_0(u) := \int \mathbb{I}(x < u) dP_0(x)$ and $G_0(u) = \int \mathbb{I}(x < u) x dP_0(x)$. Then $h(u)$ is given by

$$\begin{aligned}
h(u) &= \int |x - u| dP_0(x) = \int (u - x)\mathbb{I}(x < u) dP_0(x) + \int (x - u)\mathbb{I}(x > u) dP_0(x) \\
&= \int (u - x)\mathbb{I}(x < u) dP_0(x) + \int (x - u)(-\mathbb{I}(x < u) + 1) dP_0(x) \\
&= uF_0(u) - G_0(u) + \int (u - x)\mathbb{I}(x < u) dP_0(x) + \int (x - u) dP_0(x) \\
&= 2uF_0(u) - 2G_0(u) - u + \mu_0 \\
&= u[2F_0(u) - 1] + [\mu_0 - 2G_0(u)]
\end{aligned}$$

Therefore

$$h'(u) = 2F_0(u) - 1 \implies h'(\mu_0) = 2F_0(\mu_0) - 1$$

Now we study term 3. Note that $|(f_n - f_0)| = ||x - \bar{X}_n| - |x - \mu_0|| \leq |\mu_0 - \bar{X}_n|$. Therefore, the total variational norm of $(f_n - f_0) \leq 2|\mu_0 - \bar{X}_n|$. The WLLN says that there will exist a constant $K < \infty$ such that $|\mu_0 - \bar{X}_n| < K$ w.p. 1. Thus, the function class is bounded in total variation and is therefore Donsker. Therefore,

$$(P_n - P_0)(f_n - f_0) = o_P(n^{-1/2})$$

The result is that

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} [|X_i - \mu_0| - \psi_0 + [2F_0(\mu_0) - 1](X_i - \mu_0)] + o_P(n^{-1/2})$$

Example 9.24 (IPW Estimator is ALE)**.** *Suppose $X = (Y, \Delta, W)$ with Y the outcome of interest only observed when $\Delta = 1$. W are covariates. Suppose we wish to infer about the mean of Y. If the missingness mechanism only depends on W (MAR), the mean outcome is*

$$\psi_0 = E_0[E_0(Y|\Delta = 1, W)]$$

Let $\tilde{Q}_0(w) := E_0(Y|\Delta = 1, W = w), g_0(w) := P_0(\Delta = 1|W = w), Q_{W,0}(w) := P_0(W \leq w)$. We can now write

$$\psi_0 = E_0[\tilde{Q}_0(W)] = E_0\left[E_0\left[\frac{\Delta Y}{g_0(W)}|Y\right]\right]$$

Case 1: If g_0 is known, this motivates the following plug-in estimator:

$$\psi_{0,n} := \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i Y_i}{g_0(W_i)} = P_n f_0$$

Which is linear with influence function $\phi_{P_0}(x) : x \to \frac{\delta y}{g_0(w)} - \psi_0$.
Case 2: If the missingness probability is unknown, but is known to lie in a parametric model $\{g_\theta : \theta \in \Theta\}$ *with* $g_0 = g_{\theta_0}$. *Suppose we have an ALE* θ_n *for* θ_0 *with influence function* φ_{θ_0}. *Letting* $g_n := g_{\theta_n}$ *and* $f_n(x) := \frac{\delta y}{g_n(w)}$, *we can consider the new plug-in estimator*

$$\psi_n := \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_i Y_i}{g_n(W_i)} = P_n f_n$$

To show this estimator is asymptotically linear, examine the expansion

$$\psi_n - \psi_0 = (P_n - P_0)f_0 + P_0(f_n - f_0) + (P_n - P_0)(f_n - f_0)$$

Study term 2. First note that $g_n(w) - g_0(w) = \frac{\partial}{\partial\theta} g_\theta(w)\Big|_{\theta=\theta_0}(\theta_n - \theta) + o_P(n^{-1/2})$ *by Taylor expansion. Now Term 2 takes form*

