Contents

1 Weak convergence of Empirical Processes

1.1 Weak convergence in metric spaces

A good resource is Vdv Ch 18. We review some useful definitions regarding metric and normed spaces:

Definition 1 (Metric and Normed Spaces)**.**
Recall a **metric space** is a set $\mathbb{D}$ equipped with a metric that satisfies:

1. Symmetry: $d(x, y) = d(y, x)$
2. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$
3. $d(x, y) = 0 \iff x = y$

A *semimetric* satisfies 1 and 2 but not necessarily 3.
An *open* set is the union of open balls; a *closed* set has an open complement. The *interior* $\mathring{A}$ is the largest open set contained in A. While the *closure* $\bar{A}$ is the smallest closed set containing A.
A sequence x_n *converges* to x iff $d(x_n, x) \to 0$.
A function $f : \mathbb{D} \to \mathbb{E}$ between two metric spaces is *continuous* at a point x iff $f(x_n) \to f(x)$ for every sequence $x_n \to x$. A function is is continuous at every x iff the inverge image $f^{-1}(G)$ of every open set $G \subset \mathbb{E}$ is open in $\mathbb{D}$.
A subset of a metric space is *dense* iff its closure is the whole space. A metric space is *separable* iff it has a countable dense subset. A subset K is *totally bounded* iff for every $\epsilon > 0$, it can be covered by finitely many balls of radius ϵ. A subset of a metric space is *compact* iff it is closed and every sequence K has a converging subsequence (a subset of a semimetric space is compact iff it is totally bounded and closed).
A semimetric space is *complete* if every *Cauchy sequence*, $d(x_n, x_m) \to 0$ as $n, m \to \infty$ has a limit.

A **normed space** $\mathbb{D}$ is a vector space equipped with a *norm*, i.e., a map $|| \cdot || : \mathbb{D} \to [0, \infty)$ s.t. for every $x, y \in \mathbb{D}$ and $\alpha \in \mathbb{R}$

1. Triangle inequality: $||x + y|| \leq ||x|| + ||y||$
2. $||\alpha x|| = |\alpha| ||x||$
3. $||x|| = 0 \iff x = 0$

A seminorm satisfies 1 and 2 but not necessarily 3.

Definition 2 (Borel σ-algebra, Random element)**.** A Borel σ-algebra on a metric space $\mathbb{D}$ is the smallest σ-algebra (nonempty collection of subsets closed under complements, countable unions/intersections) that contains the open sets.
A function defined according to the metric spaces is *Borel-measurable* if it is measurable relative to the Borel σ-algebras: $f : (X, \Sigma) \to (Y, T)$ where Σ, T are the respective σ-algebras.
A borel-measurable map defined on a probability space is referred to as a **random element**.

Now we define weak convergence on metric spaces!

Definition 3 (Weak convergence, convergence in prob/a.s.)**.**
A sequence of random elements X_n with values in metric space $\mathbb{D}$ **converges weakly** to a random element

X if:

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X)) \tag{1}$$

For every bounded continuous function $f : \mathbb{D} \to \mathbb{R}$. In some cases, the random elements need not be Borel-measurable. We denote weak convergence via $X_n \rightsquigarrow X$.
An arbitrary sequence of maps $X_n : \Omega_n \to \mathbb{D}$ **converges in probability** to X if:

$$P(d(X_n, X) > \epsilon) \to 0$$

for all $\epsilon > 0$.
The sequence X_n **converges almost surely** to X if there exists a sequence of measurable random variables Δ_n s.t. $d(X_n, X) \leq \Delta_n$ and $\Delta_n \overset{a.s.}{\to} 0$.

Next we introduce the Portmanteau lemma, which provides equivalent definitions of weak convergence.

Theorem 1 (Portmanteau)**.**
For arbitrary maps $X_n : \Omega_n \to \mathbb{D}$ and random element X with values in $\mathbb{D}$, TFAE:

(i) $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ for all bounded continuous functions f

(ii) $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ for all bounded, Lipschitz functions f.

(iii) $P(X_n \in B) \to P(X \in B)$ for all Borel sets B with $P(X \subset \delta B) = 0$ (boundary prob 0)

Next we present the continuous mapping theorem, which ensures that continuous maps of convergent sequences converge to continuous map applied to the limit.

Theorem 2 (Continuous mapping theorem)**.**
Let $(\mathbb{D}, d)$ and $(\mathbb{E}, e)$ be two metric spaces. Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of $\mathbb{D}$-valued random variables and that X is $\mathbb{D}_0$-valued where $\mathbb{D}_0 \subset \mathbb{D}$. Let $f : \mathbb{D} \to \mathbb{E}$ be continuous on $\mathbb{D}_0$. Then:

1. If $X_n \rightsquigarrow X$, then $f(X_n) \rightsquigarrow f(X)$
2. If $X_n \overset{P}{\to} X$, then $f(X_n) \overset{P}{\to} f(X)$
3. If $X_n \overset{a.s.}{\to} X$, then $f(X_n) \overset{a.s.}{\to} f(X)$

We present Slutsky's theorem, which describes weak convergence proximally.

Theorem 3 (Slutsky's Theorem (HW 1.1))**.**
Suppose $(\mathbb{D}, d)$ is a metric space endowed with metric d. Also suppose that $\mathbb{D}$ is both complete (meaning every convergent sequence is Cauchy sequence) and separable (contains a countably dense subset), i.e., suppose $\mathbb{D}$ is a Polish space.

1. Suppose (X_n, Y_n) are random elements of $\mathbb{D} \times \mathbb{D}$. If $X_n \rightsquigarrow X$ and $d(X_n, Y_n) \to 0$, then $Y_n \rightsquigarrow X$.
2. Suppose (X_n, Y_n) are random elements of $\mathbb{D} \times \mathbb{D}$. If:

$$X_n \rightsquigarrow X$$
$$Y_n \rightsquigarrow Y$$

Then $(X_n, Y_n) \rightsquigarrow (X, Y)$, and continuous mapping theorem provides the application to continuous maps (sums, products, differences, quotients).

1.2 Empirical Processes

Studying the empirical process allows us to do inference on function-valued parameters, such as a cumulative distribution function! By law of large numbers and central limit theorem, we know that $F_n(t) \xrightarrow{p} F_0(t)$ and $\sqrt{n}(F_n(t) - F_0(t)) \rightsquigarrow N(0, F_0(t)(1 - F_0(t))$ for every t. Yet in this section, we improve on this result by considering $t \to F_n(t)$ as a random function.

1.2.1 Distribution Functions

The Glivenko-Cantelli theorem extends the law of large numbers and gives uniform convergence of the empirical CDF.

Theorem 4 (Glivenko-Cantelli theorem (vdV 19.1))**.** Suppose $X_1, \ldots \overset{iid}{\sim} F_0$, then $||F_n - F_0||_\infty = \sup_t |F_n(t) - F_0(t)| \overset{a.s.}{\to} 0$

The uniform/functional central limit theorem describes the convergence of the scaled empirical CDF minus the true CDF via Donsker's Theorem:

Theorem 5 (Donsker's Theorem (vdV 19.3))**.**
If $X_1, \ldots \overset{iid}{\sim} F$, the sequence of empirical processes $\sqrt{n}(F_n - F_0) \rightsquigarrow \mathbb{G}$, a mean zero Gaussian process with covariance function $F_0(\min(t_i, t_j)) - F_0(t_i)F_0(t_j)$.

Proof: If we are interested in estimating the distribution function $F_0(t)$, a natural estimator would be $P_n f(t)$ with $f \in \mathcal{F} := \{f : x \to \mathbb{1}(x \leq t) : t \in \mathbb{R}\}$ is the relevant class of functions.

Using results from 582, we know that half-line indicators in $\mathcal{F}$ have envelope function $\bar{F} = 1$ and have VC-index of 2. Thus, it holds that:

$$\sup_Q \log(N(\epsilon, \mathcal{F}, L^2(Q))) \leq C \log\left(\frac{1}{\epsilon}\right) < \infty$$

Allowing us to verify that $\mathcal{F}$ satisfies the uniform entropy integral bound in Theorem 8. Thus, $\mathcal{F}$ is Donsker, proving Donsker's Theorem.

1.2.2 Glivenko-Cantelli and Donsker (abstract)

The abstract Glivenko-Cantelli theorem make the convergence of $P_n f$ to $P_0 f$ uniform over a class of functions $\mathcal{F}$. The abstract Donsker theorem makes convergence of an empirical process evaluated at f uniform over a class of functions.

Definition 4 (Empirical process and ℓ^∞-space)**.**
We are often interested in studying the weak convergence of the empirical process evaluated at a function $f \in \mathcal{F}$:

$$\mathbb{G}f := \sqrt{n}(P_n - P_0)f$$

We are interested in studying the convergence of the stochastic process $\{\mathbb{G}f : f \in \mathcal{F}\}$. To do this, we require a metric space in which to describe stochastic convergence. A useful space for this purpose is:

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

equipped with uniform norm:

$$|| \cdot ||_{\mathcal{F}} : H \to \sup_{f \in \mathcal{F}} |H(f)|$$

Definition 5 (Glivenko-Cantelli (abstract, vdV 19.4))**.**
A class of functions $\mathcal{F}$ is Glivenko-Cantelli if:

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

Sufficient condition: a class of functions $\mathcal{F}$ with finite bracketing number $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for all $\epsilon > 0$ is Glivenko-Cantelli.

We may ask under what conditions a stochastic process $\{X_n(f) : f \in \mathcal{F}\}$ converges weakly in $\ell^\infty(\mathcal{F})$?

Theorem 6 (Weak convergence of stochastic process in $\ell^\infty(\mathcal{F})$)**.**
X_n converges in $\ell^\infty(\mathcal{F})$ to a tight random element X if and only if:

1. Convergence in distribution of marginals: for each finite collection of functions $\{f_1, \dots, f_m\} \subset \mathcal{F}$, it holds that:

$$\{X_n(f_j) : j = 1, 2, \dots, m\} \rightsquigarrow \{X(f_j) : j = 1, 2, \dots, m\}$$

2. Existence of a suitable psuedometric $\rho : \mathcal{F} \times \mathcal{F} \to [0, \infty)$ such that:
   (a) $\mathcal{F}$ not too large: $(\mathcal{F}, \rho)$ is totally bounded, i.e., $N(\epsilon, \mathcal{F}, \rho) < \infty$
   (b) X_n is smooth: X_n is asymptotically uniform equicontinuous:

   Defining: $\mathcal{F}(\delta) := \{(f_1, f_2) \in \mathcal{F}^2 \ : \ \rho(f_1, f_2) < \delta\}$
   For all positive sequences $\delta_n \to 0$, we require:

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

What about when we restrict our attention to a particular kind of stochastic process, the empirical process? When will it be $\ell^\infty(\mathcal{F})$-valued? What is its weak limit?

