2nd Year Theory Exam Study Guide, 2025

Jasper Yang August 13, 2025

Contents

1 Useful Math Facts

1.1 Fundamentals

Convexity: Let $I \subseteq \mathbb{R}$ be an interval (finite or infinite; open, closed, or half-open). A function $\phi : I \to \mathbb{R}$ is called *convex* if for all $x, y \in I$ and $t \in [0,1]$, $\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y)$.

- If ϕ is differentiable, then ϕ is convex if and only if ϕ' is increasing.
- If ϕ is twice differentiable, then ϕ is convex if and only if $\phi'' \geq 0$ everywhere.

More generally, a set $S \subseteq \mathbb{R}^n$ is convex if for all $\theta \in [0,1]$ and all $x_1, x_2 \in S$, the point $\theta x_1 + (1-\theta)x_2 \in S$. The **convex hull** of a set of points is the smallest convex set containing all those points.

Fundamental Theorem of Calculus:

- *Part 1:* If f is continuous over an interval $[a,b]$, and we define

$$F(x) = \int_a^x f(t)\,dt,$$

then $F'(x) = f(x)$ for all $x \in [a,b]$. For example, if $g(x) = \int_1^x \frac{1}{t^3+1}\,dt$, then $g'(x) = \frac{1}{x^3+1}$.
- *Part 2:* If f is continuous over $[a,b]$ and F is any antiderivative of f, then

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

Leibniz Integral Rule: If $a(x), b(x)$ are differentiable functions and $f(x,t)$ is smooth enough, then

$$\frac{d}{dx}\left(\int_{a(x)}^{b(x)} f(x,t)\,dt\right) = f(x, b(x)) \cdot b'(x) - f(x, a(x)) \cdot a'(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)\,dt.$$

Note: If the limits of integration do not depend on x, only the third term remains, corresponding to differentiation under the integral sign.

Note: This form is for Riemann integrals!

Continuity:

- **Continuous function:** A function f is continuous at x_0 if for every $\delta > 0$, there exists $\epsilon > 0$ such that $|x - x_0| < \epsilon \Rightarrow |f(x) - f(x_0)| < \delta$.
- **Lipschitz continuity:** A function $f : \mathbb{R}^n \to \mathbb{R}$ is *Lipschitz continuous* if there exists a constant $c > 0$ such that

$$|f(x) - f(y)| \leq c\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n.$$

- **Upper semicontinuity:** A function f is upper semicontinuous at x_0 if for every $\delta > 0$, there exists $\epsilon > 0$ such that $|x - x_0| < \epsilon \Rightarrow f(x) < f(x_0) + \delta$. Informally, the function cannot jump up suddenly.

- **Lower semicontinuity:** A function f is lower semicontinuous if $-f$ is upper semicontinuous.
- **d-continuity for a functional $f : \mathbb{D} \to \mathbb{R}$ in a general metric space $(\mathbb{D}, d)$:** $f : \mathbb{D} \to \mathbb{R}$ is d-continuous at $x_0 \in \mathbb{D}$ if for every sequence $x_1, x_2, ... \in \mathbb{D}$ s.t. $d(x_n, x_0) \to 0$, then $f(x_n) \to f(x)$.

Triangle Inequality: For $x, y \in \mathbb{R}$, $|x + y| \leq |x| + |y|$. This can be used, for example, by adding and subtracting a value z to get:$|x - y| = |(x - z) + (z - y)| \leq |x - z| + |z - y|$.

In $\mathbb{R}^n$, for vectors $a, b \in \mathbb{R}^n$, we have $\|a + b\| \leq \|a\| + \|b\| \quad$ (Euclidean norm). This expands to:

$$\sqrt{\sum_{i=1}^{n}(a_i + b_i)^2} \leq \sqrt{\sum_{i=1}^{n} a_i^2} + \sqrt{\sum_{i=1}^{n} b_i^2}.$$

Reverse Triangle Inequality: For all norms,$\|a - b\| \geq \big|\|a\| - \|b\|\big|$.

Cauchy–Schwarz Inequality: For $a, b \in \mathbb{R}^n$,

$$|\langle a, b\rangle|^2 \leq \|a\|^2\|b\|^2 \quad \Leftrightarrow \quad \left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right).$$

Equality holds if and only if $a = kb$ for some constant $k \in \mathbb{R}$. For random variables X, Y, the inequality becomes:

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}.$$

This tells us that

$$\mathrm{Cov}(X, Y)^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y)$$

Extreme value theorem: A function that is continuous on a compact set attains a maximum and minimum on that set (useful to show existence of MLE for example).

1.2 Linear algebra

Computing a matrix determinant

- det(2 × 2 matrix): $ad - bc$.
- det(3 × 3 matrix): $aei + bfg + cdh - ceg - bdi - afh$.
- det(triangular of diagonal matrix) = product of diagonal entries.

Matrix determinant tricks

- $\det(I_n) = 1$, where I_n is the $n \times n$ identity matrix.
- $\det(AB) = \det(A)\det(B)$ for square matrices A and B of the same size.
- $\det(A^T) = \det(A)$.
- If A is invertible, then $\det(A^{-1}) = \frac{1}{\det(A)}$.
- $\det(cA) = c^n \det(A)$ for scalar c and $n \times n$ matrix A.
- If two rows (or columns) of A are equal, then $\det(A) = 0$.

- If a matrix has a row or column of zeros, then its determinant is 0.
- $\det(A) = 0$ if and only if A is singular (not invertible).
- $\det(A)$ equals the product of the eigenvalues of A (counting multiplicities)

Orthogonal (and normalized) matrices:

- $Q^\top Q = I$ (i.e., $Q^{-1} = Q^\top$)
- Columns (and rows) of Q are orthonormal vectors
- $\|Qx\| = \|x\|$ (length-preserving)
- $\det(Q) = \pm 1$

Matrix derivatives

- $\frac{d}{dx}(a^\top x) = a \quad$ (where a is constant)
- $\frac{d}{dx}(x^\top Ax) = 2Ax \quad$ (if A is symmetric), hence $\frac{d}{dx}(x^\top x) = 2x$
- $\frac{d}{dx}(Ax) = A^\top \quad$ (Jacobian of linear map)
- $\frac{d}{dx}(x^\top A) = A \quad$ (treat x as row).
- Derivative wrt matrix A: $\frac{d}{dA}v^T Av = vv^T$
- For multivariate normal MLE: $\frac{d}{dA}\log(\det(A)) = (A^{-1})^T$

1.3 Tricks

Re-writing a max or min: We have

$$\max(a,b) = a + b - \min(a,b)$$

Derivative of inverse: We have

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

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

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

Taylor expansion: For f infinitely differentiable, the Taylor expansion about x_0 is

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

Expressions for e:

$$e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = \sum_{n=0}^{\infty} \frac{x}{n!}$$

Layer cake representation: For a non-negative RV X, we have

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

Geometric series:

- Finite: If $r \neq 1, \sum_{i=0}^{n} ar^i = a\left(\frac{1-r^{n+1}}{1-r}\right)$
- 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: $\sum_{n=0}^{\infty} p^i e^{inx} = \frac{1}{pe^{ix}-1}$

max-min inequality: $\sup_z \inf_w f(z,w) \leq \inf_w \sup_z f(z,w)$

2 Fundamentals of probability theory

2.1 True basics

Useful properties of probability, CDFs, PDFs:

i. Sub-event inequality: $A \subset B \Rightarrow P(A) \leq P(B)$
ii. Union bound: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
iii. Joint CDF expression: $P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1)$.
iv. Conditional pdf expression: $p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$,
v. Independence: For a jointly discrete or jointly continuous random vector (X, Y), $X \perp Y$ if and only if one of the following equivalent conditions hold
    - $p_{XY}(x,y) = p_X(x)p_Y(y)$, for all $(x,y) \in \Omega_{XY}$;
    - $p_{Y|X}(y|x) = p_Y(y)$, for all $(x,y) \in \Omega_{XY}$;
    - $p_{X|Y}(x|y) = p_X(x)$, for all $(x,y) \in \Omega_{XY}$.
vi. Conditional independence: $X \perp\!\!\!\perp Y|Z \Leftrightarrow P(X \leq x, Y \leq y|Z = z) = P(X \leq x|Z = z)P(Y \leq y|Z = z)$ for every x and y and P_Z-almost everywhere of z.

Bayes' Rule: Let $A_1, ..., A_k$ be a partition of Ω. If $\mathbb{P}(B) > 0$ then, for $i = 1, ..., k$:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{k} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

Bayes' Theorem for RVs: For random variables, we also have the Bayes theorem:

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

Also,

- $p_X(x) = \int p_{X,Y}(x,y)dy$
- $\int p_{X,Y}(x, Y = y_0)dx = \int\int \mathbb{I}[Y = y_0]p_{X,Y}(x,y)dxdy$

Law of total probability: If $B_1, ..., B_k$ partition the sample space Ω, then

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

Law of total expectation (Tower rule): $\mathbb{E}[g(X,Y)] = \mathbb{E}_X[\mathbb{E}_{Y|X}[g(X,Y)|X]]$.

A special case is when X is discrete: $\mathbb{E}[Y] = \sum_i \mathbb{E}[Y|X = x_i] \cdot Pr(X = x_i)$

Law of total variance:

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

Law of total covariance:

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}[\text{Cov}(X,Y)|Z] + \text{Cov}(\mathbb{E}[X|Z], \mathbb{E}[Y|Z]) \\ \text{Cov}(q(X), h(Y)) &= \text{Cov}(q(X), \mathbb{E}[h(Y)|X])\end{aligned}$$

Jensen's inequality

Let φ be a convex function. Then $\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X])$

Moment-generating function: The MGF of a RV X is

$$M_X(t) = \mathbb{E}(e^{tX}).$$

Note that M_X may not exist. However, when it exists, it admits the following interpretation in terms of the moments of RV X. First note that, for $g(t) = e^{tX}$,

$$g(t) = e^{tX} = \sum_{n=0}^{\infty} \frac{g^{(n)}(0)}{n!}t^n = 1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots .$$

Thus,

$$M_X(t) = 1 + t\mu_1 + \frac{t^2\mu_2}{2!} + \frac{t^3\mu_3}{3!} + \cdots ,$$

where $\mu_j = \mathbb{E}(X^j)$ is the j-th moment of X. Therefore,

$$\mathbb{E}(X^j) = M^{(j)}(0) = \left.\frac{d^j M_X(t)}{dt^j}\right|_{t=0},$$

where $M^{(j)}(0)$ is the j-th derivative of $M(t)$ at $t = 0$. Here you see how the moments of X is generated by the function M_X. The MGF uniquely determines the distribution of a random variable. Key properties of MGF:

- Location-scale: $M_{aX+b}(t) = \mathbb{E}(e^{(aX+b)t}) = e^{bt}\mathbb{E}(e^{atX}) = e^{bt}M_X(at)$.
- Multiplicity: $M_{X+Y}(t) = \mathbb{E}(e^{(X+Y)t}) = \mathbb{E}(e^{Xt}e^{Yt})$. Thus, $X \perp Y \Rightarrow M_{X+Y}(t) = \mathbb{E}(e^{Xt}e^{Yt}) = \mathbb{E}(e^{Xt})\mathbb{E}(e^{Yt}) = M_X(t)M_Y(t)$.

The MGF for a random vector $X = (X_1, \cdots, X_d) \in \mathbb{R}^d$ is a function of a d-dimensional argument, $t = (t_1, \cdots, t_d) \in \mathbb{R}^d$:

$$M_X(t) = \mathbb{E}(e^{t^T X}).$$

2.2 Useful properties of parametric distributions

Normal distribution:

1. The sums of normals are normal: For X, Y independent with $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, and $a_1, a_2 \in \mathbb{R}$, then $a_1X + a_2Y \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$.
2. Connection to Chi-square: If $X_1, ..., X_n \overset{iid}{\sim} N(0,1)$, then $Z_1 = X_1^2 \sim \chi_1^2$, $\sum_{i=1}^n X_i^2 \sim \chi_n^2$
3. Stein's Lemma: For $X \sim N(\theta, \sigma^2)$, we have $\mathbb{E}[g(X)(X-\theta)] = \sigma^2\mathbb{E}[g'(X)]$.

Log-normal distribution:

- A positive RV $X \sim \text{Lognormal}(\mu, \sigma^2)$ if $\log X \sim N(\mu, \sigma^2)$.
- Mean: $\exp(\mu + \sigma^2/2)$, Variance: $[\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$.

Multivariate Normal distribution:

1. The density is $f_X(x) = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\det(\Sigma)} \exp\left(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$
2. Linear transformations are MVN: $Y = CX + b \sim N_n(C\mu + b, C\Sigma C^T)$
3. Marginals are normal.
4. Independence $\Leftrightarrow$ uncorrelation: $\text{Cor}(X_i, X_j) = 0 \Leftrightarrow X_i \perp\!\!\!\perp X_j$
5. 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 $Y \sim N(\mu, \Sigma)$ with dimension n, then
$$(Y-\mu)^T\Sigma^{-1}(Y-\mu) \sim \chi_n^2$$
2. Projection property: If $Y \sim N(u, I)$ is an n-dimensional MVN and P is a rank p projection matrix, then
$$(Y-\mu)^T P(Y-\mu) \sim \chi_p^2$$
3. Special case of Gamma: $\chi_m^2 = \text{Gamma}(\frac{m}{2}, 2)$.
4. Sample variance is chi-square under normal model: $X_i \sim N(\mu, \sigma) \Rightarrow (n-1)\frac{S_n^2}{\sigma^2} \sim \chi_{n-1}^2$

Uniform distribution

1. Variance: $(b-a)^2/12$
2. Scaled minimum of uniforms is exponential: If $X_1, ..., X_n \sim \text{unif}(0,1)$, then $U = n \times \min(X_1, ..., X_n) \sim \text{Exp}(1)$.
3. Probability integral transform: If X has a continuous distribution with CDF F_X, then the RV $Y := F_X(X) \sim \text{Unif}(0,1)$.

4. $-\log(\text{Unif}(0,1)) \sim \text{Exp}(1)$.

Poisson distribution

1. Sum of Poissons are Poisson: $X_1, ..., X_n \sim \text{Pois}(\lambda_i) \Rightarrow \sum_{i=1}^n X_i \sim \text{Pois}(\sum_{i=1}^n \lambda_i)$. Can prove this by MGFs (i.e. STAT 512 HW1 Q5).
2. Poisson conditional on the sum of Poissons is multinomial. If $X_1, ..., X_n$ are independent with parameters λ_i, then $X_j | \sum_{i=j}^n X_i \sim \text{Multinomial}(\frac{\lambda_j}{\sum_{i=1}^n \lambda_i})$. (See STAT 512 HW1 Q5 for proof).
3. Connection to negative binomial: NB is a Poisson-Gamma mixture, where $X|\gamma \sim \text{Pois}(\lambda)$ and $\lambda \sim \text{Gamma}(r, p/(1-p))$

Exponential distribution

1. $f(x;\lambda) = \lambda e^{-\lambda x}\mathbb{I}[x \geq 0]$, $F(x;\lambda) = (1 - e^{-\lambda x})\mathbb{I}[x \geq 0]$
2. Memoryless property: If $X \sim \text{Exp}(\lambda)$, then $P(X > x + y | X > x) = P(X > y)$.
3. Absolute difference is exponential: If $X, Y \sim \text{Exp}(1)$, then $|X - Y| \sim \text{Exp}(1)$.
4. Minimum is exponential: $X_1, .., X_n \sim \text{Exp}\lambda$, then $\min(X_1, ..., X_n) \sim \text{Exp}(n\lambda)$.
5. Ratio is uniform: $X_1, ..., X_n \sim \text{Exp}(1)$, then $\frac{X}{X+Y} \sim \text{Unif}(0,1)$
6. $n\text{Exp}(a/n) = \text{Exp}(a)$.

Geometric distribution

1. Memoryless property: If $X \sim \text{Geo}(p)$, then $P(X > x + y | X > x) = P(X > y)$.
2. Connection to NB: Sum of r independent geometric RVs with parameter p is $NB(r,p)$.

Gamma distribution

1. Sum of gammas are gamma: For $X_1, .., X_n \sim \text{Gamma}(\alpha_i, \beta)$, $\sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.
2. Inverse gamma distribution: Has pdf $f(x, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1}$. Mean is $\frac{\beta}{\alpha-1}$ and mode is $\frac{\beta}{\alpha+1}$.
3. Note: $\Gamma(z+1) = z\Gamma(z)$. For an integer n: $\Gamma(n+1) = n!$.
4. $\text{Gamma}(1, a) = \text{Exp}(a)$.

Beta distribution:

1. Mean: $\frac{\alpha}{\alpha+\beta}$. Mode: $\frac{\alpha-1}{\alpha+\beta-2}$. Variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
2. If $X \sim \text{Beta}(\alpha, \beta)$, then $1 - X \sim \text{Beta}(\beta, \alpha)$ and $\mathbb{E}[X^{-1}] = \frac{\alpha+\beta-1}{\alpha-1}$
3. $\text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
4. $\text{Beta}(1, n) \to \text{Exp}(1)$

Multinomial distribution

1. Sum of independent multinomials (with same parameters) are multinomial: If $X \sim M_k(n; p_1, ..., p_k), Y \sim M_k(m; p_1, ..., p_k)$, then $X + Y \sim M_k(n + m; p_1, ..., p_k)$.
2. Block decomposition: If we decompose $X_1, ..., X_k$ into r blocks B_j, they are conditionally independentgiven their block sum S_j:

$$B_j | S_j \sim M_{k_{j-1}+1,...,k_j}\left(S_j; \frac{p_{k_{j-1}+1}}{\sum_{l=k_{j-1}+1}^{k_j} p_l}, ..., \frac{p_{k_j}}{\sum_{l=k_{j-1}+1}^{k_j} p_l}\right)$$

3. Negative correlation between entries: $\mathrm{Cov}(X_i, X_j) = -np_ip_j$.
4. MLE for p_j: optimize under constraint that $\sum_{i=1}^k p_i = 1$ with Lagrange multiplier to get $\hat{p}_j = \frac{X_j}{n}$.
5. Dirichlet prior is conjugate for Bayesian analysis.

2.3 Order statistics

Note: Min and max order statistics for uniform distribution converge to exponential.

Add details to this. Probably how to derive the distributions of min, max, joint from 512.

Distribution of j-th order statistic:

$$p_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} F_X(x_j)^{j-1}(1 - F_X(x_j))^{n-j} p_X(x_j)$$

Joint distribution of $X_{(i)}$ and $X_{(j)}$ for $i < j$:

$$p_{X_{(i)},X_{(j)}}(x_i, x_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(x_i)]^{i-1}p_X(x_i)[F(x_j)-F(x_i)]^{j-i-1}p_X(x_j)[1-F_X(x_j)]^{n-j}$$

Order statistics of Unif$(0, 1)$ are Beta:

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

Spacings between order statistics of Unif$(0, 1)$ are Beta:

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

If the distribution is instead $\mathrm{Unif}(0, \theta_0)$, then we divide by θ_0. So $Y_{(1)}/\theta_0 \sim \mathrm{Beta}(1, n)$.

2.4 Transformations of random variables:

PDF of a transformed RV: Let X be a continuous RV with pdf f_X. Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable and strictly monotonic function with inverse $\gamma = g^{-1}$. Then the pdf of $Y = g(X)$ is

$$p_Y(y) = \begin{cases} \frac{p_X(g^{-1}(y))}{|g'(g^{-1}(y))|} & \text{if } y \in g(\mathbb{R}), \\ 0 & \text{otherwise} \end{cases}$$

Convolution formula: Let X, Y be continuous independent RVs and let $Z = X + Y$. The pdf of Z is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx$$

Jacobian method:
Definition of the Jacobian: Consider $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ and assume that there is a 1-1

and onto mapping (a bijection) $T : \mathbb{R}^n \to \mathbb{R}^n$ for almost all x such that $y = T(x)$. We define the Jacobian matrix

$$J_T(x) = \left(\frac{\partial T(x)}{\partial x}\right) = \left(\frac{\partial y}{\partial x}\right) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

The *Jacobian* is the absolute value of the determinant of this matrix, i.e., $|\det(J_T(x))| = |\left(\frac{\partial y}{\partial x}\right)| = \left|\frac{\partial y}{\partial x}\right|$.

Jacobian method: Assume that $y = T(x)$, where T is 1-1 and onto for almost all x and the Jacobian $\det(J_T(x)) \neq 0$ for all x. Let $A, B \subset \mathbb{R}^n$ be two subsets such that $B = \{T(x) : x \in A\}$. Let f be an integrable function. Then

$$\int_A f(x)dx = \int_B f(T^{-1}(y))\,|\det(J_{T^{-1}}(y))|\,dy = \int_B f(T^{-1}(y))\left|\frac{\partial x}{\partial y}\right| dy.$$

Under the same condition, suppose X is a random variable with a PDF $p_X(x)$ and $Y = T(X)$. Then the PDF of Y is

$$\begin{aligned} p_Y(y) &= p_X(T^{-1}(y))\,|\det(J_{T^{-1}}(y)| \\ &= p_X(T^{-1}(y))\left|\frac{\partial x}{\partial y}\right|. \end{aligned}$$

2.5 Correlation, prediction, and regression

Pearson's correlation: For RVs X and Y, their (Pearson) correlation is

$$\rho_{XY} = \mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

It has three nice properties:

1) *(Symmetric property:)* $\mathrm{Cor}(X, Y) = \mathrm{Cor}(Y, X)$.
2) *(Location-scale property:)* $\mathrm{Cor}(aX + b, cY + d) = \mathrm{sign}(ac)\mathrm{Cor}(X, Y)$.
3) *(Bounded and colinearity property:)* $-1 \leq \mathrm{Cor}(X, Y) \leq 1$. $\mathrm{Cor}(X, Y) = \pm 1$ if and only if they are perfectly linear, i.e., $X = aY + b$ for some constant a, b.

Simple regression: Consider the problem of how to use data on a RV X to predict the value of another RV Y. We evaluate the strength of a predictor $g(X)$ using a loss function $L(g(x), y)$, which is typically MSE $\mathbb{E}[(g(X)-Y)^2]$. The **best predictor** for MSE is $\mathbb{E}[Y|X = x]$. For analysis, we typically decompose Y as

$$Y = \underbrace{\mathbb{E}[Y|X]}_{\text{best predictor}} + \underbrace{(Y - \mathbb{E}[Y|X])}_{\text{residuals}}.$$

Note that $\mathbb{E}$[best predictor] = $\mathbb{E}[Y]$, $\mathbb{E}$[residual] = 0, and Cov(best predictor, residual) = 0. The residual variance satisfies $\mathrm{Var}(Y - \mathbb{E}[Y|X]) = \mathbb{E}[\mathrm{Var}(Y|X)]$, and so by the LOTV, we get

$$\mathrm{Var}(Y) = \underbrace{\mathrm{Var}(\mathbb{E}[Y|X])}_{\text{Var(best predictor)}} + \underbrace{\mathbb{E}[\mathrm{Var}(Y|X)]}_{\text{average Var(residuals)}} .$$

The **best linear predictor (BLP)** is

$$m^*(x) = \alpha^* + \beta^* x = \mathbb{E}[Y] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(x - \mathbb{E}[X]) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X),$$

where $\mu_X = \mathbb{E}[X], \mu_Y = \mathbb{E}[Y], \sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y)$ and ρ_{XY} is the Pearson's correlation. If $X = (X_1, \cdots, X_p)$ is multivariate, then let $\beta = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$, $Z = (1, X_1, \cdots, X_p)^T \in \mathbb{R}^{p+1}$ be a *data vector* and $\gamma = (\alpha, \beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$ be a coefficient vector. Then the MSE has an elegant form:

$$R(\gamma) = R(\alpha, \beta) = \mathbb{E}((Y - \gamma^T Z)^2).$$

Thus, the least squares solution will be $\gamma^* = \mathbb{E}[ZZ^T]^{-1}\mathbb{E}[ZY]$. Note that $\mathbb{E}[ZZ^T]$ is a matrix and $\mathbb{E}[ZZ^T]^{-1}$ is the matrix inverse. With $\gamma^* = (\alpha^*, \beta^*)^T$, we can easily write down the BLP:

$$m^*(x) = \gamma^{*T} z = \alpha^* + \beta^{*T} x = \alpha^* + \sum_{j=1}^{p} \beta_j^* x_j.$$

Classification: In a discrete outcome setting, we often use $0 - 1$ loss: $L(c(x), y) = \mathbb{I}(y \neq c(x))$. We can then write the risk as

$$R(c) = \mathbb{E}[L(c(X), Y)] = \mathbb{E}[\mathbb{E}[L(c(X), Y)|X]]$$

which motivates the **Bayes classifier**

$$c^*(x) = \mathsf{argmax}_{y=0,1} P(y|x) = \begin{cases} 0, & \text{if } P(0|x) \geq P(1|x), \\ 1, & \text{if } P(1|x) > P(0|x). \end{cases}$$

2.6 Concentration inequalities

Concentration inequalities give finite-sample bounds for expressions of the form $P(f(X_1, ..., X_n) \geq t)$. We have learned bounds based on (a) moments, (b) MGFs, and (c) martingales.

2.6.1 Moment-based bounds

Markov's inequality: If $X \geq 0$ a.s. and $\mathbb{E}[X] < \infty$, then

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

. Many inequalities are based on this.

