Chapter 3
Lower Bounds for Estimation

Chapter 3

Lower Bounds for Estimation

1 Introduction and Examples

One of the goals of statistical theory is to describe how well we can estimate parameters of interest in principle for any given model. Since we cannot estimate parameters perfectly, what is the best we can do?

A *model* $\mathcal{P}$ is simply a collection of probability distributions for the data we observe. Consider a parameter of interest $\nu = \nu(P)$ we want to estimate. Here are some frequent goals or questions:

Question 1. Given a model $\mathcal{P}$ and a parameter of interest ν, how well can we estimate $\nu = \nu(P)$? What is our "gold standard"?
Question 2. Can we compare absolute "in principle" standards for estimation of ν in a model $\mathcal{P}$ with estimation of ν in a submodel $\mathcal{P}_0 \subset \mathcal{P}$? What is the effect of not knowing η on estimation of ν when $\theta = (\nu, \eta)$?
Question 3. For a fixed model $\mathcal{P}$ compare one or more estimators of ν to each other and to the best "in principle" bound.

The bounds we will discuss in this chapter provide some partial answers to these questions.

To indicate the scope of the questions we want to address, we begin with some examples of the models we would like to be able to handle. In all of the following examples we will suppose that we observe $X_1, \ldots, X_n$ i.i.d. as $X \sim P \in \mathcal{P}$ where $\mathcal{P}$ is the given model. Within each example the models increase in complexity: from parametric, to semiparametric, to nonparametric. For further examples see Bickel, Klaassen, Ritov, and Wellner (1993)

Example 1.1 (Survival time). Suppose that X is a non-negative random variable; think of X as a survival time.
Case A. Suppose that $X \sim \text{Exponential}(\theta)$, $\theta > 0$; thus $p_\theta(x) = \theta \exp(-\theta x) 1_{[0,\infty)}(x)$. This is a simple parametric model $\mathcal{P} = \{P_\theta : \ \theta \in \Theta = R^+\}$.
Case B. Suppose that $X \sim \text{Weibull}(\alpha, \beta)$, $\alpha > 0$, $\beta > 0$; thus

$$p_\theta(x) = (\beta/\alpha)(x/\alpha)^{\beta-1} \exp(-(x/\alpha)^\beta) 1_{[0,\infty)}(x)$$

with $\theta = (\alpha, \beta)$. This is also a simple parametric model $\mathcal{P} = \{P_\theta : \ \theta \in \Theta = R^{+2}\}$.
Case C. Suppose that $X \sim P_G$ on R^+ with density $p_G(x) = \int_0^\infty \lambda \exp(-\lambda x) dG(\lambda)$. This can be viewed as a semiparametric model, the family of all scale mixtures of exponential distributions,

$\mathcal{P} = \{P_G : G \in \mathcal{G}\}$ where $\mathcal{G}$ is the collection of all distribution functions on $[0, \infty)$.
Case D. Suppose that $X \sim P$ on R^+ with density function $p = dP/d\lambda$ assumed to be nonincreasing. This model $\mathcal{P}$ is defined only by a shape restriction on the density, and is essentially a nonparametric model.
Case E. Suppose that $X \sim P$ on R^+ with completely arbitrary distribution function F. This is simply the maximal nonparametric model on the space $\mathcal{X} = R^+$: no structure is imposed at all.

Example 1.2 (Measurement model) Suppose that X is a real-valued random variable; think of X as a measurement.
Case A. Suppose that $X \sim N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2) \in \Theta \equiv R \times R^+$. Thus

$$\mathcal{P} = \left\{P_\theta : \ P_\theta \text{ has density } p_\theta = \frac{1}{\sigma}\phi\left(\frac{\cdot - \mu}{\sigma}\right) : \theta = (\mu, \sigma^2) \in \Theta\right\}.$$

This is the most classical parametric model, the normal location - scale model. If we replace the standard normal density by some other density g_0 which is fixed and known (e.g. logistic, or Cauchy, or double exponential, or ...), the resulting model

$$\mathcal{P} = \left\{P_\theta : \ P_\theta \text{ has density } p_\theta = \frac{1}{\sigma}g_0\left(\frac{\cdot - \mu}{\sigma}\right) : \theta = (\mu, \sigma^2) \in \Theta\right\}$$

is the g_0-location - scale family.
Case B. Suppose that $X \sim P_{\theta,G}$ on R with density $p_{\theta,G}(x) = g(x - \theta)$ with G symmetric about 0 and absolutely continuous (with respect to Lebesgue measure) with density g which is itself absolutely continuous with derivative g' satisfying

$$I_g \equiv \int \frac{(g')^2}{g} d\lambda < \infty .$$

Then

$$\mathcal{P} = \{P_{\theta,G} : \theta \in R, \ G \text{ a distribution function with symmetric density } g, \ I_g < \infty\} .$$

This is a semiparametric model, the "one-sample symmetry model".

Example 1.3 (Survival time with covariates). Suppose that $X = (Y, Z)$ is a random vector on $R^+ \times R^d$: think of Y as a survival time and Z as a vector of covariates.
Case A. Suppose that $X = (Y, Z) \sim P_\theta$ with $(Y|Z = z) \sim \text{exponential}(\lambda e^{\theta' z})$; i.e. $\lambda(y|Z = z) = \lambda e^{\theta' z}$ for $y \geq 0$. This is a parametric model with parameter space $\Theta = R^+ \times R^d$.
Case B. Suppose that $X = (Y, Z) \sim P_{\theta,\lambda}$ with $\lambda(y|Z = z) = \lambda(y)e^{\theta' z}$ for $y \geq 0$ where $\theta \in R^d$ and $\lambda = \lambda(y)$ is an arbitrary "baseline" hazard function on R^+. This is a "semiparametric model", the Cox proportional hazards model for survival analysis.
Case C. Suppose that $X = (Y, Z) \sim P_{\theta,\lambda,r}$ with $\lambda(y|Z = z) = \lambda(y)e^{r(\theta' z)}$ for $y \geq 0$ where $\theta \in R^d$, $\lambda = \lambda(y)$ is an arbitrary "baseline" hazard function on R^+, and r is some unknown function from R to R. This is a more complicated variant of the Cox model.
Case D. Suppose that $X = (Y, Z) \sim P$ on $R^+ \times R^d$ where P is completely arbitrary. This is a nonparametric model. How do we define "effects" of the covariates Z on the survival time Y here?

Example 1.4 (Measurement with covariates). Suppose that $X = (Y, Z)$ is a random vector with values in $R \times R^d$; think of Y as a measurement or response and Z as a vector of covariates.
Case A. Suppose that $X = (Y, Z) \sim P_\theta$ with $Y = \theta' Z + \sigma\epsilon$ where $\theta \in R^d$, $\sigma > 0$, and $\epsilon \sim G_0$ with density g_0 is independent of Z. Here g_0 is a known density (such as the standard normal density ϕ), and $Z \sim H$ (supposed known for simplicity). This is a parametric model, the classical linear regression model (with G_0-errors).
Case B. Suppose that $X = (Y, Z) \sim P_{\theta,G}$ with $Y = \theta' Z + \epsilon$ where $\theta \in R^d$ and $\epsilon \sim G$ with density g is independent of Z, but now G (or equivalently g) is an unknown distribution. This is a semiparametric model, the linear regression model with "arbitrary" or "general" error distribution.
Case C. Suppose that $X = (Y, Z) \sim P_{\theta,\sigma,r}$,

$$Y = r(\theta' Z) + \sigma\epsilon$$

where $\theta \in R^d$, $\sigma > 0$, $\epsilon \sim G_0$ with density g_0 is independent of Z, and r is an unknown function from R to R. This is again a semiparametric model, a model for "projection pursuit" regression with G_0-errors; econometricians would call this a "single-index model".
Case D. Suppose that $X = (Y, Z) \sim P_{\sigma,r}$,

$$Y = r_1(Z_1) + \cdots + r_d(Z_d) + \sigma\epsilon$$

where $\sigma > 0$, $\epsilon \sim G_0$ with density g_0 is independent of Z, and $r = (r_1, \ldots, r_d)$ is a vector of unknown functions from R to R. This is again a semiparametric model, a model for "additive" regression with G_0-errors.
Case E. Suppose that $X = (Y, Z) \sim P_{\sigma,r}$,

$$Y = r(Z) + \sigma\epsilon$$

where $\sigma > 0$, $\epsilon \sim G_0$ is independent of Z, and r is an unknown function from R^d to R. This is still a semiparametric model, but estimation becomes increasingly problematic as the dimension d becomes even moderately large: rates of convergence of *any* estimator sequence can be no better than $n^{-p/(2p+d)}$ when r is assumed to belong to a class of functions $\mathcal{R}_p$ with bounded $p-$th order derivatives; see e.g. Stone (1982). This gives $n^{-2/16} = n^{-1/8}$ when $p = 2$ and $d = 12$, and it gives $n^{-1/22}$ when $p = 1$ and $d = 20$.
Case F. Suppose that $X = (Y, Z) \sim P$ where P is an arbitrary probability distribution on $R \times R^d = R^{d+1}$. This is a completely nonparametric version of the model. Here we need to think carefully about how to define the "effects" of the covariates Z on the response variable Y.

Of course not all problems involve independent and identically distributed data (even though it is frequently useful to put them an i.i.d. framework for theoretical analysis if possible). Here is one simple model which involves "pooling information" from three independent samples. There are many other related models, and well as models in which the independence assumption is relaxed.

Example 1.5 (Bivariate three-sample model). Suppose that we observe data as follows:
(i) The first sample is a sample of i.i.d. pairs of size n_1 from a distribution P with cumulative distribution function $H(x, y) = P(X \le x, Y \le y)$ on R^2.
(ii) The second sample of size n_2 is a sample of i.i.d. X's from the marginal distribution P_X of P (with distribution function $F(x) = P(X \le x) = H(x, \infty)$.
(iii) The third sample of size n_3 is a sample of i.i.d. Y's from the marginal distribution P_Y of P with distribution function $G(y) = P(Y \le y) = H(\infty, y)$.

How well can we estimate P (e.g. $\nu(P) = P(X \le x_0, Y \le y_0) = H(x_0, y_0)$ for a fixed point $(x_0, y_0) \in R^2$) based on all the available data?
Case A. Suppose that P is bivariate normal with mean vector μ and covariance matrix Σ. This is a parametric version of the model.
Case B. Suppose that $P_{\theta,F,G}$ where $P_{\theta,F,G}$ has distribution function given by $F_{\theta,F,G}(x, y) = C_\theta(F(x), G(y))$ for some parametric family of distribution functions C_θ on the unit square $[0, 1]^2$ with uniform marginals (such as the Morgenstern family $C_\theta(u, v) = uv(1 + \theta(1 - u)(1 - v))$. This is a semiparametric model.
Case C. Suppose that $P \in \mathcal{M}$, the collection of all distributions on R^2; this is the nonparametric version of the problem.

2 Cramér-Rao bounds for parametric models

We first discuss the elementary Cramér - Rao bound in the case of a one - dimensional parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset R$; the reader may also wish to consult Lehmann, TPE, page 115.

Here are the assumptions we will need:

Assumptions:

A. $X \sim P_\theta$ on $(\mathbb{X}, \mathcal{A})$ with $\theta \in \Theta \subset R$.

B. $p_\theta \equiv dP_\theta/d\mu$ exists where μ is $\sigma-$finite.

C. $T(X) \equiv T$ estimates $q(\theta)$ has $E_\theta|T(X)| < \infty$; set $b(\theta) \equiv E_\theta T - q(\theta) \equiv$ bias of T.

D. $q'(\theta) \equiv \dot{q}(\theta)$ exists.

Theorem 2.1 (Information bound or Cramér - Rao inequality, dimension one). Suppose that:
(C1) Θ is an open subset of the real line.
(C2) A. There exists a set B with $\mu(B) = 0$ such that: for $x \in B^c$

$$\frac{\partial}{\partial\theta} p_\theta(x) \qquad \text{exists for all} \quad \theta\,.$$

B. $A \equiv \{x : \ p_\theta(x) = 0\}$ does not depend on θ.
(C3) $I(\theta) \equiv E_\theta(\dot{\mathbf{l}}_\theta(X)^2) > 0$ where

$$\dot{\mathbf{l}}_\theta(x) \equiv \frac{\partial}{\partial\theta} \log p_\theta(x)\,;$$

here $I(\theta)$ is called the *Fisher information* for θ and $\dot{\mathbf{l}}_\theta$ is called the *score function* for θ.
(C4) $\int p_\theta(x) d\mu(x)$ and $\int T(x) p_\theta(x) d\mu(x)$ can both be differentiated with respect to θ under the integral sign.
(C5) $\int p_\theta(x) d\mu(x)$ can be differentiated twice under the integral sign.
If (C1)-(C4) hold, then

$$\begin{aligned} Var_\theta[T(X)] &\geq \frac{[\dot{q}(\theta) + \dot{b}(\theta)]^2}{I(\theta)} \qquad \text{for all} \quad \theta \in \Theta \\ &= \frac{[\dot{q}(\theta)]^2}{I(\theta)} \qquad \text{if} \quad T \quad \text{is unbiased}\,. \end{aligned}$$

Equality holds for all θ if and only if for some function $A(\theta)$ we have

$$\dot{\mathbf{l}}_\theta(x) = A(\theta)\{T(x) - E_\theta(T(X))\} \qquad \text{a.e.} \quad \mu\,.$$

If, in addition, (C5) holds, then

$$I(\theta) = -E_\theta\left\{\frac{\partial^2}{\partial\theta^2} \log p_\theta(X)\right\} = -E_\theta \ddot{\mathbf{l}}_\theta(X)\,.$$

Proof. Now

$$q(\theta) + b(\theta) = \int_{\mathbb{X}} T(x) p_\theta(x) d\mu(x) = \int_{\mathbb{X}\cap A^c \cap B^c} T(x) p_\theta(x) d\mu(x) \, ;$$

hence it follows from (C2) and (C4) that

$$\begin{aligned}
\dot{q}(\theta) + \dot{b}(\theta) &= \int_{\mathbb{X}\cap A^c \cap B^c} T(x) \frac{\partial}{\partial\theta} p_\theta(x) d\mu(x) = \int_{\mathbb{X}\cap A^c \cap B^c} T(x) \dot{\mathbf{l}}_\theta(x) p_\theta(x) d\mu(x) \\
&= E_\theta\{T(X)\dot{\mathbf{l}}_\theta(X)\} = Cov_\theta[T(X), \dot{\mathbf{l}}_\theta(X)]
\end{aligned}$$

since $\int p_\theta(x) d\mu(x) = 1$ implies, by arguing as above, that

$$0 = \int \frac{\partial}{\partial\theta} p_\theta d\mu = E_\theta[\dot{\mathbf{l}}_\theta] \, .$$

Thus, by the Cauchy -Schwarz inequality

$$[\dot{q}(\theta) + \dot{b}(\theta)]^2 = |Cov_\theta[T(X), \dot{\mathbf{l}}_\theta(X)]|^2 \le Var_\theta[T(X)] I(\theta) \, .$$

The inequality holds with equality for a fixed θ if and only if

$$\dot{\mathbf{l}}_\theta(x) = A(\theta)\{T(x) - E_\theta T(X)\} \qquad \text{a.s.} \quad P_\theta$$

for some constant $A(\theta)$. By (2.B) this implies that this holds a.e. μ. Under further regularity conditions this holds if and only if P_θ is an exponential family; see e.g. Lehmann and Casella page 121.

Finally, if (C5) holds, since

$$0 = \int \dot{\mathbf{l}}_\theta(x) p_\theta(x) d\mu(x) \, ,$$

differentiation once more (which is possible by (C5)) yields

$$\begin{aligned}
0 &= \int \ddot{\mathbf{l}}_\theta(x) p_\theta(x) d\mu(x) + \int \dot{\mathbf{l}}_\theta^2(x) p_\theta(x) d\mu(x) \\
&= \int \ddot{\mathbf{l}}_\theta p_\theta d\mu + I(\theta) \, .
\end{aligned}$$

□

Example 2.1 (Poisson(θ); an exponential family). Suppose that $X_1, \ldots, X_n$ are i.i.d. Poisson(θ) with $\Theta = (0, \infty)$; i.e. $p_\theta(x) = e^{-n\theta}\theta^{\sum_1^n x_i} / \prod_{i=1}^n x_i!$ with respect to counting measure μ^n on Z^{+n}. Then

$$\log p_\theta(X) = -n\theta + \left(\sum_{i=1}^n X_i\right) \log\theta - \sum_{i=1}^n \log(X_i!)$$

and

$$\dot{\mathbf{l}}_\theta(X) = -n + \left(\sum_{i=1}^n X_i\right)\frac{1}{\theta} = \frac{n}{\theta}(\overline{X}_n - \theta) \, .$$

Note that (1), (2), and (3) are trivial; and (4), (4') hold if $E_\theta|T(X)| < \infty$ for all θ since an absolutely convergent powerseries can be differentiated term by term. Thus the Cramér-Rao inequality hold for all T having $E_\theta|T(X)| < \infty$ for all θ. However only $q(\theta) = \theta$ (or a linear functoin of this) has a Minimum Variance Bound Unbiased (MVBU) estimator, and the MVBU estimator of θ is $\overline{X}_n$ which has variance $|\dot{q}(\theta)/A(\theta)| = \theta/n$. Thus $I_n(\theta) = n/\theta$. The bound for estimating $q(\theta) = \theta^2$ is $\dot{q}(\theta)^2/I(\theta) = (2\theta)^2/(n/\theta) = 4\theta^3/n$; but this bound cannot be achieved for $n < \infty$. In fact we know that $\sum_1^n X_i$ is a complete sufficient statistic. It is easy to check that $T^* = \overline{X}_n^2 - n^{-1}\overline{X}_n$ is unbiased; hence it is a UMVUE of θ^2. Also, its variance is $(4\theta^3)/n + (2\theta^2)/n^2 >$ the Cramér - Rao bound.

