KULLBACK-LEIBLER CONSTRAINED ESTIMATION
OF PROBABILITY MEASURES

by

Anne Sheehy

Technical Report No. 250
August 1988

Prepared under the Auspices
of
National Science Foundation
DMS 87-08083
Ingram Olkin, Project Director

Department of Statistics
Stanford University
Stanford, California
Kullback-Leibler Constrained Estimation
of Probability Measures

by

Anne Sheehy
Stanford University

July, 1988

Key Words and Phrases: constrained estimation, Kullback-Leibler divergence, weak convergence, asymptotic efficiency.
Research supported by: National Science Foundation Grant DMS-8400893 and DM-8708083
Kullback-Leibler Constrained Estimation of Probability Measures

by

Anne Sheehy
Stanford University

ABSTRACT

We consider two estimation problems in this paper. In the first, we observe \( X_1, \ldots, X_n \) i.i.d. \( P_0 \), where it is assumed known that \( E_{P_0} T = a \). In the second, we observe \( X_1, \ldots, X_n \in \mathbb{R}^d \) and \( \frac{1}{\sqrt{m}} \sum_{i=1}^{m} T(Y_i) \in \mathbb{R}^b \), where \( X_1, \ldots, X_n, Y_1, \ldots, Y_m \) are i.i.d. \( P_0 \), \( T \) is some Borel measurable function and \( n/m \to \lambda \) as \( n \wedge m \to \infty \). In both situations we consider the problem of estimating the probability measure \( P_0 \), uniformly over a class of sets \( C \). The estimators considered here are based on the minimization of the Kullback-Leibler divergence from certain collections of probability measures to the empirical measure of the \( X_i \)'s. We show that these estimators are consistent and asymptotically efficient.
§1 Introduction.

Haberman (1984) considered the problem of estimating a Euclidean-valued functional of a probability measure $P_0$, where one has observed data $X_1, \ldots, X_n$ i.i.d. $P_0$ and one knows that $P_0$ satisfies a finite number of linear constraints. He proposed an estimator for this problem based on the minimization of the Kullback-Leibler divergence from the constrained family of probability measures to the empirical measure of the observations. He also derived a number of consistency and weak convergence results for this estimator. In this paper we extend this problem and those results in a number of directions as will be described below.

We consider two kinds of estimation problems here. The first can be described as follows:-

We assume $P_0$ is a probability measure on the Borel $\sigma$-field $(\mathbb{R}^d, \mathcal{B}^d)$. Having observed data $X_1, \ldots, X_n \in \mathbb{R}^d$ i.i.d. $P_0$, we consider the problem of estimating certain linear functionals of $P_0$, given that we know $P_0$ belongs to the collection of all probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$ satisfying a finite number of linear constraints. More precisely, let $\mathcal{P}$ denote the collection of all probability measures on the Borel $\sigma$-field $(\mathbb{R}^d, \mathcal{B}^d)$ and let $T : \mathbb{R}^d \to \mathbb{R}^b$ be some Borel measurable function. We assume that $P_0$ is constrained to be in

$$\mathcal{P}_{T,a} = \{ P \in \mathcal{P} : E P T = a \}$$

for some fixed $a \in \mathbb{R}^b$. We will refer to this problem as the "constrained estimation problem". In this paper we will focus on the problem of estimating one particular functional of $P_0$, defined in the following way:-

Let $\mathcal{C}$ be a collection of sets in $\mathcal{B}^d$. We consider the problem of estimating functionals of $P_0$ of the kind

$$\nu_\mathcal{C}(P_0) \in L_\infty(\mathcal{C}) \quad \text{where} \quad \nu_\mathcal{C}(P_0)(C) = P_0(C) \quad \text{for any } C \in \mathcal{C}.$$  

In estimating $\nu_\mathcal{C}(P_0)$ we are essentially estimating the probability measure $P_0$ uniformly over $\mathcal{C}$. A closely related problem to the one described above is that
where \( P_0 \) is not in fact in \( \mathcal{P}_{T,a} \) but we want to estimate that \( Q \) in \( \mathcal{P}_{T,a} \) which is "closest" in some sense to \( P_0 \). See the example below.

In the second problem, we again observe data \( X_1, \ldots, X_n \in \mathbb{R}^d \) i.i.d. \( P_0 \). In addition we observe the \( \mathbb{R}^k \)-valued random variable \( T_m = \frac{1}{m} \sum_{i=1}^m T(Y_i) \) where \( Y_1, \ldots, Y_m \in \mathbb{R}^d \) are i.i.d. \( P_0 \), independent of the \( X_i \)'s, but are not observed and \( T(\cdot) \) is some measurable function from \( \mathbb{R}^d \) into \( \mathbb{R}^k \). We also assume that

\[
\frac{n}{m} \to \lambda \geq 0 \text{ as } n \to \infty.
\]

We will refer to this problem as the "supplementary sample moments problem". Again, we are interested in estimating functions of \( P_0 \) of the kind \( \nu_C(P_0) \in L_\infty(C) \), exactly as defined above.

The theory given here is easily adapted to the case where we are interested in estimating Euclidean valued functions of \( P_0 \) of the kind \( \nu_g(P_0) = E_{P_0} g \) where \( g : \mathbb{R}^d \to \mathbb{R}^k \) is some measurable function. For the constrained estimation problem, the analogues of the theorems given in section 3 for the problem of the estimation of \( \nu_C(P_0) \) can in fact be found in Haberman (1984), while those of section 4 can be found in Sheehy (1987).