Markov's inequality for functions: Suppose that $\mathbb{E}[X] < \infty, f : [0, \infty) \to [0, \infty)$ is a non-decreasing function for which $f(t) > 0$ for all $t > 0$, and $\mathbb{E}[f(|X - \mathbb{E}[X]|)] < \infty$. Then, it holds for all $t > 0$:

$$P(|X - \mathbb{E}[X]| > t) \leq \frac{\mathbb{E}[f(|X - \mathbb{E}[X]|)]}{f(t)}$$

Chebyshev's inequality: For k in the natural numbers

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

2.6.2 MGF-based bounds

If X has an MGF, we can get sharper tail inequalities:

Chernoff bound: The tail bound depends on the growth rate of the MGF. Suppose X has an MGF in the neighborhood of zero, i.e. $\exists$ a constant $b > 0$ such that $\mathbb{E}[\exp(\lambda X)] < \infty$ for all $\lambda \leq |b|$. Then, for all $t > 0$ and $\lambda \in (0, b]$, it is true that $P(X - \mathbb{E}[X] \geq t) \leq \frac{M_{X-\mu}(\lambda)}{e^{\lambda t}}$. So for all $t > 0$, we have

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

Sub-Gaussian RVs: An RV is called sub-Gaussian (sG) with parameter σ^2 if for all $\lambda \in \mathbb{R}$,

$$\log M_{x-\mu}(\lambda) \leq \frac{\lambda^2 \sigma^2}{2} \qquad \text{(Definition 1)}$$

Equivalently, X is sG if there exists $c > 0$ and $s > 0$ such that

$$P(|X - \mu| \geq t) \leq cP(|sZ| \geq t) \quad \forall \quad t > 0 \qquad \text{(Definition 2)}$$

where $Z \sim N(0, 1)$. This is based on Chernoff's bound applied to a Gaussian RV. The tails of a sG RV are no thicker than those of an $N(0, \sigma^2)$ RV. By Chernoff's inequality, we also get the following tail bounds:

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

Some examples of sG RVs include:

i. Any bounded RV on $[a, b]$ is sG with parameter $\sigma^2 = (b - a)^2/4$.

ii. If two zero-mean independent RVs are sG with parameters σ_1^2, σ_2^2, then $X_1 + X_2$ is sG with parameter $\sigma_1^2 + \sigma_2^2$.

iii. If two (non-independent) RVs X_1, X_2 are sG with parameter σ_1^2 and σ_2^2, then $X_1 + X_2$ is sG with parameter $\sigma_1^2 + \sigma_2^2$.

iv. Suppose $X_1, ..., X_n$ are independent with X_i having mean μ_i and beign sG with parameter σ_i^2. Then

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

Hoeffding's inequality: If the support of an RV $X \sim P$ is bounded in $[a, b]$, then X is sG with parameter $\sigma^2 = \frac{(b-a)^2}{4}$, and

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

More generally, when $X_1, ..., X_n$ are independent RVs with support in $[a, b]$, we can use the fact that $\log M_{\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])}(\lambda) = \sum_{i=1}^{n} \log M_{X_i - \mathbb{E}[X_i]}(\lambda)$ to show that for means

$$\log P(\bar{X}_n - \mathbb{E}[\bar{X}_n] \geq t) \leq \frac{-2nt^2}{(b-a)^2}$$

sub-Exponential RV: An RV X is sub-Exponential with parameters (σ^2, b) if for all $|\lambda| < \frac{1}{b}$:

$$\log M_{x-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2}. \qquad \text{(Definition 1)}$$

Equivalently, X is sub-Exponential if there exist constants $c > 0$ and $l > 0$ such that

$$P(\{X - \mu\} \geq t) \leq cP(|\varepsilon_l| \geq t) \quad \forall t > 0,$$

where $\varepsilon \sim \text{Exp}(l)$, so that $P(|\epsilon_l| \geq t) = \exp(-lt)$. By Chernoff's inequality, we have the following tail bound for a sub-Exponential (σ^2, b) RV:

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

This means that a sub-Exponential RV concentrates like a Gaussian near 0, but has thicker tails as t gets further from 0. Some examples of sub-Exponential RVs include:

i. All sG RVs are also sub-Exponential.

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

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

iv. Bounded RVs are sub-Exponential (proof uses Taylor exp, convergent geometric series).

Bernstein's inequality:

i. General form: If an RV with mean μ and variance σ^2 satisfies the Bernstein condition: $|\mathbb{E}[(X-\mu)^k]| \leq \frac{1}{2}k!\sigma^2 b^{k-2}$, then

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

ii. Bounded RV version: Suppose X is bounded so that $X - \mu| \leq b$ and $\text{Var}(X) = \sigma^2$. Whenever $|\lambda| < \frac{1}{2b}$, then $M_{X-\mu}(\lambda) \leq \exp(\lambda^2\sigma^2)$, so X is sub-exponential with parameters $(2\sigma^2, 2b)$. So

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

iii. Sample mean of bounded RVs version: Suppose $X_1, ..., X_n$ are bounded RVs so that $X_i - \mu_i| \leq b$ and $\text{Var}(X_i) = \sigma_i^2$. Then

$$P(\bar{X}_n - \mathbb{E}[\bar{X}_n] \geq t) \leq \exp\left(-\frac{nt^2}{2(\sum_{i=1}^n \sigma_i^2 + bt)}\right)$$

Comparing Bernstein's and Hoeffding's inequalities shows that if $\frac{1}{n}\sum_{i=1}^n \sigma_i^2$ is small, then Bernstein's inequality can give tighter bounds for small values of t.

2.6.3 Martingale-based bounds

The inequalities so far have focused on statistics that can be written as sample means of independent RVs. Now we consider, for arbitrary non-linear functions f of a collection of independent RVs, inequalities of the form $P(f(X_1, ..., X_n) - \mathbb{E}[f(X_1, ..., X_n)] \geq t)$.

Bounded differences property: A function $f : \mathcal{X}^n \to \mathbb{R}$ is said to satisfy the bounded differences property if, for all i, there exists a constant $c_i < \infty$ so that the following inequality holds for all $x_1, ..., x_n, x_i' \in \mathcal{X}$:

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

i.e. that the output of f does not depend too much on one of its inputs.

McDiarmid's inequality: If $X = (X_1, ..., X_n)$ is a collection of independent RVs and f satisfies the bounded differences inequality with bounds $c_1, ..., c_n$, and $\mathbb{E}[f(X)] < \infty$, then for all $t > 0$,

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

3 Convergence Theory

3.1 Types of convergence

Here we consider Borel-measurable $\mathbb{R}^d$-valued RVs.

Almost sure convergence: Consider a sequence of RVs $\{X_n\}_{n=1}^{\infty}$ and an RV X defined on a common probability space $(\Omega, \mathcal{F}, P)$. $\{X_n\}_{n=1}^{\infty}$ converges almost surely to X iff

$$P\left(\lim_{n\to\infty} \|X_n - X\| = 0\right) = P(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1.$$

This is typically denoted by $X_n \overset{a.s.}{\to} X$.

- Let $\{X_n\}$ be a sequence of random variables and X and $\tilde{X}$ be random variables defined on a common probability space. Then if $X_n \overset{a.s.}{\to} X$ and $X_n \overset{a.s.}{\to} \tilde{X}$, then $X = \tilde{X}$ a.s.

Convergence in probability: The sequence $\{X_n\}_{n=1}^{\infty}$ converges in probability to X if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\|X_n - X\| > \epsilon) = \lim_{n\to\infty} \mathbb{P}(\{\omega : \|X_n(\omega) - X(\omega)\| > \epsilon\}) = 0$$

This is often denoted by $X_n \overset{p}{\to} X$.

- Let $\{X_n\}$ be a sequence of random variables and X and $\tilde{X}$ be random variables defined on a common probability space. Then if $X_n \overset{p}{\to} X$ and $X_n \overset{p}{\to} \tilde{X}$, then $X = \tilde{X}$ a.s.
- $X_n \overset{P}{\to} X$ and $Y_n \overset{P}{\to} Y$ implies that $(X_n, Y_n) \overset{P}{\to} (X, Y)$.

Weak convergence: An $\mathbb{R}^d$-valued RV X_n converges weakly (in distribution, in law) to X if for all bounded, continuous functions $f : \mathbb{R}^d \to \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{E}[f(A_n)] = \mathbb{E}[f(A)]$$

We denote this by $X_n \Rightarrow X$, $X_n \rightsquigarrow X$, or $X_n \overset{d}{\to} X$.

- Note that weak convergence of marginals $X_n \overset{d}{\to} X$, $Y_n \overset{d}{\to} Y$ does NOT necessarily imply (X_n, Y_n) converges weakly (Stat 581 HW2).
- In the above case, we do have weak convergence of (X_n, Y_n) if $X_n \perp\!\!\!\perp Y_n$ for each n, but convergence will be $(X_n, Y_n) \overset{d}{\to} (\tilde{X}, \tilde{Y})$ for, possibly, $X \neq \tilde{X}$, $Y \neq \tilde{Y}$ since $\tilde{X} \perp\!\!\!\perp \tilde{Y}$ is required, so $\tilde{X}$ and $\tilde{Y}$ will be independent versions of X, Y (Stat 581 HW 2).

Weak convergence for a $\mathbb{D}$-valued random element X_n to X: If X_n takes values in the general metric space $(\mathbb{D}, c)$, the definition follows as above, only $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ must now hold for any bounded, $c - continuous$ functional $f : \mathbb{D} \mapsto \mathbb{R}$.

Portmanteau Lemma: Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of RVs and X be an RV. The following are equivalent (and hence definitions of weak convergence).

1. $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ for all bounded, continuous functions f.
2. CDF: For all continuity points $t \in \mathbb{R}^d$ of $P(X \leq \cdot), P(X_n \leq t) \to P(X \leq t)$ as $n \to \infty$.
3. $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ for all bounded, Lipschitz-continuous (or upper/lower semicontinuous) functions f.
4. Levy continuity (characteristic functions): For all $t \in \mathbb{R}^d$, $\mathbb{E}[\exp(it^T X_n)] \to \mathbb{E}[\exp(it^T X)]$.
5. Cramer Wold: For all $t \in \mathbb{R}^d$, $t^T X_n \Rightarrow t^T X$

3.2 Convergence theorems:

Continuous mapping theorem: Let $X_n \in \mathbb{R}^d$ be a sequence of RVs and let $g : \mathbb{R}^d \to \mathbb{R}^m$ be a continuous function at every point of a set C such that $P(X \in C) = 1$. Then

i. If $X_n \Rightarrow X$, then $g(X_n) \Rightarrow g(X)$.
ii. If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.
iii. If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.
iv. (Prelude 1 to Slutsky's lemma): If $X_n \Rightarrow X$ and $\|X_n - Y_n\| \xrightarrow{p} 0$, then $Y_n \Rightarrow X$.
v. (Prelude 2 to Slutsky's lemma): If $X_n \Rightarrow X$ and $Y_n \xrightarrow{p} c$ for some constant c, then $(X_n, Y_n) \Rightarrow (X, c)$.

See Section 6.4 for the CMT for $\mathbb{D}$-valued RVs.

Slutsky's Lemma: Let X_n be an $\mathbb{R}^d$-valued RV for which $X_n \Rightarrow X$. The following hold:

i. If the $\mathbb{R}^d$-valued RV Y_n satisfies $Y_n \xrightarrow{p} c$ for a constant $c \in \mathbb{R}^d$, then $X_n + Y_n \Rightarrow X + c$.
ii. If the $\mathbb{R}$-valued RV Y_n satisfies $Y_n \xrightarrow{p} c$ for a constant $c \in \mathbb{R}$, then $Y_n X_n \Rightarrow cX$.
iii. If the $\mathbb{R}$-valued RV Y_n satisfies $\xrightarrow{p} c$ for a constant $c \in \mathbb{R}\backslash\{0\}$, then $X_n/Y_n \Rightarrow X/c$.

Partial Slutsky's lemma for weak convergence: Let $\mathcal{F}$ be a given function class. Let $X_1, X_2, ...$ and $Y_1, Y_2, ...$ be two sequences of $\ell^\infty(\mathcal{F})$-valued RVs, and suppose that $\|X_n - Y_n\|_{\mathcal{F}} \xrightarrow{p} 0$. If X_n converges weakly to X in $\ell^\infty(\mathcal{F})$ relative to $\|\cdot\|_{\mathcal{F}}$ for some $\ell^\infty(\mathcal{F})$-valued RV X, then Y_n also converges weakly to X in $\ell^\infty(\mathcal{F})$.

Weak law of large numbers: Let $X_1, ..., X_n \overset{i.i.d.}{\sim} P$. Then, letting $\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^n X_i$, we have

$$\text{If } \mathbb{E}_P[|X_i|] < \infty, \text{ then } \bar{X}_n \xrightarrow{p} \mathbb{E}_P[X_i]$$

3.3 Central Limit Theorems

Univariate CLT: Let $X_1, ..., X_n \overset{i.i.d.}{\sim} P$. Let $\sigma_p^2 \equiv \mathrm{Var}_P(X_i)$. If $\mathbb{E}_P[X_i^2] < \infty$, then

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

Multivariate CLT: Let $X_1, ..., X_n \overset{i.i.d.}{\sim} P$, where P is a distribution with support in $\mathbb{R}^d$ that satisfies $\mathbb{E}_P[\|X_i\|^2] < \infty$. It holds that

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

where $\mu \equiv \mathbb{E}_P[X_i]$ and $\Sigma \equiv \mathbb{E}_P[(X_i - \mu)(X_i - \mu)^T]$. (Proof follows by the Cramér-Wold device, applying the univariate CLT to $t^T \bar{X}$ for an arbitrary $t \in \mathbb{R}^d$).

Lindeberg-Feller CLT: This generalizes the usual CLT to the setting if independent but not necessarily identically distributed observations. Let $\{X_{ni}\}_{i=1}^n$ be an independent collection of $\mathbb{R}$-valued RVs. Suppose that the means $\mu_{ni} \equiv \mathbb{E}[X_{ni}]$ and the variances $\sigma_{ni}^2 \equiv \mathrm{Var}(X_i)$ exist and are finite. Suppose that $\sigma_n^2 \equiv \sum_{i=1}^n \sigma_{ni}^2 > 0$ for all n. Finally, let $Y_{ni} \equiv (X_{ni} - \mu_{ni}/\sigma_n)$.

If the Lindeberg condition holds:

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

Or, alternatively if the Lyapunov condition holds:

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

then

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

3.4 Delta methods

Univariate delta method ($\mathbb{R} \to \mathbb{R}$): Suppose $f : \mathbb{R} \to \mathbb{R}$ is differentiable at ψ_0, and $r_n(\psi_n - \psi_0) \rightsquigarrow Z$ holds. Then

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

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

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

Multivariate delta method ($\mathbb{R}^d \to \mathbb{R}^p$): Suppose $f : \mathbb{R}^d \to \mathbb{R}^d$ is differentiable at $\psi_0 \in \mathbb{R}^d$, meaning there exists an $\mathbb{R}^p \times \mathbb{R}^d$ Jacobian matrix $J_f = \begin{pmatrix} \nabla f_1 \\ \vdots \\ \nabla f_p \end{pmatrix}$, so $(J_f)_{ij} = \frac{\partial}{\partial \psi_j} f_i(\psi) \Big|_{\psi = \psi_0}$. Further, suppose that $r_n(\psi_n - \psi_0) \rightsquigarrow Z$ holds. Then

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

Delta method 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 an asymptotically linear estimator for $\psi_{0,j}$ for all $j \in \{1, ..., d\}$. 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 functions equal to:

$$\phi_{P_0} : x \mapsto \langle \nabla f(\psi_0), \phi_{P_0}(x) \rangle,$$

where $\psi_{P_o}(x)$ is the influence function of ψ_n. This implies

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

Derivation of univariate Delta method: Suppose we have an RV ψ_n such that $\sqrt{n}(\psi_n - \psi_0) \xrightarrow{d} N(0, \sigma^2)$ and that we want to know the asymptotic distribution of $g(\psi_n)$, where g is differentiable at ψ_0 with $g'(\psi_0) \neq 0$ and g' continuous. Then we have by Taylor expansion

$$g(\psi_n) = g(\psi_0) + g'(\psi^*)(\psi_n - \psi_0)$$

for some ψ^* between ψ_0 and ψ_n. By the CMT, since $\psi_n \xrightarrow{p} \psi_0$ (implied by Chebyshev's applied to convergence in distribution statement above) and $\psi^* \to \psi_0$ by sandwich, we have $g'(\psi^*) \to g'(\psi_0)$. So then, by Slutsky's theorem, we get

$$\sqrt{n}(g(\psi_n) - g(\psi_0)) \xrightarrow{d} N(0, g'(\psi_0)^2\sigma^2)$$

This formulation shows how we can derive the second-order Delta method if $g'(\psi_0) = 0$ to get convergence at an n rate to $\sigma^2 \frac{g''(\psi_0)}{2}\chi^2$.

3.5 Stochastic order notation and Prokhorov's Theorem:

(Stochastic) order notation: Let $\{x_n\}_{n=1}^\infty$ and $\{r_n\}_{n=1}^\infty$ be real-valued sequences and $\{X_n\}_{n=1}^\infty$ and $\{R_n\}_{n=1}^\infty$ be sequences of RVs.

1. O: We say that $x_n = O(r_n)$ if $\limsup_{n\to\infty} \left|\frac{x_n}{r_n}\right| < \infty$. Equivalently, if there exists an $M > 0$ such that $\mathbb{I}[|x_n| \leq M|r_n|] \overset{n\to\infty}{\to} 1$.
2. o: We say that $x_n = o(r_n)$ if $\limsup_{n\to\infty} \left|\frac{x_n}{r_n}\right| = 0$. Equivalently, if for all $M > 0$, $\mathbb{I}[|x_n| \leq M|r_n|] \overset{n\to\infty}{\to} 1$.
3. O_p: O_p: We say that $X_n = O_p(R_n)$ if for all $\epsilon > 0$, there exists a constant $M > 0$ such that
$$\limsup_{n\to\infty} P\left(\left|\frac{X_n}{R_n}\right| > M\right) < \epsilon.$$
i.e. if X_n/R_n is bounded in probability. We say that $\{X_n\}_{n=1}^\infty$ is **uniformly tight** if $X_n = Op(1)$.
4. o_p : We say that $X_n = o_p(R_n)$ if for all constants $M > 0$,
$$P\{|X_n| \leq M|R_n|\} \overset{n\to\infty}{\to} 1 \Leftrightarrow P\left\{\left|\frac{X_n}{R_n}\right| > M\right\} \to 0$$
or, equivalently iff
$$\frac{X_n}{R_n} \xrightarrow{p} 0.$$
Hence $X_n = o_p(1) \Leftrightarrow X_n \xrightarrow{p} 0$.

The definitions above extend to random vectors by replacing $|\cdot|$ with $\|\cdot\|$.

o-notation for functions: If $q : \mathbb{R}^k \to \mathbb{R}$, we say that $q = o(\|u\|)$ if $\frac{q(u)}{\|u\|} \overset{u\to 0}{\to} 0$.

Prokhorov's theorem: Let X_n be random vectors in $\mathbb{R}^p$.

i. (Weak convergence implies uniform tightness): If $X_n \Rightarrow X$ for some X, then $X_n = O_p(1)$.

ii. (Uniform tight implies subsequence that converges weakly - like Bolzano-Weierstrass for RVs): If $X_n = O_p(1)$, then there exists a subsequence $\{X_{n_j}\}_{j=1}^{\infty} \subseteq \{X_n\}_{n=1}^{\infty}$ such that $X_{n_j} \Rightarrow X$ for some X.

Useful properties of stochastic order notation:

1. $X_n = o_p(R_n)$ iff $X_n = R_n Y_n$ for some $Y_n = o_p(1)$, iff $\frac{X_n}{R_n} = o_p(1)$.
2. $X_n = O_p(R_n)$ iff $X_n = R_n Y_n$ for some $Y_n = O_p(1)$, iff $\frac{X_n}{R_n} = O_p(1)$.
3. $o_p(1) + o_p(1) = o_p(1)$ and $o_p(n^{-\alpha}) + o_p(n^{-\beta}) = o_p(n^{-\min(\alpha,\beta)})$ for $\alpha, \beta > 0$.
4. $O_p(1) + O_p(1) = O_p(1)$.
5. $O_p(1)O_p(1) = O_p(1)$.
6. $O_p(1)o_p(1) = o_p(1)$.
7. $(1 + o_p(1))^{-1} = O_p(1)$.
8. $X_n = o_p(1)$ implies that $X_n = O_p(1)$.

4 Statistical decision theory

4.1 Setup: The decision problem in the parametric setting

Suppose we see data $X \in \mathcal{X}$ drawn from a distribution P_θ where $\theta \in \Theta$.

Action Space: We denote the set of actions $\mathcal{A}$, with elements $a \in \mathcal{A}$ denoted actions.

Decision rule: A decision rule is a function that maps from $\mathcal{X} \to \mathcal{A}$.

- In general, decision rules can be random. In such cases, we write the decision rule as $D(\cdot|x)$, a conditional probability distribution of an (now random variable) action A given that $X = x$. When $\mathcal{A}$ has finitely many actions, $D(a|x)$ is the probability of deciding $d = a$ given $X = x$.
- If we restrict to deterministic decision rules, we write the rule as $T : \mathcal{X} \to \mathcal{A}$.
- We generally denote $\mathcal{D}$ for the class of stochastic decision rules and $\mathcal{T}$ for the (smaller) class of deterministic rules.

Loss: A loss function quantifies the penalty that we incur if θ is true, but we choose an action a. For example, squared error loss is:

$$L(a, \theta) = (a - \theta)^2$$

Frequentist risk (sometimes just "risk"): The risk of a decision rule is the expected loss at a given value of θ:

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

If the decision rule is deterministic, then we have

$$\mathcal{R}(D,\theta) = \int_{\mathcal{X}} L(T(x),\theta) dP_\theta(x)$$

Frameworks for evaluating decision rules:

- **Admissibility:** A decision rule D is called **inadmissible** if there exists a rule D' such that $\mathcal{R}(D',\theta) \leq \mathcal{R}(D,\theta)$ for all $\theta \in \Theta$ and for which $\mathcal{R}(D',\theta) < \mathcal{R}(D,\theta)$ for some $\theta \in \Theta$, and is called admissible otherwise.
- **Minimaxity:** Under the minimax framework, we prefer D_1 to D_2 if it has lower maximal risk. That is, if

  $$\sup_{\theta\in\Theta} \mathcal{R}(D_1,\theta) < \sup_{\theta\in\Theta} \mathcal{R}(D_2,\theta).$$

  A minimax rule is a rule $D^* \in \mathcal{D}$ that is optimal according to the maximal risk criterion:

  $$\sup_{\theta\in\Theta} \mathcal{R}(D^*,\theta) = \inf_{D\in\mathcal{D}} \sup_{\theta\in\Theta} \mathcal{R}(D,\theta).$$

- **Bayesian framework:** Define a distribution Π over Θ. The **Bayes risk** of a rule D is defined as the expected risk of D over $\theta \sim \Pi$. That is:

  $$r(D,\Pi) = \int \mathcal{R}(D,\theta) d\Pi(\theta).$$

  A **Bayes rule** is a rule $D_\Pi \in \mathcal{D}$ that is optimal according to the Bayes risk criterion:

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

- **Neyman-Pearson criterion:** In hypothesis testing, a popular lens for choosing a decision rule is the Neyman-Pearson paradigm. For a specified type I error level α, it advocates choosing a decision rule D as follows (and if possible):

  $$\text{minimize} \quad \mathcal{R}(\theta, D) \text{ over } D \in \mathcal{D}$$

  subject to $\sup_{\theta\in\Theta_0} \mathcal{R}(\theta, D) \leq \alpha$. A rule that satisfies this criterion is UMP level α.

4.2 Finding Bayes rules

Posterior distribution: Let Π be a prior distribution of Θ, and let $p(\cdot|\theta)$ and π denote the densities of P_θ and Π with respect to appropriate dominating measures. Then the conditional distribution of θ given X is referred to as the posterior distribution, and has density

$$p(\theta|x) = \frac{p(x,\theta)}{p(x)} = \frac{p(x|\theta)\pi(\theta)}{\int_\Theta p(x|\theta')d\Pi(\theta')}$$

- Typically, we write $X \sim P$ (as opposed to $X \sim |_\theta$) for the marginal distribution of X. We have $P(X \in A) = \int_\Theta \int_A P(X|\theta)\pi(\theta)dxd\theta$.

concept: A **kernel** of the posterior distribution is a function $f : \Theta \times \mathcal{X} \to [0, \infty)$ such that there exists a function $c : \mathcal{X} \to [0, \infty)$ so that the density $p(\theta|x)$ satsfies $p(\theta|x) = c(x)f(\theta, x)$. Note that kernels of distributions *uniquely determine distributions* because densities must integrate to one so $\int c(x)f(\theta, x)d\theta = 1 \Rightarrow c(x) = 1/[\int f(\theta, x)d\theta]$. Note that the numerator of a typical Bayesian posterior density, $f(x, \theta) = p(x|\theta)\pi(\theta)$, is a kernel.

Conjugate prior: Suppose the $X|\Theta = \theta \sim P_\theta$, $\theta \in \Theta$. This is the equivalent to what frequentist call a statistical model. Further, suppose that the prior Π is restricted to belong to some family $\mathcal{P}_\Pi$. Then Π is called a **conjugate prior** if, for almost all x, the posterior distribution $P(\cdot|x)$ falls in $\mathcal{P}_\Pi$.

- A useful result is that if $X_1, ..., X_n|\Theta = \theta \overset{iid}{\sim} N(\theta, \sigma^2)$ and $\Theta \sim N(\mu, \tau^2)$, then

$$\Theta|X \sim N\left(\frac{\mu/\tau^2 + n\bar{x}/\sigma^2}{1/\tau^2 + n/\sigma^2}, \frac{1}{1/\tau^2 + n/\sigma^2}\right)$$

Finding Bayes rules by minimizing posterior expected loss: Suppose that $\Theta \sim \Pi, X|\Theta = \theta \sim P_\theta$ and $L(a, \theta) \geq 0$ for all $\theta \in \Theta, a \in \mathcal{A}$. If

(i) There exists a rule $D \in \mathcal{D}$ with finite Bayes risk, and
(ii) There exists $D_\Pi \in \mathcal{D}$ such that, for almost all x, $D_\pi(\cdot|x)$ minimizes

$$\mathbb{E}\left[\int_{\mathcal{A}} L(a, \theta)D(da|x)\middle| X = x\right] \quad \text{in } D(\cdot|x),$$

then D_Π is a Bayes rule. Here, "almost all" is with respect to the marginal law of X. Some Bayes rules for common loss functions are:

(a) Squared error loss: $L(a, \theta) = (a - \psi(\theta))^2$, posterior mean: $T_\Pi = \mathbb{E}[\psi(\theta)|X = x]$.
(b) Absolute deviation loss: $L(a, \theta) = |a - \psi(\theta)|$, posterior median: $T_\Pi = \text{median}(\psi(\theta)|X = x)$.
(c) Weighted squared error loss: $L(a, \theta) = w(\theta)(a - \psi(\theta))^2$, weighted posterior mean: $T_\Pi = \frac{\mathbb{E}[w(\theta)\psi(\theta)|X=x]}{\mathbb{E}[w(\theta)|X=x]}$.
(d) $0-1$ loss: $L(a, \theta) = \mathbb{I}[a \neq \psi(\theta)]$, maximum posterior probability: $T_\Pi = \text{argmax}_a Pr(\psi(\theta) = a|X = x)$.

Restriction to deterministic rules for convex loss functions: If $a \mapsto L(a, \theta)$ is convex in a for all θ, $\mathcal{D}$ is unrestricted, $\mathcal{A}$ is a convex set, and there exists a Bayes rule $D_\Pi \in \mathcal{D}$, then there exists a deterministic Bayes rule. That is, a Bayes rule in $\mathcal{D}$ such that $D(\cdot|x)$ is a degenerate distribution that places point mass at a_x for some $a_x \in \mathcal{A}$. This follows from Jensen's inequality. We $L(a, \theta)$ is *strictly* convex, then all Bayes rules are deterministic.

4.3 Finding minimax rules: 581 version (least favorable priors or constant risk)

Least favorable prior: A prior Π^* is called least favorable if

$$r(D_{\Pi^*}, \Pi^*) = \sup_\Pi r(D_\Pi, \Pi).$$

Least favorable prior - minimaxity: If Π satisfies

$$r(D_\Pi, \Pi) = \sup_{\theta \in \Theta} \mathcal{R}(D_\Pi, \theta).$$

then

(i) D_Π is minimax.
(ii) If D_Π is unique Bayes with respect to Π, then D_Π is unique minimax.
(iii) Π is least favorable.

Corollary: A Bayes rule with constant risk is minimax: Suppose that Π is a prior distribution on Θ such that $\mathcal{R}(D_\Pi, \theta)$ is constant, that is, $\mathcal{R}(D_\Pi, \theta)$ does not depend on θ. Then, D_Π is minimax.