Definition 6 (P_0-Donsker)**.**
A function class $\mathcal{F}$ is P_0-Donsker if $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F}) \iff ||\mathbb{G}_n||_\mathcal{F} \to ||\mathbb{G}||_\mathcal{F}$ where $\mathbb{G}$ is a mean-0 Gaussian process with covariance function:

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

To show a function class is Donsker, we must satsify the conditions in Theorem 6

1. $\ell^\infty(\mathcal{F})$-valued: $\mathbb{G}_n$ is $\ell^\infty(\mathcal{F})$-valued when there exists a P_0-integrable envelope function $\bar{F}$ that upper bounds $f \in \mathcal{F}$ pointwise:

$$\sup_{f \in \mathcal{F}} |f(x)| \leq \bar{F}(x) \; \forall \; x \in \mathcal{X}$$

2. Convergence of marginals: this is guaranteed by the MV-CLT.
3. Existence of psuedometric that ensures total boundedness and asymptotic uniform equicontinity. The guaranteed pseudometric for Donsker classes can always be taken as:

$$\rho_{P_0}(f_1, f_2) = ||f_1 - f_2||_{L_2(P_0)} := \left[\int (f_1(x) - f_2(x))^2 dP_0\right]^{1/2}$$

Sufficient conditions: see Theorem 8.

We don't always have to rely on proving a function class is Donsker, as we can leverage permanence properties.

Theorem 7 (Donsker Permanance Properties)**.**
If $\mathcal{F}$ and $\mathcal{G}$ are P-Donsker classes, then the following are also P-Donsker

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

Theorem 8 (Sufficient Conditions for a Class to be Donsker (vdV 19.5, 19.14))**.**

1. Satisfy finite bracketing integral: $\mathcal{F}$ is P_0-Donsker if:

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

2. Satisfy uniform integral bound with finite square integrable envelope: $\mathcal{F}$ is P_0-Donsker if it has an

envelope function $\bar{F}$ satisfying $P\bar{F}^2 < \infty$ and

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

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

Example 1 (Constructing Confidence Bands for CDF)**.**
Suppose our goal is to construct confidence bands for a CDF:

$$F_0(t) := P_0(X \leq t)$$

We will consider estimating F_0 using the class of functions $\mathcal{H} = \{x \to \mathbb{1}(x \leq t) : t \in \mathbb{R}\}$. Recall that the stochastic process $\{\sqrt{n}(F_n(t) - F_0(t)) : t \in \mathbb{R}\} = \{\mathbb{G}_n h : h \in \mathcal{H}\}$ is just the empirical process evaluated at h as h varies in $\mathcal{H}$.
Note by Donsker's theorem that $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{H})$ where $\mathbb{G}$ is a mean-0 Gaussian process.
By the continuous mapping theorem, $||\mathbb{G}_n||_{\mathcal{H}} \rightsquigarrow ||\mathbb{G}||_{\mathcal{H}}$, i.e., the supremum norm over $\mathcal{H}$.
Since our goal is to construct confidence bands for $F_0(t)$, we must construct (L_n, U_n) s.t.

$$\lim_{n\to\infty} P(L_n(t) \leq F_0(t) \leq U_n(t)) \geq 1 - \alpha \qquad \forall\, t \in \mathbb{R}$$

We can construct a valid asymptotic $(1-\alpha)$-confidence band for F_0 via:

$$L_n(t) := F_n(t) - \frac{c}{\sqrt{n}} \qquad U_n(t) := F_n(t) - \frac{c}{\sqrt{n}}$$

where c is the $(1-\alpha)$-quantile of $||\mathbb{G}||_{\mathcal{H}}$.

To show this is a valid confidence band:

$$\begin{aligned}
&\lim_{n\to\infty} P_0(L_n(t) \leq F_0(t) \leq U_n(t)) \qquad \forall t \in \mathbb{R} \\
&= \lim_{n\to\infty} P_0\left(F_n(t) - \frac{c}{\sqrt{n}} \leq F_0(t) \leq F_n(t) - \frac{c}{\sqrt{n}}\right) \qquad \forall t \in \mathbb{R} \\
&= \lim_{n\to\infty} P_0\left(-c \leq \sqrt{n}(F_0(t) - F_n(t)) \leq c\right) \qquad \forall t \in \mathbb{R} \\
&= \lim_{n\to\infty} P_0\left(\sqrt{n}|F_n(t) - F_0(t)| \leq c\right) \qquad \forall t \in \mathbb{R} \\
&= \lim_{n\to\infty} P_0\left(\sup_t \sqrt{n}|F_n(t) - F_0(t)| \leq c\right) \\
&= \lim_{n\to\infty} P_0\left(\sup_{h\in\mathcal{H}} \sqrt{n}|(P_n - P_0)h| \leq c\right) \\
&= \lim_{n\to\infty} P_0\left(||\mathbb{G}_n||_{\mathcal{H}} \leq c\right) \\
&= P_0\left(||\mathbb{G}||_{\mathcal{H}} \leq c\right) \\
&= (1-\alpha)
\end{aligned}$$

Example 2 (Examples of Donsker Classes)**.**
Listed below are several examples of function classes that are P_0-Donsker:

1. Any VC class (vdV 19.2)
    (a) Class of half-line indicators: $\mathcal{F} := \{\mathbb{1}(x \le t) \text{ for } t \in \mathbb{R}\}$ (VC=1)
    (b) Class of indicators on bounded support: $\mathcal{F} := \{\mathbb{1}(-a \le x \le b) \text{ for } a < b \in \mathbb{R}\}$ (VC=1)
    (c) Class of bivariate half-line indicators $\mathcal{F} := \{\mathbb{1}(x \le t_1) \times \mathbb{1}(y \le t_2) \text{ for } (t_1, t_2) \in \mathbb{R}^2\}$ (VC=2)
    (d) Any boolean valued function that can be computed using a finite number of arithmetic or comparison operations.
    (e) Set of polynomials of degree less than some number: $\mathcal{F} := \{\sum \lambda_i f_i \ ; \ f_1 = 1, f_2 = x, f_3 = x^2, \ldots, f_k = x^k\}$
    (f) Union, intersection, positive/negative restriction, etc. of known VC classes (permanence properties).
2. Any function class with envelope and satisfies the bracketing/uniform entropy integral (vdV 19.2)
    (a) Bounded Lipschitz functions: $\mathcal{F} := \{f : [0,1] \to [0,1]; f(x) - f(y) \le L|x - y|\}$
    (b) Functions Lipschitz in indexing parameters: $\mathcal{F} := \{g_\beta : \beta \in \mathbb{R}^p; ||\beta||_2 \le 1 \text{ and s.t. } |g_{\beta_1}(x) - g_{\beta_1}(x)| \le m(x)||\beta_1 - \beta_2|| \text{ for measurable } m(x)\}$
    (c) Sobolev Classes: $\mathcal{F} := \{f : [0,1] \to \mathbb{R}; ||f||_\infty \le 1, f^{(k-1)} \text{ abs continuous}, \int (f^{(k)})^2 dx \le 1\}$ for $k \ge 1$.
    (d) Bounded monotone functions: $\mathcal{F} := \{f : \mathbb{R} \to [-M, M] \text{ s.t. } M < \infty\}$
    (e) Functions of bounded variation: can be considered as differences of monotone increasing functions from previous bullet. $\mathcal{F} := \{f : \mathbb{R} \to \mathbb{R}; ||f||_V \le M\}$ where $||f||_V := \int |df(x)|$ is the total variation norm.
        i. In more generality: $\mathcal{F} := \{f : \mathbb{R}^m \to \mathbb{R}; ||f||_V^* \le M\}$ is Donsker where $||f||_V^*$ is the uniform sectional variation norm.
    (f) If $\mathcal{F}$ and $\mathcal{G}$ are Donsker (have finite uniform entropy integral relative to envelopes F and G), the class $\mathcal{F}\mathcal{G}$ of functions $x \to f(x)g(x)$ is Donsker.
    (g) Fixed Lipschitz transformation: a Lipschitz function $\phi(f, g)$ is donsker if f, g range over Donsker classes $\mathcal{F}, \mathcal{G}$.
3. Any function class that is not Glivenko-Cantelli cannot be Donsker.
4. Donsker preservation properties.

The following result will become very useful later on as we study the construction of efficient estimators.

Theorem 9 (Controlling empirical process terms (vdV 19.24)).
To establish $\sqrt{n}(P_n - P_0)g_n = o_P(1)$ we require:

1. $P_0 g_n^2 = o_P(1)$
2. g_n is in a Donsker class $\mathcal{F}$

Variant: to establish $\sqrt{n}(P_n - P_0)(h_n - h_0) = o_P(1)$ we require:

1. $\{h_k\}_{k=1}^\infty$ is a sequence of random functions in $L^2(P)$ s.t., $P(h_n \in \mathcal{F}) \to 1$ for Donsker class $\mathcal{F} \subset L^2(P)$
2. $P(h_n - h_0)^2 = o_P(1)$ for some $h_0 \in \mathcal{F}$

2 Asymptotic Linearity

Motivation: two statisticians are asked to estimate ψ_1 and ψ_2. Together their goal is to estimate $\psi_0 = \psi_1 + \psi_2$. Suppose that both ψ_{1n} and ψ_{2n} are $\sqrt{n}$**-consistent and asymptotically normal (CAN)**. Will $\psi_{0n} := \psi_{1n} + \psi_{2n}$ be CAN? Not necessarily, but in some cases, yes!

Definition 7 (Asymptotic Linearity, Influence Function)**.**
An estimator ψ_n of ψ_0 is **asymptotically linear** if it can be written as:

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

Where ϕ_{P_0} satisfies:

1. P_0-Mean 0: $P_0\phi_{P_0} = 0$
2. P_0-squared integrable: $P_0\phi_{P_0}^2 < \infty$

The function ϕ_{P_0} is known as the **influence function** of ψ_n. Heuristically, the influence function evaluated at X_i measures first-order contribution of observation i to the estimator.

Note that we've defined asymptotically linear in such a way that implies *consistency* and *asymptotic normality*!

$$\begin{aligned}\sqrt{n}(\psi_n - \psi_0) &= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} \phi_{P_0}(X_i) + o_P(1/\sqrt{n})\right]\\ &= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} \phi_{P_0}(X_i) - \mathbb{E}[\phi_{P_0}(X)]\right] + o_P(1)\\ &\rightsquigarrow N(0, \text{Var}[\phi_{P_0}])\end{aligned}$$

Example 3 (ALE examples)**.**
Sample mean: a finite-sample and asymptotically linear estimator. $\psi_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of $\psi_0 = \mathbb{E}_{P_0}[X]$. We can write it as:

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

Implying the influence function is $\phi_{P_0}(x) = x - \psi_0$.
Sample variance (unknown mean): a nonlinear but asymptotically linear estimator. Let $\psi_0 = \sigma_0^2$ which we estimate via $\psi_n = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}[X_i - \mu_n]^2$ with $\mu_n = \mathbb{E}_{P_n}(X)$. Letting $\mu_0 = \mathbb{E}_{P_0}(X)$, we have:

$$\begin{aligned}\psi_n &= \sigma_n^2 \equiv \text{Var}[X] \equiv \text{Var}[X - \mu_0]\\ &= \frac{1}{n}\sum_{i=1}^{n}[X_i - \mu_0]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)\right]^2\end{aligned}$$

Implying

$$\begin{aligned}\psi_n - \psi_0 &\equiv \sigma_n^2 - \sigma_0^2 \\ &= \frac{1}{n}\sum_{i=1}^n [(X_i-\mu_0)^2 - \sigma_0^2] - \underbrace{\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu_0)\right]^2}_{\star}\end{aligned}$$

And noting that $\star$ is just the square of a sample mean minus and true mean, by CLT we obtain:

$$\star = (O_p(n^{-1/2}))^2 = O_p(n^{-1}) \equiv o_p(n^{-1/2})$$

Thus, σ_n^2 is asymptotically linear with influence function $\phi_{P_0}(x) = [x-\mu_0]^2 - \sigma_0^2$.