Example 2.2 (Location with known "shape" g). Suppose that $X_1, \ldots, X_n$ are i.i.d. with density $p_\theta(x) = g(x - \theta)$ where g is a known density (such as $N(0,1)$ or Cauchy or logistic or double exponential or extreme value). Then, assuming that g' exists a.e. (Lebesgue) and the other regularity conditions hold,

$$\dot{\mathbf{l}}_\theta(x) = \frac{\partial}{\partial\theta} \log g(x - \theta) = -\frac{g'(x-\theta)}{g(x-\theta)} \equiv -\frac{g'}{g}(x - \theta)$$

so that

$$\begin{aligned} I(\theta) &= E_\theta\{\dot{\mathbf{l}}_\theta^2(X)\} = \int \left\{\frac{g'}{g}(x-\theta)\right\}^2 g(x-\theta)dx \\ &= \int \left\{\frac{g'}{g}(y)\right\}^2 g(y)dy = \int \frac{[g'(y)]^2}{g(y)} dy \equiv I_g \end{aligned}$$

and $I_n(\theta) = nI(\theta) = nI_g$. Thus for any unbiased estimator $\hat{\theta}_n$ of θ we have

$$Var_\theta(\hat{\theta}_n) \ge \frac{1}{nI_g}, \qquad \text{or} \qquad Var_\theta(\sqrt{n}(\hat{\theta}_n - \theta)) \ge \frac{1}{I_g}.$$

Example 2.3 (Scale with known shape g). Suppose that $X_1, \ldots, X_n$ are i.i.d. with density

$$p_\theta(x) = \frac{1}{\theta} g\left(\frac{x}{\theta}\right)$$

where g is a known density (such as Exponential(1) or Cauchy, or logistic, or Gamma$(5,1)$). Then, assuming that g' exists a.e. (Lebesgue) and the other regularity conditions hold,

$$\begin{aligned} \dot{\mathbf{l}}_\theta(x) &= \frac{\partial}{\partial\theta} \log\left\{\frac{1}{\theta} g\left(\frac{x}{\theta}\right)\right\} = -\frac{1}{\theta} + \frac{g'(x/\theta)}{g(x/\theta)}\left(\frac{-x}{\theta^2}\right) \\ &= \frac{1}{\theta}\left\{-1 - \frac{x}{\theta}\frac{g'}{g}(x/\theta)\right\} \end{aligned}$$

so that

$$\begin{aligned} I(\theta) &= E_\theta\{\dot{\mathbf{l}}_\theta^2(X)\} = \frac{1}{\theta^2}\int \left(-1 - \frac{x}{\theta}\frac{g'}{g}(x/\theta)\right)^2 \frac{1}{\theta} g(x/\theta)dx \\ &= \frac{1}{\theta^2}\int \left(-1 - y\frac{g'}{g}(y)\right)^2 g(y)dy \equiv \frac{1}{\theta^2} I_g(scale). \end{aligned}$$

and $I_n(\theta) = nI(\theta) = nI_g(scale)$. Thus for any unbiased estimator $\hat{\theta}_n$ of θ we have

$$Var_\theta(\hat{\theta}_n) \ge \frac{\theta^2}{nI_g(scale)}, \qquad \text{or} \qquad Var_\theta(\sqrt{n}(\hat{\theta}_n - \theta)) \ge \frac{\theta^2}{I_g(scale)}.$$

Example 2.4 (Elementrary mixture model). Suppose that f_0 and f_1 are two known $\mu-$ densities, and that

$$p_\theta(x) = \theta f_0(x) + (1-\theta) f_1(x), \qquad \text{for} \;\; \theta \in [0,1] = \Theta . \tag{1}$$

Then $\log p_\theta(x) = \log\{\theta f_0(x) + (1-\theta) f_1(x)\}$ and hence

$$\dot{\mathbf{l}}_\theta(x) = \frac{f_0(x) - f_1(x)}{\theta f_0(x) + (1-\theta) f_1(x)} .$$

Hence we calculate

$$I(\theta) = \int \frac{(f_0(x) - f_1(x))^2}{\theta f_0(x) + (1-\theta) f_1(x)} d\mu(x) .$$

Note that if $f_0 \neq f_1$ on a set of positive $\mu-$measure, then the information $I(\theta)$ is finite and positive for all $\theta \in (0,1)$, while $I(\theta)$ converges to

$$\int \frac{(f_0 - f_1)^2}{f_1} d\mu \qquad \text{as} \;\; \theta \to 0 ;$$

similarly $I(\theta)$ converges to

$$\int \frac{(f_0 - f_1)^2}{f_0} d\mu \qquad \text{as} \;\; \theta \to 1 .$$

These limiting values may be infinite. This can be viewed as an example of missing data: Suppose that the complete data is $Y = (X, \Delta)$ where Δ takes values in $\{0,1\}$, $(X|\Delta) \sim F_0^\Delta F_1^{1-\Delta}$, and $P(\Delta = 1) = \theta = 1 - P(\Delta = 0)$. Then the joint density of $Y = (X, \Delta)$ is given by

$$q_\theta(x, \delta) = f_0(x)^\delta f_1(x)^{1-\delta} \theta^\delta (1-\theta)^{1-\delta}$$

If we just observe $Y_1 = X$, then this has the marginal density given by (1). Note that the score for θ based on observation of Y is

$$\dot{\mathbf{l}}_\theta(x, \delta; \mathcal{Q}) = \frac{\delta}{\theta} - \frac{1-\delta}{1-\theta} = \frac{\delta - \theta}{\theta(1-\theta)} ,$$

so that the information for θ in the complete data is

$$I(\theta, \mathcal{Q}) = \frac{1}{\theta(1-\theta)} .$$

Also note that

$$\dot{\mathbf{l}}_\theta(x, \mathcal{P}) \;\; = \;\; \frac{f_0(x) - f_1(x)}{\theta f_0(x) + (1-\theta) f_1(x)} = E\{\dot{\mathbf{l}}_\theta(X, \Delta; \mathcal{Q}) | X = x\} .$$

It follows by the Cauchy-Schwarz or Jensen inequalities applied conditionally that

$$\begin{aligned}
I(\theta, \mathcal{P}) &= E\{\dot{\mathbf{l}}_\theta^2(X, \mathcal{P})\} \\
&= E\{[E\{\dot{\mathbf{l}}_\theta(X, \Delta; \mathcal{Q}) | X\}]^2\} \\
&\leq E\{E\{\dot{\mathbf{l}}_\theta^2(X, \Delta; \mathcal{Q}) | X\}\} = E\{\dot{\mathbf{l}}_\theta^2(X, \Delta; \mathcal{Q})\} \\
&= I(\theta, \mathcal{Q}) .
\end{aligned}$$

These relations are in fact true in considerable generality for missing data, as we will see later.

The Multiparameter Cramér - Rao inequality

Now we extend theorem 2.1 to the case in which the model is a $k-$dimensional parametric family: $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset R^k$.

Assumptions:

A. $X \sim P_\theta$ on $(\mathbb{X}, \mathcal{A})$ with $\theta \in \Theta \subset R^k$.

B. $p_\theta \equiv dP_\theta/d\mu$ exists where μ is $\sigma-$finite.

C. $T(X) \equiv T$ estimates $q(\theta)$ where $q : \Theta \to R$, and $E_\theta|T(X)| < \infty$; set $b(\theta) \equiv E_\theta T - q(\theta) \equiv$ bias of T.

D. $\dot{q}(\theta) \equiv \nabla q(\theta)$ $(k \times 1)$ exists.

Theorem 2.2 (Information inequality, $\Theta \subset R^k$). Suppose that:
(M1) Θ is an open subset of R^k.
(M2) A. There exists a set B with $\mu(B) = 0$ such that: for $x \in B^c$

$$\frac{\partial}{\partial \theta_i} p_\theta(x) \qquad \text{exists for all} \quad \theta \quad \text{and} \quad i = 1, \ldots, k\,.$$

B. $A \equiv \{x : \ p_\theta(x) = 0\}$ does not depend on θ.
(M3) The $k \times k$ matrix $I(\theta) \equiv (I_{ij}(\theta)) = E_\theta(\dot{\mathbf{l}}_\theta(X)\dot{\mathbf{l}}_\theta^T(X))$ is positive definite where

$$\dot{\mathbf{l}}_{\theta_i}(x) \equiv \frac{\partial}{\partial \theta_i} \log p_\theta(x)\,;$$

here $I(\theta)$ is called the *Fisher information matrix* for θ, $\dot{\mathbf{l}}_{\theta_i}$ is called the *score function* for θ_i, and $\dot{\mathbf{l}}_\theta$ is called the *score* for θ.
(M4) $\int p_\theta(x)d\mu(x)$ and $\int T(x)p_\theta(x)d\mu(x)$ can both be differentiated with respect to θ under the integral sign.
(M5) $\int p_\theta(x)d\mu(x)$ can be differentiated twice under the integral sign.
If (M1)-(M4) hold, then

$$\begin{aligned} Var_\theta[T(X)] &\geq \alpha^T I^{-1}(\theta)\alpha \qquad \text{for all} \quad \theta \in \Theta \\ &= \dot{q}^T I^{-1}(\theta)\dot{q}(\theta) \qquad \text{if} \quad T \quad \text{is unbiased} \end{aligned}$$

where

$$\alpha \equiv (\alpha_1, \ldots, \alpha_k)' \equiv \nabla(q(\theta) + b(\theta)) = \nabla E_\theta(T(X))\,.$$

If, in addition, (M5) holds, then

$$(2) \qquad I(\theta) = -E_\theta \ddot{\mathbf{l}}_{\theta\theta}(X) = -\left(E_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\right\}\right)\,.$$

Proof. The following proof is for the case of an unbiased estimator. Since

$$q(\theta) = E_\theta T(X) = \int T(x) p_\theta(x) d\mu(x)\,,$$

differentiating with respect to each θ_j gives

$$\begin{aligned}
\dot{q}(\theta) &= \int_{\mathbb{X}\cap B^c \cap A^c} T(x) \frac{\nabla p_\theta(x)}{p_\theta(x)} p_\theta(x) d\mu(x) \\
&= E_\theta T(X) \dot{\mathbf{l}}_\theta(X) \\
&= E_\theta\{(T(X) - E_\theta T(X))\dot{\mathbf{l}}_\theta(X)\} \\
&= Cov_\theta(T(X), \dot{\mathbf{l}}_\theta(X))\,.
\end{aligned}$$

Multiplying by $\dot{q}^T(\theta) I^{-1}(\theta)$ we find that

$$\dot{q}^T(\theta) I^{-1}(\theta) \dot{q}(\theta) = Cov_\theta(T(X), \dot{q}^T(\theta) I^{-1}(\theta) \dot{\mathbf{l}}_\theta(X))\,.$$

Hence by the Cauchy-Schwarz inequality

$$\begin{aligned}
|\dot{q}^T(\theta) I^{-1}(\theta) \dot{q}(\theta)| &= |Cov_\theta(T(X), \dot{q}^T(\theta) I^{-1}(\theta) \dot{\mathbf{l}}_\theta(X))| \\
&\leq \{Var_\theta(T(X)) \dot{q}^T(\theta) I^{-1}(\theta) I(\theta) I^{-1}(\theta) \dot{q}(\theta)\}^{1/2} \\
&= \{Var_\theta(T(X)) \dot{q}^T(\theta) I^{-1}(\theta) \dot{q}(\theta)\}^{1/2}
\end{aligned}$$

and it follows that

$$Var_\theta(T(X)) \geq \dot{q}^T(\theta) I^{-1}(\theta) \dot{q}(\theta)$$

with equality if and only if

$$T(X) = \dot{q}^T(\theta) I^{-1}(\theta) \dot{l}_\theta(X).$$ □

Corollary 1 (I.i.d. special case). When $X = (X_1, \ldots, X_n)$ with the X_i's i.i.d. $P_\theta \in \mathcal{P}$ satisfying M1-M4, then

$$\begin{aligned}
& I_n(\theta) = n I_1(\theta) \equiv n I(\theta), \\
& \dot{l}_\theta(X) = \sum_{i=1}^n \dot{l}_\theta(X_i),
\end{aligned}$$

and the conclusion can be written, for an unbiased estimator $T_n \equiv T(X_1, \ldots, X_n)$, as

$$Var_\theta(\sqrt{n}(T_n - q(\theta))) \geq \dot{q}(\theta)^T I^{-1}(\theta) \dot{q}(\theta)$$

Note that the function (for sample size $n = 1$) involved here is

$$\widetilde{l}_\nu(X_1) = \dot{q}^T(\theta) I^{-1}(\theta) \dot{\mathbf{l}}_\theta(X_1)\,;$$

we will call $\widetilde{l}_\nu$ the *efficient influence function* for estimation of $\nu(P_\theta) = q(\theta)$: that is, if T_n is an asyptotically efficient estimator of $\nu(P_\theta) = q(\theta)$, then T_n is *asymptotically linear* with influence function exactly $\widetilde{l}_\nu$:

$$\sqrt{n}(T_n - q(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{l}_\nu(X_i) + o_p(1) \to_d N(0, \dot{q}^T(\theta) I^{-1}(\theta) \dot{q}(\theta))\,.$$

This proof should be compared with the proof given of Theorem 6.6, Lehmann and Casella (1998), pages 127 - 128.

Our goal will be to interpret theorem 2.2 geometrically. But first, here is an easy example.

Example 2.5 (Weibull). If $(\mathbb{X}, \mathcal{A}) = (R^+, \mathcal{B}^+)$, the non-negative real numbers with its usual Borel $\sigma-$field, then the *Weibull* family $\mathcal{P}$ is the parametric model with densities

$$p_\theta(x) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right) 1_{[0,\infty)}(x)$$

with respect to Lebesgue measure where $\theta = (\alpha, \beta) \in (0,\infty) \times (0,\infty) \subset R^2$. For the Weibull family $\mathcal{P}$, $\log p_\theta(x)$ is differentiable at every $\theta \in \Theta$ and the scores are:

$$\begin{aligned} \dot{\mathbf{l}}_\alpha(x) &= \frac{\beta}{\alpha}\left\{\left(\frac{x}{\alpha}\right)^{\beta} - 1\right\} \\ \dot{\mathbf{l}}_\beta(x) &= \frac{1}{\beta} - \frac{1}{\beta}\log\left\{\left(\frac{x}{\alpha}\right)^{\beta}\right\}\left\{\left(\frac{x}{\alpha}\right)^{\beta} - 1\right\}. \end{aligned}$$

Thus $\dot{\mathcal{P}} \equiv [\dot{\mathbf{l}}_\theta]$ is the two-dimensional subspace of $L_2(P_\theta)$ spanned by $\dot{\mathbf{l}}_\alpha$ and $\dot{\mathbf{l}}_\beta$, and the Fisher information matrix is

$$I(\theta) = E\{\dot{\mathbf{l}}_\theta(X)\dot{\mathbf{l}}_\theta^T(X)\} = \begin{pmatrix} \beta^2/\alpha^2 & a/\alpha \\ a/\alpha & b^2/\beta^2 \end{pmatrix}$$

where, with $Y \sim$ Exponential(1), and $\gamma \equiv .577216\ldots =$ Euler's constant,

$$\begin{aligned} a &= -E\{(Y-1)^2\log(Y)\} = -(1-\gamma) \\ b^2 &= E\{[(Y-1)\log(Y) - 1]^2\} = \frac{\pi^2}{6} + (1-\gamma)^2. \end{aligned}$$

The computation of $I(\theta)$ is simplified by noting that $Y \equiv (X/\alpha)^\beta \sim$ Exponential(1). Now $det(I(\theta)) = (b^2 - a^2)/\alpha^2 = (\pi^2/6)\alpha^{-2} > 0$, so $I(\theta)$ is nonsingular. Thus by theorem 2.2, if $q(\theta) = \nu(P_\theta)$ is a real-valued function of θ (e.g. $q(\theta) = \nu(P_\theta) = \int_0^\infty x dP_\theta(x) = \alpha\Gamma(1 + 1/\beta)$, and $T = T_n$ is any estimator of $q(\theta)$ based on $X = (X_1, \ldots, X_n)$ (with X_i's i.i.d. P_θ) satisfying Assumption C and hypothesis (4), then

$$\begin{aligned} Var_\theta[T(X)] &\geq \frac{\alpha^T I^{-1}(\theta)\alpha}{n} \\ &= \frac{\dot{q}(\theta)^T I^{-1}(\theta)\dot{q}(\theta)}{n} \qquad \text{if } T \text{ is unbiased}. \end{aligned}$$

Equivalently,

$$Var_\theta[\sqrt{n}(T_n - q(\theta))] \geq \dot{q}(\theta)^T I^{-1}(\theta)\dot{q}(\theta)$$

if T is unbiased. Our goal will be to compare the information bounds for several functions $q(\theta) = \nu(P_\theta)$ when the model is $\mathcal{P}$, or $\mathcal{P}_0 = \{P_{(\alpha,\beta_0)} : \alpha > 0\}$ with β_0 fixed (and known), or $\mathcal{M}_2 = \{P \text{ on } R^+ : E_P X^2 < \infty\}$.

For the function $q(\theta) = E_\theta(X) = \alpha\Gamma(1 + 1/\beta)$,

$$\dot{q}(\theta) = (\Gamma(1+1/\beta), -\alpha\Gamma'(1+1/\beta)/\beta^2)^T = \Gamma(1+1/\beta)(1, -\alpha\psi(1+1/\beta)/\beta^2)^T,$$

where $\psi \equiv \Gamma'/\Gamma$ is the digamma function, and hence the information inequality yields, for any unbiased estimator $T = T_n$ of $q(\theta) = E_\theta(X)$,

$$\begin{aligned} Var_\theta[\sqrt{n}(T_n - q(\theta))] &\geq \dot{q}(\theta)^T I^{-1}(\theta)\dot{q}(\theta) \\ &= \frac{6\Gamma^2(1+1/\beta)}{\pi^2}\frac{\alpha^2}{\beta^2}\left\{b^2 + 2a\psi(1+1/\beta) + \psi^2(1+1/\beta)\right\} . \end{aligned}$$

Note that when $\beta = \beta_0$ is known, then $I_{11} = \beta_0^2/\alpha^2$, $I_{11}^{-1} = \alpha^2/\beta_0^2$, and the information for (unbiased) estimates of $q(\theta) = E_{\theta_0}X$ is given by $(\alpha^2/\beta_0^2)\Gamma(1+1/\beta_0)^2$; this is always less than or equal to the bound obtained in the last display when β is unknown, with equality when $\beta_0 = 1$. In fact there is very little difference between the information bounds $I^{-1}(P_\theta|\nu, \mathcal{P}_0)$, $I^{-1}(P_\theta|\nu, \mathcal{P})$, and $I^{-1}(P_\theta|\nu, \mathcal{M}_2)$, for this particular parameter $\nu(P_\theta)$. See Figures 3.1 and 3.2 for some comparisons.