Least favorable prior sequence: Let $\{\Pi_k : k = 1, 2, ...\}$ be a sequence of priors and denote the limit inferior of the Bayes risk with respect to priors in this sequence by r_0. That is,

$$\liminf_{k \to \infty} r(D_{\Pi_k}, \Pi_k) = r_0. \tag{1}$$

This sequence is called **least favorable** if, for all priors Π, $r(D_\Pi, \Pi) \leq r_0$

- If $\{\Pi_k\}$ is least favorable, then $\limsup_k r(D_{\Pi_k}, \Pi_k) \leq r_0$. Hence, the limit in (1) exists, and is equal to r_0.

Theorem: Minimax-least favorable prior sequence: Suppose that $\{\Pi_k\}$ is a prior sequence and let r_0 be as defined in (1). If $D \in \mathcal{D}$ satisfies

$$\sup_{\theta \in \Theta} \mathcal{R}(D, \theta) = r_0 \tag{2}$$

then

(i) D is minimax.
(ii) $\{\Pi_k\}$ is a least favorable prior sequence.

Lemma: Sufficient condition for minimaxity in larger models: Let $\mathcal{P}_1 \subset \mathcal{P}_2$ denote two statistical models. That is, two collections of distributions on $\mathcal{X}$. If D_1 is a minimax decision rule over the model $\mathcal{P}_1$ that satisfies

$$\sup_{P \in \mathcal{P}_1} \mathcal{R}(D_1, P) = \sup_{P \in \mathcal{P}_2} \mathcal{R}(D_1, P),$$

then D_1 is also minimax over the model $\mathcal{P}_2$, that is,

$$\sup_{P \in \mathcal{P}_2} \mathcal{R}(D_1, P) = \inf_{D \in \mathcal{D}} \sup_{P \in \mathcal{P}_2} \mathcal{R}(D, P).$$

4.4 Finding admissible rules

Unique Bayes rule: For a prior Π, a rule D_Π is called unique Bayes if, for all $\theta \in \Theta$, a rule is Bayes if and only if it is equal to D_Π a.e. P_θ. That is, any other Bayes rule under Π would have to agree with D_Π for all x with $P_\theta(x) > 0$.

Lemma: Conditions where a Bayes rule is unique: Let Π be a prior distribution and let D_Π be a Bayes rule for this prior. If

(i) the loss function $(a, \theta) \mapsto L(a, \theta)$ is squared error (in fact if it is convex).
(ii) $r(D_\Pi, \Pi) < \infty$.
(iii) for any A in the σ-field $\mathcal{A}$ on $\mathcal{X}$, $Q(A) \equiv \int P_\theta(X \in A) d\Pi(\theta) = 0$ implies that $P_\theta(X \in A) = 0$ for all $\theta \in \Theta$.
  - A sufficient condition for this to hold is: If there exists a measure ν on $(\mathcal{X}, \mathcal{A})$ such that, for all $\theta \in \Theta$, $P_\theta \ll \nu$ and $\nu \ll P_\theta$.

then D_π is unique Bayes.

Admissibility of unique Bayes rules: Any unique Bayes rule is admissible.

Unique minimax: A rule D^* is called unique minimax if, for all $\theta \in \Theta$, a rule is minimax if and only if it is equal to D^* a.e. P_θ.

Admissibility of unique minimax rules: Any unique minimax rule is admissible.

Admissibility of sample mean for univariate normal data: If $X = (X_1, ..., X_n) \overset{iid}{\sim} N(\theta, \sigma^2), \theta \in \Theta = \mathbb{R}$ with σ^2 known, then $T : x \mapsto \bar{x}_n$ is an admissible estimator of θ with respect to the mean squared error risk.

4.5 Stein's paradox

Stein's Lemma: Let $Y \sim N(\mu, \sigma^2)$ and let $g : \mathbb{R} \to \mathbb{R}$ be such that $\mathbb{E}[|g'(Y)|] < \infty$. Then $\mathbb{E}[g(Y)(Y - \mu)] = \sigma^2 \mathbb{E}[g'(Y)]$. There is also a multivariate generalization of this that follows by applying Fubini's theorem.

James-Stein Theorem: Suppose we observe $X_1, ..., X_n \overset{iid}{\to} N(\theta, \sigma^2 I_d)$ where $\theta \in \mathbb{R}^d$. The objective is to estimate θ with performance quantified via the MSE risk:

$$\begin{aligned}\mathcal{R}(T, \theta) &= \sum_{i=1}^{d} \mathbb{E}_{P_\theta}\left[(T(X)_j - \theta_j)^2\right] \\ &= \mathbb{E}_{P_\theta}\left[\|T(X) - \theta\|^2\right].\end{aligned}$$

Then the sample mean $T : x \mapsto \bar{x}_n$ is inadmissible. One estimator that dominates it is the **James-Stein estimator**, which takes the form:

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

4.6 Decision theory outside of parametric models: minimax rate optimality

More generally, suppose we observe data $W \in \mathcal{W}$ from some distribution $P \in \mathcal{P}$ where $\mathcal{P}$ is a statistical model. The typical case is when $P = Q^n$ for some distribution $Q \in \mathcal{Q}$ (i.e. P is the distribution of n iid rdraws from Q).

- Now a loss function is $L : \mathcal{A} \times \mathcal{P} \to \mathbb{R}$, and so the frequentist risk is $\mathcal{R}(T, P) = \int L(T(w), P) dP(w)$ and the Bayes risk (where a prior Π is placed on P) is $r(T, \Pi) = \int \mathcal{R}(T, P) d\Pi(P)$.
- For estimating functions we often use **mean integrated square error (MISE)**: $L(a, P) = \int [a(x) - f_Q(x)]^2 dx$.
- The definition of a minimax rule becomes $T^* \in \mathcal{T}$ such that

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

Outside of the fully parametric setting, it is very difficult to find minimax rules (i.e. no natural extension to least favorable prior approach). We do know that the risk goes to 0 as $n \to \infty$ for some rule.

Minimax rate optimality: Since we can't find the minimax rule or the minimax risk or exactly, we usually settle for a notion of "nearly minimax optimal" as $n \to \infty$. Specifically, a sequence of rules $\{T_n\}$ is minimax rate optimal if

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

where $\mathcal{T}_n$ is the collection of allowable rules for each n and $P = Q^n$. This is a $\liminf$, so to show it for a given $\{T_n\}$, we seek to upper bound the denominator (the maximum risk of T_n) and lower bound the numerator (the minimax risk of any rule).

Lower bounding the minimax risk: The following three strategies are useful for lower bounding the minimax risk:

1. **Bound with the Bayes risk:** For any decision problem, we have

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

   This is just as in STAT 581. Note that the proof follows by combining the inequality $\sup_{P \in \mathcal{P}} \mathcal{R}(T, P) \geq \sup_\pi r(T, \Pi)$ with the min-max inequality.

2. **Le Cam's method:** First, define the following:
   - **Divergence/discrepancy:** We define

$$d(P_1, P_2) = \inf_{a \in \mathcal{A}} [L(a, P_1) + L(a, P_2)].$$

     Intuitively, this is high when there is no action that is simultaneously good for both P_1 and P_2. In point estimation settings ($\psi(P)$ is parameter of interest) under squared error, we get $d(P_1, P_2) = \frac{1}{2}[\psi(P_1) - \psi(P_2)]^2$.
     Under MISE (if f_Q has finite L_2-norm) and the loss is mean integrated squared error (MISE), then $d(P_1, P_2) = \frac{1}{2} \int [f_{Q_1}(x) - f_{Q_2}(x)] dx$.

- **Testing affinity:** We define

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

  This measures distributional overlap - it is essentially the area under both f_{P_1} and f_{P_2} It is often difficult to compute. It also requires $p_1 << p_2$.
- **Total variation:**

  $$\text{TV}(P_1, P_2) = \sup_A |P_1(A) - P_2(A)| = 1 - \|P_1 \wedge P_2\|.$$

  This is the area under f_{P_1} that is not under f_{P_2}.
- **KL Divergence:**

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

  By Pinsker's inequality, $TV(P_1, P_2) \leq \sqrt{KL(P_1, P_2)/2}$, but this bound is trivial when $KL(P_1, P_2) \geq 2$. In this case, a tighter bound is $TV(P_1, P_2) \leq 1 - \frac{1}{2}\exp(-KL(P_1, P_2))$.

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

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

Loosely speaking, this bound is maximized (i.e. is tightest) when it is difficult to determine whether the data came from P_1 or P_2 (small KL-divergence), but price for wrong decision is high (large divergence).

3. **Fano's method:** In some cases, Le Cam's method gives a rate-optimal minimax lower bound, but in others only comparing two distributions is not enough (i.e. in infinite-dimensional models). In such cases, i.e. estimating a smooth density with MISE loss, we may use Fano's method which says:
   Let $\eta := \min_{j \neq k} d(P_j, P_k)$ and $\tilde{P} = \frac{1}{N}\sum_{j=1}^N P_j$, i.e. a uniform mixture, we obtain

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

5 Estimation: Part I (Parametric models)

5.1 Sufficiency, minimal sufficiency, completeness, ancillarity

Statistical model: A statistical model $(\mathcal{X}, \mathcal{P})$ consists of a sample space $\mathcal{X}$ and a family $\mathcal{P} \equiv \{P_\theta | \theta \in \Omega\}$ of possible probability probability distributions on $\mathcal{X}$, where θ is an unknown (possibly infinite-dimensional) parameter. A **statistic** is any (measurable) function of the data. We write for it $T(X)$.

Sufficient statistic: $T(X)$ is a sufficient statistic for $\mathcal{P}$ (i.e. for θ) if the conditional distribution of X given T is constant in θ. In other words, if for every event $A \subseteq \mathcal{X}$, $P_\theta[X \in A|T]$ does not depend on θ.

- Less formally: A sufficient statistic holds all information from the data that is relavant to the inference procedure.
- **Bayesian sufficiency:** Suppose θ itself is a random variable with prior distribution π of $\mathcal{P}$ (or Ω). $T(X)$ is a sufficient statistic for $\mathcal{P}$ (i.e. for θ) if, for every π, the conditional (posterior) distribution of $\theta|X$ depends on X only through the value of T(X).
- In general, if T is a sufficient statistic and $T \overset{1-1}{\Leftrightarrow} V$, the V is also sufficient. This is especially useful in applications of the UMVUE supermarket, for instance when we would rather use $(\bar{X}, s_n^2)$ instead of $(\sum X_i^2, \sum X_i)$.

Fisher-Neyman factorization theorem: Consider a (parametric or nonparametric) statistical model $(\mathcal{X}, \mathcal{P})$, where each $\mathcal{P}_\theta \in \mathcal{P}$ is determined by a pdf or pmf $f(x)$. The statistic $T \equiv T(X)$ is sufficient for $\mathcal{P}$ (i.e., for θ) if and only if $f_\theta(x)$ factors as follows:

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

where $g_\theta(T(x))$ depends on θ only through $T(x)$ and $h(x)$ does not depend on θ.

- We define the likelihood ratio (LR) as $L_{\theta_1,\theta_2}(x) = \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}$. The LR depends on X only through the value of the sufficient statistic, as we see by Fisher-Neyman factorization.

Minimial sufficient statistic: $T^*(X)$ is a minimal sufficient statistic for $\mathcal{P} \equiv \{\mathcal{P}_\theta\}$ if T^* is sufficient and if, for every other sufficient statistic $T(X)$, T^* is a reduction of T, i.e. $T * (X) = h(T(X))$ for some function h.

- The set of LRs is a minimal sufficient statistic: Let $T^{**}(X)$ be the entire family of pairwise LRs:

  $$T^{**}(X) \equiv \{L_{\theta_1,\theta_2}(X) | \theta_1, \theta_2 \in \Omega\}.$$

  When $T(X)$ is sufficient, we have already shown in 1.3 that the LR is a function of $T(X)$ $\forall$ θ_1, θ_2. Thus, $T^{**}(X)$ is a function of θ. So, by definition $T^{**}(X)$ is a minimal sufficient statistic.

Lehmann-Scheffe theorem: Suppose that X has pmf or pdf $f_\theta(x)$, $\theta \in \Omega$ and that $T^*(X)$ satisfies the following property: *for every pair of sample points* $x, y \in \mathcal{X}$*,*

$$T^*(X) = T^*(Y) \Leftrightarrow \frac{f_\theta(y)}{f_\theta(x)} \text{ is } \theta\text{-free.}$$

Then T^* is minimal sufficient for θ.

- k-parameter exponential family: Let $X_1, ..., X_n$ be an iid sample of RVs from a distribution with pdf (or pmf) that takes the form:

$$a(\theta_1, ..., \theta_k) \exp(\theta_1 T_1(x) + \cdots + \theta_k T_k(x)) \cdot h(x)$$

  where $\theta = (\theta_1, ..., \theta_k) \in \Omega$ is a k-dimensional parameter. The $X = (X_1, ..., X_n)$ has pdf

$$f_\theta(x) = [a(\theta)]^n \exp\left[\theta_1 \sum_{i=1}^n T_1(x_i) + \cdots + \theta_k \sum_{i=1}^n T_k(x_i)\right] \cdot \prod_{i=1}^n h(x_i).$$

  Thus, $(\sum T_1(X_i), ..., \sum T_k(X_i))$ is a k-dimensional sufficient statistic. Further, it is minimal sufficient *provided that the parameter space* $\Omega \subseteq \mathbb{R}^k$ *affinely spans* $\mathbb{R}^k$ by the Lehmann-Scheffe theorem since

$$\frac{f_\theta(y)}{f_\theta(x)} = \exp\left[\sum_{j=1}^k \theta_j \left(\sum_{i=1}^n T_j(y_i) - \sum_{i=1}^n T_j(x_i)\right)\right] \cdot \frac{\prod h(y_i)}{\prod h(x_i)}$$

  is θ-free if the affine span condition is met.

Ancillary statistic: A statistic $V \equiv V(X)$ on $\mathcal{X}$ is **ancillary** for $\mathcal{P} \equiv \{P_\theta\}$ if its distribution does not depend on θ; i.e. for an $A \subset \mathcal{X}$ and any integrable function g on $\mathcal{X}$, $P_\theta[V \in A]$ and $\mathbb{E}_\theta[g(V)]$ are θ-free.

- **Location family:** Let P_0 be determined by a pdf f_0 on $\mathcal{X} \equiv \mathbb{R}$. The location-parameter family of pdfs on $\mathbb{R}$ is $\{f_\mu(x) \equiv f_0(x-\mu) | \mu \in \mathbb{R}\}$. Any location-invariant statistic in a location-parameter family is ancillary (i.e. set of sample spacings $X_{(2)} - X_{(1)}, ...$, sample range, sample variance)
- **Scale family:** The scale-parameter family of pdfs is $\{f_\sigma(x) \equiv \sigma^{-1} f_0(\sigma^{-1}x) | \sigma \in \mathbb{R}_+\}$. Any scale-invariant statistic in a scale-parameter family (T-statistic, set of sample ratios $\frac{X_{(1)}}{X_{(n)}}, ...$) is ancillary.
- **Location/scale family:** The location/scale family of pdfs is given by $\{f_{\mu,\sigma}(x) \equiv \sigma^{-1} f_0(\sigma^{-1}(x-\mu)) | (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+\}$. Any location-scale invariant statistic (i.e. set of normalized sample spacings $\frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}$ and sample range/sample sd ratio $\frac{X_{(n)} - X_{(1)}}{s_n}$) is ancillary in a location-scale family (i.e. $N(\mu, \sigma^2)$).
- **1-parameter truncation family:** Let $X_1, ..., X_n$ be an i.d.d. sample from the truncation pdf

$$f_\theta(x) \equiv [B(\theta)^{-1}] I_{(a,\theta]}(x) \cdot b(x), \qquad x > a.$$

  where $a \in [-\infty, \infty)$ is specified, $\theta \in (a, \theta)$, and $B(\theta) \equiv \int_a^\theta b(x)dx < \infty \forall \theta > a$. The sufficient statistic $T(X) \equiv X_{(n)}$ is complete, thus minimal sufficient for θ.
  Similarly, the sufficient statistic $T(X) \equiv X_{(1)}$ is complete, hence minimal sufficient for θ, if $I_{(a,\theta]}(x), x > a$ is replaced by $I_{[\theta,a]}(x), x < a$.

Complete statistic: Let $(\mathcal{X}, \mathcal{P} \equiv \{P_\theta\})$ be a statistical model. A statistic $T \equiv T(X)$ is **complete** for $\mathcal{P}$ if

$$\mathbb{E}_\theta[g(T)] \text{ is } \theta\text{-free } \Rightarrow g(T) \text{ is constant a.e.}$$

or, equivalently, if

$$\mathbb{E}_\theta[g(T)] = 0, \forall\theta \Rightarrow g(T) = 0$$

- Completeness and ancillarity are antithetical properties.

Basu's Theorem (and others):

- if T is complete for $\mathcal{P}$, then no (non-constant) function of T is ancillary.
- If T is complete and sufficient for $\mathcal{P}$, then for every θ, T is independent of any ancillary statistic V.
- (other): If T is complete and sufficient, then it is minimal sufficient.

List of complete sufficient statistics in broad families:

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

$$f_\theta(x) = h(x)\exp\left(\sum_{i=1}^{k}\theta_i T_i(x) - A(\theta)\right), \qquad \theta \in \Omega \subset \mathbb{R}^k,$$

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

(ii) When X follow 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.

5.2 Minimum variance unbiased estimation (MVUE):

Uniformly minimum-variance unbiased estimator (UMVUE): An unbiased estimator $\hat{\tau}$ of τ is a **UMVUE** if, for all other unbiased estimators $\tilde{\tau}$,

$$\mathrm{Var}_\theta(\hat{\tau}) \le \mathrm{Var}_\theta(\tilde{\tau}) \qquad \forall\theta$$

Note that $\hat{\tau}$ may not exist. If it does, it has the smallest MSE among all *unbiased* estimators, but there may exist biased estimators with smaller MSE.

Rao-Blackwell-Lehmann-Scheffe (RBLS) Theorem: This provides a way to compute a UMVUE and guarantees its uniqueness. Let $T \equiv T(X)$ be complete and sufficient for θ. If there exists at least one unbiased estimator $\tilde{\tau} \equiv \tilde{\tau}(X)$ for $\tau(\theta)$ then there exists a unique UMVUE $\hat{\tau} = \hat{\tau}(T)$ for $\tau(\theta)$, namely

$$\hat{\tau}(T) \equiv \mathbb{E}[\tilde{\tau}(X)|T(X)]$$

- This results combines completeness (which gives uniqueness of unbiased estimator via uniqueness lemma) with the **Improvement lemma:** Suppose that $T \equiv T(X)$ is a sufficient statistic for θ. If $\tilde{\tau} \equiv \tilde{\tau}(X)$ is an unbiased estimator of $\tau \equiv \tau(\theta)$, then

$$\hat{\tau} \equiv \hat{\tau}(T) \equiv \mathbb{E}_\theta[\tilde{\tau}|T] = \mathbb{E}[\tilde{\tau}|T]$$

does not depend on θ by sufficiency and $\hat{\tau}$ is also unbiased for τ with smaller MSE than $\tilde{\tau}$.

UMVUE Supermarket Corollary: The easiest way to find a UMVUE. Let $T \equiv T(X)$ be complete and sufficient for θ. Then any function $\phi(T)$ is the UMVUE of it's expectation $\mathbb{E}_\theta[\phi(T)] \equiv \tau(\theta)$ (provided that the expectation is finite $\forall\theta$). So if we can find a CSS, we just need to find a function ϕ.

5.3 Maximum likelihood estimation:

Maximum likelihood estimator (MLE): Consider a statistical model $(\mathcal{X}, \mathcal{P} = \{P_\theta | \theta \in \Theta\})$ where each P_θ is determined by a pdf/pmf $f_\theta(x)$ wrt a dominating measure $d\mu(x)$ (i.e. dx). Suppose we observe $X = x$. We say $\hat{\theta} = \hat{\theta}(x)$ is the MLE of θ if

$$\max_{\theta\in\Theta} f_\theta(x) = f_{\hat{\theta}}(x)$$

Note that the MLE may not exist, and if it does it may not be unique. However, if it does exist then it does not depend on the choice of dominating measure for the family of distributions $\{P_\theta\}$, as the equation above can also be expressed in terms of likelihood ratios

$$\exists \quad \hat{\theta} \text{ s.t. } L_{\theta,\hat{\theta}}(x) \geq 1 \quad \forall \quad \theta \in \Omega$$

Thus, if the MLE exists *it is a function of the minimal sufficient statistic* T^{**}, the set of all likelihood ratios.

In the simple case where θ is a single real-valued parameter and $X = (X_1, ..., X_n)$ are i.i.d. from a regular family of pdf with common support, we have that the MLE maximizes

$$l_n(\theta) = l_n(\theta; x_1, ..., x_n) = \sum_{i=1}^{n} \log(f_\theta(x_i))$$

Because $l_n(\theta)$ is smooth, we may try to find the MLE by finding the roots of the likelihood equation (LEQ)

$$\frac{dl_n(\theta)}{d\theta} = \sum_{i=1}^{n} \frac{d\log(f_\theta(x_i))}{d\theta} = 0$$

The LEQ may have no real roots, one real root, or multiple real roots in $\Omega = (a, b)$, and such a root may correspond to a local maximum, local minimum, inflection point, (or saddle point in multivariate case). In order to find the MLE, we must find the global maximum.

Wald theorem (Existence of MLE): Suppose that θ is an identifiable parameter for the family $\{f_\theta | \theta \in \Omega\}$. Let $X = (X_1, ..., X_n)$ consist of i.i.d. observations from $f_\theta(x_i)$, so $f_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i)$. Suppose that $\theta = \theta_0$.

- If $\Omega = \{\theta_1, ..., \theta_r\}$ is finite, then the MLE $\hat{\theta}^{(n)}$ always exists, is unique for sufficiently large n, and is strongly consistent for θ_0.
- If Ω is not finite, assume that $f_\theta(x_i)$ is (upper semi-)continuous in θ and that these following global dominance and identifiability conditions are satisfied:

  - $\mathbb{E}_\theta\left[\sup_{\theta\in\Omega}\log^+\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right)\right] < \infty$
  - $\liminf_{\theta\to\partial\Omega} K(\theta_0,\theta) > 0$

  Then with P_{θ_0} - probability 1 the MLE $\hat{\theta}^{(n)}$ exists and is unique for sufficiently large n, and is $\hat{\theta}^{(n)}$ strongly consistent for θ_0.

Invariance of MLE: Let $\tau = g(\theta)$ for some function g. If $\hat{\theta}$ is the MLE of θ, then $\hat{\tau} = g(\hat{\theta})$ is the MLE of τ.

We now consider the asymptotic behavior of MLEs (they are a special case of Z-estimators, as we will see later)

Fisher information number (FIN): Consider a one-dimensional parametric model $\{P_\theta|\theta \in \mathbb{R}\}$ for $\Theta \subseteq \mathbb{R}$. The **Fisher information number** $I_X(\theta)$ is an intrinsic characteristic of the model $\{f_\theta(x)\}$: defined as

$$I_X(\theta) \equiv \mathbb{E}\left\{\left[\frac{d\log f_\theta(X)}{d\theta}\right]^2\right\} = -\mathbb{E}_\theta\left[\frac{d^2\log f_\theta(X)}{d\theta^2}\right]$$

Note that this is just the variance of the score under the true θ (since the expectation is zero). Intuitively, it tells us about curvature of the log-likelihood around this maximum point. If the curve is more peaked at that point, the estimate is probably better - if it is flat then the estimate is not much better than others nearby.

- **Fisher information under smooth transformations:** Let $\theta = \theta(\nu)$ be a smooth function of ν, so $g_\nu = f_{\theta(\nu)}$ is a smooth reparameterization of the model (but not necessarily 1-1). Then the information number $I_g(\nu)$ for the model $\{g_\nu(x)\}$ parameterized by ν is related to $I_f(\theta)$ for the model $\{f_\theta(x)\}$ by:

$$I_g(\nu) = I_f(\theta(\nu))\left(\frac{d\theta}{d\nu}\right)^2$$

- **Fisher information under i.i.d. observations:** Suppose $X = (X_1, ..., X_n)$is a set of n i.i.d. observations with $X_i \sim \{f_\theta(x_i)\}$, a regular 1-parameter family. Then $f_\theta(x) = \prod_{i=1}^n f_\theta(x_i)$, and the information number for the data is

$$I_X(\theta) = nI_{X_i}(\theta)$$

  where I_{X_i} is the information number for a single observation. Thus, the CR lower bound is $O(1/n)$.

Cramér-Rao-Fréchet lower bound for 1-D parameter: We start with the 1-parameter case where $\theta \in \mathbb{R}$. Let $T(X)$ be any real valued statistic such that $\mathbb{E}_\theta[T(X)] < \infty\forall\theta$. Assume that $I_X(\theta) > 0$. Then the **information inequality/CRF lower bound** states that

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

Equality holds iff and only if the score function is a linear function of $T(X)$ (i.e. if $\{f_\theta(x)\}$ is a 1-parameter exponential family).

CR Lower bound under smooth transformations of θ: Suppose that $T(X)$ is an unbiased estimator of a one-dimensional $\tau(\theta)$, a smooth function of θ. Then

$$\mathrm{Var}_\theta(T(X)) \geq \frac{[\tau'(\theta)]^2}{I_X(\theta)},$$

is an intrinsic lower bound that depends on the function $\tau(\theta)$ to be estimated and the model $\{f_\theta(x)\}$. Again, equality holds iff $\{f_\theta(x)\}$ is a 1-parameter exponential family.

Fisher information matrix (FIM): Now we extend these concepts to the case where $\theta = (\theta_1, ..., \theta_k)$. For any smooth function $g(\theta)$, let $\nabla_\theta g(\theta) = \left(\frac{\partial g}{\partial \theta_1}, ..., \frac{\partial g}{\partial \theta_k}\right)'$ be a $k \times 1$ gradient vector. Here, we define the **Fisher information matrix** to be

$$I_X(\theta) \equiv \{I_{ij}(\theta) | i, j = 1, ..., k\} : k \times k,$$

which is positive semi-definite because

$$I_X(\theta) = \mathbb{E}_\theta \left[[\nabla_\theta \log f_\theta(X)][\nabla_\theta \log f_\theta(X)]'\right]$$

- Note that
$$I_{ij(\theta)} = -\mathbb{E}\left[\frac{\partial^2 \log f_\theta(X)}{\partial\theta_i \partial\theta_j}\right]$$
- Information is additive for independent data! The FIM $I(\theta)$ for n i.i.d. observations is $nI_{X_1}(\theta)$. More generally, for $X = (U, V)$, $U \perp V$, $U \sim g_\theta(u)$,and $V \sim h_\theta(v)$, we have $I_X(\theta) = I_U(\theta) + I_v(\theta)$.