$$\begin{aligned}
P_0(f_n - f_0) &= \int \tilde{Q}_0(w,1) g_0(w) \left[\frac{1}{g_n(w)} - \frac{1}{g_0(w)}\right] Q_{W,0}(dw) \\
&= -\int \tilde{Q}_0(w,1) \frac{1}{g_0(w)} \left[g_n(w) - g_0(w)\right] Q_{W,0}(dw) + o_P(n^{-1/2}) \\
&= -\int \tilde{Q}_0(w,1) \frac{1}{g_0(w)} \left[\frac{\partial}{\partial\theta} g_\theta(w)\Big|_{\theta=\theta_0}(\theta_n - \theta)\right] Q_{W,0}(dw) + o_P(n^{-1/2}) \\
&= -\int \tilde{Q}_0(w,1) \frac{1}{g_0(w)} \left[\frac{\partial}{\partial\theta} g_\theta(w)\Big|_{\theta=\theta_0} \left(\frac{1}{n}\sum_{i=1}^{n} \varphi_{\theta_0}(X_i)\right)\right] Q_{W,0}(dw) + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \gamma_0 \varphi_{\theta_0}(X_i) + o_P(n^{-1/2})
\end{aligned}$$

For $\gamma_0 = -\int \tilde{Q}_0(w,1) \frac{1}{g_0(w)} \frac{\partial}{\partial\theta} g_\theta(w)\Big|_{\theta=\theta_0} Q_{W,0}(dw)$.
Term 3 requires that $(f_n - f_0)$ *falls in a Donsker class with probability approaching 1.*
Under this condition ψ_n *is asymptotically linear with influence function*

$$\phi^*_{P_0}(x) := \phi_{P_0}(x) + \gamma_0 \varphi_{\theta_0}(X_i)$$

Thus if θ_n *is an asymptotically linear (and parametric efficient) estimator of* θ_0, *we can obtain smaller variance than the Case 1 estimator!*

Example 9.25 (Absolute Mean Difference and Gini Index is ALE (Theory Exam 2020))**.** *Suppose* $X_1, \ldots, X_n \overset{iid}{\sim} P_0 \in M$ *where* M *is the set of all nonparametric distributions on* $(0, \infty)$ *with finite second moment. Define* $\delta_0 := \Delta(P_0)$ *to be the population mean absolute difference parameter for* X_1, X_2 *independent draws from* P

$$P \to \Delta(P) := E_P|X_1 - X_2|$$

Step 1: Plug-in Est ALE*: First, we show* $\Delta(P_n)$ *is an ALE and derive its distribution. We start with the associated V-statistic*

$$\Delta(P) = \int\int |x_1 - x_2| dP(x_1) dP(x_2)$$

$$V_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{n} |X_i - X_j|$$

By a linearization argument, we can show that the expansion of $V_n - \Delta(P_0)$ *is dominated by the following term*

$$\begin{aligned} &2(P_n - P_0)\int |x-u|du \\ &\implies V_n - \Delta(P_0) = \frac{1}{n}\sum_{i=1}^{n} 2\left(\int |X_i - u|dP_0(u) - \Delta(P_0)\right) + o_P(n^{-1/2}) \end{aligned}$$

However, the V-statistic is not exactly $\Delta(P_n)$ *because the statistic is defined in terms of independent* X_1, X_2*, and we have some repeated indices. Instead, we define* $\Delta(P_n)$ *as the U-statistic*

$$U_n = \Delta(P_n) = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j<i}|X_i - X_j| \equiv \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}|X_i - X_j|$$

By a similar linearization argument and Hajek projection we can show $U_n - V_n = O(n^{-1})$ *implying*

$$\begin{aligned} U_n - \Delta(P_0) &= (V_n - \Delta(P_0)) + (U_n - V_n) \\ &= 2(P_n - P_0)\int |x-u|dP_0(u) + o_p(n^{-1/2}) \end{aligned}$$

Implies $U_n = \Delta(P_n)$ *is ALE with influence function*

$$\phi(x) = 2\left(\int |x-u|dP_0(u) - \Delta(P_0)\right)$$

Therefore,

$$\begin{aligned} \sqrt{n}(\Delta(P_n) - \Delta(P_n)) &\rightsquigarrow N\left(0, 4\left[\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[|X-U||X]^2 - 2\mathbb{E}_{P_0}[|X-U||X]\delta_0 + \delta_0^2]\right]\right) \\ &\equiv N\left(0, 4\left[\mathbb{E}_{P_0}[\mathbb{E}_{P_0}[|X-U||X]^2] - \delta_0^2\right]\right) \end{aligned}$$