Suppose we knew the true mean μ_0. We might instead use the finite-sample linear estimator $\sigma_{n0}^2 := \frac{1}{n}\sum_{i=1}^n [X_i - \mu_0]^2$. However,

$$\begin{aligned}\sigma_n^2 - \sigma_{n0}^2 &= [\sigma_n^2 - \sigma_0^2] - [\sigma_{n0}^2 - \sigma_0^2] \\ &= \frac{1}{n}\sum_{i=1}^n [(X_i-\mu_0)^2 - \sigma_0^2] + o_P(n^{-1/2}) - \frac{1}{n}\sum_{i=1}^n [(X_i-\mu_0)^2 - \sigma_0^2] \\ &= o_P(n^{-1/2})\end{aligned}$$

So we don't gain much by knowing the nuisance parameter!
p-th sample quantile: Goal is to estimate $Q_0(p)$, the p-th quantile. Let P_0 have distribution function F_0 and density f_0. Let $Q_n(p)$ denote the p-th sample quantile. Then:

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

Theorem 10 (Z-estimators are ALE (vdV 5.21))**.**
No nuisance: Consider estimating a summary $\psi_0 \in \mathbb{R}$ of P_0 that is defined as the unique solution in ψ to:

$$P_0 U(\psi) = 0$$

An estimating equation-based estimator, or **Z-estimator**, ψ_n is defined as the solution to the estimating equation:

$$P_n U(\psi) = 0$$

In fact, we only require that there be a near solution.

$$P_n U(\psi) = o_P(n^{-1/2})$$

Under either of these two conditions, the Z-estimator has the form:

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^n \left(-\frac{\partial}{\partial\psi} P_0 U(\psi)\Big|_{\psi=\psi_0}\right)^{-1} U(\psi_0)(X_i) + o_P(n^{-1/2}) \tag{3}$$

where $\phi_{P_0}(x) = \left(-\frac{\partial}{\partial\psi} P_0 U(\psi)\Big|_{\psi=\psi_0}\right)^{-1} U(\psi_0)(x)$ is the influence function of ψ_n. It is clearly P_0-mean-0 because $P_0 U(\psi_0)(x) = 0$. To check that it is P_0-squared integrable, we require more information.

Nuisance (HW 2.3): suppose now that the estimating function now depends on a nuisance parameter: $U(\psi, \eta)$. Suppose that ψ_0 is the solution in ψ to the equation $P_0 U(\psi, \eta_0)$. Suppose an ALE is available for η_0, η_n with IF φ_{P_0}. Define ψ_n to be a solution or near solution in ψ to:

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

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

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

The delta-method allows us to characterize the limiting distribution of differentiable transformations of asymptotic linear estimators!

Theorem 11 (Delta Method (vdV 3.1))**.**
Suppose ψ_n is an estimator of $\psi_0 \in \mathbb{R}^d$ s.t.,

$$\sqrt{n}(\psi_n - \psi_0) \rightsquigarrow N(0, \Sigma)$$

If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then

$$f(\psi_n) - f(\psi_0) = \langle \psi_n - \psi_0, \nabla f(\psi_0)\rangle + o_P(n^{-1/2})$$

Theorem 12 (Delta Method for ALEs)**.**
Suppose $\psi_n \in \mathbb{R}^d$ is an asymptotically linear estimator of $\psi_0 \in \mathbb{R}^d$, implying $\psi_{n,j}$ is ALE for $\psi_{0,j}$ for all $j \in \{1, \ldots, d\}$. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable (at ψ_0). Then $f(\psi_n)$ is itself an asymptotically linear estimator for $f(\psi_0)$ with influence function equal to:

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

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

Example 4 (Goodness of fit statistics (vdV 19.23))**.**
Consider the Kolmogorov-Smirnov statistic: $\sqrt{n}||F_n - F_0||_\infty \rightsquigarrow ||\mathbb{G}||_\infty$ by continuous mapping theorem.

One could test whether the distribution matches an assumed $H_0 := F_0 = F^*$, construct a null 95% uniform CI, and test whether F_n calls in the interval.

One could also test against a broader null hypothesis, such as P belongs to a certain family $\{P_\theta : \theta \in \Theta\}$. It is natural that we consider discrepancy between P_n and $P_{\hat{\theta}}$ where $\hat{\theta}$ is an estimator of θ. Consider the null $H_0 : F_0$ is normal. The modified Komolgorov-Smirnov statistic for testing normality is:

$$\sup_t \sqrt{n}\left|F_n(t) - \Phi\left(\frac{t - \bar{X}}{S}\right)\right|$$

Here, the limit distribution is a misture of a Gaussian process and a drift term:

$$\begin{aligned}\sqrt{n}(P_n - P_{\hat{\theta}}) &= \sqrt{n}(P_n - P_\theta) - \sqrt{n}(P_{\hat{\theta}} - P_\theta) \\ &= \sqrt{n}(P_n - P_\theta) - \sqrt{n}(\hat{\theta} - \theta)^T \dot{P}_\theta\end{aligned}$$

If P is Frechét differentiable. If $\hat{\theta}_n$ is asymptotically linear with influence function $\psi_\theta(x)$, then:

$$\begin{aligned}\sqrt{n}(P_n - P_{\hat{\theta}})f &\rightsquigarrow \mathbb{G}_{P_\theta} f - \mathbb{G}_{P_\theta} \psi_\theta^T \dot{P}_\theta f \quad \text{Uniformly over } f \in \mathcal{F} \\ \sqrt{n}(F_n - F_{\hat{\theta}}) &\rightsquigarrow \mathbb{G}_{P_\theta} f - \mathbb{G}_{P_\theta} \psi_\theta^T \dot{P}_\theta f \quad \text{Uniformly over } f \in \mathcal{F} := \{\mathbb{1}(x \leq t)\} \\ \sup_t \sqrt{n}\left|F_n(t) - \Phi\left(\frac{t - \bar{X}}{S}\right)\right| &\rightsquigarrow ||\mathbb{G}_{P_\theta} f - \mathbb{G}_{P_\theta} \psi_\theta^T \dot{P}_\theta f||_{\mathcal{F}}\end{aligned}$$

2.1 V/U-statistics

Many parameters of interest can be written as:

$$V(P) = \int\int\cdots\int H(x_1, \ldots, x_m) dP(x_1) \ldots dP(x_m) \tag{4}$$

With the function H known as the **kernel**. Some examples include:

1. General raw moment: $V(P) = \int g(x) dP(x)$
2. Variance: $V(P) = \int\int \frac{1}{2}(x_1 - x_2)^2 dP(x_1) dP(x_2)$
3. Kendall's Tau: $V(P) = 4P(X_1 < X_2, Y_1 < Y_2) - 1$ or

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

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

$$\int\int \left[\int \{\mathbb{1}(x_1 \leq u) - F^*(u)\}\{\mathbb{1}(x_2 \leq u) - F^*(u)\} F^*(du)\right] dP(x_1) dP(x_2)$$

Definition 8 (V-statistic)**.**
For functions written in the form of Equation 4, a natural estimator is obtained by plugging in the empirical distribution P_n leading to a **V-statistic**:

$$V_n := V(P_n) = \frac{1}{n^m} \sum_{i_1=1}^{n} \ldots \sum_{i_m=1}^{n} H(X_{i_1}, \ldots, X_{i_m}) \tag{5}$$

Some examples include:

1. General raw moment: $V_n = \frac{1}{n}\sum_{i=1}^n g(X_i)$
2. Variance: $V(P) = \frac{1}{2n^2}\sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2$ or $V_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$
3. Kendall's Tau:

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

4. Cramer-von Mises GOF criterion:

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

The linearization argument presented below is useful for determining the asymptotic distribution of a V-statistic under nondegeneracy.

Theorem 13 (Linearization of V-statistic; Non-Degenerate)**.**
V-statistics can be written in a linear form. Assume V is symmetric in its arguments. If it is not, we can symmetrize it by computing an average when we permute the arguments. By inducting on m, we can show that:

$$\begin{aligned} V_n - V_0 &= (P_n^m - P_0^m)H \\ &= \sum_{k=1}^{m} \binom{m}{k} (P_n - P_0)^k H_k \end{aligned}$$

Where $H_k := (P_n - P_0)^{m-k} H = \int \dots \int H(x_1, \dots, x_k, x_{k+1}, \dots, x_m) P_0(dx_{k+1}) \dots P_0(dx_m)$ is simply the function when we've integrated out all the terms excluding the 1 through k-th terms.
Defining $\tau_k^2 := \text{Var}[H_k(X_1 \dots, X_k)]$ to be the variance of the k-variate function, and letting $a := \min(a : \tau_a^2 > 0)$, the dominant term in the above expansion is:

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

If $a = 1$, we can rewrite the above expansion as

$$V_n - V_0 = m(P_n - P_0)H_1 + \sum_{k=2}^{m} \binom{m}{k} (P_n - P_0)^k H_k \tag{6}$$

Where the remaining terms are higher orders of $(P_n - P_0)$. Thus, this expansion provides a first-order approximation for the estimation error. Under $a = 1$, the asymptotic behavior is dictated by the dominant first order term. If $0 \neq \text{Var}_0(H_1(X)) =: \tau_1^2$, i.e., the distribution is **non-degenerate**, the dominant first-order term is:

$$m(P_n - P_0)H_1 \equiv \frac{1}{n} \sum_{i=1}^{n} m(H_1(X_i) - V_0)$$

Provided $\mathcal{H} := \{x \to H(x, x_2, \dots, x_m) : (x_2, \dots, x_m) \in \mathcal{X}_0^{m-1}\}$ is P_0-Donsker, V_n is ALE with influence function $\phi_{P_0}(x) = m(H_1(x) - V_0)$, implying

$$\begin{aligned} V_n - V_0 &= \frac{1}{n} \sum_{i=1}^{n} m(H_1(X_i) - V_0) + o_P(n^{-1/2}) \\ \sqrt{n}(V_n - V_0) &\rightsquigarrow N(0, m^2 \tau_1^2) \end{aligned}$$

Note that V_n is typically biased for V_0. Consider the case of $m = 2$:

$$\begin{aligned}
n^2\mathbb{E}[V_n] &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} H(X_i, X_j)\right] \\
&= \mathbb{E}\left[\sum_{i\neq j}^{n} H(X_i, X_j)\right] + \mathbb{E}\left[\sum_{i=1}^{n} H(X_i, X_i)\right] \\
&= n(n-1)\mathbb{E}[H(X_1, X_2)] + n\mathbb{E}[H(X_1, X_1)] \\
&= n(n-1)V_0 + n\mathbb{E}[H(X_1, X_1)] \\
\implies n^2(\mathbb{E}[V_n] - V_0) &= n\left[\mathbb{E}[H(X_1, X_1)] - V_0\right] \\
\implies (\mathbb{E}[V_n] - V_0) &= \frac{1}{n}\left[\mathbb{E}[H(X_1, X_1)] - V_0\right] \neq 0
\end{aligned}$$

So V_n is biased for V_0 despite the bias decaying to 0 at a $\frac{1}{n}$ rate. But to correct for the finite sample bias, we motivate an alternative that avoids ties and leans on independent pairs of observations.