Cramér-Rao-Fréchet lower bound for $k > 1$: Assume that $I_X(\theta)$ is positive definite. For any real valued statistic $T(X)$ such that $\mathbb{E}_\theta[T(X)]^2 < \infty \quad \forall\theta$, then **information inequality/CRF lower bound** states that

$$\mathrm{Var}_\theta\left(T(X)\right) \geq \{\nabla_\theta \mathbb{E}[T(X)]\}'[I_X(\theta)]^{-1}\{\nabla_\theta \mathbb{E}[T(X)]\}.$$

If equality holds here, then T must be a linear combination of the score vector, but this does not imply exponential family here!

CR lower bound under smooth transformations of θ in higher dimensions: Suppose that $T(X)$ is an unbiased estimator of $\tau(\theta)$, a smooth function of θ. Then

$$\mathrm{Var}_\theta[T(X)] \geq [\nabla_\theta \tau(\theta)]'[I_X(\theta)]^{-1}[\nabla_\theta \tau(\theta)]$$

Nuisance parameters: Suppose that $\theta = (\theta_1, ..., \theta_k)$ but the quantity $\tau \equiv \tau(\theta_1)$ to be estimated depends only on θ_1. In this context, $(\theta_2, ..., \theta_k)$ are considered **nuisance parameters**.

- If $(\theta_2, ..., \theta_k)$ are known, then the 1-parameter CR lower bound is appropriate and takes the form
$$\mathrm{Var}_\theta(T(X)) \geq \frac{(\frac{d\tau}{d\theta_1})^2}{I_{11}(\theta)}$$
for an unbiased estimator T of τ.

- If $(\theta_2, ..., \theta_k)$ are unknown, then the k-parameters CR bound is appropriate. Here, $\nabla_\theta \tau = \left(\frac{d\tau}{d\theta_1}, 0, ..., 0\right)^T$, so the bound becomes

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

where the information matrix is now partitioned as

$$I_X(\theta) = \begin{pmatrix} I_{11}(\theta) & \underbrace{I_{12}(\theta)}_{1\times(k-1)} \\ \underbrace{I_{21}(\theta)}_{(k-1)\times 1} & \underbrace{I_{22}(\theta)}_{(k-1)\times(k-1)} \end{pmatrix}$$

and

$$I_{11\cdot 2}(\theta) \equiv I_{11}(\theta) - I_{12}(\theta)[I_{22}(\theta)]^{-1} I_{21}(\theta).$$

Relationship between information and sufficiency:

- **Information and sufficiency:** Let $X \sim \{f_\theta(x)\}$ be a regular family of pdfs of $\mathcal{X}$ (here, we mean Θ is an open set and f_θ is differentiable for almost every x). Let $T \equiv T(X)$ be a statistic with induced pdf family $\{g_\theta(t)\}$ on the range $\mathcal{T}$. Then T provides no more information about θ than does X, that is

$$I_X(\theta) \geq I_T(\theta) \qquad \text{i.e. } I_X(\theta) - I_T(\theta) \text{ is psd } \forall \theta.$$

Equality holds if and only if T is sufficient.
- Suppose that $\tau(\theta)$ is differentiable and that the model $\{f_\theta(x)\}$ satisfies the condition: *For each* $\theta \in \Omega$ there exists and open neighborhood $U(\theta) \subset \Omega$ of θ and a function $G(x;\theta)$ such that $\mathbb{E}_\theta[G(X;\theta)] < \infty$ and $\left[\frac{f_\phi(x)}{f_\phi(x)} - 1\right]^2 \leq (\phi - \theta)^2 \cdot G(x;\theta) \forall \phi \in U(\theta)$ Then,

$$\exists \lim_{\psi \to \theta} \frac{[\tau(\psi) - \tau(\theta)]^2}{A(\psi, \theta)} = \frac{[\tau'(\theta)]^2}{I_X(\theta)}$$

and thus, for any unbiased estimator $T \equiv T(X)$ of $\tau(\theta)$,

$$\mathrm{Var}_\theta(T) \geq \frac{[\tau'(\theta)]^2}{I_X(\theta)}$$

Fisher-Cramer Theorem (MLE is CANE): Let $\{\tilde{\theta}^{(n)}\}$ be any weakly consistent sequence of roots of the LEQ. Assume that for $r = 1$ and $r = 2$,

$$\mathbb{E}_\theta \left[\left| \frac{d^r \log(f_\theta(X_i))}{d\theta^r} \right|_{\theta = \theta_0} \right] < \infty,$$

and that for $r = 3$ $\exists$ an open neighborhood $U(\theta_0)$ of θ_0 such that

$$\mathbb{E}_{\theta}\left[\sup_{\theta \in U(\theta_0)} \left| \frac{d^3 \log(f_\theta(X_i))}{d\theta^3} \right| \right] = Q(\theta_0) < \infty.$$

the two above conditions are called the Cramer conditions. They are local conditions since $U(\theta_0)$ is an arbitrarily small neighborhood of θ_0. By contrast, the Wald condition is global.

If the information matrix $I(\theta_0) = I_{X_i}(\theta_0)$ is positive-definite, then

$$\sqrt{n}\left(\tilde{\theta}^{(n)} - \theta_0\right) \xrightarrow{d} N_1\left(0, [I(\theta_0)]^{-1}\right),$$

so $\{\tilde{\theta}^{(n)}\}$ is a CANE sequence of estimators of θ. Thus, if the conditions for Wald's theorem hold so that the MLE sequence $\{\hat{}^{(n)}\}$ is a consistent sequence of roots of the LEQ, then $\{\hat{}^{(n)}\}$ is CANE for θ_0. Under the multi-parameter Cramer conditions, there is a *unique* weakly consistent root of the LEQs.

Note that

- **Proposition 14.22:** If the LEQs have a unique root, then it must be the unique CANE root.
- **Exponential family log-concavity:** If the pdf $f_\theta(x)$ is strictly log-concave in θ, then the log-likelihood is strictly concave on θ, so the LEQ can have at most one root in the parameter space Ω, which must be the unique CANE root. This occurs, for example, when $\{f_\theta(x_i)\}$ is a 1-parameter exponential family.
- A k-parameter exponential pdf is strictly log-concave in its natural parameter θ.
- The LEQs in the multivariate case are equivalent to the system of equations

  $$\mathbb{E}_\theta[T_i] = T_i(x), \qquad i = 1, ..., k$$

  If this system has a solution $\hat{\theta} = \hat{\theta}(x)$ in Ω, then this solution is unique and $\hat{\theta}$ is the unique MLE of θ.
- $I_X(\theta) = \text{Cov}_\theta((T_1(X)), ..., T_k(X))^T$

Note: We later show that this complicated list of assumptions can be replaced by the **Quadratic Mean Differentiability (QMD)** assumption.

Summary: Asymptotic behavior of MLEs: Let $X = (X_1, ..., X_n)$ be i.i.d. observations from a regular 1-parameter family $\{f_\theta(x)\}$, and suppose the model is differentiable in quadratic mean (QMD) with nonsingular Fisher information $I(\theta) > 0$. Let $\hat{\theta}_n$ denote the MLE. Then:

- The MLE is consistent: $\hat{\theta}_n \xrightarrow{p} \theta$,
- If $\tau(\theta)$ is a smooth (differentiable) function, then:

  $$\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta)] \xrightarrow{d} N\left(0, \frac{[\tau'(\theta)]^2}{I(\theta)}\right)$$

- So, the MLE achieves the Cramér–Rao lower bound asymptotically (it is asymptotically efficient).
- Further, MLEs are invariant under smooth reparameterization: $\tau(\hat{\theta}_n)$ is the MLE for $\tau(\theta)$. This is called the **invariance property**.

5.4 Hypothesis testing:

Testing a simple null vs. simple alternative: The simplest hypothesis testing case considers $H_0 : X \sim P_0$ vs. $H_1 : X \sim P_1$. Here, the action space has two members, $\mathcal{A} = \{\{\text{Accept } H_0\}, \{\text{Reject } H_0\}\}$, and the parameter space has two members $\{\theta_0, \theta_1\}$ which parameterize f_0 and f_1 respectively. Unless otherwise specified, assume 0-1 loss here. The **risk function** $R = R_\phi$ is a risk *vector* in this case and $R_\phi = (R_0, R_1) = (P_0[\phi \text{ rejects } H_0], P_1[\phi \text{ accepts } H_0]) = (\pi_\phi(0), 1-\pi_\phi(1))$. where P_1 is P_{f_1}. Consider the set of all attainable risk vectors $\mathcal{R} = \{R_\phi | \phi \in \Phi\}$, which is a subset of the unit square (football shaped and concave, see PSI p. 289). The souwthwest boundary of this region in $\mathbb{R}^2$ corresponds to the risk vectors of admissible tests. Note that

- ϕ is admissible $\Rightarrow$ ϕ is a Bayes test, i.e. $\phi = \phi_\psi$ for some prior $\psi = (\psi_0, \psi_1)$.
- $\phi = \phi_\psi$ is a Bayes test $\Rightarrow$ ϕ is a **likelihood ratio** (LR) test, i.e. it has the form

$$\phi(x) = \phi_c(x) = \begin{cases} 0 \text{ (accept } H_0) & \text{if } \lambda(x) = \frac{f_1(x)}{f_0(x)} < c \\ 1 \text{ (reject } H_0) & \text{if } \lambda(x) = \frac{f_1(x)}{f_0(x)} > c \\ \gamma(x) \text{ (randomize)} & \text{if } \lambda(x) = \frac{f_1(x)}{f_0(x)} = c \end{cases}$$

  for some $c \in [0, \infty]$ and some measurable function $0 \leq \gamma \leq 1$. This comes from Bayes rule (see PSI p. 292).
- **The Neyman-Pearson (NP) criterion (simple hypothesis case):** Fix $0 < \alpha < 1$. A test ϕ is a level α test for H_0 if

$$\pi_\phi(0) = \mathbb{E}_0[\phi(X)] \leq \alpha$$

  A level α test for $H_0 : \theta \in \Omega_0$ is **most powerful (MP)** level α for H_1 if

$$\pi_\phi(1) = \mathbb{E}_1[\phi(X)] = \sup_{\phi'} \pi_{\phi'}(1),$$

  where ϕ' ranges over all level α tests for H_0.
  - Note: Since ϕ is admissible $\Rightarrow$ ϕ is a Bayes test $\Rightarrow$ ϕ is a LR test AND the risk vector of any MP level α test must lie on the SW boundary of the PSI diagram (i.e.) must minimize R_1 for a given R_0, we have that *Any MP level α test must be a LR test.*
- **Neyman-Pearson Lemma:** Let ϕ be a likelihood ratio (LR) test with $c < \infty$ and set $\alpha = \mathbb{E}_0[\phi(X)]$. Then ϕ is a MP level α test for testing H_0 vs. H_1. That is, if ϕ' is any other test such that $\mathbb{E}_0[\phi'(X)] \leq \alpha$, then $\mathbb{E}_1[\phi'(X)] \leq \mathbb{E}_1[\phi'(X)]$.
- **Consistency of MP and LR tests for i.i.d. samples.** Consider the problem of testing $H_0 : X_i \sim f_0$ vs. $H_1 : X_i \sim f_1$ based on an i.i.d. sample $X_1, ..., X_n$. For each $0 < \alpha < 1$ the MP level α test ϕ_n is **consistent**, i.e. its power at H_1 approaches 1 as $n \to \infty$.

$$P_1[\phi_n(X_1, ..., X_n) \text{ rejects } H_0] \to 1.$$

Testing a composite null hypothesis and/or a composite alternative:

- **The Neyman-Pearson (NP) criterion (composite hypothesis case):** Fix $0 < \alpha < 1$. A test ϕ is a level α test for testing $H_0 : \theta \in \Omega_0$ if

$$\sup_{\theta \in \Omega_0} \pi_\phi(0) = \mathbb{E}_0[\phi(X)] \leq \alpha$$

  A level α test ϕ for testing $H_0 : \theta \in \Omega_0 vs. H_1 : \theta \in \Omega_1$ is **uniformly most powerful (UMP)** level α for H_1 if

$$\pi_\phi(\theta) = \sup_{\phi' \text{ level } \alpha} \pi_{\phi'}(\theta) \quad \forall \quad \theta \in \Omega_1.$$

  - Note that the NP criterion requires only that the *supremum* of the Type I error probabilities be controlled for $\theta \in \Omega_0$. It does not consider the detailed behavior of the power function of Ω_0.
- **One-parameter testing with one-sided alternatives:** Consider the problem of testing $H_0 : \theta = \theta_0$ and $H_1^{>} : \theta > \theta_0$, or $H_0^{\leq} : \theta \leq \theta_0$ and $H_1^{>} : \theta > \theta_0$. UMP level α tests exist for these problems when $f_\theta(x)$ has a monotone likelihood ratio.
  - **Monotone likelihood ratio (MLR)** $\{f_\theta(x)\}$ has a strict monotone likelihood ratio (MLR) if there exists a real-valued statistic $T = T(X)$ such that for each pair $\theta_1 < \theta_2 \in \Omega$, the LR $\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}$ is a (strictly) increasing function of $T(x)$, i.e.,

$$\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = g_{\theta_1,\theta_2}(T(x)),$$

    where $g_{\theta_1,\theta_2}(t)$ is strictly increasing in t. Equivalently, $\{f_\theta(x)\}$ has MLR if for every $x_1 > x_2$,

$$\frac{f_{\theta_2}(x_1)}{f_{\theta_1}(x_1)} > \frac{f_{\theta_2}(x_0)}{f_{\theta_1}(x_0)}$$

    - Note: It follows from the Factorization criterion that $T(X)$ is a real-valued sufficient statistic for θ.
    - An i.i.d. sample from any 1-parameter exponential family has strictly MLR (PSI example 18.10), but many others do not have this property.
  - **PSI Theorem 18.20:** Let $f_\theta(x)$ have MLR in T and let $\phi = \phi(T)$ be the test

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

    where c_α and γ_α are chosen to satisfy

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

    Then ϕ is UMP level α for testing H_0 vs H_1.
- **One-parameter testing with two-sided alternatives:** Let X have pdf or pmf $f_\theta(x)$ where θ is a real parameter whose parameter space Ω is an interval (possible infinite). Consider the problems of testing $H_0 : \theta = \theta_0 \quad$ vs. $\quad H_1^{/} =: \theta \neq \theta_0$, or $H_{a,b} : a \leq \theta \leq b \quad$ vs. $H_{a,b}^c : \theta < a$ or $b < \theta$.

A key point here is that UMP level α tests *do not exist* for these problems when $f_\theta(x)$ has a monotone likelihood ratio (see proof of UMP, won't work for two sided). Instead, we aim to find UMP **unbiased** level α tests.

- **Unbiased test:** A test ϕ is unbiased for $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \in \Omega_1$ if it is more likely to reject H_0 when H_0 is false than when it is true. That is, its power function π_ϕ satisfies
$$\sup_{\theta \in \Omega_0} \pi_\phi(\theta) \leq \inf_{\theta \in \Omega_1} \pi_\phi(\theta)$$
  - UMP level α tests for one-sided alternatives usually have one-sided rejection regions and are unbiased, hence they are biased for two-sided alternatives.
  - Instead, UMPU (UMP unbiased) level α tests with two-sided rejection regions often can be constructed for two-sided alternatives, especially in 1-parameter exponential families.

5.5 General testing strategies

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 $\Theta_0 = \{\theta = (\psi, \eta) : \psi = 0\}$

(i) **Wald test:** The Wald test rejects H_0 when ψ_n is far from its value under the null, using the asymptotic normality of θ_n itself under the null to determine the rejection threshold:

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

So an asymptotic α-level test rejects H_0 when $\sqrt{n} A_\theta(\hat{\psi} - \psi)$ is outside of the $(1-\alpha)$ quantile of $\chi^2(m)$

(ii) **Likelihood ratio test:** The likelihood ratio test rejects H_0 when $\inf_{\theta_0 \in \Theta_0} D_{KL}(P_\theta \| P_{\theta_0}) = \inf_{\theta_0 \in \Theta_0} P_\theta(l_\theta - l_{\theta_0})$ is large.
The LRT test uses a test statistics based on the ERM scaled by $2n$:

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

In other words, the LRT compares the log likelihood of the unrestricted MLE to the log-likelihood of the null-restricted MLE. In a QMD model, we have the expansion

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

where $\tilde{\theta}$ is between $\hat{\theta}_0$ and $\hat{\theta}$. By asymptotic results of MLEs, we have

$$\sqrt{n} I_\theta(\hat{\theta}_0 - \hat{\theta}) = \sqrt{n} I_\theta (P_n - P_0) \left(\begin{pmatrix} 0 \\ I_{\theta,22}^{-1} \dot{\ell}_{\theta,2} \end{pmatrix} - I_\theta^{-1} \dot{\ell}_\theta \right) + o_p(1) \rightsquigarrow \begin{pmatrix} N(0, A_\theta) \\ 0 \end{pmatrix}$$

where again $A_\theta = I_{\theta,11} - I_{\theta,12} I_{\theta,22}^{-1} I_{\theta,12}^{-1}$. Combining this finding with the above gives **Wilk's theorem**, which says that under H_0 :

$$L_n \rightsquigarrow \begin{pmatrix} N(0, A_{\theta_0}) \\ 0 \end{pmatrix} \begin{pmatrix} A_{\theta_0}^{-1} & \cdots \\ \cdots & \cdots \end{pmatrix} \begin{pmatrix} N(0, A_{\theta_0}) \\ 0 \end{pmatrix} \rightsquigarrow \begin{pmatrix} \chi^2(m) \\ \cdots \end{pmatrix}$$

So $L_{n,\psi} \rightsquigarrow Z^T A_\theta^{-1} Z \sim \chi^2(m)$ where $Z \sim N(0, A_\theta)$. Thus, an α-level test rejects when L_n exceeds the $1 - \alpha$ quantile of a $\chi^2(m)$ distribution.

(iii) **Score test:** Under the null, scores have mean $P_{\theta_0} \dot{\ell}_{\theta_0}$. Roughly, a score test rejects the null if the estimate of this expectation $P_n \dot{\ell}_{\theta_0,\eta}$ is far from 0. Define

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

So the restricted MLE $\hat{\theta}_0 \in \Theta_0$ follows

$$\begin{aligned} Z_n(\hat{\theta}_0) &= Z_n(\hat{\theta}_0) + \sqrt{n} P_0 \dot{\ell}_{\hat{\theta}_0} - \sqrt{n} P_0 \dot{\ell}_{\hat{\theta}_0} \\ &= \sqrt{n}(P_n - P_0) \dot{\ell}_{\hat{\theta}_0} + \sqrt{n}(P_0 \dot{\ell}_{\hat{\theta}_0} - P_0 \dot{\ell}_{\theta_0}) \\ &= \underbrace{\sqrt{n}(P_n - P_0) \dot{\ell}_{\theta_0}}_{\text{CLT}} + \underbrace{\sqrt{n}(P_0 \dot{\ell}_{\hat{\theta}_0} - P_0 \dot{\ell}_{\theta_0})}_{\text{Delta method}} + \underbrace{\sqrt{n}(P_n - P_0)(\dot{\ell}_{\hat{\theta}_0} - \dot{\ell}_{\theta_0})}_{\text{Donsker}} \end{aligned}$$

Under conditions of the score function, the third term is $o_p(1)$. By the multivariate delta method, appealing to results about asymptotic linearity with nuisance parameters, we then have

$$P_0 \dot{\ell}(\psi_0, \hat{\eta}_0) - P_0 \dot{\ell}(\psi_0, \eta_0) = - \begin{pmatrix} I_{\theta,12} \\ I_{\theta,22} \end{pmatrix} (\hat{\eta}_0 - \eta_0) + o_p(n^{-1/2}),$$

and since $\hat{\eta}_0 - \eta_0 = I_{\theta,22}^{-1}(P_n - P_0) \dot{\ell}_{\theta,2} + o_p(n^{-1/2})$, we simplify further to

$$P_0 \dot{\ell}(\psi_0, \eta_0) - P_0 \dot{\ell}(\psi_0, \eta_0) = -(P_n - P_0) \begin{pmatrix} I_{\theta_0,12} I_{\theta_0,22}^{-1} \dot{\ell}_{\theta_0,2} \\ \dot{\ell}_{\theta_0,2} \end{pmatrix} + o_p(n^{-1/2}),$$

So plugging back into the earlier expansion gives

$$Z_n(\hat{\theta}_0) = \sqrt{n}(P_n - P_0) \begin{pmatrix} \dot{\ell}_{\theta_0,1} \\ \dot{\ell}_{\theta_0,2} \end{pmatrix} - \sqrt{n}(P_n - P_0) \begin{pmatrix} I_{\theta_0,12} I_{\theta_0,22}^{-1} \dot{\ell}_{\theta_0,2} \\ \dot{\ell}_{\theta_0,2} \end{pmatrix} + o_p(1) \rightsquigarrow \begin{pmatrix} N(0, A_{\theta_0}) \\ 0 \end{pmatrix}.$$

So by the CMT and Slutsky's lemman, we have under H_0:

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

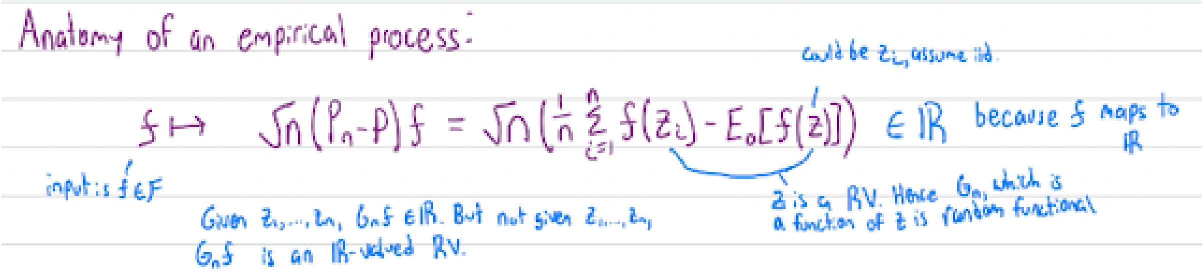
Thus, an asymptotic level α test compares the test statistic above to the $1 - \alpha$ quantile of the $\chi^2(m)$ distribution.

6 Empirical process theory

Before continuing our study of estimators, we take a detour at empirical process theory.

Empirical process: An empirical process is a stochastic process (collection of RVs) that characterizes the deviation of an empirical distribution from its expectation. For a function $f \in \mathcal{F}$, we write $\mathbb{G}_n$ for the empirical process where typically

- $$\mathbb{G}_n := \mathbb{G}_n f := \sqrt{n}(P_n - P)f = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}_0[f(Z)]\right)$$
  (a random functional: $\mathcal{F} \to \mathbb{R}$)
- And also sometimes it refers to the stochastic process itself:
  $$\mathbb{G}_n := \{\mathbb{G}_n f : f \in \mathcal{F}\},$$
  i.e. the set of RVs indexed by $\mathcal{F}$. For each $f \in \mathcal{F}$, we get an $\mathbb{R}$-valued RV.



Anatomy of an empirical process:

could be z_i, assume iid.

$$f \mapsto \sqrt{n}(P_n - P)f = \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} \underbrace{f(z_i) - E_0[f(z)]}_{\text{z is a RV. Hence } G_n \text{, which is a function of z is random functional}}\Big) \in \mathbb{R}$$ because f maps to $\mathbb{R}$

input is $f \in F$

Given $z_1, \ldots, z_n$, $G_n f \in \mathbb{R}$. But not given $z_1, \ldots, z_n$, $G_n f$ is an $\mathbb{R}$-valued RV.

Empirical process theory gives us the tools to study function-valued (i.e. infinite-dimensional) parameters. In particular, it helps us study terms of the form $P_n f$, which is the empirical measure indexed by $f \in \mathcal{F}$, uniformly over the function class $\mathcal{F}$. Often, this comes up in the form

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |(P_n - P)f|.$$

i. Showing that the above is $o_p(1)$ requires a new uniform analog to the LLN.

ii. We also are often interested in showing uniform convergence to a Gaussian process: $\{\mathbb{G}_n(f) : f \in \mathcal{F}\} := \{\sqrt{n}(P_n - P)f : f \in \mathcal{F}\} \rightsquigarrow \mathbb{G}$, which is a uniform analog to the CLT.

These are interesting for two main reasons:

1. As we will see, many statistical estimands can be written as functionals on the distribution function $\Psi(F)$. Uniform consistency and uniform convergence can guarantee consistency (when functional is continuous in supnorm) and asymptotic normality (when functional is Hadamard differentiable) of plug-in estimators $\Psi(F_n)$.
2. We are sometimes interested in function-valued estimands and summaries of such. The tools of empirical process theory allow us to build uniform (i.e. that hold uniformly over the domain) confidence sets for them.

3. There is also great interest in the study of the concentration of $\|P_n - P\|_{\mathcal{F}}$ about its mean, which can help establish performance guarantees on empirical risk minimizers (ERMS).

6.1 Empirical risk minimizers (ERMs)

Empirical risk minimzation: Suppose we observe $(X_1, ..., X_n) \overset{iid}{\sim} P$. We denote $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$ as a generic loss function (and allow elements of Θ to be possibly infinite dimensional).

- Our **goal** is to find $\hat{\theta}$ in a predefined Θ such that

$$P\ell(\cdot, \hat{\theta}) = \inf_{\theta \in \Theta} P\ell(\cdot, \theta).$$

  - So we don't care so much about approximating the true θ_0 itself, so much as finding a θ that has a similar loss.
  - Also note that θ need not uniquely determine P.
  - For convenience, we often write $\ell(\cdot, \theta) := l(\theta)$.
  - We often assume $\text{argmin}_{\theta \in \Theta} P\ell(\theta)$ is non-empty and write $\theta_0 \in \text{argmin}_{\theta \in \Theta} P(\ell(\theta))$
- **Regret:** We now use regret (rather than just loss) to measure the discrepancy between θ_0 and $\hat{\theta}$:

$$\text{Reg}(\hat{\theta}) := P\ell(\hat{\theta}) - \inf_{\theta \in \Theta}(\theta_0) = P\ell(\hat{\theta}) - P\ell(\theta_0).$$

  Note that regret is always non-negative.
- **Emprical risk minimizer:** The ERM uses the empirical distribution P_n to find $\hat{\theta}$, so

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} P_n \ell_n(\theta)$$

- **Bounding regret via empirical processes:** We can show that the regret is upper bounded by

$$\text{Reg}(\hat{\theta}) \leq 2\|P_n - P\|_{\mathcal{F}},$$

  where $\mathcal{F} := \{x \mapsto \ell(x, \theta) : \theta \in \Theta\}$ is the class of loss functions (as functions of the data, indexed by $\theta \in \Theta$).
  - Note that $\|P_n - P\|_{\mathcal{F}}$ is an RV that depends on n. By Markov/McDiarmid inequalities, bounds on $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]$ will lead to probabilistic bounds on $\|P_n - P\|_{\mathcal{F}}$ of the same order, so we often focus on $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]$.

6.2 Uniform consistency for function classes on [0,1]

An important case of empirical risk minimization is when $\mathcal{F}$ is a collection of $[0, 1]$-valued functions (i.e. 0-1 loss). In this case, $\|P_n - P\|_{\mathcal{F}}$ satisfies the bounded differences inequality (trivially with bound $1/n$), so by McDiarmid's inequality we have

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

Hence, to bound tails of $\|P_n - P\|_{\mathcal{F}}$ it suffices to bounds its expectation (if $\mathcal{F}$ isn't 0-1 valued, we can use Markov's inequality). We do this by a symmetrization argument with Rademacher complexity.