The following is an example of the first kind of estimation problem.

**Example: Known Marginals.** Suppose \( R_{0i} \) is a probability mass function for \( i = 1, \ldots, k \leq d \). Suppose \( R_{0i} \) and \( P_{0i} \)-the \( i \)th marginal distribution of the \( X \)'s—have support concentrated on \( \{\alpha_{i1} \ldots \alpha_{ib_i}\} \), \( i = 1, \ldots, k \). Let

\[
T(x) = \left( 1_{\{\alpha_{i1}\}}(x_1) - R_{01}(\alpha_{i1}), \ldots, 1_{\{\alpha_{ib_i}\}}(x_1) - R_{01}(\alpha_{ib_i}), \ldots, 1_{\{\alpha_{ki}\}}(x_k) - R_{0k}(\alpha_{ki}), \ldots, 1_{\{\alpha_{kib_k}\}}(x_k) - R_{0k}(\alpha_{kib_k}) \right) \in \mathbb{R}^b
\]

where \( b = \sum_{i=1}^k b_i \), and \( x = (x_1, \ldots, x_d)^T \).

Now let

\[
a = (0, \ldots, 0)^T \in \mathbb{R}^b.
\]
Constraining $P_0$ to be in $\mathcal{P}_{T,a}$ is equivalent to saying we know the first $k$ marginal distributions of $P_0$.

Mosteller (1968) gives some interesting examples where $\mathcal{P}_{T,a}$ is as defined in the example above and $P_0 \not\in \mathcal{P}_{T,a}$ but one wants to estimate that $Q_0$ which is in $\mathcal{P}_{T,a}$ and which is "closest" in some sense to $P_0$. One such example is the situation where we want to compare transition probabilities in social mobility tables between societies. We may first want to adjust the marginal distributions of row and column variables so that the variations in the relative sizes of social classes from society to society do not affect our results. Mosteller suggests adjusting the marginal probabilities so that they are uniformly distributed on each table. The method he proposes, Iterative Proportional Fitting, gives the same estimates as the ones we get from using the Kullback-Leibler constrained estimator which we will describe below.

Jagers et al. (1985) describe a situation in the sample survey framework where $(X_1,Y_1),\ldots,(X_n,Y_n) \in \mathbb{R}^2$ i.i.d. $P_0$ are observed and we know the marginal distribution of the $X$'s, call it $P_{0X}$. Also we know that $P_{0X}$ has finite support. They consider the problem of estimating functions of the kind $\nu_x(P_0) \equiv E_{P_0} g$, where $g$ is some Borel-measurable, real valued function as defined above. In fact they focus on functions $g$ which are functions only of the $Y$ variable. The estimator they examine is called a "post-stratified" estimator and is defined by

$$\hat{E}_{P_0} g(Y) = \sum_x \frac{P_{0x}(x)}{\nu(x)} g(Y_i) 1[X_i = x],$$

where $\nu(x)$ is the number of $(X,Y)$ pairs where $X_i = x$, and the summation is taken over all the $x$ values in the support of $P_{0X}$ with $\nu(x) > 0$. It can be shown that the Kullback-Leibler constrained estimator, described below, gives the same estimate as the post-stratified estimator in this situation.

We now describe the estimators which we will consider here. The idea is the following: In the fully non-parametric set-up, having observed $X_1,\ldots,X_n$ i.i.d. $P_0$, when no knowledge of $P_0$ is assumed, it is known that $\nu_C(\mathcal{P}_n)$, is in a sense,
an optimal estimator for $\nu_C(P_0)$ (under appropriate regularity conditions on $C$ and $P_0$), where

\begin{equation}
\mathbb{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i \in (\cdot)]}.
\end{equation}

We call $\mathbb{P}_n$ the empirical measure of the observations $X_1, \ldots, X_n$.

In the constrained estimation problem we know that $P_0$ is in a certain restricted class of probability measures and of course feel that we should take this information into account when estimating functions of $P_0$. We propose doing this by using $\nu_C(Q_n(a))$ as an estimate of $\nu_C(P_0)$, where $Q_n(a) \in \mathcal{P}_{T,a}$ and $Q_n(a)$ is the "closest" as measured by the Kullback-Leibler divergence, probability to $\mathbb{P}_n$, in $\mathcal{P}_{T,a}$.

We use a similar idea in the supplementary sample moments problem. Define

\begin{equation}
\mathbb{P}_m^2(\cdot) = \frac{1}{m} \sum_{i=1}^{m} 1_{[Y_i \in (\cdot)]}
\end{equation}

and

\begin{equation}
T_N = \frac{n}{N} \int Td\mathbb{P}_n + \frac{m}{N} \int Td\mathbb{P}_m^2,
\end{equation}

where $N = n + m$.

With these definitions, we use $\nu_C(Q_N(T_N))$ as an estimator for $\nu_C(P_0)$ where, $Q_N(T_N) \in \mathcal{P}_{T,T_N}$ is that probability measure which is the "closest", as measured by the Kullback-Leibler divergence, to $\mathbb{P}_n$.

We now give a definition of the Kullback-Leibler divergence and introduce some notation.