Definition 9 (U-statistics)**.**
A **U-statistic** averages out $H(X_{i_1}, \ldots, X_{i_m})$ over a set of *unique* indices, which eliminates the cases of matching indices which introduces bias:

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

where $\mathcal{D}_{m,n} := \{\underline{i}_m \subset \{1, \ldots, n\} := (i_1, \ldots, i_2, \ldots, i_m) : 1 \leq i_1 < \ldots < i_m \leq n\}$. We assume WLOG that the kernel H is symmetric in its arguments.

Are U_n and V_n asymptotically equivalent? And can we still obtain a nice linearization of the U-statistic?

Theorem 14 (Linearization of U-statistic via Hajék projection)**.**
In order to obtain the same linearization for a U-statistic $U_n - V_0$ where:

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

where $\mathcal{D}_{m,n} := \{\underline{i}_m \subset \{1, \ldots, n\} := (i_1, \ldots, i_2, \ldots, i_m) : 1 \leq i_1 < \ldots < i_m \leq n\}$, we can consider a basic case and use projection methods for the general case.

Case 1 (m=2): defining

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

We have

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

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

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

Thus, U_n is ALE for V_0 with IF: $\phi : x \to m[H_1(x) - V_0]$.

Case 2 (general): The idea is to find the closest approxiamtion to $U_n - V_0$ within the class of random variables of the form $\sum_{i=1}^n g_i(X_i)$ for P_0-square-integrable functions $g_i : \mathcal{X}_0 \to \mathbb{R}$. Turns out for a given mean zero function of the observations $T_n := T_n(X_1, \ldots, X_n)$, the projection of T_n onto space of random variables of the desired form is given by the *Hajék projection*:

$$\bar{T}_n := \sum_{i=1}^n \mathbb{E}_0[T_n(X_1, X_2, \ldots, X_n)|X_i]$$

Noting that:

$$\mathbb{E}_0(U_n - V_0|X_i) = \frac{m}{n}[H_1(X_i) - V_0]$$

Therefore, the Hajék projection of $(U_n - V_0)$ onto the space of functions provides the desired linearization:

$$\bar{U}_n := \frac{1}{n}\sum_{i=1}^n m[H_1(X_i) - V_0]$$

Therefore, U_n is an ALE for V_0 with IF $\phi : x \to m[H_1(x) - V_0]$ and is therefore asymptotically equivalent to V_n.

A nice property of the U-statistic is we can find its finite-sample/asymptotic variance!

Theorem 15 (Finite-sample/asymptotic variance of U-statistic)**.**
Setting $\tau_k^2 := \text{Var}_0[H_k(X_1, \ldots, X_k)]$, the variance of the k-th variate function, it can be shown:

$$\text{Var}_0[U_n] = \binom{n}{m}^{-1} \sum_{k=1}^m \binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}\tau_k^2$$

Under $a := \min\{k \geq 1 : \tau_k^2 > 0\}$, we have

$$\text{Var}_0[U_n] = \frac{a!}{n^a}\binom{m}{a}^2 \tau_a^2 + O(n^{-(a+1)})$$

implying:

$$n^a \text{Var}(U_n) \overset{n\to\infty}{\longrightarrow} a!\binom{m}{a}^2 \tau_a^2 \tag{7}$$

We end this subsection with some examples of U-statistics:

Example 5 (U-statistics)**.**
Case 1 (Sample variance):

$$\begin{aligned} H(x_1, x_2) &= \frac{1}{2}(x_1 - x_2)^2 \\ H_1(x) = \mathbb{E}_{P_0}[H(x, X)] &= \frac{1}{2}\mathbb{E}_{P_0}[(X - x)^2] = \frac{1}{2}\left(\sigma_0^2 + (\mu_0 - x)^2\right) \\ \tau_1^2 &= \mathrm{Var}_{P_0}(H_1(X)) = \frac{1}{4}\left(\mathbb{E}_{P_0}((X-\mu_0)^4) - \sigma_0^4\right) \end{aligned}$$

Provided $\tau_1^2 \neq 0$, we have that:

$$\begin{aligned} U_n - V_0 &= \frac{1}{n}\sum_{i=1}^{n} m(H_1(X_i) - V_0) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \mu_0)^2 - \sigma_0^2\right] + o_P(n^{-1/2}) \\ \implies n^{1/2}(U_n - V_0) &\rightsquigarrow N(0, 4\tau_1^2) \end{aligned}$$

Case 2 (Kendall's Tau): let $Z = (X, Y)$ denote the random data unit and let $F_L(x, y) := P_0(X < x, Y < y)$ and $F_U(x, y) := P_0(X > x, Y > y)$.

$$\begin{aligned} H(z_1, z_2) &= 2\left[\mathbb{1}(x_1 < x_2, y_1 < y_2) + \mathbb{1}(x_1 > x_2, y_1 > y_2)\right] - 1 \\ H_1(z) &= \mathbb{E}_{P_0}[H(z, Z)] = 2[P_0(X < x, Y < y) + P_0(X > x, Y > y)] - 1 \\ \tau_1^2 &= \mathrm{Var}_{P_0}[H_1(Z)] = 4\mathrm{Var}_{P_0}[F_L(X, Y) + F_U(X, Y)] \end{aligned}$$

Provided $\tau_1^2 \neq 0$, we have that:

$$\begin{aligned} U_n - V_0 &= \frac{1}{n}\sum_{i=1}^{n} m(H_1(X_i) - V_0) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} 2\cdot\left(2[F_L(X_i, Y_i) + F_U(X_i, Y_i)] - 1 - (4P_0(X_1 < X_2, Y_1 < Y_2) - 1)\right) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} 4\cdot[F_L(X_i, Y_i) + F_U(X_i, Y_i) - 2P_0(X_1 < X_2, Y_1 < Y_2)] + o_P(n^{-1/2}) \\ \implies n^{1/2}(U_n - V_0) &\rightsquigarrow N(0, 4\tau_1^2) \end{aligned}$$

Noting that under $H_0 : X \perp Y$:

$$\begin{aligned} V_0 &= 4P_0(X_1 < X_2, Y_1 < Y_2) - 1 = 4[P_0(X_1 < X_2)\cdot P(Y_1 < Y_2)] - 1 = 0 \\ F_L(X, Y) &= F_0(X)F_0(Y) \text{ and } F_U(X, Y) = (1 - F_0(X))(1 - F_0(Y)) \\ \tau_1^2 &= \mathrm{Var}_{P_0}[H_1(Z)] = \mathrm{Var}_{P_0}[2(F_0(X)F_0(Y) - (1 - F_0(X))(1 - F_0(Y))) - 1] \\ &= \mathrm{Var}_{P_0}[2UV - 2(1-U)(1-V)] \quad \text{(for U, V std unif)} \\ &= \frac{1}{9} \quad \text{(Law total var)} \\ \implies n^{1/2}(U_n) &\rightsquigarrow N(0, 4/9) \end{aligned}$$

Case 3 (Cramer von-Mises Measure): denoting $F_0(x) = P_0(X \leq x)$, consider $H_0 : F_0 = F^*$:

$$\begin{aligned} H(x_1, x_2) &= \int \{\mathbb{1}(x_1 \leq u) - F^*(u)\}\{\mathbb{1}(x_2 \leq u) - F^*(u)\} dF^*(u) \\ H_1(x) = \mathbb{E}_0[H(x, X)] &= \int \{F_0 * (u) - F^*(u)\}\{\mathbb{1}(x_2 \leq u) - F^*(u)\} dF^*(u) \overset{H_0}{=} 0 \\ \tau_1^2 &\overset{H_0}{=} 0 \end{aligned}$$

Thus, the U-statistic has degenerate order-1 asymptotic distribution. Yet $\tau_2^2 = 1/90$ suggesting. This motivates the idea of whether:

$$n^{a/2}[U_n - V_0] = n[U_n - V_0] \rightsquigarrow Y$$

Defining $V_n = \int [F_n(u) - F^*(u)]^2 dF^*(u) \equiv \int [F_n(u) - F_0(u)]^2 dF_0(u)$ under H_0:

$$\begin{aligned} nV_n &= \int [n^{1/2}(F_n(u) - F_0(u))]^2 dF_0(u) \\ &= \int_0^1 [n^{1/2}(P_n - P_0)\mathbb{1}(x \leq u)]^2 dF_0(u) \\ &= \int_0^1 [\mathbb{G}(u)]^2 du \end{aligned}$$

Which holds by Donsker Theorem and Continuous mapping theorem.
What can we say about the U-statistic? Recalling that:

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

Where $D_n := \frac{1}{n}\sum_{i=1}^n H(X_i, X_i)$, we can obtain that $D_n \xrightarrow{p} \mathbb{E}_{P_0}[H(X_1, X_1)] = \frac{1}{6}$. Thus, by Slutsky's Lemma under H_0 with $V_0 = 0$:

$$\begin{aligned} n(U_n - V_0) &= \frac{nV_n - D_n}{1 - \frac{1}{n}} = \frac{n(V_n - V_0) - D_n}{1 - \frac{1}{n}} \\ &\xrightarrow{n \to \infty} \frac{n(V_n - V_0) - D_n}{1} = \int_0^1 [\mathbb{G}(u)]^2 du - \frac{1}{6} \end{aligned}$$

We previously illustrated 1-degenerate U/V-statistics were those with first order behavior (in $P_n - P_0$) that equaled 0 by $\text{Var}_{P_0}(H_1) := \tau^2 = 0$. An example of this is estimating the Cramer von-Mises measure. What can we conclude about the asymptotic behavior of such estimators?

Theorem 16 (Asymptotic Distribution of a 1-degenerate U/V-statistic)**.**
Suppose H is a symmetric kernel and $m \geq 2$. Suppose the kernel is 1-degenerate. and $\tau_2^2 > \tau_1^2 = 0$. Then the U-statistic has asymptotic distribution:

$$n(U_n - V_0) \rightsquigarrow \sum_{k=1}^{\infty} \lambda_j (Z_j^2 - 1)$$

Where $Z_j \overset{iid}{\sim} N(0, 1)$, $Z_j^2 \sim \chi_1^2$, and $Z_j^2 - 1$ is mean-centered chi-square, and λ_j are the eigenvalues of a certain linear operator.

Under appropriate regularity conditions we also get:

$$n(V_n - V_0) \rightsquigarrow \sum_{k=1}^{\infty} \lambda_j (Z_j^2)$$

2.2 Functional Delta Method

We have seen that when an estimator ψ_n of ψ_0 is writable as a differentiable function h of another estimator θ_n of θ_0, we can use the delta method to study the asymptotic behavior of ψ_n based on the behavior of θ_n.

$$\psi_n - \psi_0 = h(\theta_n) - h(\theta_0) = h'(\theta_0)(\theta_n - \theta_0) + \text{rem}_h(\theta_n, \theta_0)$$

What if we wish to study a fixed *functional* on the empirical distribution, $\psi_n := \Psi(F_n)$. Is there an analogous representation to the delta method/Taylor expansion?

Step one to assessing the asymptotic behavior of a plug-in estimator $\psi(F_n)$ is ensuring consistency. This requires a continuity condition on the functional, and consistency follows by a continuous mapping theorem argument.