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 a value in $\{-1, 1\}$ with probability $1/2$. In this sense, it measures the degree to which a class of functions $\mathcal{F}$ can fit to random noise. It is defined

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

where ϵ_i are iid Rademacher RVs.

- Note that we call $R_n : \mathcal{F} \to \mathbb{R}$ a **Rademacher process** if $R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)$. So the Rademacher complexity is the expectation of the supnorm of the Rademacher process.

Bounding $\|P_n - P\|_{\mathcal{F}}$ via Rademacher Complexity: We can show that

$$\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \leq 2\mathbb{E}[\|R_n\|_{\mathcal{F}}]$$

VdV and Wellner Lemma 2.3 [More general upper bound]: Let Φ be any non-decreasing convex function. Then,

$$\mathbb{E}[\Phi(\|P_n - P\|_{\mathcal{F}})] \leq \mathbb{E}[\Phi(2\|R_n\|_{\mathcal{F}})]$$

Lower bound for $\|P_n - P\|_{\mathcal{F}}$: If $\mathcal{F}$ is a collection of $[0, 1]$-valued functions, then

$$\frac{1}{2}\mathbb{E}[\|R_n\|_{\mathcal{F}}] - \sqrt{\frac{\log 2}{2n}} \leq \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \leq 2\mathbb{E}[\|R_n\|_{\mathcal{F}}].$$

Moreover, with probability at least $1 - 2\exp(-2nt^2)$, we have

$$\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] - t \leq \|P_n - P\|_{\mathcal{F}} \leq \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] + t$$

Vapnik-Chervonenkis (VC) Dimension: Again, consider 0-1 loss functions $\mathcal{F} := \{x \to \{0, 1\} : x \in \mathcal{X}\}$. Define the following

- The **projection** of $\mathcal{F}$ onto points $x_1, ..., x_n$ as

  $$\mathcal{F}_{x_1,...,x_n} := \{f(x_1), ..., f(x_n) : f \in \mathcal{F}\}$$

  Note that $\mathcal{F}_{x_1,...,x_n} \subseteq \{0, 1\}^n$, so $\underbrace{|F_{x_1,...,x_n}|}_{\text{cardinality}} \leq 2^n$ for all $(x_1, ..., x_n)$.
- We say that $\mathcal{F}$ **shatters** a set of points $(x_1, ..., x_n)$ if $|F_{x_1,...,x_n}| = 2^n$.
- We then define the **growth function/shattering number** as:

  $$\Pi_{\mathcal{F}}(n) := \sup_{x_1,...,x_n} |\mathcal{F}_{x_1,...,x_n}| \qquad \text{(i.e. the largest cardinality of } \mathcal{F}_{x_1,...,x_n}\text{)}$$

- We also define the **growth function for a family of sets** (family of sets is a set of sets). Let $\mathcal{A}$ be a family of subsets of $\mathcal{X}$. Then we define

  $$\Pi_{\mathcal{A}}(n) := \Pi_{\mathcal{F}}(n), \text{ where } \mathcal{F} = \{x \mapsto \mathbb{I}[x \in \mathcal{A}] : A \in \mathcal{A}\}$$

  The properties of growth functions include (here $\mathcal{A}$ and $\mathcal{B}$ are two families of sets):

i. $\Pi_{\mathcal{A}}(n+m) \leq \Pi_{\mathcal{A}}(n)\Pi_{\mathcal{A}}(m)$
ii. $\Pi_{\mathcal{A}\cup\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n) + \Pi_{\mathcal{B}}(n)$
iii. $\Pi_{A\cup B: A\in\mathcal{A}, B\in\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n)\Pi_{\mathcal{B}}(n)$
iv. $\Pi_{A\cap B: A\in\mathcal{A}, B\in\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n)\Pi_{\mathcal{B}}(n)$

The **VC dimension** is the largest natural number n such that there exists some collection $x_1, ..., x_n$ shattered by $\mathcal{F}$.

- The VC dimension of a class of sets $\mathcal{A}$ is

$$\text{VC}(\mathcal{A}) := \sup\{n \in \mathcal{N} : \Pi_{\mathcal{A}}(n) = 2^n\}$$

- The VC dimension of a class of Boolean-valued functions $\mathcal{F}$ is

$$\text{VC}(\mathcal{F}) := \sup\{n \in \mathcal{N} : \Pi_{\mathcal{F}}(n) = 2^n\}$$

- For more general classes of $\mathbb{R}$-valued functions, the VC dimension is defined as the VC dimension of the collection of subgraphs, namely

$$\mathcal{A} := \{(x,t) \in \mathcal{X} \times \mathbb{R} : t < f(x) : f \in \mathcal{F}\}$$

In other words, the VC dimension is the cardinality of the largest set of points that the algorithm can shatter (i.e. perfectly classify).

VC index: The VS index is the VC dimension + 1.

Upper bound for VC dimension by number of operations: Consider a family of boolean-valued functions $\mathcal{F} := \{x \mapsto f(x,\theta) : \theta \in \mathbb{R}^p\}$, where for each $f : \mathbb{R}^m \times \mathbb{R}^p \mapsto \{0,1\}$. Suppose that f can be computed using no more than t operations of the following types:

i. arithmetic $(+, -, \times, /)$
ii. comparisons $(>, \geq, <, \leq, \neq, =)$

Then,

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

Finite class lemma (Massart): Let $A \subset \mathbb{R}^n$ with $|A| < \infty$ and $R := \max_{a\in A} \|a\|_2$. Also let $\epsilon_1, ..., \epsilon_n$ be iid Rademacher RVs. Then,

$$\mathbb{E}\left[\max_{a\in A}\left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i a_i\right|\right] \leq \frac{R\sqrt{2\log(2|A|)}}{n}$$

This is useful because it allows us to show the following lemma:

Lemma 4.1: Rademacher complexity upper bound: If $\mathcal{F}$ is a function class satisfying $\sup_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}} |f(x)| \leq 1$, then

$$\mathbb{E}[\|R_n\|_{\mathcal{F}}] \leq \sqrt{\frac{2\log(2\mathbb{E}[\mathcal{F}_{x_1^n}])}{n}},$$

where $x_1^n = (x_1, ..., x_n)$ and we recall $\mathcal{F}_{x_1^n} = \{(f(x_1), ..., f(x_n)) : f \in \mathcal{F}\}$. Recalling that for Boolean-valued functions $\Pi_{\mathcal{F}}(n) = \sup_{x_1,...,x_n} |F_{x_1,...,x_n}|$, we have that if $\mathcal{F}$ is Boolean-valued then

$$\frac{1}{2}\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \leq \mathbb{E}[\|R_n\|_{\mathcal{F}}] \leq \sqrt{\frac{2\log(2\Pi_F(n))}{n}}$$

- Notice that if $\Pi_{\mathcal{F}}(n) = 2^n$ then this bound is trivial, but if $\mathcal{F}$ is VC and $n > \text{VC}(\mathcal{F})$, then $\Pi_{\mathcal{F}}(n) < 2^n$ and this bound is useful.
- If $\log \Pi_{\mathcal{F}}(n) = o(n)$, then $\|R_n\|_{\mathcal{F}}$ is controlled.

This result is especially powerful when combined with Sauer's lemma

Sauer's lemma: This gives an even stronger result than Lemma 4.1 (and uses Lemma 4.1) in the case where $\text{VC}(\mathcal{F}) \leq d$. If $\text{VC}(\mathcal{F}) \leq d$, then

$$\Pi_{\mathcal{F}}(n) \leq \sum_{k=0}^{d} \binom{n}{k},$$

and consequently,

$$\Pi_{\mathcal{F}}(n) \leq \begin{cases} 2^n & \text{if } n \leq d \\ (e/d)^d n^d & \text{if } n > d \end{cases}$$

So if $\mathcal{F}$ is Boolean-valued, we can combine this result with Lemma 4.1 to get

$$\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \leq 2\sqrt{\frac{2\log 2 + 2d\log(en/d)}{n}} = O\left(\sqrt{\frac{logn}{n}}\right)$$

- This is tightest when $d = \text{VC}(\mathcal{F})$.
- Sauer's lemma tells us about how much smaller we can expect $\Pi_{\mathcal{F}}(n)$ to be than 2^n when $n > \text{VC}(\mathcal{F})$. It says the growth function attains polynomial order in this case, which leads to control of Rademacher complexity and uniform norm.

Lemma 4.3.3: Let A be a finite set and let u be a class of subsets of A. Then

$$|u| \leq |\{B \subset A | B \text{ is shattered by } A\}|$$

6.3 Uniform consistency for $\mathbb{R}$-valued function classes

The above results applied to boolean-valued function classes $\mathcal{F}$. We now seek to ensure uniform consistency of the ERM over richer classes of functions. Let $\mathcal{F}$ be a class of $\mathbb{R}$-valued functions. There are two primary approaches to show uniform consistency here. The first is to use bracketing/covering numbers and the Glivenko-Cantelli theorem. The second is to define a canonical Rademacher process and use Dudley's entropy integral. We start with the first.

6.3.1 The Glivenko-Cantelli theorem

ϵ**-bracket:** An ϵ-bracket $[\ell, u]$ is the set of all functions pointwise between ℓ and u such that $\ell, u \in L^r(P)$ and $\|u - \ell\|_{L^r(P)} \leq \epsilon$.

Bracketing number: The bracketing number is the the minimal number of ϵ-brackets needed to cover $\mathcal{F}$. That is

$$N_{[]}(\epsilon, \mathcal{F}, L^r(P)) := \min\{m : \exists\{[\ell_j, u_j] : j = 1, ..., m\} \text{ s.t. } \mathcal{F} \subseteq \cup_{j=1}^{m}[\ell_j, u_j]\}$$

ϵ**-cover:** First, recall that a **pseudometric space** (S, d) is a pairing of a set S and a pseudometric d (i.e. Euclidian distance). An ϵ**-cover** of a set T is a set $T_1 \subseteq T \subseteq S$ such that for all $\theta \in \theta_1$, there exists a $\theta_1 \in T_1$ s.t. $d(\theta, \theta_1) \leq \epsilon$.

Covering number: The covering number of a set T is $N(\epsilon, T, d) = \min\{|T_1| : T_1 \text{ is an } \epsilon\text{-cover of } T\}$.

- Here, T is an arbitrary set. It could be a function class.
- The function $\epsilon \mapsto \log N(\epsilon, T, d)$ is called the metric entropy of T.
- T is **totally bounded** if for all $\epsilon > 0, N(\epsilon, T, d) < \infty$.

Relationship between bracketing and sup-norm covering number:

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

ϵ**-packing:** An ϵ-packing is a set $T_1 \subseteq T \subseteq S$ where (S, d) is a pseudometric space, such that for all $\theta, \theta' \in T_1$, $d(\theta, \theta') \geq \epsilon$. The ϵ**-packing number** of T is $M(\epsilon, T, d) := \max\{|T_1| : T_1 \text{ is an } \epsilon\text{-packing of } T\}$.

- Fact: $M(2\epsilon) \leq N(\epsilon) \leq M(\epsilon)$.

These concepts lead to the following key theorem, which gives us uniform consistency without needing Rademacher complexity fo Markov's inequality.

(General) Glivenko-Cantelli theorem: If $\mathcal{F}$ is a class of functions for which $N_{[]}(\epsilon, \mathcal{F}, L^1(P)) < \infty$ for every $\epsilon > 0$, then $\mathcal{F}$ is P-Glivenko-Cantelli, i.e.

$$\|P_n - P\|_{\mathcal{F}} = o_p(1)$$

- For the class of non-parametric Lipschitz-continuous functions on $\mathbb{R}^d$, we have $\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) = \Theta((L/\epsilon)^d)$

6.3.2 Canonical Rademacher processes and Dudley's entropy integral

Canonical Rademacher process: We now return to Rademacher complexity, recalling that it gives us bounds on $\mathbb{E}[\|P_n - P\|]_{\mathcal{F}}$. In a slight change of setting, we consider the **canonical Rademacher process (CRP)**, defined as $\{X_\theta : \theta \in T\}$, where

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

where $T \subset \mathbb{R}^n$ is an index set and ϵ is a vector of Rademacher RVs. Note that

- The CRP is a **stochastic process**, i.e. a collection of RVs indexed by some set (in this case T).
- The CRP is **zero-mean**: $\mathbb{E}[X_\theta] = 0$ for all $\theta \in T$.
- The CRP is a **sub-Gaussian (sG) stochastic process** wrt the Euclidian distance metric. Definition: If (S, d) is a pseudometric space and $T \subset S$, then a stochastic process $\{X_\theta : \theta \in T\}$ is sG if for all $\theta, \theta' \in T$ and all $\lambda \in \mathbb{R}$,

  $$\mathbb{E}[\exp(\lambda(X_\theta - X_{\theta'}))] \leq \exp\left[\frac{\lambda^2 d(\theta, \theta^2)}{2}\right],$$

  i.e. if $(X_\theta - X_{\theta'})$ is an sG RV with parameter $\sigma^2 = d(\theta, \theta')^2$.

These properties give us a nice bound for the CRP:

$$\mathbb{E}[\|R_n\|_{\mathcal{F}}] \leq 2\delta + 2\mathbb{E}[D_{Z_1^n}]n^{-1} \sup_Q \sqrt{\log 2N(\delta, \mathcal{F}, L^2(Q))},$$

where $\delta \equiv n^{-1/2}\epsilon$, Q is the set of discrete distributions on the sample space of Z and Z_1^n is $Z_1, ..., Z_n$ from P. Further, if $\mathcal{F} := \{g(\cdot, \beta) : \beta \in B\}$ is a set of functions bounded in their index parameters, then $\mathbb{E}[\|R_n\|_{\mathcal{F}}] = O\left(L\sqrt{\frac{p\log(Ln)}{n}}\right)$. But, this is a loose bound! The $\log(n)$ term is bad!

Dudley's entropy integral: This gets us a tighter bound using a chaining argument. For any mean-zero sG process wrt pseudometric d with D as the diameter of the index set T:

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

Moreover, if $\{X_\theta : \theta \in T\}$ is seperable, then

$$\mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] \leq 8\int_0^D \sqrt{\log(N(\tilde{\epsilon}, T, d))} d\tilde{\epsilon}$$

Note that when $\log N(\epsilon) = C\epsilon^{-r}$ for some $C > 0$, then the latter integral only exists for $r < 2$. In these cases, the second bound is not informative, but the first may still be.

Application of Dudley's entropy integral to Rademacher complexity: If $\mathcal{F}$ is a real-valued function class that is closed under negations (i.e. $\mathcal{F} = -\mathcal{F}$), then

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

Note that if $\mathcal{F}$ is bounded in $[-D, D]$, we can replace ∞ with D. This tells us that if the entropy integral is finite, then $\mathbb{E}[\|P_n - P\|]_{\mathcal{F}} = O(n^{-1/2})$ and we control $\|P_n - P\|_{\mathcal{F}}$ to get a uniform LLN.

Bracketing integral bound: We can also control $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}]$ directly using a bracketing integral. There exists a constant $C > 0$ such that for any class $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ with envelope function F (i.e. $|f(z)| \le F(z) \forall z \in \mathcal{Z}$),

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

where $\|F\| = (PF^2)^{1/2}$. So if the bracketing integral is finite, then $\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] = O(n^{-1/2})$ and the empirical process $\sqrt{n}(P_n - P)f = O(1)$ converges to a tight limit process uniformly in $\mathcal{F}$.

6.4 Uniform convergence of empirical processes

6.4.1 More details

$\ell^\infty(\mathcal{F})$ metric space: In order to discuss the stochastic convergence (i.e. convergence of stochastic process $\{\mathbb{G}_n f : f \in \mathcal{F}\}$), we must define a metric space. A common one is

$$\ell^\infty(\mathcal{F}) := \{H : \mathcal{F} \to \mathbb{R}\} \text{ s.t. } \sup_{f \in \mathcal{F}} |H(f)| < \infty$$

equipped with the uniform norm $H \mapsto \|H\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |H(f)|$.

- We say a RV H is $\ell^\infty(\mathcal{F})$-valued to mean that for any event ω, the realization $H(\omega)$ of H is a (deterministic) element of $\ell^\infty(\mathcal{F})$.
- This is useful because $\mathbb{G}_n f$ is a map $\mathcal{F} \to \mathbb{R}$ – $f \in \mathcal{F}$ are not random – the randomness only arises through data.
- The data (which determine P_n) are random, so we can say $H = f(Z(\omega))$ is an $\ell^\infty(\mathcal{F})$-valued RV.

Continuous mapping theorem for $\mathbb{D}$-valued RVs: Let $(\mathbb{D}, d)$ be a metric space. Suppose that $X_1, X_2, ..., X_n$ are $\mathbb{D}$-valued RVs, and that X is a $\mathbb{D}_0$-valued RV with $\mathbb{D}_0 \subset \mathbb{D}$. If $X_n \rightsquigarrow X$ in $\mathbb{D}$, then $f(X_n) \rightsquigarrow f(X)$ in $\mathbb{R}$ for any $f : \mathbb{D} \to \mathbb{R}$ continuous on $\mathbb{D}_0$

- Note that the function $g : \ell^\infty(\mathcal{F}) \to \mathbb{R}$ is such that $g(H) = \|H\|_{\mathcal{F}}$ is continuous wrt the uniform norm (although there may be measurability issues if $\mathcal{F}$ is uncountable), since for an arbitrary sequence $H_k \to H$ in $\ell^\infty(\mathcal{F})$

  $$|g(H_k) - g(H)| = |\|H_k\|_F - \|H\|_F| \underbrace{\le}_{\text{reverse triange inequality}} \|H_k - H\|_{\mathcal{F}} \to 0,$$

  so by the CMT $g(\mathbb{G}_n) \rightsquigarrow g(\mathbb{G})$.

Asymptotically uniform ρ-equicontinuity: Let $\rho : \mathcal{F} \times \mathcal{F} \to [0, \infty)$ be a pseudometric on $\mathcal{F}$. For $\delta > 0$, define the class:

$$\mathcal{F}(\delta) := \{(f_1, f_2) : f_1, f_2 \in \mathcal{F}, \rho(f_1, f_2) < \delta\} \subset \mathcal{F} \times \mathcal{F}.$$

A sequence of stochastic processes X_n on $\mathcal{F}$ is asymptotically uniform ρ-equicontinuous if for all deterministic positive sequences $\delta_n \to 0$, it holds that

$$\sup_{(f_1, f_2) \in \mathcal{F}(\delta_n)} |X_n(f_1) - X_n(f_2)| = o_p(1)$$

Alternative characterization of weak convergence: Recall that a sequence of random elements Z_n **converge weakly in** $\mathcal{Z}$ (relative to ρ) to a random element $\mathcal{Z}$ (where both lie in a metric space $(\mathcal{Z}, \rho)$) if $\mathbb{E}[g(Z_n)] \underset{n\to\infty}{\to} \mathbb{E}[h(Z)]$ for every bounded, ρ-continuous function $h : \mathcal{Z} \to \mathbb{R}$. (By Portmanteau's lemma, we only need Lipschitz continuity).

This "characterization theorem" states that X_n converges weakly in $\ell^\infty(\mathcal{F})$ (relative to the uniform norm) to a tight random element X is and only if both

- Convergence in distribution of marginals: For each finite collection $\{f_j : j = 1, 2, .., m\} \subseteq \mathcal{F}$, it holds that $\{X_n(f_j) : j = 1, 2, ..., m\} \rightsquigarrow \{X(f_j) : j = 1, 2, ..., m\}$.
- Existence of a suitable pseudometric: $\exists$ a pseudometric ρ on $\mathcal{F}$ so that both
  - $\mathcal{F}$ is not too large: $(\mathcal{F}, \rho)$ is totally bounded, i.e. there exists an $\epsilon > 0$ such that $N(\epsilon, \mathcal{F}, \rho) < \infty$.
  - X_n is sufficiently smooth: X_n is asymptotically uniform ρ-equicontinuous.

Note: A random element is called **uniformly tight** if for all $\epsilon > 0$, there exists a compact set $K = K_\epsilon$ such that $P_0(X \in K) \geq 1 - \epsilon$.

Note: Recall $(\mathcal{F}, \rho)$ is **totally bounded** if for all $\epsilon > 0$, there exists a finite ϵ-covering of $\mathcal{F}$ under ρ, i.e. $N(\epsilon, \mathcal{F}, \rho) < \infty$.

6.4.2 Donsker and Glivenko-Cantelli classes and theorems:

We start by presenting two key results. Suppose we are interested in uniform inference about the random function $t \mapsto F_n(t)$ over its domain. We may also be interested in whether plug-in estimators $\Psi(F_n)$ are consistent and asymptotically normal (in some sense) for $\Psi(P_0)$. We now have the tools to understand and prove the following fundamental results.

(Classical) Glivenko-Cantelli Theorem (CDFs): Suppose $X_1, ..., X_n \overset{iid}{\sim} F_0$. Then $\|F_n - F_0\|_\infty = \sup_t |F_n(t) - F_0(t)| \overset{a.s.}{\to} 0$. This is a speciel case of the general G-C theorem where $\mathcal{F} = \{\mathbb{I}[X \leq t] : t \in \mathbb{R}\}$. Two corollaries of this theorem are:

- Concentration of supnorm:

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

- Plug in functionals: Any plug-in estimator $\Psi(F_n)$ for a functional $\Psi(F_0)$ is almost surely consistent provided Ψ is continuous wrt the supnorm metric.

Donsker's Theorem (CDFs): Suppose $X_1, ... \overset{iid}{\sim} F_0$. 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 examples for computing confidence bounds for this CDF.

- We also call $\mathbb{G}$ a **Brownian bridge process:** $\mathbb{G}$ is a Brownian bridge process if for all fixed collections $h_1, h_2, ..., h_m \in \mathcal{H}$, the random vector $(\mathbb{G}h_1, ..., \mathbb{G}h_n)$ follows $\text{MVN}(0, \Sigma)$ where for any j, k, $\Sigma_{jk} := F_0(\min(t_j, t_k) - F_0(t_j)F_0(t_k))$

These results of course are not exclusive to CDFs (i.e. functions of the form $t \mapsto \mathbb{I}[X \leq t]$). We now extend the ideas to arbitrary classes of functions $\mathcal{F}$.

Glivenko-Cantelli class: A class of functions $\mathcal{F}$ is P_0-G-C if

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

So the G-C (CDF) theorem above says that $\mathcal{F} = \{t \to \mathbb{I}[X \leq t] : t \in \mathbb{R}\}$ is G-C.

Donsker class: Let $\mathcal{F}$ be a collection of functions from $\mathcal{X} \to \mathbb{R}$ and let $\mathbb{G}_n$ be an empirical process with sample paths (i.e. when $\omega \in \Omega$ known) in $\ell^\infty(\mathcal{F})$. If $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$, where $\mathbb{G}_n = \sqrt{n}(P_n f - P_0 f), f \in \mathcal{F}$ for a tight random element $\mathbb{G}$, then we say $\mathcal{F}$ is P_0-Donsker. Importantly, this means (since the supremum is a continuous transformation) that $\sup_{f \in \mathcal{F}} \sqrt{n}(P_n f - P_0 f) = Op(1) \Leftrightarrow \sup_{f \in \mathcal{F}}(P_n f - P_0 f) = Op(1/\sqrt{n})$

Note: A Donsker class is also G-C, but the reverse does not necessarily hold!

Properties of Donsker class:

i. If $\mathcal{F}$ is P_0-Donsker, then $\mathbb{G}$ is a mean-zero Gaussian process with covariance function $F_0(\min(t_i, t_j) - F_0(t_i)F_0(t_j))$.

ii. If $\mathcal{F}$ has finite and P_0-integrable envelope function $\bar{F}$, i.e. a function such that $\sup_{f \in \mathcal{F}} |f(x)| \leq \bar{F}$ for all $x \in \mathcal{X}$, the sample paths of $\mathbb{G}_n$ belong to $\ell^\infty(\mathcal{F})$ since for any realization $x_1, ..., x_n$, we have

$$|\frac{1}{n} \sum_{i=1}^{n} |f(z_i)| + P_0(f) \leq \frac{1}{n} \sum_{i=1}^{n} \bar{F}(x_i) + P_0(\bar{F}) < \infty$$

iii. For a Donsker class, the pseudometric guaranteed to exist can always be taken to equal the **standard deviation pseudometric**, namely

$$\begin{aligned} \rho_{P_0} : (f_1, f_2) \mapsto \mathrm{sd}_{P_0}(f_1(X) - f_2(X)) &= \sqrt{P_0(f_1 - f_2)^2 - [P_0(f_1 - f_2)]^2} \\ &= \sqrt{\|f_1(X) - f_2(X)\|_{L_2 P_0}} \end{aligned}$$

iv. **Permanence property of Donsker class:** Let $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k$ be P-Donsker classes with $\|P\|_{\mathcal{F}_j} < \infty$ for each j. Let $\phi : \mathbb{R}^k \mapsto \mathbb{R}$ be a function for which there exists a constant $c > 0$ such that

$$|\phi \circ h(x) - \phi \circ g(x)|^2 \leq c \sum_{j=1}^{k} [h_j(x) - g_i(x)]^2$$

for every $h, g \in \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_k$ and $x \in \mathcal{X}$. Then $\phi \circ (\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k)$ is P-Donsker provided $\phi \circ (f_1, f_2, ..., f_k)$ is P-square integrable for some $f_j \in \mathcal{F}_j j = 1, 2, ..., k$.
Corollary: If $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker classes and $\|P\|_{\mathcal{F} \cup \mathcal{G}} < \infty$, then the following are also P-Donsker

(a) The pairwise infima $\mathcal{F} \wedge \mathcal{G}$ and pairwise suprema $\mathcal{F} \vee \mathcal{G}$
(b) Pairwise sums $\mathcal{F} + \mathcal{G}$ (note $\mathcal{F} + \mathcal{G} = \{x \mapsto f(x) + g(x) : f \in \mathcal{F}, g \in \mathbb{G}\}$),
(c) $\mathcal{F} \cup \mathcal{G}$
(d) If only $\mathcal{F}$ is known to be P-Donsker, then if $\mathcal{G} \subseteq \mathcal{F}$, then $\mathcal{G}$ is P-Donsker.
(e) The closures $\bar{\mathcal{F}}$ and $\bar{\mathcal{G}}$ (i.e. the set of all elements of $\mathcal{F}$ and its $L^2(P)$ limit points) is also Donsker.

Bounded variation norm implies Donkser:

- **Variation norm:** The variation norm of a function $f : \mathbb{R} \to \mathbb{R}$ is given by $\|f\|_v := \int |df(z)| = \sup_{\text{finite partitions of } \mathbb{R}} \sum |f(x_{i+1} - f(x_i))|$. It measures the total absolute deviation of f over its domain. In special cases it has a simpler form:
  - If f is differentiable w/ derivative f', the $\|f\|_v = \int \left| \frac{df(z)}{dz} \right| dz = \int |f'(z)| dz$
  - If f is a step function, then $\|f\|_v = \sum_z |\Delta f(z)|$ where $\Delta f(z) := f(z) - f(z^-)$

  For multivariate functions $f : \mathbb{R}^m \to \mathbb{R}$, we extend this idea to the **uniform sectional variation norm** defines as

$$\|f\|_v^* : -\sup_s \sup_{r_s} \|f\|_{v,r_s}$$

Now we can introduce the theorem. For each $M_0 < \infty$, $\{f : B \subseteq \mathbb{R}^m \to \mathbb{R} \text{ s.t. } \|f\|_v \leq M_0\}$ is a P_0-Donsker class for each P_0 such that $P_0 I_B = 1$.