**Definition 1.** Define the Kullback-Leibler (K-L) divergence between two probability measures $Q$ and $P$ on $(\mathbb{R}^d, \mathcal{B}^d)$ to be,

\[ K(Q, P) = \begin{cases} 
\int \log \frac{dQ}{dP} dQ, & \text{if } Q \ll P \\
+\infty & \text{otherwise}
\end{cases} \]
Notice that $K(\cdot, \cdot)$ is not a metric, in fact it is not even symmetric in its arguments. However, its value is always non-negative and

$$K(Q, P) = 0 \quad \text{if and only if} \quad \frac{dQ}{dP} = 1 \quad \text{a.e.} \quad Q.$$  

If there exists a probability measure $Q$ minimizing $K(\cdot, P_0)$ (respectively, $K(\cdot, P_n)$) over $\mathcal{P}_{T,a}$ denote that $Q$ by $Q(a)$ (respectively, $P_n(a)$).

In the constrained estimation problem, the idea of using $\nu_C(Q_n(a))$ as an estimate of $\nu_C(P_0)$ has been used extensively in the literature but usually in the context where $P_0$ represents the probabilities in a contingency table and the constraints on $P_0$ are that certain of the marginals are known (see Haberman (1984) for some appropriate references). Haberman (1984) considers the more general problem which we consider here but focuses on estimating functions of $P_0$ of the form $\nu_g(P_0) \equiv E_{P_0} g$, where $g$ is some Euclidean-valued, Borel-measurable function. In his paper Haberman does not consider the supplementary sample moments problem nor does he prove efficiency of the estimator for the constrained estimation problem.

From now on, when we talk about the estimator $\nu_C(Q_n(a))$ it should be understood that we are working with the constrained estimation problem and when we write $\nu_C(Q_n(T_N))$ we are working with the supplementary sample moments problem.

In section 2, we list conditions under which these estimators exist and show that (under certain regularity conditions) they are consistent. Csiszár(1975) and Haberman(1984) have done most of the work here.

In section 3, we show that these estimators, when properly normalized, converge weakly to limiting Gaussian processes. We also make a comparison between these estimators and the estimator $\nu_C(P_n)$. It is shown that in each of the two problems the asymptotic variance of the proposed estimator, at any $C \in C$, is less than that of the corresponding quantity for the estimator $\nu_C(P_n)(C)$. In section 4
we show that in the constrained estimation problem, \( \nu_C(Q_n(a)) \) is an asymptotically efficient estimator of \( \nu_C(P_0) \), while in the supplementary sample moments problem \( \nu_C(Q_n(T_N)) \) is an asymptotically efficient estimator of \( \nu_C(P_0) \). Ireland and Kullback (1968) have shown that in the situation where \( P_0 \) represents the probabilities in a contingency table, and the constraints are that certain marginal probabilities are known, then \( \nu_C(Q_n(a)) \) is Best Asymptotically Normal. That we can show \( \nu_C(Q_n(a)) \) is efficient in a more general context is therefore not too surprising.

The proof of the results stated in sections 2, 3 and 4 are all contained in section 5. In section 6 certain technical definitions and theorems which are needed in the proofs are stated and proved.

§2. Existence and Consistency of the Estimators.

§2.1 Introduction. We first summarize results, proved by Csiszár(1975) about the existence of \( Q(a) \) (as defined in section 1).

Define

\[ S(P_0, \infty) = \{ Q \ll P_0 : Q \in \mathcal{P}, K(Q, P_0) < \infty \}. \]

We need to know under what conditions there exists a probability measure \( Q(a) \in S(P_0, \infty) \) minimizing \( K(\cdot, P_0) \) over \( \mathcal{P}_{T,a} \).

The following result is due to Csiszár (1975):

(C1) Let

\[ \Theta(P_0) = \{ \theta \in \mathbb{R}^b : E_{P_0} \exp(\theta' T) < \infty \} \]

\[ B_{T,a,P_0} = \{ a \in \mathbb{R}^b : \mathcal{P}_{T,a} \cap S(P_0, \infty) \neq \emptyset \}. \]

If \( \Theta(P_0) \) is open, there exists a unique \( Q(a) \) in \( \mathcal{P}_{T,a} \) minimizing \( K(\cdot, P_0) \) over \( \mathcal{P}_{T,a} \) for each inner point \( a \in B_{T,a,P_0} \) and its \( P_0 \)-density is of the form

\[ \frac{dQ(a)}{dP_0}(x) = c \exp(\theta(a)' T(x)) \quad \text{for some} \quad \theta(a) \in \Theta(P_0). \]

This is Theorem 3.3 of Csiszár (1975).
We note at this point that if \( a \) is an element of the relative interior of the convex hull of the support of \( P_0 T^{-1} \) (i.e. \( a \in \text{relintcosupp} P_0 T^{-1} \)), then \( a \) is an inner point of \( B_{T,a,P_0} \) (see Propositions A.3.2 and A.3.3 of Sheehy (1987)). One consequence of this is that if \( \Theta(P_0) \) is open and \( a \in \text{relintcosupp} P_0 T^{-1} \), then there exists \( Q(a) \) minimizing \( K(\cdot, P_0) \) over \( \mathcal{P}_{T,a} \) and \( Q(a) \) has \( P_0 \)-density of the form given in (1) above.

Therefore, showing that the probability measure \( Q_n(a) \) (or \( Q_n(T_N) \) exists, with probability one for \( n \) sufficiently large, amounts to showing that \( a \in \text{relintcosupp} P_n T^{-1} \) (respectively \( T_N \in \text{relintcosupp} P_n T^{-1} \)) with probability one for \( n \) sufficiently large.