Definition 10 (ρ-continuity)**.**
Let $\mathcal{P}$ denote the statistical model, i.e., the set of all CDFs. Assume the set is convex. Let ρ denote a norm on $\mathcal{P}$. A functional $\psi : \mathcal{P} \to \mathbb{R}$ is said to be $\rho-$continuous at $\tilde{F} \in \mathcal{P}$ if for all deterministic sequences $\{\tilde{F}_k\}_{k=1}^{\infty} \subset \mathcal{P}$ s.t.

$$\rho(\tilde{F}_k - \tilde{F}) \longrightarrow 0$$

Implies

$$\psi(\tilde{F}_k) \longrightarrow \psi(\tilde{F})$$

This motivates a continuous mapping theorem.

Theorem 17 (Continuous Mapping Theorem)**.**
If ψ is a $\rho-$continuous functional at F_0 and $\rho(F_n - F_0) \xrightarrow{p} 0$, then:

$$\psi(F_n) \xrightarrow{p} \psi(F_0)$$

The next step is demonstrating the asymptotic linearity of the functional, which requires a Taylor expansion. What notion of differentiability do we need for the functional?

Definition 11 (Gâteaux Differentiability)**.**
Suppose $\mathcal{P}$ is convex, meaning for $F_1, F_2 \in \mathcal{P} \to \alpha F_1 + (1-\alpha)F_2 \in \mathcal{P}\ \forall \alpha \in (0,1)$.
By noting $F_n = F_0 + \frac{1}{\sqrt{n}}\sqrt{n}(F_n - F_0)$, we note that $\sqrt{n}(F_n - F_0) \in \mathcal{Q}(F_0) := \{c(F - F_0) : c \in \mathbb{R}, F \in \mathcal{P}\}$.

Suggests studying $\epsilon \in \mathbb{R}$ and $h \in \mathcal{Q}(F_0)$.
The *Gâteaux derivative* of Ψ at $F \in \mathcal{P}$ in the direction of $h \in \mathcal{Q}(F)$ is given by:

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

A functional Ψ is **Gâteaux differentiable** at $F \in \mathcal{P}$ if the derivative exists for all directions $h \in \mathcal{Q}(F)$ and if $h \to \dot{\Psi}(F;h)$ is a linear functional.

However, as we will see, Gâteaux differentiability is insufficient to give us the asymptotic linearity of $\Psi(F_n)$. Define the remainder of the approximation to be

$$R_{F_0,\epsilon} := \frac{\Psi(F_0+\epsilon h) - \Psi(F)}{\epsilon} - \dot{\Psi}(F_0;h)$$

Setting $\epsilon = \epsilon_n = n^{-1/2}$ and $h = h_n = \sqrt{n}(F_n - F_0)$, we have:

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

How can we show that $R_{F_0,\epsilon_n}(h_n) = o_P(1)$?

Turns out, Gâteaux differentiability only implies that $R_{F_0,\epsilon}(h) \to 0$ for some fixed direction $h \in \mathcal{Q}(F)$. However, the remainder term above depends on a random direction. We want this result to hold uniformly over all directions. Thus, we require a stronger form of functional differentiability.

Definition 12 (Hadamard Differentiablity)**.**
Suppose that the functional $\Psi : \mathcal{P} \to \mathbb{R}$ has differential $\dot{\Psi}(F,h)$ at $F \in \mathcal{P}$ and in direction $h \in \mathcal{Q}(F)$. Suppose the remainder term above tends to zero uniformly over directions $h \in H$ such that $H \in \mathcal{H} := \{\text{all compact subsets of } \{\mathcal{Q}(F), \rho\}\}$:

$$\lim_{\epsilon \to 0} \left[\sup_{h \in H} |R_{F_0,\epsilon}(h)| \right] = 0 \qquad \text{for each } H \in \mathcal{H}$$

Theorem 18 (Negligibility of the remainder under Hadamard differentiability)**.**
The remainder term $R_n := R_{F_0,\epsilon_n}(h_n)$ is $o_P(1)$ if Ψ is Hadamard differentiable at F_0 relative to $\rho = ||\cdot||_\infty$.

Theorem 19 (Functional Delta Method)**.**
Suppose Ψ is Hadamard differentiable at F_0 relative to $\rho : (h_1,h_2) \to \sup_x |h_1(x) - h_2(x)| \equiv ||\cdot||_\infty$. Letting $\epsilon_n = n^{-1/2}$ and $H_n := \sqrt{n}(F_n - F_0)$, it holds that $R_{F_0,\epsilon_n}(H_n) = o_P(1)$, implying:

$$\begin{aligned} \Psi(F_n) - \Psi(F_0) &= \dot{\Psi}(F_0;F_n - F_0) + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \dot{\Psi}(F_0;\mathbb{1}(X_i \le \cdot) - F_0) + o_p(n^{-1/2}) \end{aligned} \tag{8}$$

To ascertain whether $\Psi(F_n)$ is asymptotically linear, we must check whether the proposed IF is P_0 mean zero and finite squared integrable. Since $\dot{\Psi}$ is linear, we know it's mean 0.

The following is a useful theorem that allows us to switch the order of integration.

Theorem 20 (Integration by Parts)**.**
Let $g : [a, b] \to \mathbb{R}$ and $h : [a, b] \to \mathbb{R}$ be cadlag (continuous from right with limits on left) functions. Note: CDFs are cadlag with bounded variation. It holds that:

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

where $h(u-)$ is the left hand limits. If at least one of the two functions is continuous

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

3 Efficiency Theory

3.1 Parametric Efficiency Theory

Suppose a *statistical model* $M = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}$ is a regular parametric model and all members of M are absolutely continuous wrt Lesbegue measure (i.e., $\mu(A) = 0 \implies P_\theta(A) = 0$ for $A \in \mathcal{B}$).

Suppose we observed $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta_0}$ with $\theta_0 \in \Theta$ are are interested in estimating $\tau_0 := \tau(\theta_0)$.

The **Fisher information** for θ is defined as

$$\mathcal{I}(\theta) := P_\theta \left(\frac{\partial}{\partial \theta} \log p_\theta \right)^2$$

which measures the curvature of the log-likelihood. The curvier, the more information there is about a parameter.

Can we characterize the optimal efficiency for estimators of a target of interest, τ_0? Hájek's convolution theorem gives us results for certain classes of models (sufficiently smooth, QMD models) and certain estimators (regular estimators). We first begin with a definition of regular estimators and QMD.

Definition 13 (QMD)**.**
A statistical model is sufficiently smooth, quadratic mean differentiable, at θ if there exists a function $\dot{\ell}_\theta$ s.t.:

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

QMD-ness ensures that the score function has mean zero ($P_\theta \dot{\ell} = 0$) and the Fisher Information exists.

Definition 14 (Regular Estimator)**.**
An estimator τ_n of τ_0 is regular if $\forall h \in \mathbb{R}$, it holds that:

$$\sqrt{n} \left(\tau_n - \tau \left(\theta_0 + \frac{h}{\sqrt{n}} \right) \right) \rightsquigarrow Z$$

Where Z does not depend on h. Essentially, regular estimators have limiting distributions that are stable uniformly under small fluctuations about θ_0.

Theorem 21 (Hájek's Convolution Theorem)**.**
If (a) the statistical model M is sufficiently smooth (Quadratic mean differentiable), (b) the information is nonzero, $\mathcal{I}(\theta_0) > 0$, and (c) τ_n is a regular estimator of τ_0 with $n^{-1/2}(\tau_n - \tau_0) \rightsquigarrow Z$.

There exist two independent variables Z_0 and Δ_0 s.t. $Z \overset{d}{=} Z_0 + \Delta_0$ with $Z_0 \sim N(0, v_0)$ where

$$v_0 = v_0(M) := \left(\frac{\partial}{\partial \theta} \tau(\theta) \Big|_{\theta = \theta_0} \right)^2 \frac{1}{\mathcal{I}(\theta_0)}$$

A simple corollary is that **any regular estimator τ_n of τ_0 has asymptotic variance of $\sqrt{n}(\tau_n - \tau_0)$ no smaller than v_0.**

A regular estimator achieving this bound is said to be *efficient*.

3.2 Efficiency in more general models

Suppose that $X_1, \dots, X_n \overset{iid}{\sim} P_0 \in M$ and consider the functional $\psi : M \to \mathbb{R}$.

Our goal is to estimate $\psi_0 := \psi(P_0)$ from the observed data. What is the best variance we can achieve when M is statistical model indexed by an infinite dimensional parameter?

Guiding principle: estimation of ψ_0 in M should be at least as hard as estimating it in any (parametric) submodel $M_1 \subset M$ containing P_0.

Let $\mathcal{H}(P_0)$ denote an index set of all smooth (QMD) one-dimensional parametric submodels of M centered at P_0. In other words, for each model in the index set, $h \in \mathcal{H}(P_0)$, there exists a $\delta > 0$ s.t.

1. Submodels in M: $P_{\theta,h} \in M$ for all $\theta \in [0, \delta)$
2. P_0 is origin: $P_{\theta,h} = P_0$ when $\theta = 0$
3. QMD at origin: $M_h = \{P_{\theta,h} : \theta \in [0, \delta)\}$ is QMD at $\theta = 0$.

A *regular* estimator in an infinite dimensional model is regular with respect to all parametric submodels (regular relative to M_h for all $h \in \mathcal{H}(P_0)$).

Our objective is to find the lower bound on the variance of the limiting distribution of a regular estimator in our model M, $v_0^*(M)$.

Definition 15 (Generalized Cramer-Rao Lower Bound)**.**
It is no easier to estimate ψ_0 in an infinite dimensional model M, than over all possible submodels M_h.

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

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

This inequality is known as the **Generalized Cramer Rao Lower Bound**. Where

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

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

3.2.1 Outline and Suppositions

We saw in the definition of the generalized Cramer-Rao Lower Bound that the fisher information $\mathcal{I}_{M_h}(0)$ only depends on the direction h, i.e. the choice of QMD submodels, through the score g_h.

Under **pathwise differentiability** condition on ψ, we have that the numerator of the GCRLB also depends on h through the score g_h. Specifically, there exists a P_0-mean zero square integrable function $D(P_0) : \mathcal{X} \to \mathbb{R}$, the **gradient**, s.t. for all $h \in \mathcal{H}(P_0)$

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

If this holds, the GCRLB becomes

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

Where g is a score and $\mathcal{G}(P_0) := \{g_h : h \in \mathcal{H}(P_0)\}$ is the collection of scores of all QMD submodels centered at P_0, i.e., **tangent set** of M at P_0

When does this bound have a closed form? If we assume that $g = cD(P_0)$ for some $c \in \mathbb{R}$, i.e., that $cD(P_0) \in \mathcal{G}(P_0)$, then

$$v_0^*(M) \geq \sup_{g \in \mathcal{G}(P_0)} \frac{(P_0[D(P_0)g])^2}{P_0(g^2)} = P_0(D(P_0)^2)$$

In this case, $D(P_0)$ is referred to as the **canonical gradient/efficient influence function**.

3.2.2 Pathwise Differentiability

We focus on the condition required to ensure that the derivative of the functional ψ in the GCRLB numerator is well defined. Over a parametric model (single path), usual differentiability of real-valued functions suffices. Can we generalize this to define derivatives over all smooth parametric paths?