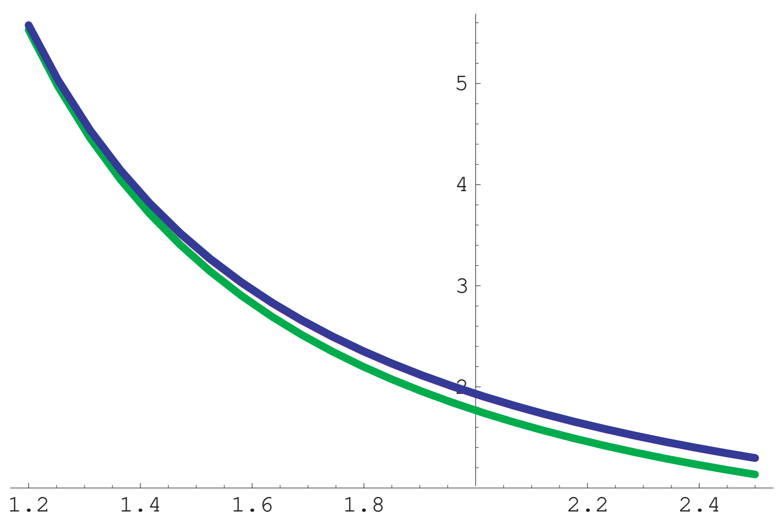

Figure 3.1: Information bounds, $\alpha = 3$, $1.2 \leq \beta \leq 2.5$; green $= I^{-1}(P|\nu, \mathcal{P}_0)$, purple $= I^{-1}(P|\nu, \mathcal{P})$, blue $= Var_\theta(X)$; (purple coincides with blue so not visible!)

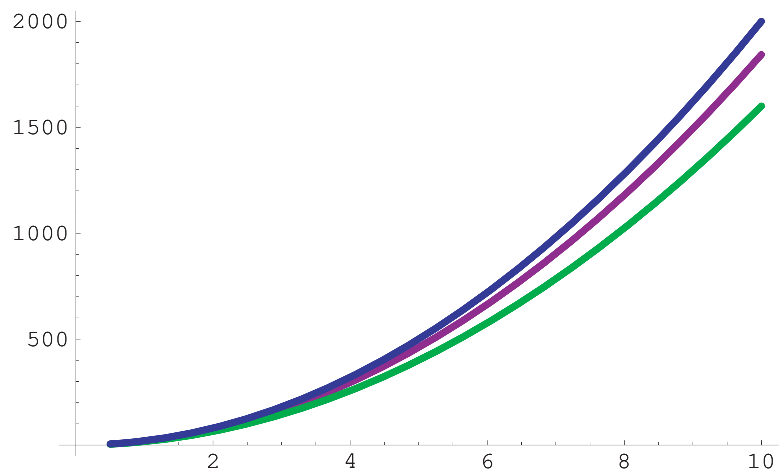

Figure 3.2: Information bounds, $.5 \leq \alpha \leq 10$, $\beta = .5$; green $= I^{-1}(P|\nu, \mathcal{P}_0)$, purple $= I^{-1}(P|\nu, \mathcal{P})$, blue $= Var_\theta(X)$

For the function

$$\nu(P_\theta) = q(\theta) = P_\theta(X \geq x_0) = \exp(-(x_0/\alpha)^\beta)$$

where $x_0 \in (0, \infty)$ is fixed, we have

$$\dot{q}(\theta) = (x_0/\alpha)^\beta \exp(-(x_0/\alpha)^\beta)(\beta/\alpha, -\log(x_0/\alpha))',$$

and hence the information bound for estimation of $q(\theta)$ is given by

$$\dot{q}(\theta)^T I^{-1}(\theta)\dot{q}(\theta) = \frac{6}{\pi^2}\left(\frac{x_0}{\alpha}\right)^{2\beta} \exp(-2(x_0/\alpha)^\beta)\left\{b^2 + 2a\beta\log(x_0/\alpha) + \frac{\beta^2}{\alpha^2}(\log(x_0/\alpha))^2\right\}.$$

When $\beta = \beta_0$ is known, then the information bound for estimation of $q(\theta) = q(\alpha, \beta_0) = \exp(-(x_0/\alpha)^{\beta_0})$ is given by

$$\left\{\frac{\beta}{\alpha}\left(\frac{x_0}{\alpha}\right)^\beta \exp(-(x_0/\alpha)^\beta)\right\}^2 \frac{\alpha^2}{\beta^2} = \left(\frac{x_0}{\alpha}\right)^{2\beta}\exp(-2(x_0/\alpha)^\beta)$$

In this case there is quite a considerable difference between the information bounds $I^{-1}(P_\theta|\nu, \mathcal{P}_0)$, $I^{-1}(P_\theta|\nu, \mathcal{P})$, and $I^{-1}(P_\theta|\nu, \mathcal{M}_2)$, for the parameter $\nu(P_\theta)$; see Figures 3.3 - 3.6 for some comparisons.

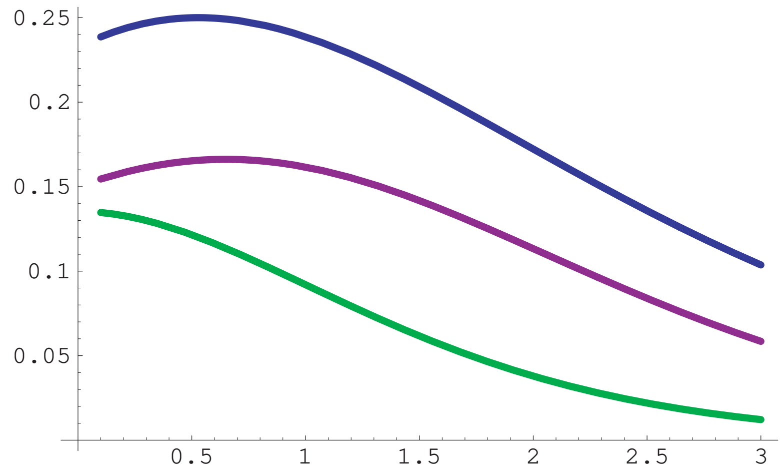

Figure 3.3: Information bounds for $\nu(P_\theta) = P_\theta(X \geq x_0)$, $\alpha = 1$, $x_0 = .5$, $.1 \leq \beta \leq 3$: green $= I^{-1}(P|\nu, \mathcal{P}_0)$, purple $= I^{-1}(P|\nu, \mathcal{P})$, blue $= nVar(\mathbb{F}_n(x_0) = P_\theta(X \geq x_0)(1 - P_\theta(X \geq x_0))$

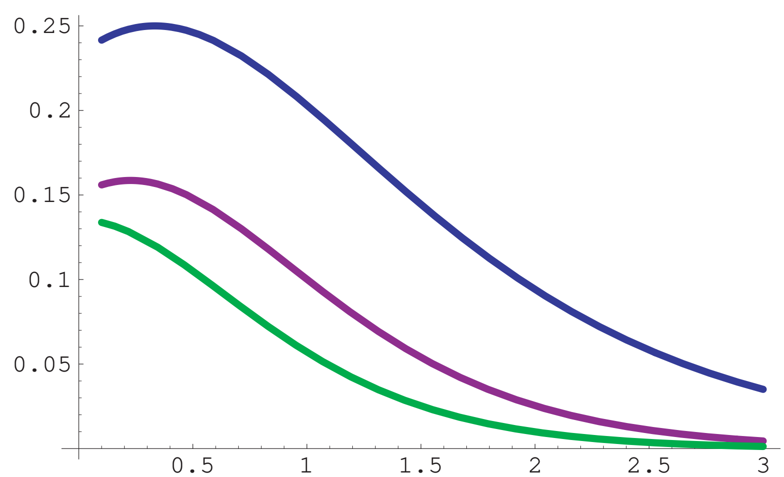

Figure 3.4: Information bounds for $\nu(P_\theta) = P_\theta(X \geq x_0)$, $\alpha = 3$, $x_0 = 1$, $.1 \leq \beta \leq 3.0$; green $= I^{-1}(P|\nu, \mathcal{P}_0)$, purple $= I^{-1}(P|\nu, \mathcal{P})$, blue $= nVar(\mathbb{F}_n(x_0)) = P_\theta(X \geq x_0)(1 - P_\theta(X \geq x_0))$

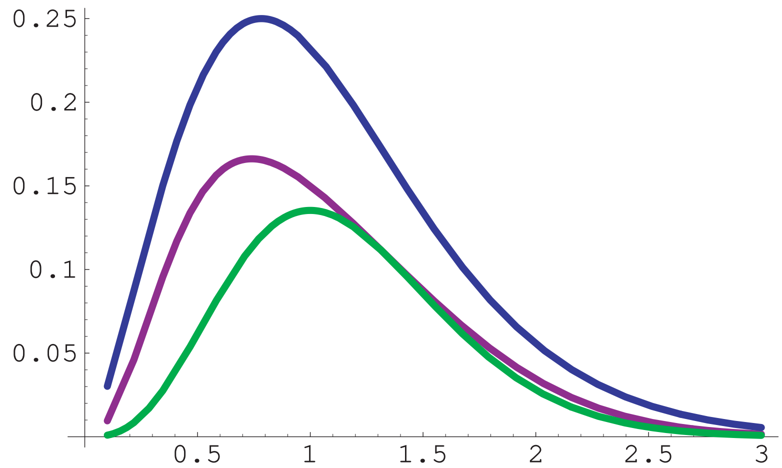

Figure 3.5: Information bounds for $\nu(P_\theta) = P_\theta(X \geq x_0)$, $\alpha = 1$, $\beta = 1.5$, $.1 \leq x_0 \leq 3.0$; green $= I^{-1}(P|\nu, \mathcal{P}_0)$, purple $= I^{-1}(P|\nu, \mathcal{P})$, blue $= nVar(\mathbb{F}_n(x_0)) = P_\theta(X \geq x_0)(1 - P_\theta(X \geq x_0))$

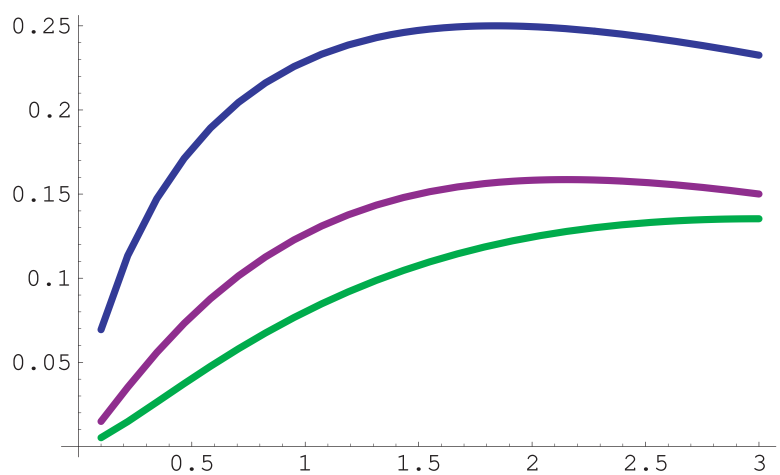

Figure 3.6: Information bounds for $\nu(P_\theta) = P_\theta(X \geq x_0)$, $\alpha = 3$, $\beta = .75$, $.1 \leq x_0 \leq 3.0$; green $= I^{-1}(P|\nu, \mathcal{P}_0)$, purple $= I^{-1}(P|\nu, \mathcal{P})$, blue $= nVar(\mathbb{F}_n(x_0)) = P_\theta(X \geq x_0)(1 - P_\theta(X \geq x_0))$

Some Geometry

The bounds given in theorems 2.1 and 2.2 lead us to the following definitions. Suppose that ν is a Euclidean parameter defined on a regular parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. We can identify

ν with the parametric function $q : \Theta \to R^m$ defined by

$$q(\theta) = \nu(P_\theta) \qquad \text{for} \quad P_\theta \in \mathcal{P}\,.$$

Fix $P = P_\theta$ and suppose that q has a total differential matrix $\dot{q}_{k\times m}$ at θ. Define

(b) $\quad I^{-1}(P|\nu,\mathcal{P}) = \dot{q}^T(\theta) I^{-1}(\theta)\dot{q}(\theta), \qquad$ the *information bound* for ν

and

(c) $\quad \widetilde{\mathbf{l}}(\cdot, P|\nu,\mathcal{P}) = \dot{q}^T(\theta) I^{-1}(\theta)\dot{\mathbf{l}}_\theta, \qquad$ the *efficient influence function* for ν.

As defined in (b) and (c), the information bound and influence function appear to depend on the parametrization $\theta \mapsto P_\theta$ of $\mathcal{P}$. However, as our notation indicates, they actually depend only on ν and $\mathcal{P}$. This is proved in the following proposition.

Proposition 2.1 The information bound $I^{-1}(P|\nu,\mathcal{P})$ and the efficient influence function $\widetilde{l}(\cdot, P|\nu,\mathcal{P})$ are invariant under smooth changes of parametrization.

Proof. We do this by formal calculation. Suppose that $\gamma \mapsto \theta(\gamma)$ is a one-to-one continuously differentiable mapping of an open subset Γ of R^k onto Θ with nonsingular differential $\dot{\theta}$. We represent $\mathcal{P} = \{Q_\gamma : \ \gamma \in \Gamma\}$ where $Q_\gamma \equiv P_{\theta(\gamma)}$. Identify ν by

$$\nu(\gamma) \equiv \nu(Q_\gamma) \equiv q(\theta(\gamma))\,.$$

Then, by the chain rule, the Fisher information matrix for γ is

$$\dot{\theta}^T(\gamma) I(\theta(\gamma))\dot{\theta}(\gamma)$$

while

$$\dot{\nu}(\gamma) = \dot{q}^T(\theta(\gamma))\dot{\theta}(\gamma)\,.$$

Substituting back into (b) gives the same answer for $\gamma \mapsto Q_\gamma$ as for $\theta \mapsto P_\theta$. A similar calculation works for $\widetilde{\mathbf{l}}$. $\square$

Now we specialize slightly: suppose that $\theta' = (\nu', \eta')$ where $\nu \in \mathcal{N} \subset R^m$, $\eta \in \mathcal{H} \subset R^{k-m}$; here ν is the *parameter of interest* and η is a *nuisance parameter*. We can think of this as $q(\theta) = q(\nu,\eta) = \nu$ so that $\dot{q}(\theta) = (I, 0)'$ is a $k \times m$ matrix; here I is the $k \times k$ identity matrix.

If $\theta_0 = (\nu_0, \eta_0) \in \Theta$, let $\mathcal{P}_1(\eta_0) \equiv \{P_\theta : \ \eta = \eta_0, \ \nu \in \mathcal{N}\}$. This is the model when $\eta = \eta_0$ is known. We want to assess the cost of not knowing η by comparing the information bounds and efficient influence functions for ν at P_{θ_0} in $\mathcal{P}_1(\eta_0)$ and $\mathcal{P}$.

We let $\langle\cdot,\cdot\rangle_0$ be the inner product in $L_2(P_{\theta_0})$, $\|\cdot\|_0$ the norm, and write E_0 for expectation under P_{θ_0}.

Suppose the model is regular and write $\dot{\mathbf{l}}$ for the score function at θ_0 and $\widetilde{\mathbf{l}} = I^{-1}(\theta_0)\dot{\mathbf{l}}$ for the efficient influence function of the parameter θ at P_{θ_0} in $\mathcal{P}$. Decompose

$$\dot{\mathbf{l}} = \begin{pmatrix} \dot{\mathbf{l}}_1 \\ \dot{\mathbf{l}}_2 \end{pmatrix}, \qquad \widetilde{\mathbf{l}} = \begin{pmatrix} \widetilde{\mathbf{l}}_1 \\ \widetilde{\mathbf{l}}_2 \end{pmatrix},$$

with $\tilde{\mathbf{l}}_1$ and $\dot{\mathbf{l}}_1$ $m-$vectors, $\tilde{\mathbf{l}}_2$ and $\dot{\mathbf{l}}_2$ $(k-m)-$vectors. Write $I(\theta_0)$ in block matrix form, suppressing dependence on θ_0, as

$$I = [I_{ij}]_{i,j=1,2} = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

with I_{11} $m \times m$, I_{12} $m \times (k-m)$, I_{21} $(k-m) \times m$, I_{22} $(k-m) \times (k-m)$, and similarly decompose $I^{-1}(\theta_0)$ into I^{ij}, $i,j = 1,2$. By well-known block matrix forms of matrix inverses we have

$$(4) \qquad I^{-1}(\theta) = [I^{ij}]_{i,j=1,2} = \begin{pmatrix} I_{11\cdot 2}^{-1} & -I_{11\cdot 2}^{-1} I_{12} I_{22}^{-1} \\ -I_{22\cdot 1}^{-1} I_{21} I_{11}^{-1} & I_{22\cdot 1}^{-1} \end{pmatrix}$$

where

$$(5) \qquad I_{11\cdot 2} \equiv I_{11} - I_{12} I_{22}^{-1} I_{21}\,, \qquad\qquad I_{22\cdot 1} \equiv I_{22} - I_{21} I_{11}^{-1} I_{12}\,.$$

By (b) and (c), the information bound for estimating ν in $\mathcal{P}$ is $I^{11} = I_{11\cdot 2}^{-1}$ and the efficient influence for ν in $\mathcal{P}$ is

$$(6) \qquad \begin{aligned} \tilde{\mathbf{l}}_1 &= I^{11}\dot{\mathbf{l}}_1 + I^{12}\dot{\mathbf{l}}_2 \\ &= I_{11\cdot 2}^{-1}(\dot{\mathbf{l}}_1 - I_{12} I_{22}^{-1}\dot{\mathbf{l}}_2) \qquad \text{by } (b) \\ &\equiv I_{11\cdot 2}^{-1}\mathbf{l}_1^*\,. \end{aligned}$$

Since

$$\begin{aligned} I_{11\cdot 2} &= E_0(\dot{\mathbf{l}}_1 - I_{12} I_{22}^{-1}\dot{\mathbf{l}}_2)(\dot{\mathbf{l}}_1 - I_{12} I_{22}^{-1}\dot{\mathbf{l}}_2)' \\ &= E_0(\mathbf{l}_1^* \mathbf{l}_1^{*'})\,, \end{aligned}$$

we see that (6) has the same form as $\tilde{\mathbf{l}} = I^{-1}(\theta_0)\dot{\mathbf{l}}$ with $\tilde{\mathbf{l}}$ replaced by $\tilde{\mathbf{l}}_1$, $I(\theta_0) = E_0(\dot{\mathbf{l}}\dot{\mathbf{l}}')$ replaced by $I_{11\cdot 2} = E_0(\mathbf{l}_1^* \mathbf{l}_1^{*T})$, and $\dot{\mathbf{l}}$ replaced by

$$(7) \qquad \mathbf{l}_1^* \equiv \dot{\mathbf{l}}_1 - I_{12} I_{22}^{-1}\dot{\mathbf{l}}_2\,.$$

We therefore call $\mathbf{l}_1^*$ the *efficient score function* for ν in $\mathcal{P}$, and call $I_{11\cdot 2}$ the *information for* ν in $\mathcal{P}$.