In the constrained estimation problem, results are stated both for the case where \( E_{P_0} T = a \) and the case where \( E_{P_0} T \neq a \), but \( a \in \text{relintcosupp} P_0 T^{-1} \) and there exists \( Q(a) \) minimizing \( K(\cdot, P_0) \) over \( \mathcal{P}_{T,a} \) with \( P_0 \)-density as given in (1).

If \( E_{P_0} T \neq a \), then \( Q(a) \) is the "closest"—as measured by the Kullback-Leibler divergence—probability measure to \( P_0 \) having \( E_{P_0} T = a \). It is of interest to study the behavior of this estimator when \( E_{P_0} T \neq a \) for two reasons. First, because it is important to know how the estimator behaves when the constraint does not hold. And second, because sometimes, as in Example 1 above, we may know that \( E_{P_0} T \neq a \), but may still be interested in estimating \( Q(a) \).

Notice that when \( \text{Var}_{P_0} T(\equiv E_{P_0}(T - E_{P_0} T)(T - E_{P_0} T)' ) \) is singular, there exist non-zero vectors \( u \) such that \( u'T = c \) a.s. \( P_0 \), \( c \) a constant. When this is the case we see that the vector \( \theta(a) \) defined in (1) will not be unique. Indeed if

\[
H = \{ u : u'T = c, \text{ a.s. } P_0, \text{ for some constant } c \}
\]

and \( \theta^* = \theta(a) + u, u \in H \) then

\[
\frac{dQ_{\theta^*}}{dP_0}(x) = \frac{\exp(\theta^*T(x))}{\int \exp(\theta^*T(x))dP_0(x)}
= \frac{\exp(\theta(a)'T(x))}{\int \exp(\theta(a)'T(x))dP_0(x)} \text{ a.s. } P_0.
\]
It is clear then that $\theta(a)$ is unique only up to its projection on $H^\perp$. From now on we will define $\theta(a)$ to be that projection.

§2.2 Statements of Results. Propositions 1, 1' and 2 below are Glivenko-Cantelli type results for the estimators which we consider here. We need to restrict the collection of sets $C$ to some pointwise separable, Vapnik-Červonenkis (VC) class of sets (see Definitions 8 and 9). Classes of sets which are pointwise separable, VC classes include the classes of all closed balls, all ellipsoidal regions, and all convex hulls of at most $k$ points ($k$ fixed) in $\mathbb{R}^d$. Classes of sets which are not VC include the class of all convex sets in $\mathbb{R}^d$, $d \geq 2$ and the class of all lower layers in $\mathbb{R}^d$, $d \geq 2$. The pointwise separability condition is needed to ensure the measurability of the quantities in which we are interested.

**Proposition 1.** Suppose $a \in \text{relintcosupp} P_0 T^{-1}$ and $\Theta(P_0)$ is open, so that there exists $Q(a)$ minimizing $K(\cdot, P_0)$ over $P_{T,a}$ with $P_0$ density as given in (1) above. Suppose $C$ is a pointwise separable, VC class of sets in $\mathcal{B}^d$. Then with probability one, for $n$ sufficiently large, there exists $Q_n(a)$ minimizing $K(\cdot, P_n)$ over $P_{T,a}$ where

$$
(i) \quad \frac{dQ_n(a)}{dP_n}(x) = c_n \exp(\theta_n(a)'T(x)), \quad \theta_n(a) \in H^\perp
$$

$$
(ii) \quad \theta_n(a) \rightarrow_{a.s.} \theta(a) \quad \text{as} \quad n \rightarrow \infty
$$

and

$$
(iii) \quad \|\nu_C(Q_n) - \nu_C(Q(a))\|_{L_\infty(C)} \rightarrow_{a.s.} 0 \quad \text{as} \quad n \rightarrow \infty.
$$

**Proposition 1'.** Suppose $E_{P_0} T = a$ and $0 \in \text{int} \Theta(P_0)$. Then with probability one, for $n$ sufficiently large, there exists $Q_n(a)$ minimizing $K(\cdot, P_n)$ over $P_{T,a}$ where

$$
(i) \quad \frac{dQ_n(a)}{dP_n}(x) = c_n \exp(\theta_n(a)'T(x)), \quad \theta_n(a) \in H^\perp
$$
(ii) \( \theta_n(a) \rightarrow_{a.s.} 0 \quad \text{as} \quad n \rightarrow \infty. \)

and

(iii) \( \| \nu_C(Q_n) - \nu_C(P_0) \|_{L_\infty(C)} \rightarrow_{a.s.} 0 \quad \text{as} \quad n \rightarrow \infty. \)

**Proposition 2.** Suppose \( 0 \in \text{int } \Theta(P_0) \). Suppose \( C \) is a pointwise separable, VC class of sets in \( B^d \). Then with probability one, for \( n \) sufficiently large, there exists \( Q_n(T_N) \) minimizing \( K(\cdot, P_n) \) over \( \mathcal{P}_{T,T_N} \) with

(i) \( \frac{dQ_n(T_N)}{dP_n}(x) = c_n \exp(\theta_n(T_N)T(x)), \quad \theta_n(T_N) \in H^\perp, \)