Definition 16 (Pathwise differentiability)**.**
A parameter ψ is **pathwise differentiable** (at P_0 relative to M) if for all $h \in \mathcal{H}(P_0)$ (i.e., collection of QMD submodels) there exists a P_0-mean zero function $D(P_0)$ known as the **gradient** such that

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

Note that pathwise differentiability connects the derivative of the functional to the score under the submodel.

Now we present an example with a nice choice of QMD submodel.

Example 6 (Example gradient of generalized moment functional)**.**
Suppose M is a nonparametric model. Consider the generalized moment functional $\psi = Pf$ where f is a fixed function s.t. $Pf^2 < \infty$. We will consider various choices of $h \in \mathcal{H}(P_0)$, 1-dimensional QMD parametric submodels that reduce to P_0 when the parameter $\theta = 0$.

Submodel 1: Consider a submodel with bounded score $g : \mathcal{X} \to \mathbb{R}$ that takes the form

$$p_\theta(x) = [1 + \theta g(x)]p_0(x)$$

Note that our goal is to study the derivative of $\psi(P_\theta) = P_\theta f$ over this new perturbed distribution.

$$\begin{aligned} \psi(P_\theta) &= \int f(x)p_\theta(x)d\mu(x) \\ &= \int f(x)p_0(x)d\mu(x) + \theta \int f(x)g(x)p_0(x)d\mu(x) \\ &= \psi(P_0) + \theta P_0[fg] \end{aligned}$$

So the derivative of $\psi(P_\theta)$ evaluated at $\theta = 0$ over this choice of model h is:

$$\begin{aligned}\frac{\partial}{\partial\theta}\psi(P_\theta)\Big|_{\theta=0} &= P_0[fg] \\ &\equiv P_0\left[(f - P_0(f))\,g\right] \qquad \text{(Bc } P_0(f) \text{ is constant and } g \text{ is } P_0\text{-mean zero)}\end{aligned}$$

Note that $f - P_0(f) = D(P_0)$ is a gradient because it is P_0-mean zero.

Submodel 2: Consider a different submodel also with bounded score $g : \mathcal{X} \to \mathbb{R}$ that takes the form

$$p_\theta(x) = \frac{\exp(\theta g(x))p_0(x)}{c_g(\theta)}$$

where $c_g(\theta)$ is a normalizing constant. Note that $[c_g(0)]^{-1} = 1$ implying that $\frac{d}{d\theta}\log c_g(\theta)\Big|_{\theta=0} = \frac{\theta}{c_g(0)}\Big|_{\theta=0} = 0$. Thus,

$$\frac{d}{d\theta}\log p_\theta(x)\Big|_{\theta=0} = \frac{d}{d\theta}\left(\theta g(x) + \log p_0(x) - \log c_g(\theta)\right)\Big|_{\theta=0} = g(x)$$

Now we inspect the form of the functional

$$\begin{aligned}\frac{\partial}{\partial\theta}\psi(P_\theta)\Big|_{\theta=0} &= \frac{\partial}{\partial\theta}\int f(x)p_\theta(x)d\mu(x)\Big|_{\theta=0} \\ &= \frac{\partial}{\partial\theta}\int f(x)\exp(\log p_\theta(x))d\mu(x)\Big|_{\theta=0} \\ &= \int f(x)\underbrace{\left[\frac{\partial}{\partial\theta}\log p_\theta(x))\Big|_{\theta=0}\right]}_{g(x)}\underbrace{\exp\left[\log p_0(x)\right]}_{p_0(x)}d\mu(x) \qquad \text{(Move deriv inside \& chain rule)} \\ &= P_0[fg] \\ &= P_0[(f - P_0(f))g]\end{aligned}$$

Where $D(P_0) = f - P_0(f)$ is the gradient of ψ at P_0 relative to M.

For a pathwise differentiable parameter, does a gradient always exist? Yes, according to the Riez Representation Theorem.

Theorem 22 (Riez Representation Theorem)**.**
If $\psi : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional, there exists a unique element $h_0 \in \mathcal{H}$ such that

$$\psi(h) = \langle h, h_0\rangle \qquad \forall h \in \mathcal{H}$$

3.2.3 Tangent Sets, Tangent Spaces, Gradients

Next, we turn our focus to under what conditions we have that the score $g = cD(P_0)$ for some $c \in \mathbb{R}$, or under what conditions $D(P_0) \in G(P_0)$, i.e., the tangent set or the collection of scores for all the QMD submodels.

We formalize idea of indexing submodels by their scores at $\theta = 0$ to describe the set of possible "directions" we can perturb P_0 while staying in our model M via the tangent set and tangent space.

Definition 17 (Tangent Set, Tangent Space, Locally nonparametric model)**.**

The **tangent set** of M at P_0, denoted $G(P_0) = \{g_h : h \in \mathcal{H}(P_0)\}$ is the collection of scores of all QMD submodels centered at $\theta = 0$.

The **tangent space**, denoted $T_M(P)$ corresponds to the $L^2(P)$-closure of the linear span of the tangent set.

In most cases, the two are equivalent.

In an unrestricted **(locally) nonparametric model** M, for each $P \in M$, $T_M(P) = L_0^2(P)$ which is infinite dimensional.

A parametric model has a finite dimensional $T_M(P)$ for each $P \in M$. A semiparametric model has infinite dimensional $T_M(P)$ for each $P \in M$ yet does not equal $L_0^2(P)$.

Shown below are a few examples of deriving forms of tangent set. We also assess whether the gradient lies in the tangent set.

Example 7 (Tangent set for nonparametric model, GCRLB form)**.**
We claimed above that $\mathcal{G}(P_0) = L_0^2(P_0) := \{g \in L^2(P_0) : P_0 g = 0\}$. To show this, we pursue mutual containment. Fix a given $g \in L_0^2(P_0)$ and define the path

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

The score is equal to

$$\begin{aligned} \frac{d}{d\theta} \log p_{\theta,g}(x)\Big|_{\theta=0} &= \frac{d}{d\theta} \log[1 + \theta g(x)] + \log p_0(x)\Big|_{\theta=0} \\ &= g(x) \end{aligned}$$

Hence, an element of $g \in L_0^2(P_0)$ is also in the tangent set $\mathcal{G}(P_0)$. Hence $L_0^2(P_0) \subseteq \mathcal{G}(P_0)$.
However, all scores in QMD submodels must be P_0-mean 0 and finite variance. Therefore $\mathcal{G}(P_0) \subseteq L_0^2(P_0)$.
Therefore, $\mathcal{G}(P_0) = L_0^2(P_0)$ in a nonparametric model.

Does the gradient lie in the tangent set, $D(P_0) \in \mathcal{G}(P_0)$. If M is nonparametric and ψ is pathwise differentiable, the gradient $D(P_0)$ is a P_0-mean-zero and squared-integrable function. Then by definition, $D(P_0) \in L_0^2(P_0) = \mathcal{G}(P_0)$.

Thus, in nonparametric models M, for pathwise differentiable parameters, the GCRLB is

$$v_0^*(M) \geq P_0(D(P_0)^2)$$

Example 8 (Derive form of tangent set for parametric model)**.**
Suppose a parametric model indexed by a finite-dimensional parameter: $M = \{P_\beta : \beta \in \mathbb{R}^q\}$. Let $P_0 = P_{\beta_0}$ for some β_0.

We will show that the tangent set $\mathcal{G}(P_0)$ is equal to

$$\tilde{\mathcal{G}}(P_0) := \{x \to u^T s_0(x) : u \in \mathbb{R}^q\}$$

and s_0 is the score for β at β_0. Note this is a linear span in $\mathbb{R}^q$, implying the tangent set is finite dimensional. We will show the result via mutual containment.

($\tilde{\mathcal{G}}(P_0) \subseteq \mathcal{G}(P_0)$): Fix a $u \in \mathbb{R}^q$. We will show there exists a submodel $M_u := \{P_{\beta(\theta)} : \theta \in [0, \delta)\} \subset M$ such that $\beta(0) = \beta_0$ and has score $x \to u^T s_0(x)$ for θ at $\theta = 0$.
Let $\beta(\theta) := \beta_0 + \theta u$ denote a submodel. Under regularity conditions (Lemma 7.6 in vdV), we can think about the score as

$$\begin{aligned}\frac{\partial}{\partial\theta} \log p_{\beta(\theta)}(x)\Big|_{\theta=0} &= \left[\frac{\partial}{\partial\theta}\beta(\theta)\big|_{\theta=0}\right] \left[\nabla_\beta \log p_\beta(x)\big|_{\beta=\beta_0}\right] \\ &= u^T s_0(x)\end{aligned}$$

Hence we've exhibited a submodel of M with score $u^T s_0(x)$. Since $u \in \mathbb{R}^q$ was arbitrary, we have that $\tilde{\mathcal{G}}(P_0) \subseteq \mathcal{G}(P_0)$.

($\mathcal{G}(P_0) \subseteq \tilde{\mathcal{G}}(P_0)$): Consider a generic submodel of M, $\{P_{\beta(\theta)} : \theta \in \mathbb{R}^q\}$ with $\beta(0) = \beta_0$ and with score of g at $\theta = 0$.
Under regularity conditions $g(x) = \frac{\partial}{\partial\theta} \log p_{\beta(\theta)}(x)$ and assuming $\beta(\theta)$ is differentiable at $\theta = 0$, by chain rule we have

$$g(x) = \left(\frac{\partial}{\partial\theta}\beta(\theta)\big|_{\theta=0}\right) \left(\nabla_\beta \log p_{\beta(\theta)}(x)\big|_{\beta=\beta_0}\right)$$

Define the first term as u^T. The second term is $s_0(x)$ byt definition. Hence, for any submodel of M, we showed that if $g \in \mathcal{G}(P_0)$, then $g \in \tilde{\mathcal{G}}(P_0)$.
By mutual containment, $\mathcal{G}(P_0) = \tilde{\mathcal{G}}(P_0)$.

Does the gradient lie in the tangent set, $D(P_0) \in \mathcal{G}(P_0)$. Not in this case!
Consider estimating the generalized moment $\psi(P) = Pf$ and suppose $M = \{P_\theta = N(\theta, 1) : \theta \in \mathbb{R}\}$. The score for θ at θ_0 is $s_0(x) = x - \theta_0$.
By the above, the tangent set for the parametric model is

$$\mathcal{G}(P_0) = \{x \to c(x - \theta_0) : c \in \mathbb{R}\}$$

We previously showed that the gradient $D(P_0) : x \to f - P_0 f$.
Can we find a $c \in \mathbb{R}$ s.t. $cD(P_0) \in \mathcal{G}(P_0)$?

1. In general we cannot. Consider $f(x) = x^2$. We can never scale a quadratic (form of $D(P_0)$) to look like a linear function (the tangent set $\mathcal{G}(P_0)$).
2. If $f(x) = x$ is the identity, then yes! Then $D(P_0) = x - \theta_0$. This is because when $f = x$, the functional Pf corresponds to θ in the model.

This raises a challenge. To obtain a closed form for the GCRLB, we require that $cD(P_0) \in \mathcal{G}(P_0)$. This is always true for nonparametric models, but is not always true for semiparametric and parametric models. How can we move forward when we can't access the variance lower bound? As explained in the next subsection, we replace the gradient $D(P_0)$ by $D*(P_0)$, its projection onto the linear span of the tangent set, the tangent space.