If on the other hand $\eta = \eta_0$ is treated as known, the information bound for ν in $\mathcal{P}_1(\eta_0)$ is I_{11}^{-1} and the corresponding efficient influence function for ν in $\mathcal{P}_1(\eta_0)$ is just

$$(8) \qquad I_{11}^{-1}\dot{\mathbf{l}}_1\,.$$

From the block matrix formulas relating $[I_{ij}]$ and $[I^{ij}]$, we can derive some important relations between these quantities. First note from (4) and (5) that

$$(9) \qquad (I^{11})^{-1} = I_{11\cdot 2} = I_{11} - I_{12} I_{22}^{-1} I_{21}\,,$$

so not knowing η decreases the information for ν by $I_{12} I_{22}^{-1} I_{21}$. Similarly,

$$I_{11}^{-1} = I^{11} - I^{12}(I^{22})^{-1} I^{21}$$

or

$$(10) \qquad I^{11} = I_{11\cdot 2}^{-1} = I_{11}^{-1} + I^{12}(I^{22})^{-1} I^{21}\,,$$

so not knowing η increases the information bound (inverse information) by $I^{12}(I^{22})^{-1}I^{21}$. Moreover, from (9),

$$(11) \qquad I_{11\cdot 2} = I_{11} \qquad \text{and} \qquad I_{11\cdot 2}^{-1} = I_{11}^{-1}$$

if and only if

$$(12) \qquad I_{12} = 0\,.$$

In this case it also follows from (6), (7), and (11) that

$$(13) \qquad \tilde{\mathbf{l}}_1 = I_{11}^{-1}\dot{\mathbf{l}}_1 \qquad \text{and} \qquad \mathbf{l}_1^* = \dot{\mathbf{l}}_1\,.$$

Definition 2.1 $\{\widehat{\nu}_n\}$ is an *adaptive estimator* of ν in the presence of η if $\widehat{\nu}_n$ is regular on $\mathcal{P}$ and efficient for each of the models $\mathcal{P}_1(\eta)$ for $\eta \in \mathcal{H}$.

If an adaptive estimate exists we can do as well not knowing η as knowing it. By (4) and (13), a necessary condition for the existence of adaptive estimates in regular parametric models is

$$(14) \qquad I_{12}(\theta) = 0 \qquad \text{for all} \quad \theta\,.$$

Adaptation is very much a feature of the parametrization, as the following examples show.

Example 2.6 (Gaussian location - scale). Suppose that

$$\mathcal{P} = \{P_\theta :\ p_\theta = \phi((\cdot - \nu)/\eta)/\eta,\ \nu \in R,\ \eta > 0\}\,,$$

the usual normal location - scale model. Note that

$$\dot{\mathbf{l}}_\nu(x) = \frac{x-\nu}{\eta^2}\,, \qquad \dot{\mathbf{l}}_\eta(x) = \frac{1}{\eta}\left\{\frac{(x-\nu)^2}{\eta^2} - 1\right\}\,,$$

and the information matrix $I(\theta)$ is given by

$$I(\theta) = \begin{pmatrix} 1/\eta^2 & 0 \\ 0 & 2/\eta^2 \end{pmatrix} = \frac{1}{\eta^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\,.$$

Thus we can estimate the mean equally well whether we know or do not know the variance.

Example 2.7 (Reparametrization of Gaussian location-scale). Now suppose that $\mathcal{P}$ is the same as in the preceding example, but we reparametrize as follows:

$$P_\theta = N(\nu, \eta^2 - \nu^2), \qquad \eta^2 > \nu^2\,.$$

Then easy calculation using (2 shows that

$$I_{12}(\theta) = -\frac{\nu\eta}{(\eta^2-\nu^2)^2}\,,$$

Thus lack of knowledge of η in this parameterization does change the information bound for estimation of ν.

We can think of $\mathbf{l}_1^*$ as the $\dot{\mathbf{l}}_1$ corresponding to the reparametrization $(\nu, \eta) \mapsto (\nu, \eta + I_{22}^{-1}(\theta_0)I_{21}(\theta_0)(\nu - \nu_0))$. With this reparametrization, adaptation at θ_0 becomes possible since $\dot{\mathbf{l}}_2$ is unchanged and condition (4) is satisfied. If we can paste together these local reparametrizations and find $(\nu, \eta) \mapsto (\nu, \gamma(\nu, \eta))$ such that

$$\gamma(\nu, \eta) - \gamma(\nu_0, \eta_0) = \eta - \eta_0 + I_{22}^{-1}I_{21}(\theta_0)(\nu - \nu_0) + o(|\theta - \theta_0|)$$

for every $\theta_0 = (\nu_0, \eta_0)$, then under this reparametrization the necessary condition for adaptation holds. For instance in example 2.7 we can take $\gamma(\nu, \eta) = \eta^2 - \nu^2$. These remarks have little practical significance since the initial parametrization is usually natural and the reparametrization is not.

The efficient influence function $\widetilde{\mathbf{l}}_1$ and efficient score function $\mathbf{l}_1^*$ can be interpreted geometrically in the Hilbert space $L_2(P_\theta)$; see BKRW sections A.1 and A.2. for elementary Hilbert space theory. First suppose that $m = 1$. Let $[\dot{\mathbf{l}}_2]$ be the linear span of the components of $\dot{\mathbf{l}}_2$ in $L_2(P_{\theta_0})$. Then by BKRW Example A.2.1, $I_{12}I_{22}^{-1}\dot{\mathbf{l}}_2$ is the projection of $\dot{\mathbf{l}}_1$ of $[\dot{\mathbf{l}}_2]$, and by (7) the efficient score function $\mathbf{l}_1^*$ is the projection of $\dot{\mathbf{l}}_1$ on the orthocomplement of $[\dot{\mathbf{l}}_2]$.

We can also relate the efficient influence functions $\widetilde{\mathbf{l}}_1$ and $I_{11}^{-1}\dot{\mathbf{l}}_1$ for ν in $\mathcal{P}$ and $\mathcal{P}_1(\eta_0)$. In particular, $I_{11}^{-1}\dot{\mathbf{l}}_1$ is the projection of $\widetilde{\mathbf{l}}_1$ on $[\dot{\mathbf{l}}_1]$. We need only check that $\widetilde{\mathbf{l}}_1 - I_{11}^{-1}\dot{\mathbf{l}}_1 = (I^{11} - I_{11}^{-1})\dot{\mathbf{l}}_1 + I^{12}\dot{\mathbf{l}}_2$ is orthogonal to $\dot{\mathbf{l}}_1$, and this follows easily from $I^{11}I_{11} + I^{12}I_{21} = 1$.

If $m > 1$ these relationships continue to hold if projection is interpreted componentwise. The following basic proposition can be viewed as providing the rationale for two different approaches to computing information bounds in semiparametric models which are presented in Chapter 3 of BKRW (1993).

Proposition 2.2 (Efficient Score and Efficient Influence Functions)
A. The efficient score function $\mathbf{l}_1^*(\cdot, P_{\theta_0}|\nu, \mathcal{P})$ is the projection of the score function $\dot{\mathbf{l}}_1$ on the orthocomplement of $[\dot{\mathbf{l}}_2]$ in $L_2(P_{\theta_0})$.
B. The efficient influence function $\widetilde{\mathbf{l}}(\cdot, P_{\theta_0}|\nu, \mathcal{P}_1(\eta_0))$ is the projection of the efficient influence function $\widetilde{\mathbf{l}}_1$ on $[\dot{\mathbf{l}}_1]$ in $L_2(P_{\theta_0})$.

See Figures 1 and 2.

Here is another relationship between the influence and score functions of $\mathcal{P}_1(\eta_0)$ and its companion $\mathcal{P}_2(\nu_0) \equiv \{P_{\nu_0,\eta} : \ \eta \in \mathcal{H}\}$ We use the subscript 2 for score and influence functions in the companion model. The efficient influence function $\widetilde{\mathbf{l}}_1$ can be written as

(15) $\quad \widetilde{\mathbf{l}}_1 = I_{11}^{-1}\dot{\mathbf{l}}_1 - I_{11}^{-1}I_{12}\widetilde{\mathbf{l}}_2 \, .$

This relationship was implicit in section 4 of Begun, Hall, Huang, and Wellner (1983). It appears in the context of semiparametric models (with ν infinite-dimensional and η finite-dimensional in section 5.4 of BKRW (1993). Note that (15) provides an immediate proof, by orthogonality of $\widetilde{\mathbf{l}}_2$ to $[\dot{\mathbf{l}}_1]$, of the formula

(16) $\quad I_{11\cdot 2}^{-1} = I_{11}^{-1} + I_{11}^{-1}I_{12}I_{22\cdot 1}^{-1}I_{21}I_{11}^{-1} \, ,$

which is another way of writing (10).
Proof of (15): From (6),

$$\begin{aligned}
\widetilde{\mathbf{l}}_1 + I_{11}^{-1}I_{12}\widetilde{\mathbf{l}}_2 &= I^{11}\dot{\mathbf{l}}_1 + I^{12}\dot{\mathbf{l}}_2 + I_{11}^{-1}I_{12}(I^{21}\dot{\mathbf{l}}_1 + I^{22}\dot{\mathbf{l}}_2) \\
&= I_{11}^{-1}\left\{(I_{11}I^{11} + I12I^{21})\dot{\mathbf{l}}_1 + (I_{11}I^{12} + I_{12}I^{22})\dot{\mathbf{l}}_2\right\} \\
&= I_{11}^{-1}\dot{\mathbf{l}}_1 \, ,
\end{aligned}$$

Table 3.1: Efficient Scores and Influence Functions

name	**notation**	$\mathcal{P}$	$\mathcal{P}_1(\eta_0)$
efficient score	$\mathbf{l}_1^*(\cdot, P\vert\nu, \cdot)$	$\mathbf{l}_1^* = \dot{\mathbf{l}}_1 - I_{12}I_{22}^{-1}\dot{\mathbf{l}}_2$	$\dot{\mathbf{l}}_1$
information	$I(P\vert\nu, \cdot)$	$E\mathbf{l}_1^*\mathbf{l}_1^{*T} = I_{11} - I_{12}I_{22}^{-1}I_{21} \equiv I_{11\cdot 2}$	I_{11}
efficient influence function	$\widetilde{\mathbf{l}}_1(\cdot, P\vert\nu, \cdot)$	$\widetilde{\mathbf{l}}_1 = I^{11}\dot{\mathbf{l}}_1 + I^{12}\dot{\mathbf{l}}_2$ $= I_{11\cdot 2}^{-1}\mathbf{l}_1^*$ $= I_{11}^{-1}\dot{\mathbf{l}}_1 - I_{11}^{-1}I_{12}\widetilde{\mathbf{l}}_2$	$I_{11}^{-1}\dot{\mathbf{l}}_1$
information bound	$I^{-1}(P\vert\nu, \cdot)$	$I^{11} = I_{11\cdot 2}^{-1}$ $= I_{11}^{-1} + I_{11}^{-1}I_{12}I_{22\cdot 1}^{-1}I_{21}I_{11}^{-1}$	I_{11}^{-1}

and rearranging yields (15). $\square$

The following table summarizes the efficient score functions, efficient influence functions, information, and inverse information for the two models $\mathcal{P}$ and $\mathcal{P}_1(\eta_0)$.

Proposition 2.2 can be put in a broader context.

Proposition 2.3 Suppose that $m = 1$ and that T_n is an asymptotically linear estimator of ν with influence function ψ. Then:
A. T_n is Gaussian regular if and only if

(4) $$\psi - \widetilde{\mathbf{l}}_1 \perp \dot{\mathcal{P}} = [\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]\,,$$

or, equivalently, if and only if both

(5) $$\langle \psi, \dot{\mathbf{l}}_1 \rangle_0 = 1$$

and

(6) $$\psi \perp [\dot{\mathbf{l}}_2]\,.$$

B. If T_n is regular, then $\psi \in \dot{\mathcal{P}} = [\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]$ if and only if $\psi = \widetilde{l}_1$.

Note that (5) and (6) are asymptotic versions of the equations leading to the Cramér-Rao information bound. Consider the problem of minimizing $\Sigma(P_{\theta_0}, T) = E_0\psi^2$ subject to (5) and (6). For simplicity take $k = 2$. If we write

$$\psi = c\dot{\mathbf{l}}_1 + d\dot{\mathbf{l}}_2 + \Delta$$

where $\Delta \perp [\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]$, then (6) holds if and only if

$$\psi = c(\dot{\mathbf{l}}_1 - I_{12} I_{22}^{-1} \dot{\mathbf{l}}_2) + \Delta = c\mathbf{l}^* + \Delta\,,$$

while (5) forces

$$c = \|\dot{\mathbf{l}}_1 - I_{12} I_{22}^{-1} \dot{\mathbf{l}}_2\|_0^{-2}\,.$$

Finally,

$$\|\psi\|_0^2 = \|\mathbf{l}_1^*\|_0^{-2} + \|\Delta\|_0^2\,.$$

Therefore the minimizing $\Delta = 0$ and as expected the minimizing ψ is the efficient influence function. This argument makes clear the characterizing features of the efficient influence function implied in proposition 2.2, part B:

(i) $\widetilde{\mathbf{l}}_1$ and all other influence functions are orthogonal to $[\dot{\mathbf{l}}_2]$.
(ii) $\widetilde{\mathbf{l}}_1$ is the unique influence function belonging to $[\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]$.
(iii) $\widetilde{\mathbf{l}}_1$ can be obtained by projecting any influence function ψ corresponding to a regular estimator for ν on $[\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]$.

Here is a slight generalization of proposition 2.3 to a general function $\nu(P_\theta) = q(\theta)$.

Proposition 2.4 (Characterization of Gaussian regular estimators). Suppose that T_n is an asymptotically linear estimator at θ_0 of $\nu(P_\theta) = q(\theta)$ with influence function ψ where $q : \Theta \to R^m$. Then: Then:

A. T_n is Gaussian regular estimator at θ_0 if and only if $q(\theta)$ is differentiable at θ_0 with derivative $\dot{q}(\theta)$ and, with $\widetilde{\mathbf{l}}_\nu \equiv \widetilde{\mathbf{l}}(\cdot, P_{\theta_0}|\nu, \mathcal{P})$,

$$\psi - \widetilde{\mathbf{l}}_\nu \perp \dot{\mathcal{P}} = [\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]\,, \tag{7}$$

where (7) is equivalent to

$$E_0(\psi \dot{\mathbf{l}}) = \langle \psi, \dot{\mathbf{l}} \rangle_0 = \dot{q}(\theta_0)\,. \tag{8}$$

B. If T_n is regular, then $\psi \in \dot{\mathcal{P}}^m$ if and only if

$$\psi = \widetilde{\mathbf{l}}_\nu = \dot{q}^T(\theta_0) I^{-1}(\theta_0) \dot{\mathbf{l}}_\theta\,. \tag{9}$$

Proof. By asymptotic linearity of T_n and corollary 4 of Le Cam's second lemma, it follows that

$$\text{(a)}\qquad \begin{pmatrix} \sqrt{n}(T_n - q(\theta_0)) \\ L_n(\theta_0 + t_n/\sqrt{n}) - L_n(\theta_0) \end{pmatrix} \to_d N\left(\begin{pmatrix} 0 \\ -\Sigma_{22} \end{pmatrix}, \Sigma \right) \qquad \text{under } P_{\theta_0}$$

where $t_n \to t$ and

$$\text{(b)}\qquad \Sigma = [\Sigma_{ij}], \quad \Sigma_{11} = E_0(\psi\psi^T), \quad \Sigma_{12} = E_0(\psi \dot{\mathbf{l}}^T)t\,, \quad \Sigma_{22} = t^T I(\theta_0) t\,.$$

Consequently, by Le Cam' third lemma, (Lemma 3.4)

$$\text{(c)}\qquad \sqrt{n}(T_n - q(\theta_0)) \to_d N(\Sigma_{12}, \Sigma_{11}) \qquad \text{under } P_{\theta_0 + n^{-1/2} t_n}$$

Now assume that T_n is regular. Then

$$\text{(d)}\qquad \sqrt{n}(T_n - q(\theta_0 + t_n/\sqrt{n})) \to N(0, \Sigma_{11}) \qquad \text{under } P_{\theta_0 + n^{-1/2} t_n}$$

and from (c) and (d) we conclude that

$$\text{(e)} \qquad \sqrt{n}(q(\theta_0 + n^{-1/2}t_n) - q(\theta_0)) \to \Sigma_{12} = E_0(\psi \dot{\mathbf{l}}^T)t\,.$$

But this implies that q is differentiable at θ_0 with derivative $\dot{q}(\theta_0)$ satisfying (8) and hence (7).