- For $m = 1$, this takes care of all indicators for half-lines ($(-\infty, z]$ for $z \in \mathbb{R}$ - this is Donsker's theorem. Variation norm is bounded by 1 here), all univariate bounded and monotone functions (even though this class is infinite dimensional unlike half-lines, which are indexed by t), and all differentiable univariate functions defined over a bounded region and with uniformly bounded derivative (variation norm bounded by derivative times region width)
- For $m > 1$, this takes care of all indicators of quarter-planes in $\mathbb{R}^2$ and beyond, all primitives (i.e. anti-derivatives) of integrable multivariate functions, and all differentiable multivariate functions defined over a bounded region and with uniformly bounded derivative.

Strategies to show that a class is Donsker:

1. Try the characterization formula for weak convergence (usually hard).
2. Show $\mathcal{F}$ satisfies the finite bracketing integral property (VdV 19.5): For $\delta > 0$, define the bracketing integral

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

   . $\mathcal{F}$ is P-Donsker if $J_{[]}(\delta, \mathcal{F}, L^2(P)) < \infty$.
3. Show $\mathcal{F}$ satisfies the uniform entropy bound (VdV 19.14): $\mathcal{F}$ is P-Donsker if it has envelope $\bar{F}$ satisfying $P\bar{F} < \infty$ and

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

   where the supremum is over all discrete probability measures Q on $\mathcal{Z}$ and $\|\bar{F}\|_{Q,2} = Q\bar{F}^2$.
4. Use a permanence of Donsker class property.
5. Show bounded variation norm (see examples - this is most common).
6. Use partial Slutsky's theorem for weak convergence (STAT 583 HW 1).

Van der Vaart Lemma 19.24 (Conditions to show $\sqrt{n}$ asymptotic negligibility): To establish that $\sqrt{n}(P_n - P_0)g_n \xrightarrow{p} 0$, it suffices to establish that

1. $P_0 g_n^2 \to 0$ and
2. $g_n$falls in a P_0-Donsker class (w/ prob → 1).

This is a critical result! We often study terms like $(P_n - P_0)g_n$ for some random function g_n. The fact that P_n and g_n both depend on n complicates things, but if $g_n \in \mathcal{F}$ for some function class $\mathcal{F}$, then we can instead study $\sup_{f\in\mathcal{F}} |(P_n - P_0)f|$. We have shown that

$$(P_n - P_0)g_n \leq \sup_{f\in\mathcal{F}} |(P_n - P_0)f| = \begin{cases} o_p(1) & \text{if } \mathcal{F} \text{ is G-C} \\ O_p(n^{-1/2}) & \text{if } \mathcal{F} \text{ is } P_0\text{-Donsker.} \end{cases}$$

VdV Lemma 19.24 tells us that if $\mathcal{F}$ is P_0-Donsker and $P_0 g_n^2 \xrightarrow{p} 0$, then we further have that

$$|(P_n - P_0)g_n| = o_p(n^{-1/2}),$$

which means it is asymptotically negligible at a $\sqrt{n}$ rate.

7 Estimation: Part II

7.1 M and Z estimation:

7.1.1 M estimation

M-estimation: Suppose $X_1, ..., X_n \overset{iid}{\sim} P_0 \equiv P_{\theta_0}$, where P_{θ_0} belongs to a model $\mathcal{M} \equiv \{P_\theta : \theta \in \Theta\}$. Unless otherwise specified, we do not place any restrictions on Θ (could be finite dimensional parameter or functional etc.).

Suppose our aim is to estimate some function of θ_0, i.e. $\phi_0 = \Phi(\theta_0)$ for some $\Phi : \Theta \to \mathbb{R}^k$. The simplest case is $\Phi(\theta) = \theta$. Consider a user defined $S \supseteq \text{IM}(\Phi)$. Then, if $\{m_\phi : \phi \in S\}$ is a collection of $\mathcal{X} \to \mathbb{R}$ functions, then suppose

$$\phi_0 \in \text{argmax}_\phi P_0 m_\phi = \text{argmax}_\phi \mathbb{E}_{P_0}[m_\phi(X)]$$

An **M-estimator** replaces P_0 with the empirical measure.

$$\phi_n = \text{argmax}_\phi P_n m_\phi = \text{argmax}_\phi \frac{1}{n}\sum_{i=1}^{n} m_\phi(X_i)$$

Technically, we don't require that ϕ_n maximizes an expectation. We can more generally suppose that $\{M_\theta : \theta \in \Theta\}$ is a collection of $S \to \bar{\mathbb{R}}$-valued functions satisfying $\Phi(\theta) \in \text{argmax} M_\theta(\phi)$ for all $\theta \in \Theta$. Then, letting M_n denote an estimator of $M_0 \equiv M_{\theta_0}$, then $\phi_n \in \text{argmax} M_n(\phi)$ so that in our typical case $M_0(\phi) = P_0 m_\phi(X)$ and $M_n(\phi) = P_n m_\phi(X)$.

Consistency of M-estimators (VdV 5.21): In the univariate setting, we may be able to use the WLLN to prove that $\phi_n \xrightarrow{p} \phi_0$. But more generally, we require the following:

i. A near maximizer for M_n is available: The sequence of estimators ϕ_n satisfies $M_n(\theta_n) \geq \sup_\phi M_n(\theta) - o_p(1)$.

ii. ϕ_n is a well-separated maximum of M_0: For all $\epsilon > 0$, $M_0(\phi_0) > \sup_{\phi:\|\phi-\phi_0\|>\epsilon} M_0(\phi)$.

iii. M_n is uniformly consistent: $\sup_{\phi\in S} |M_n(\theta) - M_0(\theta)| \xrightarrow{p} 0$. More generally, we require that $\{m_\phi : \phi \in S\}$ is a **Glivenko-Cantelli class**: meaning $\sup_\phi |(P_n - P_0)m_\phi| = o_p(1)$. A sufficient condition for this is that the function maximized is continuous in x, § has compact support, and is dominated by an integrable function. This implies a finite bracketing number, which gives Glivenko-Cantelli.

Asymptotic linearity and normality of M-estimators (VdV 5.23): Suppose ϕ_n is a near-maximizer of $M_n := P_n m_\phi$ (i.e. $P_n m_{\theta_n} \geq \sup_\phi P_n m_\theta - o_p(n^{-1})$). Under the following conditions

i. $\phi_n \xrightarrow{p} \phi_0$ (conditions above).

ii. m_ϕ is differentiable: $\phi \mapsto m_\phi(x)$ is differentiable at ϕ_0 for P_0-almost every x with derivative $\dot{m}_\phi(x)$.

iii. For every $\phi, \tilde{\phi}$ in a neighborhood of ϕ_0 and a function $G : \mathcal{X} \to \mathbb{R}$ with $P_0 G^2 < \infty$, it holds that

$$|m_\phi(x) - \tilde{m}_\phi(x)| \leq \|\phi - \tilde{\phi}\| G(x) \text{ for all } x \in \mathcal{X}$$

iv. m_θ is sufficiently smooth: Uniform convergence under local alternatives: There exists a non-singular matrix V_{ϕ_0} such that

$$\lim_{\epsilon\to 0} \sup_{h\in\mathbb{R}^d:\|h\|=1} \frac{|P_0 m_{\phi_0+\epsilon h} - P_0 m_{\phi_0} - \frac{1}{2}\epsilon^2 h^t v_{\phi_0} h|}{\epsilon^2} = 0.$$

Note that this assumption is equivalent to assuming that $\left\{ \frac{m_\phi - m_{\phi_0} - (\phi-\phi_0)^T \cdot m_{\phi_0}}{\|\phi-\phi_0\|} : \|\phi - \phi_0\| < \epsilon \right\}$ forms a Donsker class. We can replace this hard to verify assumption with two conditions:

- Condition 1: Assume that $P_0 m_\phi$ admits a second-order Taylor expansion at ϕ_0:

$$P_0 m_\phi = P_0 m_{\phi_0} + \frac{1}{2}(\phi - \phi_0)^T V_{\phi_0}(\phi - \phi_0) + o(\|\phi - \phi_0\|^2)$$

where V_{ϕ_0} is the matrix of second derivatives of m_ϕ at ϕ_0.

- Condition 2: Lipschitz-continuity for all $x \in \mathcal{X}$ and every ϕ_1, ϕ_2 in a neighborhood of ϕ_0:

$$|m_{\phi_1}(x) - m_{\phi_2}(x)| \leq \cdot m(x)\|\phi_1 - \phi_2\|.$$

Under these conditions, recalling V_{ϕ_0} is the matrix of second derivatives of m_θ at θ_0, we have

$$\sqrt{n}(\phi_n - \phi_0) = -V_{\phi_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\phi_0}(x_i) + o_p(1) \rightsquigarrow N(0, V_{\phi_0}^{-1} P_0(\dot{m}_{\phi_0} \dot{m}_{\phi_0}^T V_{\phi_0}^{-1})$$

7.1.2 Z estimation

Z estimation: Suppose $X_1, ..., X_n \overset{iid}{\sim} P_0 \equiv P_{\theta_0}$, where P_{θ_0} belongs to a model $\mathcal{M} \equiv \{P_\theta : \theta \in \Theta\}$. Unless otherwise specified, we do not place any restrictions on Θ (could be finite dimensional parameter or functional etc.).

Suppose our aim is to estimate some function of θ_0, i.e. $\phi_0 = \Phi(\theta_0)$ for some $\Phi : \Theta \to \mathbb{R}^k$. The simplest case is $\Phi(\theta) = \theta$. Consider a user defined $S \supseteq \text{IM}(\Phi)$. Then, if $\{z_\phi : \phi \in S\}$ is a collection of $\mathcal{X} \to \mathbb{R}$ *estimating functions*, then

$$\phi_0 \text{ is a solution to } P_0 z_\phi = \mathbb{E}_{P_0}[z_{\phi_0}(X)] = 0$$

An **Z-estimator** replaces P_0 with the empirical measure.

$$\phi_n \text{ is a solution to } P_n z_\phi = \frac{1}{n}\sum_{i=1}^{n} z_\phi(X) = 0$$

Technically, we don't require that ϕ_n solves an expectation. We can more generally define $\{Z_\theta : \theta \in \Theta\}$ as a collection of $S \to \bar{\mathbb{R}}^b$-valued functions for which, for all $\theta \in \Theta$, ϕ_0 is a solution in ϕ to $Z_\theta(\phi) = 0$. Then, letting Z_n be an estimator of $Z_0 \equiv Z_{\theta_0}$, then the Z-estimator ϕ_n is the solution in ϕ to $Z_n(\phi) = 0$ so that in our typical case $Z_\theta(\phi) = P_0 z_\phi(X)$ and $Z_n(\theta) = P_n z_\phi(X)$.

Relationship between Z and M-estimators:

- $M \Rightarrow Z$: If for each θ, M_ϕ is differentiable at all ϕ, then we can take

  $$Z_\theta : \phi \to \nabla M_\theta(\phi)$$

  so that $Z_\theta(\phi) = 0$ provided ϕ falls in the interior of S. But there may be multiple solutions ϕ_n, even if M_θ has a global maximum.
- $Z \Rightarrow M$: If we let

  $$M_\theta : \phi \mapsto -\|Z_\theta(\phi)\|$$

  then the set of maximizers of $M_\theta(\phi)$ is the same as the set of roots to $Z_\theta(\phi) = 0$.

Consistency of Z-estimators in the 1-D case (VdV 5.10): Let $\phi_0 \in \mathbb{R}$. Suppose that for each $\phi, Z_n(\phi) \xrightarrow{p} Z_0(\phi)$. Suppose also that at least one of the following conditions holds:

i. Each $\phi \mapsto Z_n(\phi)$ is continuous and has exactly one zero ϕ_n.
ii. $\phi \mapsto Z_n(\theta)$ is non-decreasing and ϕ_n satisfies $Z_n(\phi_n) = o_p(1)$.

If ϕ_0 is a point such that, for all $\epsilon > 0$, $Z_0(\phi_0 - \epsilon) < 0 < Z_0(\phi_0 + \epsilon)$, then $\phi_n \xrightarrow{p} \phi_0$.

Note: When $Z_0(\phi_0) = P_0 z_\phi$ and $Z_n(\phi) = P_n z_\phi$, then this follows by the WLLN if $P_0|z_\phi| < \infty \ \forall \phi$. Also, requiring $Z_n(\phi) \xrightarrow{p} Z_0(\phi)$ is much weaker than requiring uniform consistency.

Consistency of Z-estimators in the general case (VdV 5.9): We now require the following, which are essentially the conditions for consistency of the M-estimator applied to the Z estimator:

i. A near solution for Z_n is available: The sequence of estimators ϕ_n satisfies $P_n z_{\theta_n} = Z_n(\phi_n) = o_p(1)$
ii. ϕ_n is a well-separated minimizer: For all $\epsilon > 0$:

$$0 = -\|Z_0(\phi_0)\| > -Z_0(\phi)\| \quad \forall\ \phi : \|\phi - \phi_0\| > \epsilon$$

iii. Uniform consistency of estimating equations across all $\phi \in S$:

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

Equivalently, we require that the class of estimating functions $\{z_{\phi,j} : \phi \in S, j = 1, ..., k\}$ (where k is dimension) lies in a **Glivenko-Cantelli class**: meaning $\sup_\phi |(P_n - P_0)m_\phi| = o_p(1)$. A sufficient condition for this is that the estimating function is continuous in x, § has compact support, and is dominated by an integrable function. This implies a finite bracketing number, which gives Glivenko-Cantelli.

Asymptotic linearity and normality of Z-estimators: Suppose ϕ_n and ϕ_0 are the (near) solutions in ϕ to $P_n z_\phi = o_p(n^{-1/2})$ and $P_0 z_\phi = 0$. If

i. $\phi_n \xrightarrow{p} \phi_0$ (conditions above).
ii. Conditions on estimating function: Suppose the estimating function is squared integrable, $P\|z_{\phi_0}\|^2 < \infty$ and that the function $\phi \to P_0 z_\phi$ is differentiable at a zero ϕ_0 with nonsingular Jacobian matrix V_{ϕ_0} (we sometimes call this $\dot{Z}_0$ or $\dot{u}_0$).
iii. Suppose the class of estimating functions $\{z_\phi : \phi \in S\}$ is a Donsker class. A sufficient condition is that the estimating functions are Lipschitz in their indexing parameters:

$$\|z_{\phi_1} - z_{\phi_2}\| \leq \dot{z}(x)\|\phi_1 - \phi_2\|$$

.

Then,

$$\sqrt{n}(\phi_n - \phi_0) = -V_{\phi_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{\theta_0}(X_i) + o_p(1) \rightsquigarrow N\left(0, V_{\phi_0}^{-1}P_0[z_{\phi_0}z_{\phi_0}^T](V_{\phi_0}^{-1})\right)$$

7.2 Kernel Density estimation

Kernel density estimation is a useful tools for estimating functionals that depend on local features of the data generating distribution (i.e. density, regression function). Some canonical examples are the average density and the value of a density at a point.

Kernel: A kernel is a function $f : \mathbb{R} \to \mathbb{R}$ that satisfies $\int K(u)du = 1$.

An s-**order** kernel satisfies

- $\int u^r K(u)du = 0$ for all $r \in \{1, ..., s-1\}$ and
- $|\int u^s K(u)du| < \infty$.

Kernel density estimator (KDE): A KDE is an estimator $\hat{f}$ taking the form

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

Note the following remarks:

i. If K is symmetric (i.e. $K(u) = K(-u)$), then K is at least a 2nd order kernel.

ii. Higher order ($S \geq 3$) kernels can lead to estimators with lower bias.

iii. Common choices of K include a uniform kernel $K(U) = \frac{1}{2}\mathbb{I}[|u| \leq 1]$, Epanechikov kernel $\frac{3}{4}(1-u^2)\mathbb{I}[|u| \leq 1]$ and Gaussian kernel $K(U) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$.

iv. For certain theoretical guarantees, we often user bounded kernels with bounded supports, althought the Gaussian kernel has unbounded support.

v. **Lemma:** If K is non-negative, then for any $h > 0$, $\hat{f}$ is a pdf (proof by integration with change of variable).

Bounding the MSE and MISE of KDEs: We may be interested in estimating the density at a point, in which case we can use squared error loss, or we can use integrated squared error loss. The MISE is defined as

$$\text{MISE}(\hat{f}) := \int \mathbb{E}_f[\hat{f}(x) - f(x)^2]dx.$$

In 582 Lecture and HW2, we show how we can bound the $MSE = \text{bias}^2 + \text{variance}$ of KDEs of any order at a fixed point x_0 (and that can be easily extended to MISE) using Taylor expansions if we assume that f is smooth enough. We often do this by assuming f lies in a Holder class.

Holder(β, L) class: The Holder (β, L) class of functions is the set of all $(\beta-1)$-times differentiable real-valued functions whose derivative $f^{(\beta-1)}$ satisfies

$$|f^{(\beta-1)}(x_1) - f^{(\beta-1)}(x_2)| \leq L|x_1 - x_2| \text{ for all } x_1, x_2.$$

Finding the optimal bandwidth for 2nd order kernels:

Using the Taylor expansion arguments in lecture, we get that for estimating $f(x_0)$ for a fixed x_0 using a 2nd order kernel in a Holder class with $\beta = 2$,

$$\text{MSE}(\hat{f}(x_0)) \leq \underbrace{h^4 L^2 \sigma_k^4}_{\text{bias}^2} + \underbrace{\frac{1}{nh}R(k)}_{\text{var}} + \underbrace{c/h}_{\text{constant}\to 0}$$

Only c and h depend on h, and we can pick a c that holds uniformly over any range $h \in (0, m]$. So we can optimize this bound by picking $h \propto n^{-1/5}$ so that

$$h^4 L^2 \sigma_k^4 \asymp \frac{1}{nh}R(k)$$

Finding the optimal bandwidth for 2nd order kernels:

In the more general case of a β-th order kernel, we apply a $(\ell - 1)$-th order Taylor expansion (where ℓ is the largest integer strictly less than β) at x_0 with the MVT and find that the general optimal bandwidth is $h \asymp n^{-1/(2\beta+1)}$ so that $\text{MSE} \asymp n^{-2\beta/(2\beta+1)}$

7.3 Asymptotic Linearity:

Asymptotic linearity: An estimator ψ_n of $\psi_0 \in \mathbb{R}$ based on $X_1, X_2, ... \overset{iid}{\sim} P_0 \in \mathcal{M}$ is said to be **asymptotically linear** if there exists a function $x \mapsto \phi_{P_0}(x)$ (called the **influence function**) such that under sampling from P_0:

1. $\mathbb{E}[\phi_{P_0}(X)] = 0$ and $\mathrm{Var}(\phi_{P_0}(X)) < \infty$
2. The estimator ψ_n admits the asymptotic representation

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

Asymptotically linear estimators are consistent and asymptotically normal with limiting distribution $N(0, \mathrm{Var}(\phi_{P_0}))$.

Examples of asymptotically linear estimators:

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

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

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

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

iii. p-th sample quantile: The p-th sample quantile $Q_n(p) = \psi_n$ is an asymptotically linear estimator for the p-th population quantile $Q_0(p) = \psi_0$:

$$Q_n(p) - Q_0(p) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{F_0(Q_0(p)) - \mathbb{I}[X_i \leq Q_0(P)]}{f_0(Q_0(p))}\right] + o_p(n^{-1/2})$$

iv. Estimating equations (no nuisance): Let ψ_0 be the solution in ψ to $P_0 U(\psi) = 0$ and ψ_n be a near-solution to the sample version so $P_n U(\psi) = o_p(n^{-1/2})$. If $\psi_n \xrightarrow{p} \psi_0$, and the class of estimating equations $\{U(\psi) : \psi \in S\}$ forms a P_0-Donsker class, then ψ_n satisfies:

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n}\left[\dot{u}(\psi_0)\right]^{-1} U(\psi_0)(x) + o_p(n^{-1/2})$$

where $\dot{u}(\psi_0) = \frac{d}{d\psi} P_0 U(\psi)|_{\psi=\psi_0}$.

v. Estimating equations (with nuisance): Consider the setup above, but now suppose the estimating function U also depends on a nuisance parameter η, so it is written $U(\psi, \eta)$. Then ψ_0 is the solution in ψ to $P_0 U(\psi, \eta_0) = 0$. Now suppose η_n is an ALE for η_0 with IF φ_{P_0}. Let ψ_n be the 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 (and, that $\{U(\psi,\eta)\}$ is a P_0-Donsker class), we have that ψ_n is an asymptotically linear estimator for ψ_0 with influence function:

$$\phi_{P_0}(x) := -\left(\frac{\partial}{\partial\psi}P_0U(\psi,\eta_0)|_{\psi=\psi_0}\right)^{-1}\left[U(\psi_0,\eta_0)(x) + \frac{\partial}{\partial\eta}P_0U(\psi_0,\eta)|_{\eta=\eta_0}\varphi_{P_0}(x)\right]$$

Importantly, this shows us that the asymptotic behavior of ψ_n is unchanged by the introduction of nuisance parameter estimation in the special case that $\frac{\partial}{\partial\eta}P_0U(\psi_0,\eta)|_{\eta=\eta_0} = 0$.

Delta method for influence curves: Let ψ_n be an asymptotically linear estimator of $\psi_0 \in \mathbb{R}^p$ with influence curve ϕ_{P_0}. Suppose that $h : \mathbb{R}^p \to \mathbb{R}$ is differentiable at ψ_0 with derivative $h'(\psi_0) \neq 0$ (i.e. not the 0 vector). Then, we have

$$h(\psi_n) = h(\psi_0) + \frac{1}{n}\sum_{i=1}^{n} h'(\psi_0)^T\phi_{P_0}(X_i) + o_p(n^{-1/2})$$

So $h(\psi_n)$ is an asymptotically linear estimator for $h(\psi_0)$ with influence function

$$x \mapsto h'(\psi_0)^T\phi_{P_0}(x)$$

7.4 V/U statistics

Many parameters of interest can be written as a statistical functional

$$V(P) = \int\int\cdots\int H(x_1,...,x_m)dP(x_1)\cdots dP(x_m)) := P^mH$$

for some **symmetric** function $H : \mathcal{X}^m \to \mathbb{R}$, which we call the kernel function.

- For a given m-variate function f, we use the notation P_0f to refer to the function $P_0f : (x_1,x_2,...,x_{m-1}) \mapsto \int f(x_1,x_2,...,x_m)dP(x_m)$. We recursively define $P_0^{m-k}H$ as the mapping $(x_1,x_2,...,x_k) \mapsto \int\int\cdots\int f(x_1,x_2,...,x_k,x_{k+1},...,x_m)dP_0(x_m)\cdots dP_0(x_{k+1})$, i.e. integrating out the last $m-k$ entries.
- We also use the notation H_k to refer to the k-variate function $P_0^{m-k}H$.
- We can always assume H is symmetric, since if it is not we can symmetrize it my making it invariant to permutations in its arguments. Symmetrized version is $\tilde{H}(x_1,x_2) = \frac{H(x_1,x_2)+H(x_2,x_1)}{2}$. With symmetry, we have important fact that $(P_n-P_0)P_0H = P_0(P_n-P_0)H$

Suppose we are interested in estimating $V_0 := V(P_0)$ using $X_1,X_2,...,X_n \overset{iid}{\sim} P_0$. Assume $P^mH^2 < \infty$.

V-statistic: We define the natural plug-in estimator of $V(P_0)$ to be the V-statistic:

$$V_n := V(P_n) = P_n^mH = \frac{1}{n^m}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_m=1}^{n} H(X_{i_1},X_{i_2},...,X_{i_m}).$$

In other words, V_n is the average of H over the cartesian product of the sample with itself m times.

Some examples of V-statistics include:

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

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

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

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

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

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

and V-statistic

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

iv. Cramer-von Mises goodness-of-fit criterion: $V(P) = \int [F_P(x) - F_P^*(x)]^2 F^*(dx)$ for a 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

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

Asymptotic representation of V-statistics: Suppose H is symmetric in its arguments (for a non-symmetric H, we can always symmetrize it). Then for any $m \in \{1, 2, ...\}$, we have the representation

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

where $H_k := P_0^{m-k}H$. Letting $\tau_k^2 = \mathrm{Var}(H_k(X_1, ..., X_k))$ 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 say that the V-statistic is **non-degenerate** and get the sum is dominated by the $m(P_n - P_0)H_1$ term, and so (provided the kernel $H \in \mathcal{H}$ is in a Donsker class) V_n is asymptotically linear for V_0 with influence function $m(H_1(x) - V_0)$.

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

U-statistic: In general, V-statistics are biased estimators of V_0 in finite samples. This is because the V statistics sum over indices that are not always unique. A U-statistic is just like a V-statistic, only H is averaged out only over unique sets of indices:

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

where $\mathcal{D}_{m,n} := \{\underset{\sim}{i}_m \subset \{1, ..., n\} := (i_1, ..., i_2, ..., i_m) : 1 < i_1 \cdots < i_m < n\}$.

When $\text{Var}(H_1(X)) > 0$ (non-degenerate case), V and U statistics are asymptotically equivalent.

Asymptotic representation of U-statistics: Consider the $m = 2$ case. Define

$$V_n := \frac{1}{n^2}\sum_{i,j} H(X_i, X_j), \qquad U_n := \frac{1}{n(n-1)}\sum_{i\neq j} H(X_i, X_j), \qquad D_n := \frac{1}{n}\sum_{i=1}^{n} H(X_i, X_j).$$

Then notice that

$$\begin{aligned} V_n &= \left(1 - \frac{1}{n}\right) U_n + \frac{1}{n} D_n \\ \Rightarrow U_n - V_n &= \frac{1}{n}(U_n - D_n) \\ \Rightarrow n^{1/2}(U_n - V_n) &= n^{-1/2}(U_n - D_n) = O_p(n^{-1/2}) \qquad \text{(WLLN)} \end{aligned}$$

So $U_n = V_n + O_p(n^{-1})$, implying $U_n = V_n + o_p(n^{-1/2})$ and hence

$$U_n - V_0 = (V_n - V_0) + (U_n - V_n) = m(P_n - P_0)H_1 + o_p(n^{-1/2}),$$

i.e. V and U statistics are asymptotically equivalent in the non-degeneracy case.

7.5 Functional Delta method

Under our previous delta method, we saw that if θ_n was asymptotically linear for θ_0, then we could study the asymptotic behavior of an estimator $h(\theta_n)$ for $h(\theta_0)$ for a *fixed, differentiable function* $h : \mathbb{R}^p \to \mathbb{R}$. Assuming the conditions were met, it gave that

$$h(\theta_n) - h(\theta_0) = \dot{h}(\theta_0)(\theta_n - \theta_0) + o_p(\|\theta_n - \theta_0\|),$$

where $\|\cdot\|$ can be any norm since all norms are equivalent in finite-dimensional spaces.