(ii) \( \theta_n(T_N) \rightarrow_{a.s.} 0 \quad \text{as} \quad n \rightarrow \infty, \)

and

(iii) \( \| \nu_C(Q_n(T_N)) - \nu_C(P_0) \|_{L_\infty(C)} \rightarrow_{a.s.} 0 \quad \text{as} \quad n \rightarrow \infty. \)

In Lemmas 1 and 2 and Theorem 1 we show that, under appropriate regularity conditions, if \( \{a_n\} \) is a sequence in \( \mathbb{R}^b \), converging (a.s.) to some constant \( a \), then, with probability one, for \( n \) sufficiently large, there exists \( Q_n(a_n) \) minimizing \( K(\cdot, P_n) \) over \( \mathcal{P}_{T,a_n} \) and that

\( \| \nu_C(Q_n(a_n)) - \nu_C(P_0) \|_{L_\infty(C)} \rightarrow 0 \quad \text{a.s. as} \quad n \rightarrow \infty. \)

From this result we can deduce the preceding propositions.

**Lemma 1.** If \( a \in \text{relintcosupp} P_0 T^{-1} \), \( \{a_n\} \subset \text{relintcosupp} P_0 T^{-1} \) and \( a_n \rightarrow_{a.s.} a \) then with probability one, for \( n \) sufficiently large, there exists a probability measure \( Q_n(a_n) \) minimizing \( K(\cdot, P_n) \) over \( \mathcal{P}_{T,a_n} \) and

(3) \( \frac{dQ_n(a_n)}{dP_n}(x) = c_n \exp(\theta_n(a_n)T(x)) \)
where
\[
\theta_n(a_n) \in H^\perp.
\]

**Lemma 2.** Suppose the hypotheses of Lemma 1 hold. Suppose \( Q(a) \) minimizes \( K(\cdot, P_0) \) over \( \mathcal{P}_{T,a} \) and

\[
\frac{dQ(a)}{dP_0}(x) = c \exp(\theta(a)'T(x))
\]

with
\[
\theta(a) \in H^\perp \cap \text{int} \Theta(P_0).
\]

Then
\[
\theta_n(a_n) \to_{a.s.} \theta(a) \quad \text{as} \quad n \to \infty
\]

with \( \theta_n(a_n) \) as defined in (3) above.

**Theorem 1.** Suppose the hypotheses of Lemma 2 hold. Suppose \( C \) is a pointwise separable, VC class of sets in \( B^d \). Then

\[
\|\nu_C(Q_n(a_n)) - \nu_C(Q(a))\|_{L_\infty(C)} \to_{a.s.} 0 \quad \text{as} \quad n \to \infty.
\]

§3. Weak Convergence of the Estimators.

§3.1 Statements of Main Results. In this section we examine the asymptotic behavior of the estimators \( \nu_C(Q_n(a)) \) and \( \nu_C(Q_n(T_N)) \).

We will need the following notation. For any function \( f \in L_2(P_0) \) write

\[
XX_n(f) = \int f d(\sqrt{n}(P_n - P_0)).
\]

If \( \mathcal{F} \) is a collection of functions in \( L_2(P_0) \) then we can regard \( XX_n \) as an element of \( L_\infty(\mathcal{F}) \) and we call \( XX_n \) the empirical process on \( \mathcal{F} \). If \( f, g \in \mathcal{F} \) then

\[
(XX_n(f), XX_n(g))^T \Rightarrow N(0, \Gamma(f,g))
\]
where 
\[
\Gamma(f, g) = \begin{pmatrix}
\text{Var}_{P_0} f & \text{Cov}_{P_0}(f, g) \\
\text{Cov}_{P_0}(f, g) & \text{Var}_{P_0} g
\end{pmatrix}.
\]

In Theorem 2 we show that the estimator \( \nu_C(Q_n(a)) \) has influence function, 
\( \psi_C(\cdot, P_0)(\cdot) \) where \( \psi_C(C, P_0)(\cdot) \in L_2(P_0) \) for any \( C \in \mathcal{C} \), and \( \psi_C(\cdot, P_0)(x) \in L_\infty(\mathcal{C}) \). In Theorem 3 we prove an analogous result for the estimator \( \nu_C(Q_n(T_N)) \).

**Notation.** If \( \Sigma \) is a \( b \times b \) matrix, define \( \Sigma^- \) to be the Moore-Penrose generalized inverse of \( \Sigma \).

**Theorem 2.** Let \( \mathcal{C} \) be a pointwise separable, VC class of sets in \( \mathcal{B}^d \). Suppose \( \Theta(P_0) \) is open and \( a \in \text{relintcosupp}PT^{-1} \). Suppose in addition that

\[
(i) \quad \int \|T\|^2 \exp(2\theta(a)'T)dP_0 < \infty \quad \text{and} \quad \int \exp(2\theta(a)'T)dP_0 < \infty
\]

where \( \theta(a) \) was defined in (2.1) and \( \theta(a) \in H^\perp \cap \text{int} \Theta(P_0) \). Then

\[
\mathcal{Z}_n(C) \equiv \sqrt{n}(\nu_C(Q_n(a)) - \nu_C(Q(a)))(C) 
= \int [(1_C - Q(a)(C)) - \text{Cov}_{Q(a)}(1_C, T)(\text{Var}_{Q(a)}T)^{-1}(T - a)] \frac{dQ(a)}{dP_0} dX_n + O_P(1)
\]

where the \( O_P(1) \) term is uniformly small over \( C \).