On the other hand, if q is differentiable and (8) holds, then (e) is valid, which together with (c) implies (d) and hence Gaussian regularity. The proof of A is complete.

To prove B, note that A implies that $\dot{q}$ and hence $\widetilde{\mathbf{l}}_\nu$ are well-defined and that (7) holds. Since $\widetilde{\mathbf{l}}_\nu \in \dot{\mathcal{P}}^m$, (7) yields $\psi \in \dot{\mathcal{P}}^m$ if and only if $\psi - \widetilde{\mathbf{l}}_\nu = 0$. $\square$

Choosing $q(\theta) = q(\nu, \eta) = \nu$ in proposition 2.4 immediately yields a generalization of proposition 2.3 to $m > 1$. Now (8) becomes

$$\text{(10)} \qquad E_0(\psi \dot{\mathbf{l}}_1^T) = J_{m \times m}\,,$$

$$\text{(11)} \qquad E_0(\psi \dot{\mathbf{l}}_2^T) = 0$$

where J is the identity. In particular if $m = k$ we obtain that the influence function of any linear and Gaussian regular estimate of θ has

$$\text{(12)} \qquad E_0(\psi \dot{\mathbf{l}}^T) = J_{k \times k}\,.$$

3 Regular Estimates and Superefficiency

If $X_1, \ldots, X_n$ are i.i.d. P_θ, an estimator T_n is unbiased for estimating $q(\theta)$, and the conditions of the information inequality (theorem 2.1) hold, then

$$(1) \qquad Var_\theta[T_n] \ge \frac{[\dot{q}(\theta)]^2}{nI(\theta)}\,.$$

If

$$(2) \qquad \sqrt{n}(T_n - q(\theta)) \to_d N(0, V^2(\theta))\,,$$

then it follows (from Fatou and Skorokhod, recall corollary 2.3.1) that

$$(3) \qquad V^2(\theta) \le \liminf_{n\to\infty}\{nVar_\theta[T_n]\}\,.$$

If T_n is unbiased and

$$(4) \qquad V^2(\theta) = \lim_{n\to\infty}\{nVar_\theta[T_n]\}\,,$$

then (1) implies

$$(5) \qquad V^2(\theta) \ge \frac{[\dot{q}(\theta)]^2}{I(\theta)}\,.$$

Does the inequality in (5) hold under restrictions on p_θ alone? The answer to this question is *no*, as is shown by the following example due to Hodges.

Example 3.1 (Hodges superefficient estimator). Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, 1)$ so that $I(\theta) = 1$. Let $|a| < 1$, and define

$$(6) \qquad T_n = \begin{cases} \overline{X}_n & \text{if} \quad |\overline{X}_n| > n^{-1/4} \\ a\overline{X}_n & \text{if} \quad |\overline{X}_n| \le n^{-1/4}\,. \end{cases}$$

Then

$$(7) \qquad \sqrt{n}(T_n - \theta) \to_d N(0, V^2(\theta))$$

where

$$(8) \qquad V^2(\theta) = \begin{cases} 1 & \text{if} \quad \theta \ne 0 \\ a^2 & \text{if} \quad \theta = 0\,. \end{cases}$$

Thus $V^2(\theta) \ge 1/I(\theta)$ fails at $\theta = 0$ if $|a| < 1$, and T_n is a *superefficient estimator* of θ at $\theta = 0$.

Proof of (7). Since $\sqrt{n}(\overline{X}_n - \theta) \stackrel{d}{=} Z \sim N(0,1)$ for all $n \ge 1$ and all θ,

$$\begin{aligned}
\sqrt{n}(T_n - \theta) &= \sqrt{n}(\overline{X} - \theta)1_{[|\overline{X}_n| > n^{-1/4}]} + \sqrt{n}(a\overline{X}_n - \theta)1_{[|\overline{X}_n| \le n^{-1/4}]} \\
&= \sqrt{n}(\overline{X} - \theta)1_{[\sqrt{n}|\overline{X}_n - \theta + \theta| > n^{1/4}]} \\
&\qquad + \{a\sqrt{n}(\overline{X}_n - \theta) + \sqrt{n}\theta(a-1)1_{[\sqrt{n}|\overline{X}_n - \theta + \theta| \le n^{1/4}]} \\
(9) \qquad &\stackrel{d}{=} Z1_{[|Z + \sqrt{n}\theta| \ge n^{1/4}]} + \{aZ + \sqrt{n}\theta(a-1)\}1_{[|Z + \sqrt{n}\theta| \le n^{1/4}]} \\
&\to_{a.s.} \left\{ \begin{array}{lll} Z & \text{if} & \theta \ne 0 \\ aZ & \text{if} & \theta = 0 \end{array} \right\} \sim N(0, V^2(\theta))\,.
\end{aligned}$$

Note that $V^2(\theta)$ is a discontinuous function of θ. If $\theta \equiv \theta_n = cn^{-1/2}$, then from (9), under P_{θ_n} we have

$$\begin{aligned}\sqrt{n}(T_n - \theta_n) &\stackrel{d}{=} Z 1_{[|Z+c|>n^{1/4}]} + \{aZ + c(a-1)\} 1_{[|Z+c|\le n^{1/4}]} \\ &\to aZ + c(a-1) \sim N(c(a-1), a^2)\,.\end{aligned}$$

Note that this limiting distribution depends on c, and hence Hodges' superefficient estimator is not locally regular in the following sense.

Definition 3.1 (Locally regular estimator). $T = \{T_n\}$ is a *locally regular estimator* of θ at $\theta = \theta_0$ if, for every sequence $\{\theta_n\} \subset \Theta$ with $\sqrt{n}(\theta_n - \theta_0) \to t \in R^k$, under P_{θ_n}

(10) $\quad \sqrt{n}(T_n - \theta_n) \to_d \mathbb{Z} \qquad \text{as} \quad n \to \infty$

where the distribution of $\mathbb{Z}$ depends on θ_0 but not on t. Thus the limit distributiion of $\sqrt{n}(T_n - \theta_n)$ (under sampling from P_{θ_n}) does not depend on the direction of approach t of θ_n to θ_0.

This will turn out to be a key hypothesis in the formulation of Hájek's convolution theorem in the next section.

Contiguity Theory: Le Cam's four lemmas and LAN

Consider a sequence of statistical problems (with only two sequences of probability measures) with

measure spaces:	$(\mathbb{X}_n, \mathcal{A}_n, \mu_n)$
probability measures:	$P_n << \mu_n, \quad Q_n << \mu_n$
densities:	$p_n = \frac{dP_n}{d\mu_n}, \quad q_n = \frac{dQ_n}{d\mu_n}$
likelihood ratios:	$L_n \equiv \begin{cases} q_n/p_n & \text{if} \quad p_n > 0 \\ 1 & \text{if} \quad q_n = p_n = 0 \\ n & \text{if} \quad q_n > 0 = p_n\,. \end{cases}$

Definition 3.2 (Contiguity). The sequence $\{Q_n\}$ is *contiguous* to $\{P_n\}$ if for every sequence $B_n \in \mathcal{A}_n$ for which $P_n(B_n) \to 0$ it follows that $Q_n(B_n) \to 0$.

Thus contiguity of $\{Q_n\}$ to $\{P_n\}$ means that Q_n is "asymptotically absolutely continuous" with respect to P_n in the sense of domination of measures. We therefore denote contiguity of $\{Q_n\}$ to $\{P_n\}$ by $\{Q_n\} \lhd \{P_n\}$, a notation due to Witting and Nölle (1970). Two sequences are contiguous to each other if both $\{Q_n\} \lhd \{P_n\}$ and $\{P_n\} \lhd \{Q_n\}$, and we then write $\{P_n\} \lhd \rhd \{Q_n\}$.

Definition 3.3 (Asymptotic orthogonality). The sequence $\{Q_n\}$ is *asymptotically orthogonal* to $\{P_n\}$ if there exists a sequence $B_n \in \mathcal{A}_n$ such that $Q_n(B_n) \to 1$ and $P_n(B_n) \to 0$.

Lemma 3.1 (Le Cam's first lemma). Suppose that $\mathcal{L}(L_n|P_n) \to \mathcal{L}(L)$ and $E(L) = 1$. Then $\{Q_n\} \lhd \{P_n\}$.

Corollary 2 (Normal log - likelihood). If $\mathcal{L}(\log L_n|P_n) \to \mathcal{L}(\log L) = N(-\sigma^2/2, \sigma^2)$, then $\{Q_n\} \lhd \rhd \{P_n\}$.

Proof. Note that $\mathcal{L}(L) = \mathcal{L}(e^{\sigma Z - \sigma^2/2})$ where $\mathcal{L}(Z) = N(0,1)$ and hence $E(L) = 1$. $\square$

Definition 3.4 A sequence of random variables $\{X_n\}$ (with X_n defined on $(\mathbb{X}_n, \mathcal{A}_n, P_n)$ is *uniformly integrable* if

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} E_n(|X_n| 1_{[|X_n|\geq\lambda]}) = 0\,.$$

Proposition 3.1 (Condition for Uniform integrability). $\{X_n\}$ is uniformly integrable if and only if both of the following hold

$$\sup_{n\geq 1} E_n|X_n| < \infty\,. \tag{11}$$

$$B_n \in \mathcal{A}_n \ \text{ with } \ P_n(B_n) \to 0 \ \text{ implies } \ E_n(|X_n| 1_{B_n}) \to 0\,. \tag{12}$$

Proof. See e.g. Billingsley (1968) page 34 or Chow and Teicher (1978) pages 92 - 93. $\square$

Lemma 3.2 (Le Cam's fourth lemma, Hall and Loynes (1977)). $\{Q_n\} \lhd \{P_n\}$ if and only if $\{L_n\}$ is uniformly integrable with respect to $\{P_n\}$ and $Q_n([p_n = 0]) \to 0$.

Now suppose that $\underline{X}_n = (X_1, \ldots, X_n) \in \mathbb{X}_n$ and that

$$\begin{aligned} p_n(\underline{x}_n) &= \prod_{i=1}^n f_{ni}(x_i), \qquad P_n \equiv \prod_{i=1}^n P_{ni}\,, \\ q_n(\underline{x}_n) &= \prod_{i=1}^n g_{ni}(x_i), \qquad Q_n \equiv \prod_{i=1}^n Q_{ni}\,, \end{aligned}$$

so that

$$\log L_n = \sum_{i=1}^n \log\left(\frac{g_{ni}}{f_{ni}}(X_i)\right) \quad \begin{cases} < \infty & \text{a.s. } P_n \\ > -\infty & \text{a.s. } Q_n\,. \end{cases} \tag{13}$$

Suppose that the summands in (13) satisfy the *uniform asymptotic negligibility* (UAN) condition

$$\max_{1\leq i\leq n} P_n\left(\left|\frac{g_{ni}}{f_{ni}}(X_i) - 1\right| > \epsilon\right) \to 0 \qquad \text{for all } \ \epsilon > 0\,. \tag{14}$$

To get random variables with finite variance (to which classical central limit theorems may be applied), let

$$W_n \equiv 2\sum_{i=1}^n \left\{\frac{g_{ni}^{1/2}}{f_{ni}^{1/2}}(X_i) - 1\right\} \equiv \sum_{i=1}^n T_{ni}\,, \tag{15}$$

and note that

$$Var\left(\frac{g_{ni}^{1/2}}{f_{ni}^{1/2}}(X_i)\right) \leq E\left(\frac{g_{ni}}{f_{ni}}(X_i)\right) = \int 1_{[f_{ni}=0]} g_{ni} d\mu_n \leq 1\,.$$

The following lemma reduces the proof of asymptotic normality of $\log L_n$ to the problem of establishing asymptotic normality of W_n.

Lemma 3.3 (Le Cam's second lemma). Suppose that the UAN condition (10) holds and $\mathcal{L}(W_n|P_n) \to N(-\sigma^2/4, \sigma^2)$. Then

$$\log L_n - (W_n - \sigma^2/4) = o_{P_n}(1) \tag{16}$$

and hence

$$\mathcal{L}(\log L_n|P_n) \to N(-\sigma^2/2, \sigma^2)\,. \tag{17}$$

The proof of lemma 3.3 involves a long truncation argument, and is therefore deferred to the end of the section.

Corollary 3 (LAN under differentiability). If f_n is a sequence of densities such that

$$\|\sqrt{n}(f_n^{1/2} - f^{1/2}) - \delta\|_2 \to 0 \qquad \text{as} \;\; n \to \infty$$

where $\|\cdot\|_2$ is the $L_2(\mu)$–metric and $\delta \in L_2(\mu)$, then with $p_n(\underline{x}) \equiv \prod_{i=1}^n f(x_i)$ and $q_n(\underline{x}) \equiv \prod_{i=1}^n f_n(x_i)$, it follows that

$$\log L_n - \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{2\delta}{f^{1/2}}(X_i) - \frac{1}{2}\|2\delta\|_2^2\right) = o_{P_n}(1) \tag{18}$$

and hence

$$\mathcal{L}(\log L_n|P_n) \to N(-\sigma^2/2, \sigma^2) \tag{19}$$

with

$$\sigma^2 = \|2\delta\|_2^2 = 4\int \delta^2 d\mu\,. \tag{20}$$

Corollary 4 (Hellinger-differentiable parametric model). Suppose that $\mathcal{P} = \{P_\theta :\; \theta \in \Theta \subset R^k\}$ is a regular parametric model dominated by a sigma-finite measure μ in the sense that

$$\|\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h^t \dot{\mathbf{l}}_\theta \sqrt{p_\theta}\|_\mu = o(|h|)\,.$$

Then, with $\theta_n \equiv \theta + n^{-1/2}h \in \Theta$, $h \in R^k$, $p_n(\underline{x}) \equiv \prod_{i=1}^n p_\theta(x_i)$, and $q_n(\underline{x}) \equiv \prod_{i=1}^n p_{\theta_n}(x_i)$, it follows that

$$\log L_n - \left(\frac{h^T}{\sqrt{n}}\sum_{i=1}^n \dot{\mathbf{l}}_\theta(X_i) - \frac{1}{2}h^T I(\theta) h\right) = o_{P_n}(1)\,,$$

and hence

$$\mathcal{L}(\log L_n|P_n) \to N(-\sigma^2/2, \sigma^2)$$

with $\sigma^2 = h^T I(\theta) h$.

Proof. This follows immediately from corollary 3 with the identification $\delta = \dot{\mathrm{l}}$. $\square$

Now suppose that under

$$(21) \quad P_n : \quad X_{n1}, \ldots, X_{nn} \qquad \text{are i.i.d. } f\,,$$

and under

$$(22) \quad Q_n : \quad X_{n1}, \ldots, X_{nn} \qquad \text{are independent with densities } f_{n1}, \ldots, f_{nn}$$

with respect to μ. Assume that $a_{n1}, \ldots, a_{nn}$, $n \geq 1$, are constants which satisfy

$$(23) \quad \max_{1 \leq i \leq n} \frac{a_{ni}^2}{a_n^T a_n} = \frac{\max_{1 \leq i \leq n} a_{ni}^2}{\sum_{i=1}^n a_{ni}^2} \rightarrow 0 \qquad \text{as} \qquad n \rightarrow \infty\,,$$

and suppose there exists $\delta \in L_2(F)$ such that

$$(24) \quad \sum_{i=1}^n \|(f_{ni}^{1/2} - f^{1/2}) - \frac{a_{ni}}{\sqrt{a_n^T a_n}} \delta\|_2^2 \rightarrow 0 \qquad \text{as} \qquad n \rightarrow \infty\,.$$

Corollary 5 (LAN, regression setting) Suppose that (21) - (24) hold. Then

$$(25) \quad \log L_n - \left(Z_n - \frac{1}{2} \|2\delta\|_2^2 \right) = o_{P_n}(1)$$

where

$$(26) \quad Z_n \equiv 2 \sum_{i=1}^n \frac{a_{ni}}{\sqrt{a_n^T a_n}} \frac{\delta}{f^{1/2}}(X_{ni})$$

and

$$(27) \quad \mathcal{L}(Z_n) \rightarrow N(0, \|2\delta\|_2^2) \qquad \text{as } \; n \rightarrow \infty\,.$$

Proof. See Shorack and Wellner (1986), page 154 and 163 - 165. Note that corollary 3 is the special case of corollary 5 with all $a_{ni} = 1$ and $f_{ni} = f_n$ for all $i = 1, \ldots, n$. $\square$

Lemma 3.4 (Le Cam's third lemma). Suppose that a statistic T_n satisfies

$$(28) \quad \begin{aligned} \mathcal{L}\left((T_n, \log L_n)^T | P_n \right) \quad &\rightarrow \quad \mathcal{L}((T, \log L)^T) \\ &\sim \quad N_2\left(\left(\begin{array}{c} \mu \\ -\sigma^2/2 \end{array} \right), \left(\begin{array}{cc} \tau^2 & c \\ c & \sigma^2 \end{array} \right) \right). \end{aligned}$$

Then

$$(29) \quad \begin{aligned} \mathcal{L}\left((T_n, \log L_n)^T | Q_n \right) \quad &\rightarrow \quad \mathcal{L}((T + c, \log L + \sigma^2)^T) \\ &\sim \quad N_2\left(\left(\begin{array}{c} \mu + c \\ +\sigma^2/2 \end{array} \right), \left(\begin{array}{cc} \tau^2 & c \\ c & \sigma^2 \end{array} \right) \right). \end{aligned}$$

Remark 3.1 If T_n is asymptotically linear and $\log L_n$ is asymptotically linear (which is often a consequence of the second lemma), then verification of (28) is straightforward via the multivariate central limit theorem.