Step 2: Show Gini Index is ALE*: let* $\psi_0 := \Psi(P_0)$ *be the Gini Index functional*

$$P \to \Psi(P) := \frac{E_P|X_1 - X_2|}{2E_P(X)}$$

Show plug-in Gini index $\Psi(P_n)$ *is asymptotically linear. We use the delta method for ALEs to accomplish this. Define* $f(a,b) = \frac{a}{b}$*. We can write*

$$\Psi(P_n) = f(\Delta(P_n), P_n(X)) = \frac{\Delta(P_n)}{P_n X}$$

Which is asymptotically linear by the delta method for ALEs. The influence function is as follows

$$\begin{aligned} \tilde{\phi}(x) &= \langle \nabla f(\delta_0, P_0(X)), \phi_{P_0}\rangle \\ &= \frac{2\left(\int |x-u|dP_0(u) - \Delta(P_0)\right)}{P_0(X)} - \frac{\delta_0(x - P_0(X))}{P_0(X)^2} \\ &= 2\left[\frac{E_{P_0}|X-x| - (x+\mu_0)\psi_0}{\mu_0}\right] \end{aligned}$$

Step 3 (Derive Tangent Space)*: Suppose the true mean outcome is known.* M *is known such that* $M_0 := \{P \in M : E_P(X) = \mu_0\}$*. To derive the tangent space of a model under a moment restriction we take the following steps.*

(a) Propose a submodel: I elect the linear one for simplicity $p_\theta = [1 + \theta h]p_0(x)$

(b) Write the moment restriction: $g_0(X) := X - \mu_0$ *such that* $E_P(g_0) = 0$

(c) Evaluate the scores h that allow us to remain in our model M

$$\begin{aligned} P_\theta(g_0) = 0 \\ \implies \int g_0(1+\theta h)p_0(x)dx = \int (x-\mu_0)(1+\theta h)p_0(x)dx = 0 \\ \implies \int xp_0(x)dx + \theta\int xhp_0(x)dx - \mu_0\int (1+\theta h)p_0(x)dx = 0 \\ \implies \mu_0 + \theta\int xhp_0(x)dx - \mu_0 = 0 \\ \implies P_0(Xh(X)) = 0 \end{aligned}$$

Thus, the tangent space takes the form $T_{M_0}(P_0) := \{h \in L_0^2(P_0) : \int xh(x)dP_0(x) = 0\}$ at P_0.

Example 9.26 (Robust Mean is ALE (P6 Theory Exam 2021))**.** *Let $X_1, \dots, X_n \overset{iid}{} P_0 \in M$ where M is the nonparametric model with finite second moment and strictly positive density on the nonnegative real numbers. We wish to estimate $\psi_0 = \psi(F_0)$*

$$\psi(F) := \mathbb{E}_F[X\mathbb{I}(X \le Q_\beta(F))] = \int_0^{Q_\beta(F)} udF(u)$$

Where $Q_\beta(F)$ is the β-quantile of F. Let $\mu_0 := \mu(F_0), \mu_n := \mu(F_n)$ and $q_0 := Q_\beta(F_0)$. Also note that the Gateaux derivative of Q_β at F in direction h is given by $\dot{Q}_\beta(F;h) = \frac{-h(Q_\beta(F))}{f(Q_\beta(F))}$.

(a) ***Calculate Gauteaux derivative of*** *ψ. Using the fundamental theorem of calculus and product rule, the Gauteaux derivative is defined as*