**Theorem 2'.** Let \( \mathcal{C} \) be a pointwise separable, VC class of sets in \( \mathcal{B}^d \). Suppose \( 0 \in \text{int} \Theta(P_0) \) and \( E_{P_0}T = a \). Then

\[
\mathcal{Z}_n(C) = \sqrt{n}(\nu_C(Q_n(a)) - \nu_C(P_0))(C) 
= \int [(1_C - P_0(C)) - (\text{Cov}_{P_0}(1_C, T)(\text{Var}_{P_0}T)^{-1}(T - a)]dX_n + O_P(1)
\]

where the \( O_P(1) \) term is uniformly small over \( C \).

For any function \( f \in L_2(P_0) \) write

\[
XX_m^2(f) = \int fd(\sqrt{m}(P_m^2 - P_0)).
\]
Theorem 3. Let $\mathcal{C}$ be a pointwise separable, VC class of sets in $\mathcal{B}^d$. Suppose $0 \in \text{int}\Theta(P_0)$ and $\frac{n}{m} \to \lambda$ as $n \to \infty$. Then

$$Z_N(C) = \sqrt{n}(\nu_C(Q_n(T_N)) - \nu_C(P_0))(C)$$

$$= \int \frac{\sqrt{\lambda}}{1 + \lambda}(\text{Cov}_{P_0}(1_C, T)(\text{Var}_{P_0} T)^{-1}(T - a)dX_m^2$$

$$+ \int [(1_C - P_0(C))] - \frac{1}{1 + \lambda}(\text{Cov}_{P_0}(1_C, T)(\text{Var}_{P_0} T)^{-1}(T - a)dX_n$$

$$+ o_P(1)$$

where the $o_P(1)$ term is uniformly small over $\mathcal{C}$.

Corollaries 1 and 2 below are weak convergence results for the processes $Z_n$ and $Z_N$ defined in Theorems 2 and 3 above. We use the definition of weak convergence used in Pollard(1984). Before their statements we need the following definitions.

Define for any $A, B \in \mathcal{B}^d$, and any probability measure $Q \in \mathcal{P}$ for which $E_Q||T||^2 < \infty$

$$\Gamma_Q(A, B) = \text{Cov}_Q(1_A, 1_B) - \text{Cov}_Q(1_A, T)(\text{Var}_Q T)^{-1}(\text{Cov}_Q(1_B, T)$$

and

$$\Gamma^\lambda_{P_0}(A, B) = \text{Cov}_{P_0}(1_A, 1_B) - \frac{1}{1 + \lambda}(\text{Cov}_{P_0}(1_A, T)(\text{Var}_{P_0} T)^{-1}(\text{Cov}_{P_0}(1_B, T).$$

Definition 2. For any collection of functions $\mathcal{F} \in L_2(P_0)$ define $B_{\mathcal{F}, P_0}$ to be the smallest $\sigma$-field on $L_\infty(\mathcal{F})$ which

(i) makes all finite-dimensional projections measurable and

(ii) contains all closed balls with centers in $\mathcal{C}(\mathcal{F}, P_0)$, the collection of all uniformly continuous functions in $L_\infty(\mathcal{F})$ (w.r.t. the $L_2(P_0)$ norm).

Corollary 1. Suppose the hypotheses of Theorem 2 hold. Then

$$Z_n(\cdot) = \sqrt{n}(Q_n - Q(\cdot))(\cdot) \Rightarrow Z_{01}(\cdot)$$
as random elements of \((L_\infty(C), B_C, P_0)\), where \(Z_{01}(\cdot)\) is a mean-zero, tight, Gaussian random element of \(L_\infty(C)\) with sample paths in \(C(C, P_0)\), and covariance function \(\Gamma_{Q_{01}}(\cdot, \cdot)\).

Corollary 1'. Suppose the hypotheses of Theorem 2' hold. Then

\[ Z_n(\cdot) = \sqrt{n}(Q_n - P_0)(\cdot) \Rightarrow Z_{02}(\cdot) \]

as random elements of \((L_\infty(C), B_C, P_0)\), where \(Z_{02}(\cdot)\) is a mean-zero, tight, Gaussian random element of \(L_\infty(C)\) with sample paths in \(C(C, P_0)\) and covariance function \(\Gamma_{P_0}(\cdot, \cdot)\).

Corollary 2. Suppose the hypotheses of Theorem 3 hold. Then

\[ Z_N(\cdot) = \sqrt{n}(Q_n(T_N) - P_0)(\cdot) \Rightarrow Z_{03}(\cdot) \]

as random elements of \(L_\infty(C), B_C, P_0\), where \(Z_{03}(\cdot)\) is a mean-zero, tight, Gaussian random element of \(L_\infty(C)\) with sample paths in \(C(C, P_0)\), and covariance function \(\Gamma_{P_0}(\cdot, \cdot)\).