Proofs

Proof. (lemma 3.1, Le Cam's first lemma). Let $B_n \in \mathcal{A}_n$ with $P_n(B_n) \to 0$. By the Neyman-Pearson lemma there is a critical function $\phi_n \equiv 1\{L_n > k_n\} + \gamma_n 1\{L_n = k_n\}$ such that $E_{P_n}(\phi_n) = \alpha_n \equiv P_n(B_n) \to 0$ and

$$Q_n(B_n) \le Q_n(\phi_n)\,.$$

But for any fixed $0 < y < \infty$,

$$\begin{aligned} Q_n(B_n) &\le Q_n(\phi_n) = Q_n(\phi_n 1\{L_n \le y\}) + Q_n(\phi_n 1\{L_n > y\} \\ &\le yP_n(\phi_n) + Q_n(1\{L_n > y\}) \\ &= yP_n(\phi_n) + 1 - Q_n(L_n \le y) \\ \text{(a)} \qquad &= yP_n(\phi_n) + 1 - P_n(L_n 1\{L_n \le y\})\,. \end{aligned}$$

Let $\epsilon > 0$ and choose y to be a continuity point of $\mathcal{L}(L)$ such that $1 - E(L1\{L \le y\}) < \epsilon/2$; this is possible since $E(L) = 1$ by hypothesis. Then $\mathcal{L}(L_n|P_n) \to \mathcal{L}(L)$ implies that $P_n(L_n 1\{L_n \le y\}) \to E(L1\{L \le y\})$ and hence $1 - P_n(L_n 1\{L_n \le y\}) < \epsilon$ for $n \ge N_1$. Since $P_n(B_n) \to 0$ we also have $yP_n(B_n) < \epsilon$ for $n \ge$ some N_2, and hence it follows from (a) that $Q_n(B_n) < 2\epsilon$ for $n \ge \max\{N_1, N_2\}$. $\square$

Proof. (lemma 3.2, Le Cam's fourth lemma). First note that for $B_n \in \mathcal{A}_n$ we have

$$\begin{aligned} Q_n(B_n) &= \int 1_{B_n} dQ_n \\ &= \int 1_{B_n \cap [p_n = 0]} dQ_n + \int 1_{B_n \cap [p_n > 0]} L_n dP_n \\ &= \int 1_{B_n \cap [p_n = 0]} dQ_n + \int 1_{B_n} L_n dP_n \\ \text{(a)} \qquad &\le Q_n(p_n = 0) + \int 1_{B_n} L_n dP_n \\ \text{(b)} \qquad &\ge \int 1_{B_n} L_n dP_n\,. \end{aligned}$$

Thus if $P_n(B_n) \to 0$, L_n is uniformly integrable and $Q_n(p_n = 0) \to 0$, then $Q_n(B_n) \to 0$ by (a) and proposition 3.1, so $\{Q_n\} \lhd \{P_n\}$.

Conversely, if $\{Q_n\} \lhd \{P_n\}$ so that $Q_n(B_n) \to 0$, then (b) implies that $\int 1_{B_n} L_n dP_n \to 0$, so (ii) of proposition 3.1 holds. Part (i) of proposition 3.1 holds trivially since $P_n(L_n) = \int L_n dP_n = \int 1\{p_n > 0\} dQ_n \le 1$, and therefore $\{L_n\}$ is uniformly integrable with respect to $\{P_n\}$ by proposition 3.1. Since $P_n(p_n = 0) = 0$, contiguity implies that $Q_n(p_n = 0) \to 0$. $\square$

Proof. (lemma 3.3, Le Cam's second lemma). The following proof is from Hájek and Sidák (1967). For any function h with second derivative h'' we have

$$h(x) = h(x_0) + (x - x_0)h'(x_0) + \frac{1}{2}(x - x_0)^2 \int_0^1 2(1 - \lambda)h''(x_0 + \lambda(x - x_0))d\lambda$$

by integration by parts. thus for $h(x) = \log(1+x)$

$$\log(1+x) = x - \frac{1}{2}x^2 \int_0^1 \frac{2(1-\lambda)}{(1+\lambda x)^2} d\lambda\,.$$

Thus, with T_{ni} as in (15)

$$\log\left(\frac{g_{ni}}{f_{ni}}(X_i)\right) = 2\log(1+T_{ni}/2) = T_{ni} - \frac{1}{4}T_{ni}^2 \int_0^1 \frac{2(1-\lambda)}{1+\lambda T_{ni}/2)^2} d\lambda$$

and

$$\text{(a)} \qquad \log(L_n) = W_n - \frac{1}{4}\sum_{i=1}^n T_{ni}^2 \int_0^1 \frac{2(1-\lambda)}{(1+\lambda T_{ni}/2)^2} d\lambda\,.$$

Now we truncate: set $T_{ni}^\delta \equiv T_{ni} 1_{[|T_{ni}| \le \delta]}$ for $\delta > 0$. From the normal convergence criteria (see e.g. Loéve (1963), page 316), $\mathcal{L}(W_n|P_n) \to N(-\sigma^2/4, \sigma^2)$ and the UAN condition (14) holds if and only if

$$\text{(b)} \qquad \sum_{i=1}^n P_n(|T_{ni}| > \delta) \to 0\,,$$

$$\text{(c)} \qquad \sum_{i=1}^n E(T_{ni}^\delta) \to -\frac{1}{4}\sigma^2\,,$$

and

$$\text{(d)} \qquad \sum_{i=1}^n Var(T_{ni}^\delta) \to \sigma^2$$

where all expectations and variances are under P_n. Note that $\int_0^1 2(1-\lambda)d\lambda = 1$ and

$$P_n\{\max_{1\le i\le n} |T_{ni}| > \delta\} \le \sum_{i=1}^n P_n(|T_{ni}| > \delta) \to 0 \qquad \text{by (c)}\,.$$

Let $S_n \equiv \{\max_{1\le i\le n} |T_{ni}| \le \eta\}$. Thus for any $0 < \eta < 1$ there is an $N = N(\eta)$ such that, for $n \ge N$, $P_n(S_n) > 1 - \eta$. It follows that, on S_n

$$\sup_\lambda \max_{1\le i\le n} \left|(1+\lambda T_{ni}/2)^{-1} - 1\right| \le 8\eta$$

and hence

$$\max_{1\le i\le n} \Big| \int_0^1 \frac{2(1+\lambda)}{(1+\lambda T_{ni}/2)^2} d\lambda - 1\Big| \le 8\eta\,.$$

Also, since $T_{ni} = T_{ni}^\eta$ for $i = 1, \ldots, n$ on the event S_n,

$$\Big|\sum_{i=1}^n T_{ni}^2 \int_0^1 \frac{2(1+\lambda)}{(1+\lambda T_{ni}/2)^2} d\lambda - \sum_{i=1}^n T_{ni}^2\Big| \le 8\eta \sum_{i=1}^n T_{ni}^2 = 8\eta \sum_{i=1}^n (T_{ni}^\eta)^2$$

so that

$$\left| \frac{\sum_{i=1}^n T_{ni}^2 \int_0^1 \frac{2(1+\lambda)}{(1+\lambda T_{ni}/2)^2} d\lambda}{\sum_{i=1}^n (T_{ni}^\eta)^2} - 1 \right| \le 8\eta \qquad \text{on } \ S_n \, .$$

Thus in order to prove the lemma it suffices to show that

$$\text{(e)} \qquad \sum_{i=1}^n (T_{ni}^\eta)^2 \to_{P_n} \sigma^2 \, .$$

To prove (e) it suffices, by Chebychev's inequality, to show that

$$\text{(f)} \qquad \sum_{i=1}^n E[(T_{ni}^\eta)^2] \to \sigma^2$$

and

$$\text{(g)} \qquad \lim_{\eta \to 0} \limsup_{n\to\infty} \sum_{i=1}^n Var[(T_{ni}^\eta)^2] = 0 \, .$$

But by virtue of (d), (f) is equivalent to

$$\text{(h)} \qquad \sum_{i=1}^n [E(T_{ni}^\eta)]^2 \to 0 \, .$$

We first prove (h) and hence (f): if $\eta > 2$, then $T_{ni}^\eta \le T_{ni}$ since $T_{ni} \ge -2$ a.s. by definition of T_{ni}. Therefore

$$E(T_{ni}^\eta \le ET_{ni} = 2E \left\{ \frac{g_{ni}^{1/2}}{f_{ni}^{1/2}}(X_i) \right\} - 2 \le 0$$

by Jensen's inequality. Thus for $\eta > 2$

$$\sum_{i=1}^n (-ET_{ni}^\eta)^2 \le \max_{1 \le i \le n} (-ET_{ni}^\eta) \sum_{i=1}^n (-ET_{ni}^\eta) \to 0$$

since

$$\sum_{i=1}^n (-ET_{ni}^\eta) \to \frac{1}{4}\sigma^2 \qquad \text{by } \ (c)$$

and

$$\max_{1\le i\le n} (-ET_{ni}^\eta) \to 0 \qquad \text{by the UAN condition (14)} \, .$$

Now note that if (h) holds for any $\eta > 2$, it holds for all $\eta > 0$: since

$$\sum_{i=1}^n E[(T_{ni}^\eta)^2] \le \sum_{i=1}^n E[(T_{ni}^\gamma)^2] \qquad \text{for all } \ \eta < \gamma$$

and, by (d) both $\sum_{i=1}^n Var[T_{ni}^\eta] \to \sigma^2$ and $\sum_{i=1}^n Var[T_{ni}^\gamma] \to \sigma^2$, it follows that

$$\begin{aligned}
\sum_{i=1}^n [E(T_{ni}^\eta)]^2 &= \sum_{i=1}^n \{E[(T_{ni}^\eta)^2] - Var(T_{ni}^\eta)\} \\
&\le \sum_{i=1}^n \{E[(T_{ni}^\gamma)^2] - Var(T_{ni}^\gamma) + Var(T_{ni}^\gamma) - Var(T_{ni}^\eta)\} \\
&= \sum_{i=1}^n [E[(T_{ni}^\gamma)]^2 + \sum_{i=1}^n Var(T_{ni}^\gamma) - \sum_{i=1}^n Var(T_{ni}^\eta) \\
&\to 0 + \sigma^2 - \sigma^2 = 0\,,
\end{aligned}$$

completing the proof of (h) and hence (f).

To prove (g), note that

$$\sum_{i=1}^n Var[(T_{ni}^\eta)^2] \le \sum_{i=1}^n E[(T_{ni}^\eta)^4] \le \eta^2 \sum_{i=1}^n E[(T_{ni}^\eta)^2]\,.$$

Then, by (f)

$$\limsup_{n\to\infty} \sum_{i=1}^n Var[(T_{ni}^\eta)^2] \le \eta^2\sigma^2\,,$$

and hence (g) holds. $\Box$

Proof. (corollary 3). Let

$$T_{ni} \equiv 2\left\{\frac{f_n^{1/2}(X_i)}{f^{1/2}(X_i)} - 1\right\}, \qquad i = 1, \ldots, n\,, \text{ and } n \ge 1\,.$$

Note that (all expectations and variances being calculated under P with density f)

$$E(T_{ni}) = 2\left\{\int f_n^{1/2} f^{1/2} d\mu - 1\right\} = -2H^2(f_n, f), \qquad \text{for } i = 1, \ldots, n$$

where

$$H^2(f_n, f) \equiv \frac{1}{2}\int (f_n^{1/2} - f^{1/2})^2 d\mu = \frac{1}{2}\|f_n^{1/2} - f^{1/2}\|_2^2$$

and

$$\begin{aligned}
Var(T_{ni}) &= 4\int (f_n^{1/2} - f^{1/2})^2 d\mu - [E(T_{ni})]^2 \\
&= 8H^2(f_n, f) - 4H^4(f_n, f)\,.
\end{aligned}$$

Therefore, since the hypothesis $\|\sqrt{n}(f_n^{1/2} - f^{1/2}) - \delta\|_2 \to 0$ implies $n\|f_n^{1/2} - f^{1/2}\|_2^2 \to \|\delta\|_2^2$ and $\|f_n^{1/2} - f^{1/2}\| \to 0$, the random variable W_n of lemma 2 has

(a) $E(W_n) = -2nH^2(f_n, f) \to \|\delta\|_2^2$

and

$$Var(W_n) = 8nH^2(f_n, f) - 4nH^2(f_n, f)H^2(f_n, f) \to 4\|\delta\|_2^2\,.$$

Note that since

$$\begin{aligned}
\epsilon P\left(\left|\frac{f_n}{f}(X_i)-1\right|\geq\epsilon\right) &\leq E\left|\frac{f_n}{f}(X_i)-1\right| \\
&= E\left(\left|\frac{f_n^{1/2}}{f^{1/2}}(X_i)-1\right|\left|\frac{f_n^{1/2}}{f^{1/2}}(X_i)-1\right|\right) \\
&\leq \|f_n^{1/2}-1\|_2\|f_n^{1/2}+1\|_2\to 0
\end{aligned}$$

uniformly in $1\leq i\leq n$ as $n\to$, the UAN condition (14) holds. Furthermore,

$$\begin{aligned}
&Var\left\{W_n-\frac{2}{\sqrt{n}}\sum_{i=1}^n\frac{\delta}{f^{1/2}}(X_i)\right\} \\
&= 4nVar\left\{\frac{f_n^{1/2}}{f^{1/2}}(X_i)-1-n^{-1/2}\frac{\delta}{f^{1/2}}(X_1)\right\} \\
&= 4n\|(f_n^{1/2}-f^{1/2}-n^{-1/2}\delta\|_2^2-4nH^2(f_n,f)H^2(f_n,f) \\
&= 4\|\sqrt{n}(f_n^{1/2}-f^{1/2})-\delta\|_2^2-4nH^4(f_n,f) \\
\text{(b)}\qquad &\to 0
\end{aligned}$$

and hence

$$\begin{aligned}
&E\left\{W_n-\|\delta\|_2^2-\frac{2}{\sqrt{n}}\sum_{i=1}^n\frac{\delta}{f^{1/2}}(X_i)+2\|\delta\|_2^2\right\}^2 \\
&= Var\left\{W_n-\frac{2}{\sqrt{n}}\sum_{i=1}^n\frac{\delta}{f^{1/2}}(X_i)\right\} \\
&\qquad+\left\{E\left(W_n-\frac{2}{\sqrt{n}}\sum_{i=1}^n\frac{\delta}{f^{1/2}}(X_i)+\|\delta\|_2^2\right)\right\}^2 \\
&= o(1)+(E(W_n)+\|\delta\|_2^2)^2 \qquad \text{by } (b) \\
&\to 0+0=0 \qquad \text{by } (a)\,.
\end{aligned}$$

Thus

$$\text{(c)}\qquad W_n-\|\delta\|_2^2-\left\{\frac{2}{\sqrt{n}}\sum_{i=1}^n\frac{\delta}{f^{1/2}}(X_i)-2\|\delta\|_2^2\right\}=o_p(1)\,.$$

Since

$$\mathcal{L}\left(\frac{2}{\sqrt{n}}\sum_{i=1}^N\frac{\delta}{f^{1/2}}(X_i)-2\|\delta\|_2^2\middle|P\right)\to N(-(1/2)\|2\delta\|_2^2,\|2\delta\|_2^2)$$

it follows that

$$\mathcal{L}(W_n-\|\delta\|_2^2|P)\to N(-(1/2)\|2\delta\|^2,\|2\delta\|_2^2)\,,$$

and hence, by lemma 2 that

$$\log L_n=W_n-\|\delta\|_2^2+o_p(1)=\frac{2}{\sqrt{n}}\sum_{i=1}^n\frac{\delta}{f^{1/2}}(X_i)-2\|\delta\|_2^2+o_p(1)\,.$$

□

Proof. **(lemma 3.4, Le Cam's third lemma)**. Since $\mathcal{L}(\log L_n|P_n) \to \mathcal{L}(\log L) = N(-\sigma^2/2, \sigma^2) = \mathcal{L}(\sigma Z - \sigma^2/2)$ where $Z \sim N(0,1)$, it follows that $E(L) = E\exp(\sigma Z - \sigma^2/2) = 1$, and thus $\{Q_n\} \lhd \{P_n\}$ by Le Cam's first lemma. Hence by lemma 3.2 (Le Cam's fourth lemma), L_n is uniformly integrable and $Q_n(p_n = 0) \to 0$ as $n \to \infty$.

Now let $f : R^2 \to R$ be bounded and continuous. Then

$$
\begin{aligned}
& E_{Q_n} f(T_n, \log L_n) &=& \; E_{Q_n} f(T_n, \log L_n)\{1_{[p_n>0]} + 1_{[p_n=0]}\} \\
\text{(a)} \quad & &=& \; E_{P_n} f(T_n, \log L_n) L_n + E_{Q_n} f(T_n, \log L_n) 1_{[p_n=0]} \\
\text{(b)} \quad & &\to& \; E[f(T, \log L) L] \\
\text{(c)} \quad & &=& \; E f(T + c, \log L + \sigma^2)
\end{aligned}
$$

where (b) holds since $f(T_n, \log L_n)L_n$ is uniformly integrable by uniform integrability of L_n and boundedness of f, and since the second term in (a) is bounded by $\|f\|_\infty Q_n(p_n = 0) \to 0$. It remains only to establish (c).

To verify (c), note that (28) implies that

$$
\text{(d)} \qquad \mathcal{L}(T|\log L) = \mathcal{L}(\frac{c}{\sigma^2}(\log L + \frac{1}{2}\sigma^2) + \tilde{Z})
$$

where $\mathcal{L}(\tilde{Z}) = N(\mu, \sigma^2(1-\rho^2))$, $\rho = c/(\sigma\tau)$, is independent of $\log L$, and hence

$$
\begin{aligned}
\mathcal{L}(T + c|\log L) &= \mathcal{L}(c + \frac{c}{\sigma^2}(\log L + \frac{1}{2}\sigma^2) + \tilde{Z}) \\
&= \mathcal{L}(\frac{c}{\sigma^2}(\log L + \sigma^2 + \frac{1}{2}\sigma^2) + \tilde{Z})\,.
\end{aligned}
$$

Furthermore,

$$
L = \exp(\log L) = \frac{\text{density of } N(\sigma^2/2, \sigma^2)}{\text{density of } N(-\sigma^2/2, \sigma^2)} \qquad \text{at} \quad \log L\,.
$$

Therefore

$$
\begin{aligned}
E f(T, \log L) L &= E\{E(f(T, \log L) e^{\log L}|\log L)\} \\
&= E\{e^{\log L} E(f(T, \log L)|\log L)\} \\
&= E\left\{e^{\log L} E\left(f(\frac{c}{\sigma^2}(\log L + \sigma^2/2) + \tilde{Z}, \log L)|\log L\right)\right\} \qquad \text{by } (d) \\
&= E\{E\left(f(c + \frac{c}{\sigma^2}(\log L + \sigma^2/2) + \tilde{Z}, \log L + \sigma^2)|\log L\right)\} \\
&= E f(T + c, \log L + \sigma^2) \qquad \text{by } (e)\,,
\end{aligned}
$$

which completes the proof of (c). Hence (29) holds. □

4 The Hájek - Le Cam convolution and LAM theorems

Now we give statements of several convolution and local asymptotic minimax theorems. The key hypotheses involved in virtually all the different formulations of these theorems are as follows:

A. *Local Asymptotic Normality* (LAN) of the local likelihood ratios of the model. A sufficient condition for this is differentiability of the model in an appropriate sense; recall corollaries 3 and 4.