$$\begin{aligned} \dot{\psi}(F;h) &= \frac{d}{d\epsilon}\psi(F+\epsilon h)\big|_{\epsilon=0} \\ &= \frac{d}{d\epsilon}\left[\int_0^{Q_\beta(F+\epsilon h)} udF(u) + \epsilon\int_0^{Q_\beta(F+\epsilon h)} udh(u)\right]\Big|_{\epsilon=0} \\ &= \frac{d}{d\epsilon}\left[\int_0^{Q_\beta(F+\epsilon h)} uf(u)du + \epsilon\int_0^{Q_\beta(F+\epsilon h)} udh(u)\right]\Big|_{\epsilon=0} \\ &= Q_\beta(F+\epsilon h)f(Q_\beta(F+\epsilon h))\cdot\dot{Q}_\beta(F;h) + \int_0^{Q_\beta(F+\epsilon h)} udh(u) + \epsilon\left(Q_\beta(F+\epsilon h)\dot{Q}_\beta(F;h)\right)\Big|_{\epsilon=0} \\ &= Q_\beta(F+\epsilon h)\left(f(Q_\beta(F+\epsilon h))\right)\left(\frac{-h(Q_\beta(F))}{f(Q_\beta(F))}\right) + \int_0^{Q_\beta(F+\epsilon h)} udh(u) + \epsilon\left(Q_\beta(F+\epsilon h)\dot{Q}_\beta(F;h)\right)\Big|_{\epsilon=0} \\ &= \int_0^{Q_\beta(F)} udh(u) - Q_\beta(F)h(Q_\beta(F)) \end{aligned}$$

(b) ***Asymptotic Linearity and Influence Function****. By the Functional Delta method, we know that $\psi_n = \psi(F_n)$ is asymptotically linear with influence function equal to the Gauteax derivative (under Hadamard differentiability wrt supremum norm):*

$$\psi(F_n) - \psi(F_0) = \frac{1}{n}\sum_{i=1}^n \dot{\psi}(F_0; \mathbb{I}(X_i \le \cdot) - F_0) + o_P(n^{-1/2})$$

To calculate the influence function, we look to part (a). Recalling $q_0 = Q_\beta(F_0)$:

$$\begin{aligned}\dot{\psi}(F_0; \mathbb{I}(X_i \leq \cdot) - F_0) &= \int_0^{Q_\beta(F_0)} u\ (\mathbb{I}(X_i \leq \cdot) - F_0)(du) - Q_\beta(F_0)\left[\mathbb{I}(X_i \leq Q_\beta(F_0)) - F_0(Q_\beta(F))\right] \\ &= \int_0^{q_0} u\ (\mathbb{I}(X_i \leq \cdot) - F_0)(du) - q_0\left[\mathbb{I}(X_i \leq q_0) - \beta\right] \\ &= \int_0^{q_0} u\ d(\mathbb{I}(X_i \leq u)) - \psi_0 - q_0\mathbb{I}(X_i \leq q_0) + \beta q_0 \\ &= (x - q_0)\mathbb{I}(X_i \leq q_0) - \psi_0 + \beta q_0\end{aligned}$$

(c) Show that $\sqrt{n}(\mu_n - \mu_0)$ where μ_n is the sample mean. We solve for the variance of X using the law of total variance with $A_1 = \mathbb{I}(x \leq q_0), A_2 = \mathbb{I}(x > q_0)$ which partition the outcome space.

$$\begin{aligned}Var(X) &= \sum_{i=1}^{2} P(A_i) \cdot Var(X|A_i) + \left[\sum_{i=1}^{2} \mathbb{E}[X|A_i]^2[1 - P(A_i)][P(A_i)]\right] - 2\mathbb{E}[X|A_1]P(A_1)\mathbb{E}[X|A_2]P(A_2) \\ &= \beta\, Var(X|X \leq q_0) + (1-\beta)\, Var(X|X > q_0) + \beta(1-\beta)\left(\mathbb{E}[X|X \leq q_0] - \mathbb{E}[X|X > q_0]\right)^2\end{aligned}$$

(d) Compare the asymptotic variances of $\sqrt{n}(\psi_n - \psi_0)$ to $\sqrt{n}(\mu_n - \mu_0)$ via the asymptotic relative efficiency: comparing the squares of each of the influence functions:

$$\begin{aligned}&\frac{\mathbb{E}\left(\left[(x - q_0)\mathbb{I}(x \leq q_0)\right]^2 - 2(x - q_0)\mathbb{I}(x \leq q_0)(\psi_0 - \beta q_0) + (\psi_0 - \beta q_0)^2\right)}{\mathbb{E}\left(\left[x - \mu_0\right]^2\right)} \\ &= \frac{\mathbb{E}\left(\{x\mathbb{I}(x \leq q_0) - \psi_0 - (q_0\mathbb{I}(x \leq q_0) - \beta q_0)\}^2\right)}{\mathbb{E}\left(\left[x - \mu_0\right]^2\right)} \\ &= \frac{\mathbb{E}\left(\{x\mathbb{I}(x \leq q_0)\ - \psi_0\}^2 - 2\{x\mathbb{I}(x \leq q_0)\ - \psi_0\}\{q_0\mathbb{I}(x \leq q_0) - \beta q_0)\} + \{q_0\mathbb{I}(x \leq q_0) - \beta q_0)\}^2\right)}{\mathbb{E}\left(\left[x - \mu_0\right]^2\right)} \\ &= \frac{Var(X|X \leq q_0) - 2\psi_0 q_0 + 2\psi_0\beta q_0 + 2\psi_0\beta q_0 - 2\psi_0\beta q_0 + q_0^2\beta - 2\beta^2 q_0^2 + \beta^2 q_0^2}{Var(X)} \\ &= \frac{Var(X|X \leq q_0) - 2\psi_0 q_0(1-\beta) + q_0^2\beta(1-\beta)}{Var(X)} = \frac{Var(X|X \leq q_0) - q_0(1-\beta)(2\psi_0 - q_0\beta)}{Var(X)}\end{aligned}$$

Thus, as long as $(2\psi - q_0\beta) > 0$ then we are assured a reduction in variance compared to the sample mean. This makes sense because the influence function of the trimmed mean is bounded and therefore is robust to outliers.

9.7 Efficient Estimators

Example 9.27 (Efficient Estimators Under Moment Restriction (P7 Theory Exam 2021))**.** *Suppose $X_1, \ldots, X_n \overset{iid}{\sim} P_0 \in M$ where M is the nonparametric model of each distribution P satisfying $Pf_0^2 < \infty$ with support in $(-B, +B)$ for a fixed f_0. Suppose we wish to estimate the mean of f_0: $\psi_0 = P_0 f_0$.*

Consider a multivariate $g_0 : \mathbb{R}^m \to \mathbb{R}$ that is bounded and consider the model containing the collection of distributions with the moment restriction based on the multivariate function $M_0 := \{P \in M : P^m g_0 = 0\}$.

The tangent space of M_0 at P is given by

$$T_{M_0}(P) := \{h \in L_0^2(P) : \int h(x)\bar{g}_P(X)dP(x) = 0\}$$

Where $\bar{g}_P = P^{m-1}g_0$.

(a) ***Derive form of projection onto tangent space***. *Consider an arbitrary element $s^* \in L_0^2(P)$. The projection $\Pi(s|T_M)$ onto the tangent space satisfies the following property.*

$$\langle s - \Pi(s|T_M), a\bar{g}_p(X)\rangle = 0$$

This property ensures that the set of allowable scores $s - \Pi(s|T_M)$ satisfies the desired moment restriction. Note that the model space is a linear span, so $\Pi(s|T_M)$ is given by

$$\begin{aligned} a &= \underset{a}{argmin}||s - a\bar{g}_p||^2_{L^2(P)} \\ \implies &\frac{\partial}{\partial a}\left[P(s^2) - 2aP(s\bar{g}_p) + a^2P(\bar{g}_p^2)\right] = 0 \\ \implies a^* &= \frac{\int s\bar{g}_P dP}{\int \bar{g}_P^2 dP} \end{aligned}$$

Taken together implying the form of the projection onto the tangent space is

$$\begin{aligned} s^* &= s(x) - a^*\bar{g}_p \\ &= s(x) - \frac{\int s\bar{g}_P dP}{\int \bar{g}_P^2 dP}\bar{g}_p \end{aligned}$$

The last condition we need to check is that s^* actually lives in the tangent space. To verify this

$$\begin{aligned} \int s^*(x)\bar{g}_P(x)dP(x) &= \int \left(s(x) - \frac{\int s\bar{g}_P dP}{\int \bar{g}_P^2 dP}\bar{g}_P\right)\bar{g}_P(x)dP(x) \\ &= \int (s\bar{g}_P(x) - s\bar{g}_P)dP = 0 \end{aligned}$$