Remark. At this point it is instructive to think about the behavior of the alternative estimator \(\nu_C(P_n)\) in each of the estimation problems. By well known results (see e.g. Pollard (1984)) if \(C\) is a pointwise-seperable, VC class of sets and \(P_0\) satisfies the assumptions of Corollary 1 or 2, then

\[ ||\nu_C(P_n) - \nu_C(P_0)||_{L_\infty(C)} \to a.s. 0 \quad \text{as } n \to \infty. \]

and

\[ \sqrt{n}(\nu_C(P_n) - \nu_C(P_0)) \Rightarrow XX \]

where \(XX\) is a mean-zero, tight, Gaussian process in \(C(C, P_0)\) having covariance function, \(\tilde{\Gamma}: C \times C \to \mathbb{R}^1\), given by

\[ \tilde{\Gamma}(A, B) = P_0(A \cap B) - P_0(A)P_0(B) \quad \text{for } A, B \in C. \]
Notice that for any $A \in \mathcal{C}$,
\begin{equation}
\tilde{\Gamma}(A, A) = \text{Var}(X(A)) \geq \text{Var}(Z_{C2}(A)) = \Gamma_{P_0}(A, A).
\end{equation}

In fact
\[ \tilde{\Gamma}(A, A) - \Gamma_{P_0}(A, A) = \text{Cov}_{P_0}(1_A, T)(\text{Var}_{P_0} T)^{-1}\text{Cov}_{P_0}(1_A, T). \]

If $\text{Var}_{P_0} T$ is non-singular this difference will be strictly positive whenever $\text{Cov}_{P_0}(1_A, T) \neq 0$. What (4), (5) and (6) are saying is that when $P_0$ satisfies the assumptions of Corollary 1', $\nu_C(\mathbb{P}_n)$ is, like $\nu_C(Q_n(a))$, a consistent estimator of $\nu_C(P_0)$. However the variance of the limiting distribution of $\sqrt{n}(\nu_C(\mathbb{P}_n) - \nu_C(P_0))(A)$, for any set $A \in \mathcal{C}$, is greater than the variance of the limiting distribution of $\sqrt{n}(\nu_C(\mathbb{Q}_n(a)) - \nu_C(P_0))(A)$. If $P_0 \not\in \mathcal{P}_{T,a}$, then $\nu_C(\mathbb{Q}_n(a))$ and $\nu_C(\mathbb{P}_n)$ estimate different measures, and comparison of the two estimators is more difficult.

Also notice that
\begin{equation}
\tilde{\Gamma}(A, A) - \Gamma^\lambda_{P_0}(A, A) = \frac{1}{1 + \lambda}\text{Cov}(1_A, T)^{-1}(\text{Var}_{P_0} T)^{-1}\text{Cov}_{P_0}(1_A, T).
\end{equation}

From (7) we see that by taking the information about the supplementary moments into account and using $\nu_C(Q_N(T_N))$ as an estimator of $\nu_C(P_0)$ rather than $\nu_C(\mathbb{P}_n)$ we reduce the asymptotic variance of the estimator of $\nu_C(P_0)(A) \equiv P_0(A)$ by the quantity on the right hand side of (7).

In section 4 we will show that asymptotically, in a certain sense, $\nu_C(Q_n(T_N))$ and $\nu_C(\mathbb{Q}_n(a))$ are the "best" estimators for these two situations.

As one would expect
\[ \lim_{\lambda \to 0} \Gamma^\lambda_{P_0}(A, B) = \Gamma_{P_0}(A, B). \]

That is to say that if $\frac{n}{m} \to 0$, so that we have "infinitely" more information from the $Y_i$'s than we do from the $X_i$'s, asymptotically, it is as if we knew $E_{P_0} T$. If however $\frac{n}{m} \to \infty$, then $\lambda = \infty$ and
\[ \lim_{\lambda \to \infty} \Gamma^\lambda_{P_0}(A, B) = \text{Cov}_{P_0}(1_A, 1_B) = \tilde{\Gamma}(A, A). \]
so that, asymptotically, it is as if we had no extra information from the supplementary sample moments.

§4. A Convolution Theorem for Estimators of $\nu_C(P_0)$.

§4.1 Introduction. In this section we use results of Van der Vaart (1987) and Bickel, Klaassen, Ritov and Wellner (1989) to derive a convolution theorem for “regular estimators” of $\nu_C(P_0)$ in the constrained estimation problem and in the supplementary sample moments experiment. We use the Hellinger metric to define distances between probability measures. For any $g$ and $h$ in $L_2(P_0)$ let

$$
\langle g, h \rangle_{P_0} = \int ghdP_0 \quad \text{and} \quad \|g\|_{P_0}^2 = \int g^2dP_0.
$$

If $x \in \mathbb{R}^d$, $\|x\|_d$ denotes the usual Euclidean norm.

We consider the experiment where we have observed data $X_n$, a random element of the measure space $(\mathcal{A}_n, \Omega_n)$ where $X_n = F(Z_1, \ldots, Z_N)$ for some Euclidean-valued, Borel measurable function $F$ and $Z_1, \ldots, Z_N \in \mathbb{R}^d$ are i.i.d. $P_0$ and $N \equiv N(n) \to \infty$. On the basis of $X_n$ we want to estimate some functional of $P_0$ where we know $P_0 \in \mathcal{P}_0$ where $\mathcal{P}_0$ is some collection of probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$. We call $\mathcal{P}_0$ the “model”. In the constrained estimation problem $X_n \equiv (X_1, \ldots, X_n)$ where the $X_i$'s $\in \mathbb{R}^d$ are i.i.d. $P_0$, $\mathcal{A}_n \equiv (\mathbb{R}^d)^n$, $\Omega_n \equiv (\mathcal{B}^d)^n$, $N \equiv n$ and $\mathcal{P}_0 \equiv \mathcal{P}_{T,a}$. In the supplementary sample moments problem $X_n \equiv (X_1, \ldots, X_n, \frac{1}{\sqrt{m}} \sum_{i=1}^m T(Y_i))$, where $X_1, \ldots, X_n, Y_1, \ldots, Y_m \in \mathbb{R}^d$ are i.i.d. $P_0$, $\mathcal{A}_n \equiv (\mathbb{R}^d)^n \times \mathbb{R}^b$, $\Omega_n \equiv (\mathcal{B}^d)^n \times \mathcal{B}^b$, $N \equiv n + m$ and $\mathcal{P}_0 \equiv \mathcal{P}$. $\mathcal{P}_0 \equiv \mathcal{P}$, the collection of all probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$.