3.2.4 Projections onto Hilbert Spaces

We briefly review the concept of a Hilbert Space.

Definition 18 (Hilbert Space, $L^2(P)$, Orthogonality, Projection)**.**

A real-valued **Hilbert space**, $\mathcal{H}$, is a vector space (closed under addition and scalar multiplication) equipped with an inner product $\langle \cdot, \cdot \rangle$ (satisyfing positive definiteness, symmetry, and linearity) and is complete (every Cauchy seq has limit) relative to the norm $||h|| := \langle h, h \rangle^{1/2}$.

Consider the Hilbert Space $L^2(P)$ is the collection of functions defined on the support of $P \in M$ such that $Pf^2 < \infty$. The inner product in this case is $\langle f_1, f_2 \rangle := P(f_1 f_2)$. The norm is $||f||_{L^2(P)} := \sqrt{Pf^2}$. We'll be interested in two closed subspaces of $L^2(P)$.

1. $L_0^2(P) = \{f \in L^2(P) : Pf = 0\}$: the tangent set in a nonparametric model.
2. $T_M(P)$: the tangent space, or the $L^2(P)$-closure of the linear span of the tangent set.

The **orthogonal complement** $\mathcal{H}_0^\perp$ of a subspace $\mathcal{H}_0 \subset \mathcal{H}$ is defined as:

$$\mathcal{H}^\perp := \{h \in \mathcal{H} : \langle h, h_0 \rangle = 0 \ \forall \ h_0 \in \mathcal{H}_0\}$$

The **projection** of $h_1 \in \mathcal{H}$ onto a closed subspace $\mathcal{H}_0$ is defined as:

$$\Pi_{\mathcal{H}_0}(h_1) := \underset{h \in \mathcal{H}_0}{\operatorname{argmin}} \ ||h - h_1||$$

The projection $\Pi_{\mathcal{H}_0}(h_1)$ is the unique element of $h_0 \in \mathcal{H}_0$ such that $h_0 \in \mathcal{H}_0$ and $h_1 - h_0 \in \mathcal{H}_0^\perp$ (residual in orthogonal complement).

We offer an equivalent (by Reiz Representation Theorem) definition of pathwise differentiability that exploits the fact that $L^2(P)$ is a Hilbert space. It also offers a path forward to examine the lower bound on the variance when the gradient $D(P_0) \notin T_M(P_0)$.

Definition 19 (Equivalent Def of Pathwise differentiability)**.**
A parameter $\psi : M \to \mathbb{R}$ is **pathwise differentiable** at P_0 iff there exists a continuous linear map $\dot{\psi}_{P_0} : L_0^2(P_0) \to \mathbb{R}$ s.t. for all $g \in T_M(P_0)$

$$\frac{\partial}{\partial \theta} \psi(P_\theta)\Big|_{\theta=0} = \dot{\psi}_{P_0}(g)$$

This definition is equivalent to the original definition that requires the existence of a gradient $D(P_0) \in L_0^2(P_0)$ s.t. for all $h \in \mathcal{H}(P_0)$,

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

Definition 20 (GCRLB in Hilbert Space, Canonical Gradient)**.**

Let's reconsider the GCRLB in the Hilbert space.

$$v_0^*(M) \geq \sup_{g \in T_M(P_0)} \frac{[\langle D(P_0), g \rangle]^2}{P_0 g^2}$$

When $D(P_0)$ is not in the tangent space $T_M(P_0)$, we can't obtain an explicit form of the GCRLB by Cauchy-Schwarz. Instead, we consider the projection of $D(P_0)$ onto the tangent space:

$$D^*(P_0) = \Pi_{T_M(P_0)}(D(P_0))$$

Then we can write the GCRLB as

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

Where cancellation of the first term holds because $D(P_0) - D^*(P_0) \perp T_M(P_0)$ by orthogonality.
Some properties of $D^*(P_0)$

1. $D^*(P_0)$ is a gradient: since $\langle D(P_0), g \rangle = \langle D^*(P_0), g \rangle$ for all $g \in T_M(P_0)$
2. $D^*(P_0)$ is the unique gradient that belongs to $T_M(P_0)$

We term $D^*(P_0)$ the **canonical gradient** or **efficient influence function**.

The following strategy gives us guidance on how to calculate a gradient for a particular model.

Strategy 1 (Identifying a gradient)**.**
For a given pathwise differentiable parameter $\psi : M \to \mathbb{R}$, how do we identify a gradient?

1. Take a QMD parametric submodel $\{P_\theta : \theta \in [0, \delta)\} \subseteq M$ with $P_{\theta=0} = P_0$ and score $g \in T_M(P_0)$. Choose nice submodels like the linear submodel

$$p_\theta = [1 + \theta g(x)] p_0$$

2. Compute $\frac{\partial}{\partial \theta} \psi(P_\theta) \Big|_{\theta=0}$ analytically over your chosen submodel.
3. Write $\frac{\partial}{\partial \theta} \psi(P_\theta) \Big|_{\theta=0}$ as $P_0[\tilde{D}(P_0) g]$ for some $\tilde{D}(P_0) \in L^2(P_0)$. Note: $\tilde{D}(P_0)$ can't depend on choice of g.
4. Recenter $\tilde{D}(P_0)$ so it is mean 0, just by subtracting its mean:

$$D(P_0) : x \to \tilde{D}(P_0)(x) - P_0 \tilde{D}(P_0)$$

The following theorem ensures we have a nice submodel to work with in step 1 of the strategy above.

Theorem 23 (Gradients in nested models)**.**
Let $M_1 \subseteq M_2$ be two models. Suppose $P \in M_1$ and $\psi : M_2 \to \mathbb{R}$ is pathwise differentiable at P relative to M_2. Then ψ is also pathwise differentiable at P relative to M_1 and

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

Where $\mathrm{Grad}_M(P) = \{D_0(P) + q(P) : q(P) \in T_M(P)^\perp\}$ is the collection of all gradients of the model T_M. This means it is okay to pick a bigger model, find a gradient in the bigger model, and applying it to the smaller model, because the gradients of the smaller model are larger than the bigger model.

The Theorem above suggests that we perform the 4 step strategy above by **extending to the nonparametric model**. Then start with easy submodel of the nonparametric model, find a gradient, and that gradient is guaranteed to be a gradient in the smaller model.

Example 9 (Gradient of Average Density Parameter)**.**
Consider the average density functional, $\psi : M \to \mathbb{R}$

$$\psi(P) = \int p(u)^2 du$$

Step 1: consider the nonparametric model and define the linear submodel

$$p_\theta(x) = [1 + \theta g(x)]p_0(x)$$

Step 2: analytically compute the derivative over the path along the chosen submodel. Step 3: write in terms of $\tilde{D}(u)$ uncentered.

$$\begin{aligned}\frac{\partial}{\partial\theta}\psi(P_\theta)\Big|_{\theta=0} &= \frac{\partial}{\partial\theta}\int [1 + \theta g(u)]^2 p_0(u)^2 du\Big|_{\theta=0} \\ &= 2\int g(u)p_0(u)^2 du \\ &= \int \underbrace{2p_0(u)}_{\tilde{D}(u)} g(u)dP_0\end{aligned}$$

Step 4: mean-center $\tilde{D}(u)$

$$\begin{aligned}\frac{\partial}{\partial\theta}\psi(P_\theta)\Big|_{\theta=0} &= \int 2p_0(u)g(u)dP_0 \\ &= \int \underbrace{2\left(p_0(u) - \psi(P_0)\right)}_{D(P_0)} g(u)dP_0\end{aligned}$$

Thus, $D(P_0) = 2\left(p_0(u) - \psi(P_0)\right)$ is the gradient.

3.2.5 Relationships between gradients and influence functions

Definition 21 (RAL Estimators, Gradients $\iff$ Influence Functions)**.**
Estimators that are both regular (satisysfing Def 14) and asymptotically linear (satisysfing Def 7) are called

regular asymptotic linear (RAL) estimators.
Key result 1: influence functions are gradients. If ψ_n is a asymptotic linear estimator of $\psi(P_0)$ with influence function ϕ_{P_0}, then TFAE

1. ψ is pathwise differentiable at P_0 with gradient ϕ_{P_0}.
2. ψ_n is regular at P_0

This result implies that RAL estimators exist only for pathwise differentiable parameters. Studying the pathwise derivative for our parameter is also critical.
Key result 2: gradients are influence functions. If $D(P_0)$ is a gradient, under regularity conditions, TFAE

1. An asymptotically linear estimator with influence function $D(P_0)$ exists.
2. It's possible to estimate $\psi(P_0)$ consistently.

This result implies that computing the IF of a known RAL estimator can be a way to find a gradient.
This result also implies if ψ is pathwise differentiable and can be estimated consistently, there exists an ALE meaning we can estimate it root-n consistently. (We explore constructing efficient estimators in next section)
And if we can't estimate a parameter root-n consistently, it's probably not pathwise differentiable.

The link between influence functions and gradients helps establish efficiency bounds in arbitrary models and helps us characterize efficient estimators.

Definition 22 (Efficient Estimator, EIF)**.**
By the key result 2 above, under conditions, there exists an asymptotic linear estimator ψ_n with influence function equal to the canonical gradient, $D^*(P_0)$.
By key result 1, we have that ψ_n is regular. By CLT we have that $\sqrt{n}(\psi_n - \psi_0)$ has asymptotic variance $P_0[D^*(P_0)^2]$.
Recalling $v_0^*(M)$ is the smallest variance of any regular estimator in the model M, and that ψ_n is a specific regular estimator.

$$v_0^*(M) \leq P_0[D^*(P_0)^2]$$

However we also showed in Def 20 that the GCRLB reduces to

$$v_0^*(M) \geq P_0[D^*(P_0)^2]$$

Thus, $v_0^*(M) = P_0(D^*(P_0)^2)$, meaning our estimator ψ_n achieves the lowest asymptotic variance of any regular estimator of ψ_0. We term this estimator, **efficient**.
We term the $D^*(P_0)$ the **efficient influence function**.

Example 10 (Semiparametric Efficiency Examples)**.**
Example 1 (General moment under independence): Suppose $(Y, Z) \sim P_0 \in M$ where M is the collection of all bivariate distributions where Y and Z are independent. Suppose we wish to estimate $\psi_0 := P_0 f$ for fixed bivariate function f.

1. Start with a gradient (influence function) in the nonparametric model: $D(P) := f - P(f)$.
2. Derive the form of the tangent space of the semiparametric model $T_M(P)$. Recall that a model for

$P_{Y,Z} \in M$ can be written as $P_{Y,Z} = P_Y P_Z$ implying

$$M = M_Y \otimes M_Z$$

Because Y and Z are independent of each other. Thus, the tangent space can be written as

$$\begin{aligned} T_M(P) &= T_{M_Y}(P) + T_{M_Z}(P) \\ &= L_0^2(P_Y) + L_0^2(P_Z) \end{aligned}$$

3. Project the gradient on the nonparametric model onto the tangent space of the model with independence condition to get the EIF

$$\begin{aligned} \Pi_{T_M(P)}(D(P)) &= \Pi_{T_{M_Y}(P)}(D(P)) + \Pi_{T_{M_Z}(P)}(D(P)) \\ &= \mathbb{E}_P[f(Y,Z)|Z=z] - \psi(P) + \mathbb{E}_P[f(Y,Z)|Y=y] - \psi(P) \\ &= \mathbb{E}_P[f(Y,z)] + \mathbb{E}_P[f(y,Z)] - 2\psi(P) = D^*(P) \end{aligned}$$

This is the EIF of the ψ relative to M.