We now consider a more general setup where $\psi_n = \Psi(F_n)$, where Ψ is a **fixed functional**: $\mathcal{P} \to \mathbb{R}$ for a rich class of distribution functions $\mathcal{P}$ and F_n is the empirical distribution function.

Gauteaux Derivative: Suppose $\mathcal{P}$ is a (typically very rich) convex collection of distribution functions, and denote for any $F \in \mathcal{P}$ the **cone** (i.e. space of directions) $Q(F) := \{a(F_1 - F) : F_1 \in \mathcal{P}, a > 0\}$. The Gauteaux derivative of $\Psi : \mathcal{P} \to \mathbb{R}$ at $F \in \mathcal{P}$ in the direction of $h \in Q(F)$ is:

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

provided this limit exists.

- We can represent an arbitrary $h \in Q(F)$ as $h = c(F_1 - F)$ for some $c > 0, F_1 \in \mathcal{P}$. Then, we note that $F + \epsilon h = F + \epsilon c(F_1 - F) = F(1 - \epsilon c) + \epsilon c F_1 \in \mathcal{P}$, making clear that as $\epsilon \to 0$, we walk on the line of convex combinations of F and F_1 towards F, and this idea is well defined for $0 \leq \epsilon \leq c$.

- Note $\dot{\Psi}(F;h) \in \mathbb{R}$. In plain words, this is the rate of change in the value (in $\mathbb{R}$) of the functional Ψ as the (function-valued) input gets arbitrarily close to F on the line connecting F_1 and F.

Gauteaux differentiability: A functional Ψ is said to be Gauteaux differentiable at $F \in \mathcal{P}$ if

i. The derivative $\dot{\Psi}(F;h)$ exists for each $h \in Q(F)$.
ii. $h \mapsto \dot{\Psi}(F;h)$ is a linear functional, i.e. $\dot{\Psi}(F; ah_1 + bh_2) = a\dot{\Psi}(F;h_1) + b\dot{\Psi}(F;h_2)$ for all $a, b \in \mathbb{R}$, $h_1, h_2 \in Q(F)$.

We can also define Gauteaux differentiability by considering for a given ϵ, the remainder of the Gauteaux derivative:

$$R_{F,\epsilon}(h) = \frac{\Psi(F+\epsilon h) - \Psi(F)}{\epsilon} - \dot{\Psi}(F;h) = \frac{\Psi(F+\epsilon h) - \Psi(F) - \Psi(F;\epsilon h)}{\epsilon}$$

Then, we can note

$$\Psi(F_n) - \Psi(F_0) = \Psi(F_0 + \frac{1}{\sqrt{n}}\sqrt{n}(F_n - F_0)) - \Psi(F_0)$$

and define $\epsilon_n = 1/\sqrt{n}$, $h_n = \sqrt{n}(F_n - F_0) \in Q(F)$ and show (see STAT 583 Notes p. 28-29)

$$\Psi(F_n) - \Psi(F_0) = \epsilon_n \underbrace{\left(\frac{\Psi(F_0 + \epsilon_n h_n) - \Psi(F_0)}{\epsilon_n} - \dot{\Psi}(F_0; h_n)\right)}_{R_{F_0,\epsilon_n}(h_n) := R_n} + \frac{1}{n}\sum_{i=1}^{n} \dot{\Psi}(F_0; \delta_{X_i} - F_0) \tag{3}$$

where $\delta_{x_i}(Y) = \mathbb{I}[Y \leq x_i]$ is the CDF of a point mass at x_i.

In view of the above, Gauteaux differentiability only says that $R_n \overset{\epsilon \to 0}{\to} 0$ for any fixed direction $h \in Q(F)$. But for this results to be a useful functional Delta method, need the some sense of uniform convergence to zero over all possible directions. Hence, Gauteaux differentiability alone is not enough for the functional delta method we desire, and we instead need stronger notions of differentiability.

Now define $\mathcal{H}$ as a collection of sets $H \subseteq Q(P)$ over which the remainder term above R_n goes to 0 *uniformly*, i.e.

$$\lim_{\epsilon \to 0}[\sup_{h \in H} |R_{F,\epsilon}(h)|] = 0 \text{ for each } H \in \mathcal{H} \tag{4}$$

Then, Gauteaux differentiability refers to the case where $\mathcal{H} = \{\text{all singleton subsets of } Q(F)\}$ in Equation (4).

Hadamard differentiability: Hadamard differentiability refers to the case where $\mathcal{H} = \{\text{all } \rho\text{-compact subsets of } Q(F)\}$ in Equation (4).

An alternative definition that is more useful when trying to prove Hadamard differentiability is: For any sequence $\epsilon_j \to 0$ and $\{h, h_1, h_2, ...\} \in Q(F)$ such that $\rho(h_j - h) \to 0$ and $F + \epsilon_j h_j \in \mathcal{P}$, Ψ is Hadamard differentiable iff:

$$\lim_{j \to \infty}\left[\frac{\Psi(F + \epsilon_j h_j) - \Psi(F)}{\epsilon_j} - \dot{\Psi}(F; h_j)\right] = \lim_{j \to \infty} R_{F,\epsilon_j}(h_j) = 0$$

Frechet differentiability: Frechet differentiability refers to the case where $\mathcal{H} = \{$all ρ-bounded subsets of in Equation (4).

An alternative definition that is more useful when trying to prove Frechet differentiability is: For any sequence $\{F_1, F_2, ...\} \in \mathcal{P}$ such that $\rho(P_j - P) \to 0$, setting $\epsilon_j = \rho(F - F_j)$ and $h_j := (F_j - F)/\rho(F_j - F)$, Ψ is Frechet differentiable iff:

$$\lim_{j\to\infty} \left[\frac{\Psi(F_j) - \Psi(F) - \dot{\Psi}(F; F_j - F)}{\rho(F_j - F)} \right] = \lim_{j\to\infty} R_{F,\epsilon_j}(h_j) = 0$$

Note this looks just like the Hadamard differentiability definition, but the ϵ_j's and h_j's are defined differently here, so that h_j's can move around.

Note: Frechet differentiability $\Rightarrow$ Hadamard differentiability $\Rightarrow$ Gauteaux differentiability.

It turns out that Hadamard differentiability of Ψ at F_0 relative to $\rho = \|\cdot\|_\infty$ is exactly what we need for our functional delta method in to work, i.e. Equation (3) to hold with $R_n = o_p(1)$ above!

Shao (2005) Theorem 5.5: In from Equation (3), $R_n - o_p(1)$, the remainder $\epsilon_n R_n = o_p(n^{-1/2})$ if letting $\epsilon_n = 1/\sqrt{n}$ if either:

i. Ψ is Hadamard differentiable at F_0 relative to $\rho = \|\cdot\|_\infty$ or
ii. Ψ is Frechet differentiable at F_0 relative to a norm ρ if $\rho(F_n - F_0) = O_p(n^{-1/2})$ (i.e. $\rho(h_n) = O_p(1)$).

Functional delta method: Let $\Psi : \mathcal{P} \to \mathbb{R}$ be a fixed functional. If $h \mapsto \dot{\Psi}(F; h)$ is a linear functional which satisfies Hadamard differentiability or Frechet differentiability by Shao's theorem above, then we have (letting $\epsilon_n = 1/\sqrt{n}$ and $h_n = \sqrt{n}(F_n - F_0)$)

$$\Psi(F_n) - \Psi(F_0) = \epsilon_n \underbrace{\left(\frac{\Psi(F_0 + \epsilon_n h_n) - \Psi(F_0)}{\epsilon_n} - \dot{\Psi}(F_0; h_n) \right)}_{R_{F_0,\epsilon_n}(h_n) := R_n} + \epsilon_n \dot{\Psi}(F_0; h_n)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \dot{\Psi}(F_0; \delta_{X_i} - F_0) + o_p(n^{-1/2})$$

hence giving asymptotic linearity of $\Psi(F_n)$ for $\Psi(F_0)$ under the above conditions, with influence function

$$x \mapsto \dot{\Psi}(F_0; \delta_x - F_0)$$

Note:

- $\mathbb{E}[\dot{\Psi}(F_0; \delta_x - F_0)] = \dot{\Psi}(F_0; \mathbb{E}[\delta_x] - \mathbb{E}F_0) = 0$
- Sometimes showing Hadamard differentiability is difficult. In some cases, it is easier to just show $R_n = o_p(1)$ directly, (i.e. via VdV lemma 19.24 in convolution function example)

Functional chain rule: Consider two functionals Ψ and Φ which are Hadamard diferentiable at F_0 and $\Psi(F_0)$ respectively. Then $\Phi \circ \Psi$ is also Hadamard differnetiable at F_0 with derivative (IF) given by

$$\dot{\Phi}_{\Psi(F_0)} \circ \dot{\Psi}(F_0) = \dot{\Phi}(\Psi(F_0; \dot{\Psi}(F_0; h_n)))$$

8 Efficiency Theory

8.1 Parametric efficiency theory

Suppose $\mathcal{M} \equiv \{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^k$. Based on $X_1, ..., X_n \overset{iid}{\sim} P_{\theta_0} \equiv P_0$, our aim is to estimate $\Phi(\theta_0)$ for some differentiable map $\Phi : \mathbb{R}^k \to \mathbb{R}^d$. To do this we use an estimator

$$T_n \equiv t_n(X_1, ..., X_n), \text{ where } t_n : \mathcal{X}^n \mapsto \mathbb{R}^d$$

Typically, a bare minimum requirement we have for an estimator is that it is consistent for $\Phi(\theta)$ for all $\theta \in \Theta$.

In general, parametric asymptotic theory (i.e. VdV 5.39) applies to models $\mathcal{M}$ such that

1. P_θ is absolutely continuous wrt some measure μ for all θ, with density $p_\theta = dP_\theta/d\mu$
2. The model $\mathcal{M}$ is **differentiable in quadratic mean (QMD)**. Really, QMD is a property of the *parameterization* of the model, not the model itself.

We focus on $\sqrt{n}$-consistent estimators because by VdV Theorem 8.9 (almost sure convolution), if a model $\mathcal{M}$ is QMD and the Fisher information is non-singular, then $\sqrt{n}$ convergence is the best we can do (except for on a set of θs with Lebesgue measure 0).

8.1.1 Differentiability in Quadratic Mean (QMD)

Hilbert space: Given a measure λ on a measurable space $(\mathcal{Y}, \mathcal{D})$, we let $L_2(\lambda)$ denote the Hilbert space of functions $g : \mathcal{Y} \to \mathbb{R}$ satisfying $\int g(y)^2 d\lambda(y) < \infty$ endowed with the inner product

$$\langle g, g' \rangle_{L_2(\lambda)} \equiv \int g(y)g'(y)d\lambda(y) \Rightarrow \|g\|^2_{L_2(\lambda)} = \int (g(y))^2 d\lambda(y)$$

Differentiability in quadratic mean (QMD): The map $\theta \mapsto b_\theta := \sqrt{p_\theta}$ is QMD at θ if there exists a function $\dot{\ell}_\theta \in L_2(P_\theta)$ satisfying

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

or, equivalently

$$\lim_{\|h\|^2 \to 0} \frac{1}{\|h\|^2} \int \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2} h^T \dot{\ell}_\theta(x) \sqrt{p_\theta(x)} \right]^2 d\mu(x) = 0$$

A **model $\mathcal{M}$ is QMD at** θ if $\theta \mapsto b_\theta := \sqrt{p_\theta}$ is QMD at θ. The **model itself $\mathcal{M}$ is QMD** if it is QMD at all $\theta \in \Theta$.

- Here $\dot{\ell}_\theta(x)$ is the technical definition of the **score function**. Note that this may still be defined when $\theta \mapsto \sqrt{p_\theta(x)}$ and $\theta \mapsto \log p_\theta(x)$ are non-differentiable at some x (although the former must be differentiable at almost all x).

- Under P_θ, we have $\dot{\ell}_\theta(x)$ has mean zero under P_θ (see VdV Theorem 7.2 below).
- Based on the form of QMD given above, it should come as no surprise that QMD and Frechet differentiability are closely related. QMD essentially says that $\theta \mapsto \sqrt{p_\theta}(x)$ is Frechet differentiable AND the Frechet derivative is $\frac{1}{2}\dot{\ell}_\theta(x)\sqrt{p_\theta(x)}$ a.s -μ.
- In summary, if the Frechet derivative exists and θ is an interior point, then we have QMD.

Van der Vart Theorem 7.6 (Sufficient condition for QMD): For every θ in an open subset of $\mathbb{R}^d$, let $p_\theta = \frac{dP_\theta}{d\mu}$. Suppose that for every x,

i. The map $\theta \to \sqrt{p_\theta(x)}$ is continuously differentiable on Θ.
ii. The elements of the $k \times k$ matrix

$$I_\theta = \int \frac{\dot{p}_\theta(x)\dot{p}_\theta(x)^T}{p_\theta(x)^2} dP_\theta(x) = \int \left[\frac{\frac{d}{d\theta}p_\theta(x)}{p_\theta(x)}\right]\left[\frac{\frac{d}{d\theta}p_\theta(x)}{p_\theta(x)}\right]^T dP_\theta(x)$$

exist and are finite.
iii. The map $\theta \mapsto I_\theta$ is continuous on Θ.

Then the model $\mathcal{M} \equiv \{P_\theta : \theta \in \Theta\}$ (where $\Theta \in \mathbb{R}^k$ and Θ is open) is QMD and the score is

$$\dot{\ell}_\theta = \frac{\frac{d}{d\theta}p_\theta}{p_\theta}$$

Note, of course, that by the chain rule this is the derivative of the log-likelihood. Some applications of this theorem include

- **Exponential family QMD conditions:** Suppose $\mathcal{M} \equiv \{P_\theta : \theta \in \Theta\}$, where $\Theta \in \mathcal{R}$ is open and each P_θ has density p_θ wrt a dominating measure and takes the form

  $$p_\theta(x) = d(\theta)h(x)\exp(\underbrace{Q(\theta)^T}_{\text{natural param}} t(x))$$

  for some vector of functions $g : \mathcal{X} \to \mathbb{R}_{\geq 0}, Q : \Theta \to \mathbb{R}^k, t : \mathcal{X} \to \mathbb{R}^k$ where $d(\theta)$ is a normalizing constant. If
  i. Q is continuously differentiable on Θ.
  ii. The range of Q is included in the natural parameter space

  $$\text{NPS} \equiv \{\lambda \in \mathbb{R}^k : \int h(x)\exp(\lambda' t(x)d\mu(x)) < \infty\}$$

  Then $\mathcal{M}$ is QMD.
- Similar logic for location family.

VdV 7.2 (Existence of mean-zero score and Fisher information): Suppose that Θ is an open subset of $\mathbb{R}^d$ and that $\{P_\theta : \theta \in \Theta\}$ is QMD at θ. Then $P_\theta\dot{\ell}_\theta = 0$ and the Fisher information matrix $I_\theta = P_\theta\dot{\ell}_\theta\dot{\ell}_\theta^T$ exists.

VdV 5.39 (QMD implies asymptotic linearity): Suppose that the model $\mathcal{M} \equiv \{P_\theta : \theta \in \Theta\}$ satisfies:

- $\mathcal{M}$ is QMD at an inner point θ_0 of $\Theta \subset \mathbb{R}^d$.
- There exists a measurable function G with $P_0 G^2 < \infty$ such that, for every θ_1 and θ_2 in a neighborhood of θ_0,

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq G(x)\|\theta_1 - \theta_2\|.$$

- I_{θ_0} is non-singular.
- $\hat{\theta}_n$ is consistent

Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) + o_p(1) \rightsquigarrow N(0, I_{\theta_0}^{-1}),$$

i.e. $\hat{\theta}_n$ is asymptotically linear for θ_0 with influence function $\frac{\dot{\ell}_{\theta_0}(X_i)}{I_{\theta_0}^{-1}}$.

Fisher information: A more concise definition of Fisher information is:

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

8.1.2 Regularity

Regular estimator: Let $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^k$ and $\Phi : \Theta \to \mathbb{R}^d$. An estimator $T_n = T_n(X_1, ..., X_n)$ of $\Phi(\theta)$ is regular at θ_0 if for every h

$$\sqrt{n}(T_n - \Psi(\theta_0 + h/\sqrt{n})) \overset{P^n_{\theta_0 + h/\sqrt{n}}}{\rightsquigarrow} L_{\theta_0}$$

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

Regular estimators are invariant (the shape, that is, since the location wil shift by $h/\sqrt{n}$) under local perturbations of the data-generating distribution. Regular estimators are NOT necessarily asymptotically linear.

We care about regular estimators because

- The Hodges Estimator: Is an example of a superefficient estimator that takes takes the form

$$T_n^* = \begin{cases} T_n & \text{if } |T_n - \Phi(\theta^*)| > n^{-1/4} \\ \Phi(\theta^*) & \text{otherwise} \end{cases}$$

  where T_n is a given $\sqrt{n}$ consistent estimator for $\Phi(\theta)$. We see that if $\theta = \theta^*$, then T_n^* will be asymptotically better than T_n and if $\theta = \theta^*$ then it will be asymptotically equivalent to T_n. However, it performs very poorly in *neighborhoods* of θ^*. For an efficiency theory that doesn't prefer T_n^*, we need to assess estimators on the basis of a limiting distribution under a data generating process that can change with n. Regularity fills this role.
- When we restrict to regular estimators, the **Almost Everywhere Convolution Theorem** says that the best possible limiting distribution is $N(0, \dot{\Phi}(\theta)^T I_\theta^{-1} \dot{\Phi}(\theta))$, except on set of parameters with Lebesgue measure zero.

- Regular estimators are locally asymptotically minimax, meaning their risk over $N(0, \dot{\Phi}(\theta)^T I_\theta^{-1}\dot{\Phi}(\theta))$ lower bounds the asymptotic minimax risk of any estimator in a neighborhood of θ_0.
- The CR-lower bound applies to regular estimators:

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

  and for the problem of estimating $\tau(\theta_1)$ for $\theta = (\theta_1, ..., \theta_k)$ (so $\theta_2, ..., \theta_k$ are nuisances), we have

$$\text{Var}_\theta(T(X)) \geq \frac{\left(\frac{d\tau}{d\theta_1}\right)^2}{I_{11\cdot 2}(\theta)}$$

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

8.1.3 Convergence to a Normal Experiment: Local Asymptotic Normality and Contiguity

Local asymptotic normality (LAN): LAN is a property of a model, and it deals with likelihood ratios. Asymptotic linearity (and regularity) is a property of estimators. If a model is LAN then the log-likelihood statistic for testing

$$H_{0,n} : (X_1, ..., X_n) \sim P_\theta^n \quad \text{vs.} \quad H_{1,n} : (X_1, ..., X_n) \sim P_{\theta+h/\sqrt{n}}^n$$

which takes the form

$$Q_{h,n} = \log\left(\frac{\prod_{i=1}^n p_{\theta+h/\sqrt{n}}(X_i)}{\prod p_\theta(X_i)}\right) = \sum_{i=1}^n \log(p_{\theta+h/\sqrt{n}}(X_i)) - \sum_{i=1}^n \log(p_\theta(X_i))$$

behaves, in large samples and under the null model $H_{0,n}$ like the test statistic T, which is a log-likelihood ratio of normals, i.e.

$$T = \log\left(\frac{f_h(z)}{f_0(z)}\right) = \log(f_h(z)) - \log(f_0(z))$$

does under the null model $H_0 : Z \sim N(0, I_\theta^{-1})$ vs. $H_A : Z \sim N(h, I_\theta^{-1})$, where f_0, f_h are the respective normal densities and we observe one sample Z.

LAN is a property of the null distribution of $Q_{h,n}$. It hence only tells us about Type I errors in large samples. But we later use contiguity and Le Cam's First and Third lemmas to show that is also holds under the alternatives $H_{1,n}$ and H_A, so all aspects of the test (i.e. power) are also equivalent.

Conditions for LAN: If

i. A model is QMD. Specifically, if $\theta \mapsto \sqrt{p_\theta}$ is QMD.
ii. θ is an interior point of Θ.

Then $Q_{h,n}$, as defined above, is LAN.

Contiguity preliminaries: Contingity is the asymptotic analog to absolute continuity. First we define some preliminaries

- **Absolute continuity:** Recall that if Q and P are two probability measures on the same measurable space $(\mathcal{Z}, \mathcal{G})$, then Q is **absolutely continuous** with respect to P, denoted $Q << P$, iff

$$P(A) = 0 \Rightarrow Q(A) = 0 \quad \text{for all } A \in \mathcal{G}.$$

- **Singularity (for measures):** Q and P are singular, denoted $Q \perp P$ if and only if there exists two disjoint sets A_p and A_Q in $\mathcal{G}$ such that $P(A_P) = 1, Q(A_P) = 0, P(A_Q) = 0$, and $Q(A_Q) = 1$. That is, $P \perp Q$ means their supports are disjoint.
- **Lebesgue decomposition:** For any two probability measures P and Q, there exists (unique) measures Q^a and $Q^\perp$ such that

$$Q = Q^a + Q^\perp \quad \text{with } Q^a << P \text{ and } Q^\perp \perp P$$

- **Equivalent expressions for absolute continuity:** The following are equivalent:
  i. $Q << P$
  ii. $\int V(z) dP(z) = 1$ where $V = dQ^a/dP$
  iii. $Q = Q^a$
- **Radon-Nikodym derivative:** If P and Q are measures defined on the measurable space $(\mathcal{X}, A)$ and $Q << P$, then there exists a measurable function $g : \mathcal{X} \to [0, \infty)$ such that for any set $\mathcal{X}_1 \in A$,

$$\int_{\mathcal{X}_1} dQ(x) = \int_{\mathcal{X}_1} g(x) dP(x)$$

where $g(x) = \frac{dQ(x)}{dP(x)}$. If P, Q are probability measures and absolutely continuous with respect to the Lebesgue measure with Lebesgue densities p and q and $Q << P$, then

$$\frac{dQ}{dP}(x) = \frac{q(x)}{p(x)},$$

i.e. the Radon-Nikodym derivative is the likelihood ratio.

Contiguity: For each $n = 1, 2, ...$, let Q_n and P_n be probability measures defined on a measurable space $(\mathcal{Z}_n, \mathcal{G}_n)$. We say that Q_n is **contiguous** with respect to P_n, denoted $Q_n \triangleleft P_n$, if and only if for all sequences $\{A_n : A_n \in \mathcal{G}_n\}_{n \geq 1}$ it holds that

$$P_n(A_n) \overset{n \to \infty}{\to} 0 \quad \text{implies that} \quad Q_n(A_n) \overset{n \to \infty}{\to} 0$$

Le Cam's first lemma: gives equivalent statements for contiguity:

i. $Q_n \triangleleft P_n$

ii. Let $V_n = \frac{dQ_n}{dP_n}$. If for some $\{n_k\}_{k \geq 1}$, $V_{n_k} \overset{P_{n_k}}{\rightsquigarrow} V$, then $\mathbb{E}[V] = 1$.

iii. Let $U_n = \frac{dP_n}{dQ_n}$. If for some $\{n_k\}_{n\geq 1}$, $U_{n_k} \overset{Q_{n_k}}{\rightsquigarrow} U$, then $\mathbb{E}[I[U>0]] = P[U>0] = 1$.

iv. For any Borel measurable $T_n : Z_n \to \mathbb{R}^k$, $T_n(Z_n) = o_{P_n}(1) \Rightarrow T_n(Z_n) = o_{Q_n}(1)$, where $H_n = o_{P_n}(1)$ means $P_n(\{Z_n : |H_n(Z_n)| > \epsilon\}) = P_n(|H_n(Z_n)| > \epsilon) \overset{n\to\infty}{\to} 0$ for all $\epsilon > 0$.

Corollaries to Le Cam's first lemma are:

- Corollary 1: Suppose $\log V_n = \log(dQ_n/dP_n) \overset{P_n}{\rightsquigarrow} N(\mu, \sigma^2)$ Then $Q_n \triangleleft\triangleright P_n$ iff $\mu = -\sigma^2/2$.
- Let $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ for $\Theta \in \mathbb{R}^k$. If θ_0 is an interior point of Θ, and $\theta \mapsto \sqrt{p_\theta}$ is Frechet differentiable at θ_0 w/ score $\dot{\ell}_{\theta_0}$ (i.e. QMD conditions) then $P^n_{\theta_0+h/\sqrt{n}} \triangleleft P^n_{\theta_0}$.

Le Cam's 3rd Lemma: For each $n = 1, 2, \ldots$ let Q_n and P_n be probability measures defined on a measurable space $\mathcal{Z}_n, \mathcal{G}_n$. Let $T_n : \mathcal{Z}_n \to \mathbb{R}^d$ be Borel measurable and suppose $Q_n \triangleright P_n$, and let $V_n = dQ_n/dP_n$ be the R-N derivative (note it is a measurable function and hence an RV). If

$$(T_n, V_n) \overset{P_n}{\rightsquigarrow} (T, V)$$

then $T_n \overset{Q_n}{\rightsquigarrow} L$ where L is a probability law on $(\mathbb{R}^d, \mathcal{B})$, satisfying for all $A \in \mathcal{B}$ $L(A) = \mathbb{E}[\mathbb{I}[T \in A]V]$.

This allows us to compute the weak limit of T_n under sampling from Q_n using the weak limit of T_n under P_n.

Corollary to Le Cam's Third Lemma (Normal Case): Let $V = dQ_n/dP_n$ as above. If

$$(T_n, \log V_n) \overset{P_n}{\rightsquigarrow} (T, \log V) \sim N\left(\begin{pmatrix} \mu \\ -\sigma^2/2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right),$$

then

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

In other words, under Q_n, T_n is shifted by $\tau = \mathrm{Cov}(T_n, \log V_n)$

Key corollary/implication of Le Cam's Third Lemma: (combining with First Lemma corollary 2) Let $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ for $\Theta \in \mathbb{R}^k$. If θ_0 is an interior point of Θ, and $\theta \mapsto \sqrt{p_\theta}$ is Frechet differentiable at θ_0 w/ score $\dot{\ell}_{\theta_0}$ (i.e. QMD conditions) then if

$$(T_n, \log V_n) \overset{P^n_{\theta_0}}{\rightsquigarrow} (T, \log(V)) \sim N\left(\begin{pmatrix} \mu \\ -h^T I_\theta h/2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & h^T I_\theta h \end{pmatrix}\right)$$

then

$$T_n \overset{P^n_{\theta_0+h/\sqrt{n}}}{\rightsquigarrow} N(\mu + \tau, \Sigma)$$

8.1.4 Putting it all together: Properties of RAL estimators

Le Cam's 4th Lemma (Conditions where an AL estimator is also regular): This follows from Le Cam's Third Lemma. Let $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ and let $\Phi : \Theta \to \mathbb{R}$ (same idea but vectors if $\mathbb{R}^d$). be differentiable at θ_0. Suppose that $T_n = T_n(X_1, \ldots, X_n)$ is an asymptotically linear estimator of $\Phi(\theta)$ at θ_0 with influence function $t_{\theta_0}(X)$. Then T_n is regular at θ_0 if and only if

$$\frac{d}{d\theta}\Phi(\theta)|_{\theta=\theta_0} = \mathbb{E}_{\theta_0}[t_{\theta_0}(X)\dot{\ell}_{\theta_0}(X)]$$

Properties of RAL estimators

- The Cramer-Rao lower bound:

$$\frac{d}{d\theta}\Phi(\theta)|_{\theta=\theta_0}^T I_{\theta_0}^{-1}\frac{d}{d\theta}\Phi(\theta)|_{\theta=\theta_0}$$

  is a lower bound for the variance of the limit law of RAL estimators.
- Under regularity conditions, the MLE is RAL and meets the CR lower bound so is asymptotically efficient among RAL estimators.
- The **Convolution Theorem** shows that $N(0, \frac{d}{d\theta}\Phi(\theta)|_{\theta=\theta_0}^T I_{\theta_0}^{-1}\frac{d}{d\theta}\Phi(\theta)|_{\theta=\theta_0})$ is actually the best limit law for all regular estimators, not just those contrained to be AL.