B. For the convolution theorems we will also hypothesize *regularity* of the estimators: the only estimators considered will be those for which the local limiting distributions do not depend on the direction or (magnitude) of the approach of the local parameter point to the fixed point under consideration.

C. *Pathwise differentability* of the parameter being estimated as a function of the underling $P \in \mathcal{P}$ metrized by the Hellinger metric. This amounts to Hadamard differentiability along the model $\mathcal{P}$.

Our goal in this section will be to explain the basic hypotheses require in different settings, and to discuss several useful refinements and extensions of the basic theorems. For complete proofs we refer the reader to the original articles by Hájek (1970), (1972), Le Cam (1972), Ibragimov and Has'minskii (1981), van der Vaart (1988), Millar (1983), Le Cam (1986), and Bickel, Klaassen, Ritov and Wellner (1993).

Convolution and LAM theorems for finite-dimensional parameter spaces

Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^k$ is a Hellinger differentiable parametric model. Set $l(x;\theta) \equiv \log p(x;\theta)$, and let

$$l_n(\theta) = \sum_{i=1}^n l(X_i;\theta)$$

denote the log-likelihood of $X_1, \ldots, X_n$, a sample from $P_{\theta_0} \equiv P_0 \in \mathcal{P}$. Then by corollary 4 of Le Cam's second lemma we know that

$$(1) \qquad l_n(\theta_0 + n^{-1/2}t) - l_n(\theta_0) = t^T S_n(\theta_0) - \frac{1}{2} t^T I(\theta_0) t + o_{P_0}(1)$$

where

$$S_n(\theta_0) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\mathbf{l}}(X_i;\theta_0)$$

is the score for θ at θ_0 (based on the entire sample $X_1, \ldots, X_n$) and $I(\theta_0)$ is the Fisher information matrix defined in section 2. It follows that

$$(2) \qquad l_n(\theta_0 + n^{-1/2}t) - l_n(\theta_0) \to_d N\left(-\frac{1}{2} t^T I(\theta_0) t, t^T I(\theta_0) t\right)$$

under P_0. This is sometimes called the *Local Asymptotic Normality*, or LAN condition. It is one key ingredient of the Hájek convolution theorem. The second key ingredient is the following definition of regularity of an estimator sequence T_n.

Definition 4.1 $T = \{T_n\}$ is a *locally regular estimator* of θ at $\theta = \theta_0$ if, for every sequence $\{\theta_n\} \subset \Theta$ with $\sqrt{n}(\theta_n - \theta_0) \to t \in R^k$, under P_{θ_n}

$$\sqrt{n}(T_n - \theta_n) \to \mathbb{Z} \qquad \text{as } \ n \to \infty$$

where the distribution of $\mathbb{Z}$ depends on θ_0 but not on t. Thus the limit distribution of $\sqrt{n}(T_n - \theta_n)$ does not depend on the direction of approach t of θ_n to θ_0.

With these two basic ingredients, we can state a simplified version of Hájek's (1970) convolution theorem.

Theorem 4.1 Suppose that (2) holds with $I(\theta_0)$ nonsingular and that $\{T_n\}$ is a regular estimator of θ at θ_0. Then

$$(3) \qquad \mathbb{Z} \stackrel{d}{=} \mathbb{Z}_0 + \Delta_0$$

where $\mathbb{Z}_0 \sim N(0, I^{-1}(\theta_0))$ is independent of Δ_0.

Hájek (1970) proved a somewhat more general theorem based on just the LAN hypothesis (2) using a method based on "Bayesian considerations". This method of proof is developed further in van der Vaart (1989). A different proof using characteristic functions due to Peter Bickel is given in Roussas (1972) and also in Bickel, Klaassen, Ritov and Wellner (1993). This latter type of proof was exploited and developed by R. Beran (1977a, 1977b) in more general settings.

In words, theorem 4.1 says that the limiting distribution of any regular estimator T_n of θ must be at least as "spread out" as the $N(0, I^{-1}(\theta_0))$ distribution of $\mathbb{Z}_0$. Thus an *efficient estimator* is a regular estimator for which the limiting distribution is exactly equal to $\mathbb{Z}_0$. Another way to say this is in terms of the following asymptotic optimality theorem.

Corollary 1 (Hájek, 1970). Suppose that $\{T_n\}$ is a locally regular estimator of θ at θ_0 and that $l : R^k \to R^+$ is bowl-shaped: i.e.

$$(i) \qquad l(x) = l(-x)\,,$$
$$(ii) \qquad \{x : l(x) \le c\} \qquad \text{is convex for every } \ c \ge 0\,.$$

Then

$$\liminf_{n\to\infty} E_{\theta_0} l(\sqrt{n}(T_n - \theta_0)) \ge El(\mathbb{Z}_0)$$

where $\mathbb{Z}_0 \sim N(0, I^{-1}(\theta_0))$.

If a supremum over θ in a local neighborhood of θ_0 is added to the left side of (4), then the same type of statement holds for an arbitrary (not necessarily regular) estimator T_n of θ. This is the Hájek - Le Cam asymptotic minimax theorem due to Hájek (1971), and, in a more abstract form to Le Cam (1971).

Theorem 4.2 (Hájek, 1971). Suppose that (2) holds, that T_n is any estimator of θ, and that l is bowl-shaped. Then

$$(4) \qquad \lim_{\delta\to\infty} \liminf_{n\to\infty} \sup_{\theta:\sqrt{n}|\theta-\theta_0|\le\delta} E_\theta l(\sqrt{n}(T_n - \theta)) \ge El(\mathbb{Z}_0)\,.$$

5 A Basic Inequality

First we need two lemmas.

Lemma 5.1 Let P, Q be two probability measures on a measurable space $(\mathbb{X}, \mathcal{A})$ with densities p, q with respect to a $\sigma-$finite dominating measure μ. Then

$$(1 - H^2(P,Q))^2 \le 1 - \left\{1 - \int (p \wedge q)\, d\mu\right\}^2 \le 2 \int (p \wedge q)\, d\mu\,.$$

Proof. The second inequality is trivial. To prove the first inequality, note that by Exercise 2.1.6

$$\begin{aligned}
&(1 - H^2(P,Q))^2 + (1 - \int p \wedge q d\mu)^2 \\
&= \left(\int \sqrt{pq} d\mu\right)^2 + \left(\frac{1}{2}\int |p - q| d\mu\right)^2 \\
&= \left(\int \sqrt{pq} d\mu\right)^2 + \frac{1}{4}\left(\int |\sqrt{p} - \sqrt{q}||\sqrt{p} + \sqrt{q}| d\mu\right)^2 \\
&\le \left(\int \sqrt{pq} d\mu\right)^2 + \frac{1}{4}\int (\sqrt{p} - \sqrt{q})^2 d\mu \int (\sqrt{p} + \sqrt{q})^2 d\mu \\
&= 1\,.
\end{aligned}$$

□

Lemma 5.2 If P and Q are two probability measures on a measurable space $(\mathbb{X}, \mathcal{A})$ with densities p and q with respect to a $\sigma-$finite dominating measure μ and P^n and Q^n denote the corresponding product measures on $(\mathbb{X}^n, \mathcal{A}_n)$ (of $X_1, \ldots, X_n$ i.i.d. as P or Q respectively), then

$$\rho(P^n, Q^n) = \rho(P,Q)^n\,. \qquad (1)$$

Proof. Note that

$$\begin{aligned}
\rho(P^n, Q^n) &= \int \cdots \int \sqrt{\prod_{i=1}^n p(x_i) \prod_{i=1}^n q(x_i)}\, d\mu(x_1) \cdots d\mu(x_n) \\
&= \int \cdots \int \sqrt{p(x_1)q(x_1) \cdots p(x_n)q(x_n)} d\mu(x_1) \cdots d\mu(x_n) \\
&= \int \sqrt{p(x_1)q(x_1)}\, d\mu(x_1) \cdots \int \sqrt{p(x_n)q(x_n)} d\mu(x_n) \\
&= \rho(P,Q) \cdots \rho(P,Q) = \rho(P,Q)^n\,.
\end{aligned}$$

□

Remark 5.1 Note that (1) implies that

$$H^2(P^n, Q^n) = 1 - \rho(P^n, Q^n) = 1 - \rho(P, Q)^n = 1 - (1 - H^2(P, Q))^n$$

by using exercise 2.1.5 (chapter 2, page 10) twice.

With these two lemmas in hand we can prove our basic inequality.

Proposition 5.1 Let $\mathcal{P}$ be a set of probability measures on a measurable space $(\mathbb{X}, \mathcal{A})$, and let ν be a real-valued function defined on $\mathcal{P}$. Moreover, let $l : [0, \infty) \to [0, \infty)$ be an increasing convex loss function with $l(0) = 0$. Then, for any $P_1, P_2 \in \mathcal{P}$ such that $H(P_1, P_2) < 1$ and with

$$E_{n,i} f(X_1, \ldots, X_n) = E_{n,i} f(X) = \int f(x) dP_i^n(x) \equiv \int f(x_1, \ldots, x_n) dP_i(x_1) \cdots dP_i(x_n)\,,$$

for $i = 1, 2$, it follows that

$$\begin{aligned} (2) \qquad & \inf_{T_n} \max \{E_{n,1} l(|T_n - \nu(P_1)|),\ E_{n,2} l(|T_n - \nu(P_2)|)\} \\ & \geq\ l\left(\frac{1}{4}|\nu(P_1) - \nu(P_2)|\{1 - H^2(P_1, P_2)\}^{2n}\right)\,. \end{aligned}$$

Proof. By Jensen's inequality

$$E_{n,i} l(|T_n - \nu(P_i)|) \geq l(E_{n,i}|T_n - \nu(P_i)|)\,, \qquad i = 1, 2\,,$$

and hence the left side of (2) is bounded below by

$$l\left(\inf_{T_n} \max\{E_{n,1}|T_n - \nu(P_1)|\,,\ E_{n,2}|T_n - \nu(P_2)|\right)\,.$$

Thus it suffices to prove the proposition for $l(x) = x$. Let $p_1 \equiv dP_1/(d(P_1+P_2)$, $p_2 = dP_2/d(P_1+P_2)$, and $\mu = P_1 + P_2$ (or let p_i be the density of P_i with respect to some other convenient dominating measure μ, $i = 1, 2$). Now

$$\begin{aligned} & \max \{E_{n,1}|T_n - \nu(P_1)|,\ E_{n,2}|T_n - \nu(P_2)|\} \\ & \geq\ \frac{1}{2}\{E_{n,1}|T_n - \nu(P_1)| + E_{n,2}|T_n - \nu(P_2)|\} \\ & =\ \frac{1}{2}\left\{\int |T_n(x) - \nu(P_1)| \prod_{i=1}^n p_1(x_i) d\mu(x_1) \cdots d\mu(x_n)\right. \\ & \qquad \left. +\ \int |T_n(x) - \nu(P_2)| \prod_{i=1}^n p_2(x_i) d\mu(x_1) \cdots d\mu(x_n)\right\} \\ & \geq\ \frac{1}{2}\left\{\int [|T_n(x) - \nu(P_1)| + |T_n(x) - \nu(P_2)|] \prod_{i=1}^n p_1(x_i) \wedge \prod_{i=1}^n p_2(x_i) d\mu(x_1) \cdots d\mu(x_n)\right\} \\ & \geq\ \frac{1}{2}|\nu(P_1) - \nu(P_2)| \int \prod_{i=1}^n p_1(x_i) \wedge \prod_{i=1}^n p_2(x_i) d\mu(x_1) \cdots d\mu(x_n) \\ & \geq\ \frac{1}{4}|\nu(P_1) - \nu(P_2)|\{1 - H^2(P_1^n, P_2^n)\}^2 \qquad \text{by Lemma 5.1} \\ & =\ \frac{1}{4}|\nu(P_1) - \nu(P_2)|\{1 - H^2(P_1, P_2)\}^{2n} \qquad \text{by Lemma 5.2}\ . \end{aligned}$$

□

Example 5.1 (Regular parametric model). Suppose that $\mathcal{P} = \{P_\theta : \ \theta \in \Theta \subset R^k\}$ with $P_\theta << \mu$ for all θ so that $p_\theta = dP_\theta/d\mu$ exists for all $\theta \in \Theta$. Suppose that p_θ is differentiable at $\theta_0 \in \Theta$ in the following sense: there is a function $\dot{\mathbf{l}}_\theta$ such that

$$(3) \qquad \int \{\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^T \dot{\mathbf{l}}_\theta \sqrt{p_{\theta_0}}\}^2 \, d\mu = o(|\theta - \theta_0|^2) \,.$$

Let $\theta_n = \theta_0 + n^{-1/2}h$ so that $\sqrt{n}(\theta_n - \theta_0) = h$. Note that (3) implies that

$$\begin{aligned} nH^2(P_{\theta_n}, P_{\theta_0}) &= \frac{n}{2}\int [\sqrt{p_{\theta_n}} - \sqrt{p_{\theta_0}}]^2 \, d\mu = \frac{1}{2}\int [\sqrt{n}(\sqrt{p_{\theta_n}} - \sqrt{p_{\theta_0}})]^2 \, d\mu \\ &\to \frac{1}{8}\int h^T \dot{\mathbf{l}}_\theta \dot{\mathbf{l}}_\theta^T h \, p_{\theta_0} d\mu \\ &= \frac{1}{8} h^T E_{\theta_0}\{\dot{\mathbf{l}}_\theta(X) \dot{\mathbf{l}}_\theta^T(X)\} h = \frac{1}{8} h^T I(\theta_0) h \,. \end{aligned}$$

This implies that

$$\left(1 - H^2(P_{\theta_n}, P_{\theta_0})\right)^{2n} = \left(1 - \frac{nH^2(P_{\theta_n}, P_{\theta_0})}{n}\right)^{2n} \to \exp(-(1/4) h^T I(\theta_0) h) \,.$$

Hence if we take $\nu(P_\theta) = c^T\theta$, we have

$$|\nu(P_{\theta_n}) - \nu(P_{\theta_0})| = n^{-1/2} c^T h \,,$$

and it follows from Proposition 5.1 that for any convex increasing function l we have

$$\begin{aligned} (4) \qquad & \inf_{T_n} \max\{E_{\theta_n} l(|T_n - \nu(P_{\theta_n})|), \, E_{\theta_0} l(|T_n - \nu(P_{\theta_0})|)\} \\ & \geq \ l\left(\frac{1}{4}|\nu(P_{\theta_n}) - \nu(P_{\theta_0})|\{1 - H^2(P_{\theta_n}, P_{\theta_0})\}^{2n}\right) . \\ (5) \qquad & = \ l\left(\frac{1}{4} n^{-1/2}|c^T h|\{1 - \frac{nH^2(P_{\theta_n}, P_{\theta_0})}{n}\}^{2n}\right) . \end{aligned}$$

For example, with $l(x) = x$, this yields

$$\begin{aligned} & n^{1/2} \inf_{T_n} \max\{E_{\theta_n}|T_n - \nu(P_{\theta_n})|, \, E_{\theta_0}|T_n - \nu(P_{\theta_0})|\} \\ & \geq \ \frac{1}{4}|c^T h| \left\{1 - \frac{nH^2(P_{\theta_n}, P_{\theta_0})}{n}\right\}^{2n} \to \frac{1}{4}|c^T h| \, \exp(-h^T I(\theta_0) h/4) \,. \end{aligned}$$

By choosing $h = aI^{-1}(\theta_0)c$ this bound becomes

$$\frac{1}{4} a |c^T I^{-1}(\theta_0) c| \, \exp(-a^2 c^T I^{-1}(\theta_0) c/4) = |c^T I^{-1}(\theta_0) c|^{1/2} \left\{e^{-1/4}/4\right\}$$

by taking $a = \{c^T I^{-1}(\theta_0) c\}^{-1/2}$.