(b) ***Canonical gradient***: *Note that ψ_0 has nonparametric influence function $\phi(x) = f_0(x) - \psi_0$ where $\mu_0 := \mathbb{E}[X]$. Using the fact above, we have that the canonical gradient/EIF is obtained by projecting $\phi(x)$ onto T_M.*

$$\begin{aligned} \phi^*(x) &= f_0(x) - \psi_0 - \frac{\int (f_0(x) - \psi_0)\bar{g}_P dP}{\int \bar{g}_P^2(x)dP}\bar{g}_p(x) \\ &= f_0(x) - \psi_0 - \frac{\int f_0(x)\bar{g}_P dP - \cancelto{0}{\int \psi_0 \bar{g}_P dP}}{\int \bar{g}_P^2(x)dP}\bar{g}_p(x) \\ &= f_0(x) - \psi_0 - \frac{\int f_0(x)\bar{g}_P dP}{\int \bar{g}_P^2(x)dP}\bar{g}_p(x) \end{aligned}$$

Where the cancellation occurred because $\int \bar{g}_P dP = 0$ in M_0.

(c) ***Efficient One-Step Estimator***: *an asymptotically efficient estimator in M_0 can be obtained by taking the plug-in estimator and adding the empirical mean of the EIF.*

$$\begin{aligned} \psi^*(P_n) &= \psi(P_n) + P_n\phi^*(x) \\ &= P_n f_0(X_i) + \cancelto{0}{P_n(f_0(X_i) - \psi(P_n))} - P_n\left(\frac{P_n(f_0(X)P_n^{m-1}g_0)}{P_n(P_n^{m-1}g_0)^2}P_n^{m-1}g_0\right) \\ &= P_n f_0(X_i) - \frac{P_n(f_0(X)\bar{g}_n)}{P_n(\bar{g}_n)^2}P_n(P_n^{m-1}g_0) \end{aligned}$$

Let's inspect the last term. Recognizing g_0 is P_0-mean-zero, by linearization we have:

$$\begin{aligned} P_n^m g_0 &= (P_n^m - P_0^m)g_0 \\ &= m(P_n - P)\bar{g}_P(x) + o_P(n^{-1/2}) = mP_n\bar{g}_P(x) + o_P(n^{-1/2}) \\ \implies P_n\bar{g}_P &= \frac{P_n^m g_0}{m} \end{aligned}$$

Substituting in the original expression, we have an

$$\psi^*(P_n) = P_n f_0(X_i) - \frac{1}{m}\frac{P_n(f_0(X)\bar{g}_n)}{P_n(\bar{g}_n)^2}P_n^m g_0$$

Which is the efficient one-step estimator.

(d) ***Example:*** *If $\mathbb{E}[(X-\mu_0)^3] = 0$, show the sample mean is efficient for the population mean in a population with known variance.*
We know the sample mean P_nX has nonparametric influence function of $\phi(x) = x - \mu_0$. To show it is efficient in M_0, we must show that the second term in the EIF derived above is 0. Define $g_0(x_1, x_2) = \frac{1}{2}(x_1 - x_2) - \sigma^2$ (known) such that $P^2 g_0 = 0$.

$$\begin{aligned}
&\int (x_1-\mu_0)\left[\int \frac{1}{2}(x_1-x_2)^2 - \sigma^2 dP(x_2)\right] dP(x_1) \\
&= \int (x_1-\mu_0)\left[\int \frac{1}{2}((x_1-\mu_0)-(x_2-\mu_0))^2 dP(x_2)\right] dP(x_1) - \sigma^2 \cancelto{0}{\int (x_1-\mu_0 dP(x_1)} \\
&= \frac{1}{2}\int\int (x_1-\mu_0)\left((x_1-\mu_0)^2 - \cancelto{0}{2(x_1-\mu_0)(x_2-\mu_0)} + (x_2-\mu_0)^2\right) dP(x_1)dP(x_2) \\
&= \frac{1}{2}\left(\int (x_1-\mu_0)^3 dP(x_1) + \int \left(\cancelto{0}{\int (x_1-\mu_0) dP_1(x_1)}\right)(x_2-\mu_0)^2 dP(x_2)\right) \\
&= 0
\end{aligned}$$

Where the last step holds because $\mathbb{E}[(X-\mu_0)^3] = 0$. Thus, the nonparametric influence function equals the EIF under this model, so the sample mean is efficient for the population mean.