Hajek-Le Cam Convolution Theorem: Suppose $\mathcal{M}$ is QMD, the Fisher information

$$J(\theta) = P_\theta\left[\left(\frac{d}{d\theta}\log p_\theta\right)^2\right] = \mathbb{E}\left[\left(\frac{d}{d\theta}\log p_\theta\right)^2\right] = \text{Var}\left(\frac{d}{d\theta}\log p_\theta\right)$$

satisfies $J(\theta) > 0$ and τ_n is a regular estimator of τ_0, with $\sqrt{n}(\tau_n - \tau_0) \rightsquigarrow Z$. The Z has the same distribution as $Z_0 + \Delta_0$, where Δ_0 is some unspecified RV independent of Z and

$$Z \sim N(0, V_0), \quad \text{where } V_0 = \frac{\dot{\tau}(\theta_0)^2}{J(\theta_0)}$$

Hence, for any regular estimator τ_n of τ_0, the asymptotic variance of $\sqrt{n}(\tau_n - \tau_0)$ is lower bounded by V_0, which is the well-known CR-lower bound.

8.2 General efficiency theory

We now consider an arbitrary model $\mathcal{M}$ and i.i.d. sampling of $X_1, ..., X_n$ from a fixed $P_0 \in \mathcal{M}$. Our goal is to estimate $\psi_0 := \Psi(P_0)$

Generalized Cramer-Rao (GCR) Lower bound: Define $\mathcal{H}_0(P_0)$ to be the index set for $S(P_0)$, the collection of all sufficiently smooth (QMD at $\theta = \theta_0 = 0$) 1-D parameteric submodels of $\mathcal{M}$, each centered at P_0 (i.e. so WLOG P_0 is element corresponding to index $\theta = 0$ in each). Denote $V_0(M_h)$ for the C-R lower bound for estimating $\Psi(P_0)$ within the parametric sub-model $\mathcal{M}_h := \{P_{\theta,h} : \theta \in \Theta\}$ corresponding to a given $h \in \mathcal{H}_0(P_0)$. The Generalized Cramer-Rao (GCR) lower bound is given by:

$$\sup_{h\in\mathcal{H}_0(P_0)} V_0(\mathcal{M}_h) = \sup_{h\in\mathcal{H}(P_0)} \frac{[\frac{d}{d\theta}\Psi(P_{\theta,h})|_{\theta=0}]^2}{J_{\mathcal{M}_h}(0)}$$

where $P_{\theta,h}$ is an arbitrary element of $\mathcal{M}_h$ and $J_{\mathcal{M}_h}$ is the Fisher information about $\theta = 0$ in a single observation X from P_0 in $\mathcal{M}_h$.

Some key observations:

- Note each $\mathcal{M}_h$ is a correctly specified parametric model by construction.
- Note that the denominator can be written as

$$J_{\mathcal{M}_h}(0) = \int [g_h(x)]^2 dP_0(x)$$

  where $g(x)$ is the score for θ at $\theta = 0$. So all that matters about $\mathcal{M}_h$ is the score - we really only need to take the supremum over all possible score functions that arise in 1-D models.
- The GCR idea is that the estimation of $\Psi(\theta)$ in $\mathcal{M}$ can be no easier than it is in the least-favorable parametric sub-model.

Hilbert space: A (real) Hilbert space $\mathcal{H}$ is a (real) vector space that is equipped with an inner product $(h_1, h_2) \mapsto \langle h_1, h_2 \rangle$ and is complete relative to the norm $h \mapsto \|h\| = \langle h, h \rangle^{1/2}$.

$L_0^2(P)$: We denote by $L_0^2(P)$ the collection of real-valued functions f defined on the support of P such that

i. $\int f(x)dP(x) = 0$
ii. $\int f(x)^2 dP(x) < \infty$

equipped with inner product

$$(f_1, f_2) \mapsto \langle f_1, f_2 \rangle := \int f_1(x) f_2(x) dP(x) = \mathrm{Cov}_P(f_1(x), f_2(x))$$

which gives induced norm

$$\|f\|_P := \left(\int f(x)^2 dP(x) \right)^{1/2}.$$

With this inner product, $L_0^2(P)$ is a Hilbert space.

Some notes on Hilbert spaces:

- The orthogonal compliment of a subspace $\mathcal{H}_0$ of $\mathcal{H}$ is defined as

$$\mathcal{H}_0^\perp := \{h \in \mathcal{H} : \langle h, h_0 \rangle = 0 \ \forall \ h_0 \in \mathcal{H}_0\}$$

  $\mathcal{H}_0^\perp$ is necessarily a closed subspace of $\mathcal{H}$, and $\mathcal{H}_0^\perp \cap \mathcal{H}_0$ contains only the zero element.
- **Projection:** If $\mathcal{H}_*$ is a closed subspace of $\mathcal{H}$, the projection of $h_0 \in \mathcal{H}$ into $\mathcal{H}_*$, denoted by $\Pi_{\mathcal{H}_*} h_0 = \Pi[h_0|\mathcal{H}_*]$ is the unique element $h_* \in \mathcal{H}_*$ such that

$$\|h_0 - h_*\| = \min\{\|h_0 - h\| : h \in \mathcal{H}_*\}.$$

  In addition to h_* being in $\mathcal{H}_*$, we must also have that the residual $h_0 - h_*$ is orthogonal to $\mathcal{H}_*$, i.e. $\langle h_0 - h_*, h \rangle = 0$ for all $h \in \mathcal{H}_*$.
- If $\mathcal{H}_1$ and $\mathcal{H}_2$ are orthogonal, i.e. if $\langle h_1, h_2 \rangle = 0$ for all $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$, then

$$\Pi[\cdot|\mathcal{H}_1 + \mathcal{H}_2] = \Pi[\cdot|\mathcal{H}_1] + \Pi[\cdot|\mathcal{H}_2]$$

- If $\mathcal{H}_0 = \{cf : c \in \mathbb{R}\}$ for some $f \in \mathcal{H}$, then $\Pi[h|\mathcal{H}_0] = \frac{\langle h,f \rangle}{\langle f,f \rangle}$ for any $h \in \mathcal{H}$. This follows by the projection of $h \in \mathcal{H}$ into $\mathcal{H}_0$ being able to be written as $h^* = c(h)f$ for some $c(h)$. So to get $c(h) = \langle h - h^*, f \rangle = 0$, we must have $c(h) = \frac{\langle h,f \rangle}{\langle f,f \rangle}$.

Riesz representation theorem: If $\Phi : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional, then there exists a unique element $h_0 \in \mathcal{H}$ such that $\Phi(h) = \langle h, h_0 \rangle$ for each $h \in \mathcal{H}$.

Tangent set: The tangent set of $\mathcal{M}$ at P_0 is the set of all score functions at $\theta = 0$ for sufficiently smooth 1-D parametric sub-models of $\mathcal{M}$ traversing P_0 at $\theta = 0$. Clearly, it is contained in $L_0^2(P_0)$.

Tangent space: The tangent space $T_{\mathcal{M}}(P_0) \subseteq L_0^2(P_0)$ of $\mathcal{M}$ at $P_0 \in \mathcal{M}$ is defined as the closure of the set of all possible linear combinations (i.e. linear span) of elements in the tangent set.

- The tangent space is a Hilbert space. The tangent set may not necessarily be since it may not be closed under addition.

Model classifications: A model $\mathcal{M}$ for P_0 is said to be a

- *Non-parametric model* if at each $P \in \mathcal{M}$, $T_{\mathcal{M}}(P) = L_0^2(P)$ and $L_0^2(P)$ is infinite-dimensional.
- *Parametric model* if at each $P \in \mathcal{M}$, $T_{\mathcal{M}}(P)$ is finite-dimensional (i.e. has finite basis).
- *Semi-parametric* otherwise.

Pathwise differentiability: A parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is said to be pathwise differentiable at $P_0 \in \mathcal{M}$ relative to $\mathcal{M}$ is for each $s \in T_{\mathcal{M}}(P_0)$ and each regular one-dimensional parametric sub-model $\mathcal{M}_s = \{P_{\theta,s} : \theta \in \Theta\} \in \mathcal{H}_0(P_0)$ with score s for θ at $\theta = 0$, the pathwise derivative

$$\frac{d}{d\theta}\Psi(P_{\theta,s})|_{\theta=0} = \dot{\Psi}_{P_0}(s)$$

1. Exists and does not depend on the choice of sub-model (here $\mathcal{M}_s$), except through the score s.
2. The pathwise derivative functional $\dot{\Psi}_{P_0} : T_{\mathcal{M}}(P_0) \to \mathbb{R}$ (defined above) is linear and continuous (equivalently, linear and bounded) relative to the $L_0^2(P_0)$ norm.

Some notes:

- Pathwise differentiability ensures the numerator of the GCR bound exists and is computable.
- It formalized the idea that the numerator only depends on the sub-model through the score function.
- It requires a certain smoothness of $\Psi(P)$ as a function of P. In a larger model where P can be perturbed in many ways is harder to be pathwise differentiable.
- We call a given 1-D sub-model through P at the origin a "path".

Gradient: Any element $D(P)$ (i.e. $D_P : \mathcal{X} \to \mathbb{R}) \in L_0^2(P)$ such that the pathwise derivative can be represented as

$$\dot{\Psi}_P(s) = \langle D(P), s \rangle_P = P[D(P)s] = \mathbb{E}_P[D(P)(X)s]$$

for each $s \in T_{\mathcal{M}}(P)$ is called the **gradient** of Ψ at P relative to $\mathcal{M}$.

Some key facts:

i. There is a unique canonical gradient.
ii. Denote $\mathcal{G}_{\mathcal{M}}(P) \subset L_0^2(P)$ for the collection of all gradients of Ψ at P_0 relative to $\mathcal{M}$. Let $D_0(P)$ be any given gradient. Then:

$$\mathcal{G}_{\mathcal{M}}(P) = D_0(P) + T_{\mathcal{M}}^{\perp}(P_0) \equiv \{D_0(P_0) + q(P) : q(P) \in T_{\mathcal{M}}^{\perp}(P_0)\}$$

holds for any gradient $D_0(P_0)$, where $T_{\mathcal{M}}^{\perp}(P_0) := \{q \in L_0^2(P) : \langle q, q' \rangle = 0 \,\forall\, q' \in T_{\mathcal{M}}(P_0)\}$ is the orthogonal compliment of $T_{\mathcal{M}}(P)$. **key:** The canonical gradient can be found by projecting and gradient $D(P)$ into $T_{\mathcal{M}}(P)$, i.e. $D^*(P) = \Pi[D(P)|T_{\mathcal{M}}(P_0)]$ for any $D(P) \in \mathcal{G}_{\mathcal{M}}(P)$.
iii. If $\mathcal{M}$ is non-parametric then there is only one gradient, i.e. $|\mathcal{G}_{\mathcal{M}}(P)| = 1$.
iv. If $\mathcal{M}_1 \subseteq \mathcal{M}_2$ (i.e. $\mathcal{M}_1$ is nested in $\mathcal{M}_2$), then $\mathcal{G}_{\mathcal{M}_2}(P_0) \subseteq \mathcal{G}_{\mathcal{M}_1}(P_0)$. That is, any gradient in $\mathcal{M}_2$ is also a gradient in $\mathcal{M}_1$.
v. The gradient is the representation of the pathwise derivative permitted by the Riesz representation theorem, and it exists by condition (ii.) of pathwise differentiability.

Equivalence of influence functions for RAL estimators and gradients:

First, recall the definition of regularity: An estimator ψ_n of $\psi_0 := \Psi(P_0)$ is locally regular at P_0 if for any $g \in T_{\mathcal{M}}$ and any path $\{P_\theta\}$ through P_0 at $\theta = 0$ with score g for θ at $\theta = 0$,

$$\sqrt{n}(\psi_n - \psi_{0n}) \text{ and } \sqrt{n}(\psi_n - \psi_0)$$

have the same limit distribution under sampling from $P_{n^{-1/2}}$ and P_0 respectively. Here we denote $\psi_{0n} = \Psi(P_{n^{-1/2}})$. If ψ_n is locally regular uniformly over $\mathcal{M}$, then ψ_n is regular.

The two directions of this theorem below ensure that the set of gradients is equivalent to the set of IFs corresponding to RAL estimators of ψ_0 under $\mathcal{M}$:

- Direction 1: **Influence functions are gradients** [Pfanzagl 1988, 2000; VdV 1991]: Suppose that ψ_n is an asymptotically linear estimator of $\psi_0 := \Psi(P_0)$ relative to a model $\mathcal{M}$ with influence function ϕ_{P_0}. Then the following statements are equivalent:
  1. Ψ is pathwise differentiable at P_0 and ϕ_{P_0} is the gradient of Ψ at P_0.
  2. The estimator ψ_n is regular at P_0.
- Direction 2: **Gradients are influence functions** [Klaasen 1987]: Suppose that Ψ is pathwise differentiable with gradient $D(P_0)$ at P_0 wrt $\mathcal{M}$. Then, under certain regularity conditions, the following statements are equivalent:
  1. There exists an asymptotically linear estimator ψ_n with influence function $D(P_0)$.
  2. The function $D(P_0)$ can be estimated sufficiently well (i.e.) consistently from given data

This theorem tells us that we must restrict ourselves to pathwise differentiable parameters to construct RAL estimators. And it says that gradients can be found by computing the IF of a known RAL estimator.

Asymptotically efficient estimator: An RAL estimator psi_n of ψ_0 is efficient wrt $\mathcal{M}$ if and only if

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

In other words, ψ_n is asymptotically linear for ψ_0 with influence function given by the canonical gradient. In this case, we call the canonical gradient the **efficient influence function**.

9 Strategies

9.1 Decision theory

9.1.1 Find a Bayes rule

Strategy 1 (Compute posterior): Compute a posterior and then choose rule which minimizess expected loss over posterior.

9.1.2 Prove a rule is admissible

Strategy 1 (contradiction): Assume that another rule uniformly dominates the given rule and then find a contradiction.

Strategy 2 (show unique Bayes or unique minimax): Unique Bayes and unique minimax rules are admissible. Apply results of 4.4.

Strategy 3 (Connect to squared error loss): If you have an admissible estimator under a certain loss (i.e. squared error loss) and you want to assess admissibility under a related (i.e. weighted) loss, assume it is not admissible and massage inequalities to be in terms of loss for which you have admissibility.

- Also: note that in STAT 513 HW7 Q3, we showed admissibility with respect to weighted squared error loss is equivalent to admissibility wrt squared error loss, because dividing both sides by $\theta(1-\theta)$ will preserve inequalities in risk.

9.1.3 Prove a rule is minimax

Strategy 1 (Submodel approach): This approach is best for demonstrating minimaxity over semi/non-parametric models. If D_1 is minimax over $\mathcal{P}_1$ for $\mathcal{P}_1 \subseteq \mathcal{P}_2$, 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 example, 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 (Least favorable prior or constant risk:) Appeal to theorems in 4.3 and show that the rule corresponds to a Bayes rule under a least favorable prior or has a risk that is constant in θ.

9.2 Asymptotic behavior of estimators

9.2.1 Prove consistency of an estimator

Strategy 1 (WLLN): If it can be written as a sample mean, then it is consistent by the WLLN.

Strategy 2 (CMT): If it can be written as a continuous function of a continuous estimator, then apply the CMT.

Strategy 3 (By definition of convergence in probability, often via Concentration inequalities
For instance to show MLE for uniform distribution is consistent.

i. Hoeffding bound for bounded RVs.
ii. Chebyshev inequality for RVs with finite expectations and shrinking variance.
iii. Chernoff bound if we have access to MGF. Can also use this if we know the RV to be sG or sE
iv. For non-sample means, we rely on the bounded differences/McDiarmid inequality.

Some other facts:

- If X_n, X are bounded in $[-B, B]$, then $X_n \xrightarrow{p} X \Leftrightarrow \mathbb{E}[|X_n - X|] \to 0$, i.e. if $X_n \overset{L^1(P)}{\to} X$. (STAT 513 HW5, Q4)

Strategy 4 (Plug-in estimator): 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 supremum norm metric. See 6.4.2.

Strategy 5 (Random function/ERM - Prove $\mathcal{F}$ is G-C):

i. 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.
ii. If the loss function identifies the true (possibly infinite dimensional) parameter θ_0 and $\hat{\theta}$ is an ERM, then controlling the regret ensures that $\hat{\theta} \xrightarrow{p} \theta_0$.
iii. Some GC classes include
    - Any VC class. VC permanence properties include unions, intersections, addition, multiplication, compositions.
    - If we are dealing with an $\mathbb{R}$-valued function class that forms a vector space (i.e. polynomials of degree no greater than n), then the VC dimension is the dimension of the vector space.

9.2.2 Show that a class is Donsker

1. Try the characterization formula for weak convergence (usually hard).
2. Show $\mathcal{F}$ satisfies the finite bracketing integral property (VdV 19.5): For $\delta > 0$, define the bracketing integral

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

   . $\mathcal{F}$ is P-Donsker if $J_{[]}(\delta, \mathcal{F}, L^2(P)) < \infty$.
3. Show $\mathcal{F}$ satisfies the uniform entropy bound (VdV 19.14): $\mathcal{F}$ is P-Donsker if it has envelope $\bar{F}$ satisfying $P\bar{F} < \infty$ and

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

where the supremum is over all discrete probability measures Q on $\mathcal{Z}$ and $\|\bar{F}\|_{Q,2} = Q\bar{F}^2$.

4. Use a permanence of Donsker class property.
5. Show bounded variation norm (see examples - this is most common).
6. Use partial Slutsky's theorem for weak convergence (STAT 583 HW 1).

9.2.3 Find the asymptotic distribution of an RV/estimator:

Strategy 1 (For RVs):

i. Work from first principles by writing and manipulating CDFs.
ii. Is it a sum of variables with known CDFs?
iii. Are they transformations of RVs with known distributions? Then see 2.4.

Strategy 2 (CLT for sample means):

- Apply Levy CLT for univariate sample means, multivariate CLT for multivariate iid data.
- Lindeberg-Feller CLT for independent but not identically distributed data.

Strategy 3 (Convergence theory - Slutsky's and Portmanteau):

- Apply Slutsky's if sum, difference, product, quotient of two RVs where one convergence in probability, the other converges weakly.
- Portmanteau lemma for other versions of weak convergence, especially if we can work out CDF.

Strategy 4 (Delta method): If it is a differentiable function of a statistic with a known distribution, then we can use one of the delta methods of 3.4.

Strategy 5 (M/Z estimators): M and Z estimators are consistent and asymptotically normal under conditions (G-C and Donsker respectively) on the loss functions (for M-estimators) or the estimating functions (Z-estimators).

9.2.4 Establish asymptotic linearity

Strategy 1 (Classic expansion): When $\psi_n = P_n f_n$ and $\psi_0 = P_0 f_0$, we have

$$\psi_n - \psi_0 = P_n f_n - P_0 f_0 = (P_n - P_0)f_0 + P_0(f_n - f_0) + (P_n - P_0)(f_n - f_0)$$

Then

- Red term: Linear, or the normal term. this piece we want to isolate.
- Purple term: The Taylor expansion term (see estimating equations example).
- Blue term: Intuitively, this should go to zero double fast. To establish $(P_n - P_0)(\psi_n - \psi_0) = o_p(n^{-1/2})$, we usually appeal to VdV 19.24:
  i. $\{h_k\}_{k=1}^{\infty}$ is a sequence of random functions in $L^2(P)$ s.t. $P(h_n \in \mathcal{F}) \to 1$ for a Donsker class $\mathcal{F} \subset L^2(P)$.
  ii. $P_0(h_n - h_0)^2 = o_p(1)$ for some $h_0 \in \mathcal{F}$.

Strategy 2 (Delta method for influence functions) See 3.4.

Strategy 3 (Functional delta method): Best used via the chain rule (see theorems at end of 7.5). Used to show asymptotic linearity of an estimator that takes the form of a Hadamard differentiabel functional Ψ applied to the empirical distribution F_n.

Strategy 4 (U/V statistic): Does the functional depend on two (or more) independent draws from the same distribution $X_1, X_2 \sim P$? If so, see 7.4.

9.3 Asymptotic efficiency

9.3.1 Computing the gradient of a pathwise differentiable parameter

Consider the task of computing a gradient of a pathwise differentiable functional $\Psi(P)$ at P_0 relative to $\mathcal{M}$.

Strategy 1 (from scratch): Recall that if $\mathcal{M}_1 \subseteq \mathcal{M}_2$ then any gradient of $\mathcal{M}_2$ is also a gradient of $\mathcal{M}_1$. So we can always find a gradient in a large nonparametric model and apply it to the sub-model.

i. Choose a simple parametric submodel centered at P_0. A simple choice is the submodel $\{p_{\theta,s} : \theta \in \Theta\}$ through P_0 at $\theta = 0$ with density

$$p_{\theta,s}(x) = (1 + \theta s(x))p_0(x)$$

relative to some common dominating measure.

ii. Compute the pathwise derivative of the parameter along this path

$$\frac{d}{d\theta}\Psi(\theta)|_{\theta=0}$$

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

$$\frac{d}{d\theta}\Psi(\theta)|_{\theta=0} = \langle \tilde{D}(P_0), g \rangle$$

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

iv. Re-center the gradient to be mean 0 so that it lies in the (nonparametric) tangent space:

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

Strategy 2 (Use known RAL estimator): If we have a known RAL esitmator, then its influence function is a gradient

9.3.2 Derive the tangent space

Strategy 1 (consider the model type): The tangent space takes the following forms depending on the model type:

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

- Nonparametric model: The tangent space is $L_0^2(P)$.
- Semiparametric model: Need to do some more work. More restrictive models will have smaller tangent spaces, which will still be a subset of $L_0^2(P)$.

Strategy 2 (under a moment restriction): We saw in lecture the example of deriving the tangent space of a model $\mathcal{M}$ subject to a moment restriction such as $P(g_0) = 0$. We use the linear sub-model with elements defined by density $p_{\theta,s}(x) = (1+\theta s(x))p_0(x)$ and note

$$\int g_0(x)(1+\theta s(x))f_0(x)dx \Rightarrow P_0(g_0 s) = 0,$$

Then we the projections of elements $s \in L_0^2(P)$ into $T_\mathcal{M}$ are given by

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

(can verify this). So the EIF of $\Psi(P) = Pf$ in this model is

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

Strategy 3 (Break into variationally independent components:) In lecture notes, we saw the example of the bivariate model that we could decompose into variationally independent parts. Suppose $X := (Y, Z) \sim P \in \mathcal{M}$. We can write

$$\mathcal{M} = \mathcal{M}_Z \otimes \mathcal{M}_{Y|Z}$$

where $\mathcal{M}_Z$ and $\mathcal{M}_{Y|Z}$ are models for P_Z and $P_{Y|Z}$ that are variationally independent, i.e. knowledge in one model does not restrict the other. In this case the tangent space can be written as the sum of two orthogonal subspaces

$$T_\mathcal{M}(P) = T_{\mathcal{M}_Z}(P) \oplus T_{\mathcal{M}_{Y|Z}}(P)$$

where $T_{\mathcal{M}_Z}(P) \subseteq L^2_{0,Z}(P_0)$ and $T_{\mathcal{M}_{Y|Z}}(P) \subseteq L^2_{0,Y|Z}(P_0)$ are tangent spaces generated by scores for $P_{Z,0}$ and $P_{Y|Z,0}$ respectively. This follows heuristically from fact that $\log p_{Y,Z,\theta} = \log p_{z,\theta} + \log p_{Y|Z,\theta}$.

The projection onto $\mathcal{M}$ can then be obtained by projections onto each subspace.

Strategy 4 (orthogonal nuisance): Suppose $P = AB$ and the parameter ψ depends on P only through A, so B is an orthogonal nuisance. Then we have

$$T_\mathcal{M}(P) = T_{\mathcal{M}_A}(P) + \underbrace{T_{\mathcal{M}_B}(P)}_{=0}.$$

Projection of the gradient onto $T_\mathcal{M}$ is the same as projection onto $T_{\mathcal{M}_A}$, since the pathwise derivative doesn't change as P_B fluctuates. Here, the EIF is contained entirely in $T_{\mathcal{M}_A}$, and so restricting the form of $T_{\mathcal{M}_B}$ (even assuming it is known exactly) does not impact the efficiency of the estimator!

9.3.3 Computing projections onto the tangent space

Strategy 1 (guess form of projection): In simple cases it may be possible to guess/intuit the form of the projection. Recall that a projection must (a) lie in $L^2(P_0)$, (b) lie in the tangent space, and (c) the residual must lie in the orthogonal compliment of the tangent space.

Strategy 2 (variationally independent components): In lecture we considered the case where $X := (Y, Z) \sim P_0 \in \mathcal{M}$. As shown above, the tangent space is $T_{\mathcal{M}} = T_{\mathcal{M}_Z} + T_{\mathcal{M}_{Y|Z}}$. If M_Z and $M_{Y|Z}$ are non-parametric then

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

and so the projections onto the componenets of the tangent space are

$$\begin{aligned} \Pi[s|T_{\mathcal{M}_Z}(P)] &= \mathbb{E}_P[s(Y, Z)|Z = z] \\ \Pi[s|T_{\mathcal{M}_{Y|Z}}(P)] &= s(y, z) - \mathbb{E}_P[s(Y, Z)|Z = z] \end{aligned}$$

Strategy 3 (independent components): If Y and Z above are independent (so $P(Y|Z) = P(Y)$) and each is modelled nonparametrically then we have

$$\begin{aligned} \Pi[s|T_{\mathcal{M}_Z}(P)] &= \mathbb{E}_P[s(Y, Z)|Z = z] \\ \Pi[s|T_{\mathcal{M}_Y}(P)] &= \mathbb{E}_P[s(Y, Z)|Y = y] \end{aligned}$$

Strategy 4 (Parametric model): If $\mathcal{M}$ is parametric so that $\mathcal{M} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^P\}$, then the tangent space is a finite-dimensional subspace corresponding to the linear span of the score.

$$T_{\mathcal{M}} = \left\{u^T : \frac{d}{d\theta} \log p_\theta(x) : u \in \mathbb{R}^q\right\}$$

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

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