Consider estimating the bivariate CDF under joint independence, $F_0(y_0, z_0) = P_0 \mathbb{1}(Y_i \leq y_0, Z_i \leq z_0)$. Let's postulate a simple plug-in estimator under joint independence:

$$\begin{aligned} \psi_n &= P_n[\mathbb{1}(Y \leq y_0, Z \leq z_0)] \\ &= P_n[\mathbb{1}(Y \leq y_0)] P_n[\mathbb{1}(Z \leq z_0)] \\ &= F_n(y_0) F_n(z_0) \end{aligned}$$

Let's consider correcting for possible bias via the one-step estimation method. The EIF for this parameter is

$$D^*(P) = \mathbb{1}(Z_i \leq z_0) F_P(y_0) - \psi(P) + \mathbb{1}(Y_i \leq y_0) F_P(z_0) - \psi(P)$$

Now let's find the efficient one-step estimator

$$\begin{aligned} \psi_{os} &= \psi_n - P_n[D^*(P_n)] \\ &= \psi_n - \underbrace{P_n\left[\mathbb{1}(Z_i \leq z_0) F_P(y_0) - \psi(P_n) + \mathbb{1}(Y_i \leq y_0) F_P(z_0) - \psi(P_n)\right]}_{=0} \\ &= \psi_n \end{aligned}$$

This tells us that the one step and plug-in estimators are equal, therefore the plug-in estimator is efficient. **Example 2 (Variation Independence**: suppose $P = Qg$ with $\psi : M \to \mathbb{R}$ depending on P through Q alone and that $M = M_Q \otimes M_g$, i.e., Q and g are variationally independent.

1. Figure out the form of the tangent space:

$$T_M(P) = T_{M_Q}(P) + T_{M_g}(P)$$

2. Shrinking $T_{M_g}(P)$ by putting more restrictions on g generally enlarges $T_M^\perp(P)$, increasing the number of gradients.

3. However, the EIF is not affected by shrinking $T_{M_g}(P)$ because it lies in $T_{M_Q}(P)$.

Example 3 (Estimating mean of MAR outcome): suppose $X_i = (Y_i, \Delta_i, W_i) \overset{iid}{\sim} P_0$. Consider the estimator:

$$\psi_n := \frac{1}{n} \sum_{k=1}^{n} \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}(\Delta_i = 1, W_i = W_k)}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\Delta_i = 1, W_i = W_k)}$$

which is an estimator of $\psi_0 := \mathbb{E}(\mathbb{E}(Y|\Delta = 1, W))$.
We can write $P_X = P_{Y|\Delta,W} P_{\Delta|W} P_W$. Noting that ψ_0 does not depend on $P_{\Delta|W}$ which is an *orthogonal nuisance parameter*. Thus, restrictions on the model $P_{\Delta|W}$ do not change the EIF.
Consider the nonparametric model M_0 which restricts $P_{\Delta|W} = g_0$ to be known. In this model, use the IPW estimator to estimate $\psi = \mathbb{E}(Y)$:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{Y_i \Delta_i}{g_0(W_i)}$$

Which has influence function

$$D(P)(x) = \frac{y\delta}{g_0(w)} - \psi_0(P)$$

Now we project the gradient from the nonparametric model into the tangent space of the model M

$$T_M(P) = T_{M_{Y|\Delta,W}}(P) + T_{M_W}(P)$$

Which produces

$$\Pi(D(P)|T_M(P))(x) = \frac{\delta}{P(\Delta = 1|W = w)}[y - \mathbb{E}(Y|\Delta = 1, W = w)] + \mathbb{E}(Y|\Delta = 1, W = w) - \psi_0(P)$$

is the EIF.

4 Constructing Efficient Estimators

4.1 Undersmoothing

Consider parameters that rely on "local" information such that plug-in estimators can't be well-defined (e.g., densities, regression functions). For example consider the average density parameter

$$\psi(P) = \int p(x)^2 dx$$

The parameter is pathwise differentiable, so there should exist an ALE. A plug-in estimator of this parameter is undefined because we can't define an empirical density. Thus, we must often use smoothing tools. We could plug-in a different estimator other than the empirical estimator.
Consider a kernel density estimator (KDE) of p:

$$p_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right)$$

Assume that K is a symmetric kernel and p_0 is sufficiently smooth such that the KDE achieves optimal rate in terms of MSE and MISE.
The plug-in estimator based on the KDE is

$$\psi_n = \psi(P_{n,h_n}) = \int (p_{n,h_n})^2 dx$$

Theory tells us that the optimal bandwidth for estimating the density using a KDE is $h = n^{-1/5}$. However, this bandwidth is not optimal for estimation of ψ_0, as we'll see that bias inherited from estimating the intermediate object, the nuisance p_0, is too large to yield an ALE.
To see this, consider the bias of ψ_n

$$\begin{aligned}
\text{bias}(\psi_n) &= \mathbb{E}[\psi_n - \psi_0] \\
&= \mathbb{E}\left[\int (p_{n,h_n})^2 dx - \int (p_0)^2 dx\right] \\
&= \mathbb{E}\left[2\int p_0(x)\left(p_{n,h_n}(x) - p_0(x)\right) dx\right] + \mathbb{E}\left[\int (p_{n,h_n}(x) - p_0)^2 dx\right] \\
&= 2\int p_0(x) \underbrace{\mathbb{E}\left[p_{n,h_n}(x) - p_0(x)\right]}_{\text{Bias } p_{n,h_n}} dx + \int \underbrace{\mathbb{E}\left[(p_{n,h_n}(x) - p_0)^2\right]}_{\text{MSE } p_{n,h_n}} dx
\end{aligned}$$

Recalling that $MSE = \text{bias}^2 + \text{variance}$, and $\text{Bias}(p_{n,h_n}) = \mathcal{O}(h_n^2)$ and $\text{Variance}(p_{n,h_n}) = \mathcal{O}((nh_n)^{-1})$, then

$$\text{bias}(\psi_n) = \mathcal{O}\left(h_n^2 + \frac{1}{nh_n}\right)$$

Where we've omitted the higher order bias^2 term. If we plug in the optimal bandwidth for estimating p_0, $h_n = n^{-1/5}$, we see that

$$\text{bias}(\psi_n) = \mathcal{O}(n^{-2/5})$$

However, an ALE implies

$$n^{1/2}\text{bias}(\psi_n) \overset{n\to\infty}{\longrightarrow} 0$$

However, for our estimator above

$$n^{1/2}\text{bias}(\psi_n) = \mathcal{O}(n^{1/10}) \to \infty$$

We must consider a new bandwidth to achieve asymptotic linearity. Turns out th optimal choice of bandwidth is $h_n = n^{-1/3}$. Choosing the smaller bandwidth means we allow our estimate of the density to be **undersmoothed**, i.e. more wiggly, allowing the bias to converge to 0 quickly enough and achieve an AL estimate of ψ_0.
However, this is not a general framework and there are no sensible guidelines for practically doing undersmoothing.
The next two subsections offer more general approaches.

4.2 Estimating Equations Framework

Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} P_0 \in M$ and we wish to estimate $\psi_0 = \psi(P_0)$.

From homework 2 question 3, if ψ_n is consistent for ψ_0 where ψ_n is a near solution in ψ to

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

Then ψ_n is asymptotically linear with the form

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

Where $a_0^{-1} := \left(\frac{\partial}{\partial\psi} P_0 U(\psi, \eta_n)\Big|_{\psi=\psi_0}\right)^{-1}$ and $b_0 := \frac{\partial}{\partial\eta} P_0 U(\psi_0, \eta)\Big|_{\eta=\eta_0}$.
Therefore, up to a scaling constant and an additive contribution from the nuisance, **the influence function of the estimator is the estimating function itself**, especially when $a_0 = -1$ and $b_0 = 0$. When is this the case?

Theorem 24 (Neyman Orthogonality ($b_0 = 0$))**.** Suppose that for all $P \in M$, we have a gradient D that depends on P through $\psi(P)$ and a nuisance η. Suppose that

1. Variation independence of $\psi(P)$ and η.
2. $L^2(P)$ continuity: $D(P_\theta) \to D(P_0)$ in $L^2(P_0)$ as $\theta \to 0$.
3. $b_0 := \frac{\partial}{\partial\eta} P_0 U(\psi_0, \eta)\Big|_{\eta=\eta_0}$ is well-defined.

Then $b_0 = 0$

Gradients are desirable estimating functions because they are

1. Pre-Neyman-Orthogonalized ($b_0 = 0$): estimating of the nuisance parameter does not impact the first order behavior of the estimator.
2. Pre-normalized ($a_0 = -1$): in that the estimating and influence functions are equal.

The estimating equations framework is easy to describe but can only be used when the EIF is an estimating function for ψ_0. Also requires root finding and can fail to be robust by falling out of the parameter space.

4.3 One step correction

Definition 23 (von-Mises Expansion)**.**
The von-Mises expansion is just a first order expansion of ψ:

$$\psi(P) - \psi(P_0) = -P_0 D(P) + R(P, P_0)$$

where $R(P, P_0) := \psi(P) - \psi(P_0) + P_0 D(P)$. This remainder term is often second order in worked examples.

Example 11 (von-Mises Expansion of average density)**.**
$\psi(P) := \int p^2(x)dx$ has gradient $D(P) = 2(p(x) - \psi(P))$. Therefore

$$\begin{aligned} R(P, P_0) &:= \psi(P) - \psi(P_0) + P_0[2(p(x) - \psi(P))] \\ &= -\int [p(x) - p_0(x)]^2 dx \end{aligned}$$

Which is second order in $p(x) - p_0(x)$.

Definition 24 (One-step estimator)**.**
The von-Mises expansion of ψ about P_n and P_0 is written as

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

Where D is the EIF and

1. Term 1: is a linear term
2. Term 2: is the source of excess bias of $\psi(P_n)$
3. Term 3: is an empirical process term that is negligible under certain conditions.
4. Term 4: is a second order remainder term.

We can move the excess source of bias to the LHS and define the one step estimator:

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

Under the conditions where

1. $R(P_n, P_0) = o_P(n^{-1/2})$: typically a second order term in the nuisance, typically allowing us to flexibly estimate it at $n^{1/4}$ rates. We want P_n to be in our model as well.
2. Continuity condition in $L^2(P)$: $P_0[D(P_n) - D(P_0)]^2 = o_P(1)$
3. There exists a P_0-Donsker class s.t. $D(P_n) \in \mathcal{F}$ w.p. tending to 1. Note: this condition can be removed if we use cross-fitting.

Then Term 3 and Term 4 are both $o_P(n^{-1/2})$ and we obtain

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

Implying that $\psi_{os,n}$ is ALE for ψ_0 with influence function $D(P_0)$. This means that

$$\sqrt{n}(\psi_{os,n} - \psi_0) \rightsquigarrow N(0, P_0[D(P_0)]^2)$$

further implying that the one-step estimator is asymptotically efficient.