With $l(x) = x^2$ we obtain

$$\begin{aligned} & \inf_{T_n} \max\left\{E_{\theta_n}\{n|T_n - \nu(P_{\theta_n})|^2\}, \, E_{\theta_0}\{n|T_n - \nu(P_{\theta_0})|^2\}\right\} \\ & \geq \ \frac{1}{16}|c^T h|^2 \left\{1 - \frac{nH^2(P_{\theta_n}, P_{\theta_0})}{n}\right\}^{4n} \to \frac{1}{16}|c^T h|^2 \, \exp(-h^T I(\theta_0) h/2) \,. \end{aligned}$$

By choosing $h = aI^{-1}(\theta_0)c$ this bound becomes

$$\frac{1}{16}a^2(c^TI^{-1}(\theta_0)c)^2 \exp(-a^2c^TI^{-1}(\theta_0)c/2) = c^TI^{-1}(\theta_0)c\left\{e^{-1/2}/16\right\}$$

by taking $a^2 = \{c^TI^{-1}(\theta_0)c\}^{-1}$. Thus we conclude that for the choice $h = I^{-1}(\theta_0)c/\{c^TI^{-1}(\theta_0)c\}^{1/2}$ we have

$$\begin{aligned}
&\liminf_{n\to\infty}\inf_{T_n}\max\left\{E_{\theta_n}\{n|T_n-\nu(P_{\theta_n})|^2\},\, E_{\theta_0}\{n|T_n-\nu(P_{\theta_0})|^2\}\right\}\\
&\quad\geq\ c^TI^{-1}(\theta_0)c\left\{\frac{1}{16}\exp(-1/2)\right\} = E[N(0,c^TI^{-1}(\theta_0)c)]^2\left\{\frac{1}{16}\exp(-1/2)\right\}.
\end{aligned}$$

Example 5.2 (Uniform$(0,\theta)$). Suppose that $X_1,\ldots,X_n$ are i.i.d. Uniform$(0,\theta)$ with densities $p_\theta(x) = \theta^{-1}1_{[0,\theta]}(x)$ for $\theta\in(0,\infty)$. Fix $\theta_0>0$ and let $\theta_n = \theta_0 + cn^{-1}$. Then

$$\begin{aligned}
\rho(P_{\theta_0},P_{\theta_n}) &= \int_0^{\theta_0\wedge\theta_n}\frac{1}{\sqrt{\theta_0\theta_n}}dx\\
&= \frac{\theta_0\wedge\theta_n}{\sqrt{\theta_0\theta_n}}\\
&= \begin{cases}\sqrt{\theta_0/\theta_n}, & \text{if } \theta_0\leq\theta_n\\ \sqrt{\theta_n/\theta_0}, & \text{if } \theta_0\geq\theta_n\end{cases}\\
&= \begin{cases}\sqrt{1/(1+(c/\theta_0)/n)}, & \text{if } \theta_0\leq\theta_n\\ \sqrt{1+(c/\theta_0)/n}, & \text{if } \theta_0\geq\theta_n.\end{cases}
\end{aligned}$$

Thus

$$\begin{aligned}
(1-H^2(P_{\theta_0},P_{\theta_n}))^{2n} &= \rho(P_{\theta_0},P_{\theta_n})^{2n}\\
&= \begin{cases}1/(1+(c/\theta_0)/n)^n, & \text{if } c\geq 0\\ (1+(c/\theta_0)/n)^n, & \text{if } c\leq 0.\end{cases}\\
&\to \begin{cases}e^{-c/\theta_0}, & \text{if } c\geq 0\\ e^{-|c|/\theta_0}, & \text{if } c\leq 0\end{cases}\\
&= \exp(-|c|/\theta_0).
\end{aligned}$$

Thus if $\nu(P_\theta)=\theta$ and $l(x)=x$, it follows that

$$\begin{aligned}
&\inf_{T_n}\max\left\{E_n\{n|T_n-\nu(P_{\theta_n})|\},\, E_0\{n|T_n-\nu(P_{\theta_0})|\}\right\}\\
&\quad\geq\ \frac{1}{4}n\frac{|c|}{n}\rho(P_{\theta_0},P_{\theta_n})^{2n}\\
&\quad\to\ \frac{1}{4}|c|\exp(-|c|/\theta_0) = \frac{1}{4}\theta_0(|c|/\theta_0)\exp(-|c|/\theta_0).
\end{aligned}$$

The right side is maximized by the choice $|c|=\theta_0$, and for this choice we conclude that

$$\liminf_{n\to\infty}\inf_{T_n}\max\left\{E_n\{n|T_n-\nu(P_{\theta_n})|\},\, E_0\{n|T_n-\nu(P_{\theta_0})|\}\right\}\geq\frac{e^{-1}}{4}\theta_0.$$

Note that the particular estimator $T_n = \frac{n+1}{n} \max_{1\le i\le n} X_i$ (which is the unbiased modification of the MLE) satisfies

$$
\begin{aligned}
E_\theta\{n|T_n-\theta|\} &= E_\theta n\Big|\frac{n+1}{n}X_{(n)}-\theta\Big| \\
&= (n+1)E_\theta\Big|X_{(n)}-\frac{n}{n+1}\theta\Big| \\
&= (n+1)\int_0^\theta \Big|x-\frac{n}{n+1}\theta\Big| n(x/\theta)^{n-1}dx/\theta \\
&= \theta n(n+1)\int_0^1 \Big|u-\frac{n}{n+1}\Big|u^{n-1}du \\
&= \theta n(n+1)\left\{\int_0^{n/(n+1)}\left(\frac{n}{n+1}-u\right)u^{n-1}du+\int_{n/(n+1)}^1\left(u-\frac{n}{n+1}\right)u^{n-1}du\right\} \\
&= \theta n(n+1)\frac{2}{(n+1)^2}\left(\frac{n}{n+1}\right)^n \to 2e^{-1}\theta
\end{aligned}
$$

since $P_\theta(X_{(n)}\le x) = (x/\theta)^n$

Example 5.3 (Monotone densities on R^+). Suppose that

$$\mathcal{P} = \{P \text{ on } R^+ :\ dP/d\lambda = p \text{ is monotone nonincreasing}\}.$$

Suppose that we want to estimate $\nu(P) = p(x_0)$ for a fixed $x_0 \in (0,\infty)$ on the basis of a sample $X_1,\ldots,X_n$ from $P_0 \in \mathcal{P}$. Let p_0 be the density corresponding to P_0, and suppose that $p_0'(x_0) < 0$. To apply Proposition 5.1 we need to construct some density p_n that is "near" p_0 in the sense that

$$nH^2(p_n,p_0) \to\ A$$

for some constant A, and

$$|\nu(P_n)-\nu(P_0)| = b_n^{-1}$$

where $b_n\to\infty$. Hence we will try the following choice of p_n. For $c>0$, define

$$
p_n(x) = \begin{cases} p_0(x) & \text{if } \ x\le x_0-cn^{-1/3} \ \text{ or } \ x > x_0+cn^{-1/3}, \\ p_0(x_0-cn^{-1/3}) & \text{if } \ x_0-cn^{-1/3} < x \le x_0, \\ p_0(x_0+cn^{-1/3}) & \text{if } \ x_0 < x\le x_0+cn^{-1/3}. \end{cases}
$$

It is easy to see that

(6) $$n^{1/3}|\nu(P_n)-\nu(P_0)| = |n^{1/3}(p_0(x_0-cn^{-1/3})-p_0(x_0))| \to |p_0'(x_0)|c$$

On the other hand we calculate

$$
\begin{aligned}
H^2(p_n,p_0) &= \frac{1}{2}\int_0^\infty [\sqrt{p_n(x)}-\sqrt{p_0(x)}]^2\,dx \\
&= \frac{1}{2}\int_0^\infty \frac{[\sqrt{p_n(x)}-\sqrt{p_0(x)}]^2[\sqrt{p_n(x)}+\sqrt{p_0(x)}]^2}{[\sqrt{p_n(x)}+\sqrt{p_0(x)}]^2}\,dx \\
&= \frac{1}{2}\int_0^\infty \frac{[p_n(x)-p_0(x)]^2}{[\sqrt{p_n(x)}+\sqrt{p_0(x)}]^2}\,dx
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\int_{x_0-cn^{-1/3}}^{x_0} \frac{[p_0(x_0-cn^{-1/3})-p_0(x)]^2}{[\sqrt{p_n(x)}+\sqrt{p_0(x)}]^2}\,dx \\
&\qquad + \frac{1}{2}\int_{x_0}^{x_0+cn^{-1/3}} \frac{[p_0(x_0+cn^{-1/3})-p_0(x)]^2}{[\sqrt{p_n(x)}+\sqrt{p_0(x)}]^2} \\
&= \frac{1}{2}\int_{x_0-cn^{-1/3}}^{x_0} \frac{[p_0'(x_n^*)(x_0-cn^{-1/3}-x)]^2}{[\sqrt{p_n(x)}+\sqrt{p_0(x)}]^2}\,dx \\
&\qquad + \frac{1}{2}\int_{x_0}^{x_0+cn^{-1/3}} \frac{[p_0'(x_n^{**})(x_0+cn^{-1/3}-x)]^2}{[\sqrt{p_n(x)}+\sqrt{p_0(x)}]^2} \\
&\sim \frac{1}{2}\frac{p'(x_0)^2}{(2\sqrt{p_0(x_0)})^2}\int_{x_0-cn^{-1/3}}^{x_0}(x_0-cn^{-1/3}-x)^2dx \\
&\qquad + \frac{1}{2}\frac{p'(x_0)^2}{(2\sqrt{p_0(x_0)})^2}\int_{x_0}^{x_0+cn^{-1/3}}(x_0+cn^{-1/3}-x)^2dx \\
&= \frac{p_0'(x_0)^2}{4p_0(x_0)}\frac{c^3}{3n}\,.
\end{aligned}
$$

Now we can combine these two pieces with Proposition 5.1 to find that, for any estimator T_n of $\nu(P) = p(x_0)$ and the loss function $l(x) = |x|$ we have

$$
\begin{aligned}
&\inf_{T_n}\max\left\{E_n n^{1/3}|T_n-\nu(P_n)|,\ E_0 n^{1/3}|T_n-\nu(P_0)|\right\} \\
&\geq \frac{1}{4}|n^{1/3}(\nu(P_n)-\nu(P_0))|\left\{1-\frac{nH^2(P_n,P_0)}{n}\right\}^{2n} \\
&= \frac{1}{4}|n^{1/3}(p_0(x_0-cn^{-1/3})-p_0(x_0))|\left\{1-\frac{nH^2(P_n,P_0)}{n}\right\}^{2n} \\
&\to \frac{1}{4}|p_0'(x_0)|c\exp\left(-2\frac{p_0'(x_0)^2}{12p_0(x_0)}c^3\right) = \frac{1}{4}|p_0'(x_0)|c\exp\left(-\frac{p_0'(x_0)^2}{6p_0(x_0)}c^3\right)
\end{aligned}
$$

We can choose c to maximize the quantity on the right side. It is easily seen that the maximum is achieved when

$$c = c_0 \equiv \left(\frac{2p_0(x_0)}{p_0'(x_0)^2}\right)^{1/3}.$$

This yields

$$\liminf_{n\to\infty}\inf_{T_n}\max\left\{E_n n^{1/3}|T_n-\nu(P_n)|,\ E_0 n^{1/3}|T_n-\nu(P_0)|\right\} \geq \frac{e^{-1/3}}{4}\left(2|p_0'(x_0)|p_0(x_0)\right)^{1/3}.$$

This bound has the appropriate structure in the sense that the (nonparametric) MLE of p, $\widehat{p}_n$ converges at rate $n^{-1/3}$ and the same constant is involved in its limiting distribution:

$$n^{1/3}(\widehat{p}_n(x_0)-p_0(x_0)) \to_d (|p_0'(x)|p_0(x)/2)^{1/3}(2\mathbb{Z})$$

where $\mathbb{Z} = \text{argmin}\{W(t)+t^2\}$ and W is a standard Brownian motion process started at 0, as has been shown by Prakasa Rao (1969) and Groeneboom (1985).

Example 5.4 (Interval censoring or "current status" data). Suppose that $T, T_1, \ldots, T_n$ are i.i.d. F on R^+ and $Y, Y_1, \ldots, Y_n$ are i.i.d. G and independent of the X's. Suppose that we are not able to observe the T_i's, but instead we can only observe $X_i \equiv (Y_i, 1_{[T_i \le Y_i]}) \equiv (Y_i, \Delta_i)$, $i = 1, \ldots, n$. Note that with $X = (Y, 1\{T \le Y\}) \equiv (Y, \Delta)$ we have

$$(\Delta | Y = y) \sim \text{Bernoulli}(F(y))\,.$$

It follows that if G has density g with respect to Lebesgue measure λ on R^+, then the observations $X, X_1, \ldots, X_n$ have density

$$p_F(y, \delta) = F(y)^\delta (1 - F(y))^{1-\delta} g(y)$$

for $y \in R^+$ and $\delta \in \{0, 1\}$ with respect to the product μ of counting measure and Lebesgue on $\{0, 1\} \times R^+$.

Suppose that we want to estimate $\nu(P_F) = F(x_0)$ for some fixed $x_0 \in (0, \infty)$. We would like to find a lower bound for estimation of this parameter. We will proceed much as in the previous example: Fix a distribution function F_0, and suppose that F_0 has a positive derivative at x_0: $F_0'(x_0) = f_0(x_0) > 0$. For $c > 0$, define

$$F_n(x) = \begin{cases} F_0(x) & \text{if} \quad x \le x_0 - cn^{-1/3} \quad \text{or} \quad x > x_0 + cn^{-1/3}, \\ F_0(x_0 - cn^{-1/3}) & \text{if} \quad x_0 - cn^{-1/3} < x \le x_0, \\ F_0(x_0 + cn^{-1/3}) & \text{if} \quad x_0 < x \le x_0 + cn^{-1/3}. \end{cases}$$

Then it is easily seen that

$$n^{1/3} |\nu(P_n) - \nu(P_0)| = |n^{1/3}(F_0(x_0 - cn^{-1/3}) - F_0(x_0))| \to |F_0'(x_0)| c = f(x_0) c\,. \tag{7}$$

On the other hand we calculate, letting $p_n \equiv p_{F_n}$,

$$\begin{aligned} H^2(p_{F_n}, p_{F_0}) &= \frac{1}{2} \int [\sqrt{p_n(x)} - \sqrt{p_0(x)}]^2 \, d\mu(x) \\ &= \frac{1}{2} \int \frac{[\sqrt{p_n(x)} - \sqrt{p_0(x)}]^2 [\sqrt{p_n(x)} + \sqrt{p_0(x)}]^2}{[\sqrt{p_n(x)} + \sqrt{p_0(x)}]^2} \, d\mu(x) \\ &= \frac{1}{2} \int \frac{[p_n(x) - p_0(x)]^2}{[\sqrt{p_n(x)} + \sqrt{p_0(x)}]^2} \, d\mu(x) \\ &= \frac{1}{2} \int_{x_0 - cn^{-1/3}}^{x_0} \frac{[F_0(x_0 - cn^{-1/3}) - F_0(x)]^2}{[\sqrt{F_n(x)} + \sqrt{F_0(x)}]^2} \, g(x) dx \\ &\qquad + \frac{1}{2} \int_{x_0}^{x_0 + cn^{-1/3}} \frac{[F_0(x_0 + cn^{-1/3}) - F_0(x)]^2}{[\sqrt{F_n(x)} + \sqrt{F_0(x)}]^2} \, g(x) dx \\ &\qquad + \frac{1}{2} \int_{x_0 - cn^{-1/3}}^{x_0} \frac{[1 - F_0(x_0 - cn^{-1/3}) - (1 - F_0(x))]^2}{[\sqrt{1 - F_n(x)} + \sqrt{1 - F_0(x)}]^2} \, g(x) dx \\ &\qquad + \frac{1}{2} \int_{x_0}^{x_0 + cn^{-1/3}} \frac{[1 - F_0(x_0 + cn^{-1/3}) - (1 - F_0(x))]^2}{[\sqrt{1 - F_n(x)} + \sqrt{1 - F_0(x)}]^2} \, g(x) dx \\ &= \frac{1}{8 F_0(x_0)(1 - F_0(x_0))} \int_{x_0 - cn^{-1/3}}^{x_0} [F_0(x_0 - cn^{-1/3}) - F_0(x)]^2 g(x) dx \\ &\qquad + \frac{1}{8 F_0(x_0)(1 - F_0(x_0))} \int_{x_0}^{x_0 + cn^{-1/3}} [F_0(x_0 + cn^{-1/3}) - F_0(x)]^2 \, g(x) dx \end{aligned}$$

$$
\begin{aligned}
&\quad + o(n^{-1}) \\
\sim &\quad \frac{g(x_0) f_0(x_0)^2}{4F_0(x_0)(1-F_0(x_0))} \frac{c^3}{3n}
\end{aligned}
$$

much as in the preceding example.

Combining these two pieces with Proposition 5.1 we find that, for any estimator T_n of $\nu(P_F) = F(x_0)$ and the loss function $l(x) \equiv |x|$ we have

$$
\begin{aligned}
&\inf_{T_n} \max \left\{ E_n n^{1/3} |T_n - \nu(P_{F_n})|, \, E_0 n^{1/3} |T_n - \nu(P_{F_0})| \right\} \\
&\geq \quad \frac{1}{4} |n^{1/3}(\nu(P_n) - \nu(P_0))| \left\{ 1 - \frac{nH^2(P_n, P_0)}{n} \right\}^{2n} \\
&= \quad \frac{1}{4} |n^{1/3}(F_0(x_0 - cn^{-1/3}) - F_0(x_0))| \left\{ 1 - \frac{nH^2(P_n, P_0)}{n} \right\}^{2n} \\
&\to \quad \frac{1}{4} f_0(x_0) c \exp \left(- \frac{g(x_0) f_0(x_0)^2}{6F_0(x_0)(1-F_0(x_0))} c^3 \right) .
\end{aligned}
$$

We can choose c to maximize the quantity on the right side. It is easily seen that the maximum is achieved when

$$
c = c_0 \equiv \left(\frac{6F_0(x_0)(1-F_0(x_0))}{3g(x_0) f_0(x_0)^2} \right)^{1/3} .
$$

This yields

$$
\begin{aligned}
&\liminf_{n\to\infty} \inf_{T_n} \max \left\{ E_n n^{1/3} |T_n - \nu(P_n)|, \, E_0 n^{1/3} |T_n - \nu(P_0)| \right\} \\
&\geq \quad \frac{e^{-1/3}}{4} \left(\frac{2F_0(x_0)(1-F_0(x_0)) f_0(x_0)}{g(x_0)} \right)^{1/3} .
\end{aligned}
$$

This bound again has the appropriate structure in the sense that the (nonparametric) MLE of F, $\widehat{F}_n$ converges at rate $n^{-1/3}$ and the same constant is involved in its limiting distribution:

$$
n^{1/3}(\widehat{F}_n(x_0) - F_0(x_0)) \to_d \left(\frac{F_0(x_0)(1-F_0(x_0)) f_0(x_0)}{2g(x_0)} \right)^{1/3} (2\mathbb{Z})
$$

where $\mathbb{Z} = \text{argmin}\{W(t) + t^2\}$ and W is a standard Brownian motion process started at 0; this was shown by Groeneboom and Wellner (1992).