We begin with a heuristic explanation of how the results in this section are derived. The notion of a tangent space to the model $\mathcal{P}_0$ at $P_0$ is central to this discussion.

**Definition 3.** A subspace $\mathbf{T}(P_0)$ of $L_2(P_0)$ is called a tangent space at $P_0 \in \mathcal{P}_0$ if for all $h \in \mathbf{T}(P_0)$ there exists $\{P_t\} \subset \mathcal{P}_0$ with

$$
(1) \quad \int [t^{-1}((dP_t)^{1/2} - (dP_0)^{1/2}) - \frac{1}{2} h(dP_0)^{1/2}\]t^2 \to 0 \text{ as } t \to 0.
$$
Formula (1) is interpreted as follows. Let $P_t$ and $P_0$ have densities $p_{tt}$ and $p_t$ respectively with respect to an arbitrary $\sigma$-finite measure $\mu_t$ dominating $P_t + P_0$. Then we have

$$\int [t^{-1}(p_{tt}^{1/2} - p_t^{1/2}) - \frac{1}{2} hypnot_{t}^{1/2}]^2 d\mu_t \to 0 \text{ as } t \to 0.$$ 

Basically, $T(P_0)$ being a tangent space means that it is a subspace in $L_2(P_0)$ made up of a collection of “directions” in which we can approach $P_0$ in the model $P_0$. We will use the notation $T_m(P_0)$ to denote a maximal tangent space in the model $P_0$ at $P_0$. The tangent space $T_m(P_0)$ is maximal if its closure, $\overline{T_m(P_0)}$ (in the $L_2(P_0)$ norm) contains any other tangent space $T(P_0)$.

Suppose $\nu(P_0) \in \mathbb{R}^1$. We say that $\nu(\cdot)$ is differentiable with respect to $T_m(P_0)$ at $P_0$ with derivative $\dot{\nu}(P_0) \in \overline{T_m(P_0)}$, if given any $h \in T_m(P_0)$, there exists a sequence $\{P_t\} \subset P_0$ with

$$\int [t^{-1}(dP_t)^{1/2} - (dP_0)^{1/2}) - \frac{1}{2} h(dP_0)^{1/2}]^2 \to 0 \text{ as } t \to 0$$

and

$$(2) \quad |t^{-1}(\nu(P_t) - \nu(P_0)) - \langle \dot{\nu}(P_0), h \rangle_{P_0}| \to 0 \text{ as } t \to 0.$$ 

Notice that $\dot{\nu}(P_0)$ is defined uniquely only up as far as elements orthogonal to $\overline{T_m(P_0)}$. We can think of $\langle \dot{\nu}(P_0), h \rangle_{P_0}$ as being the rate of change of the parameter $\nu$ at $P_0$ in the direction $h$. If $S$ is some closed subspace of $L_2(P_0)$ let $\Pi(\dot{\nu}(P_0)|S)$ denote the projection of the quantity $\dot{\nu}(P_0)$ onto $S$. If we observe $X_1, \ldots, X_n$ i.i.d. $P_0 \in P_0$ and $\dot{\nu}_n$ is a “regular estimator” of the parameter $\nu$ at $P_0$ (this concept will be defined later) where $\nu(\cdot)$ is differentiable with respect to $T_m(P_0)$ at $P_0$ (for some $T_m(P_0)$), and $\sqrt{n}(\dot{\nu}_n - \nu(P_0))$ has limit law $L_{P_0}(Z)$, it can be shown that

$$L_{P_0}(Z) = L_{P_0}(Z_0 + W)$$
where

\[ Z_0 \sim N(0, \| \Pi(\hat{\nu}(P_0) | \overline{T_m(P_0)})\|^2) \]

and

\[ W \text{ is independent of } Z_0. \]

Roughly then, the greater the rate of change of the parameter \( \nu \) in the neighborhood of \( P_0 \), the larger the variance of the random variable \( Z_0 \), as one would expect.

The proof of this result goes along the following lines: Suppose \( \overline{T_m(P_0)} \) is itself a tangent space. Suppose also that for every \( h \in \overline{T_m(P_0)} \), we can find a sequence \( \{P_{th}\} \) satisfying (1) and (2). We regard \( \mathcal{P}_{th} = \{P_{th}\} \) as a one-dimensional submodel of \( \mathcal{P}_{T,a} \). Using the Hajek-Le Cam Convolution Theorem (Hajek, 1970), we can then show that if \( \hat{\nu}_n \) is a "regular estimator with respect to \( \overline{T_m(P_0)} \)" of \( \nu(P_0) \) in the model \( \mathcal{P}_{th} \), with limit law \( \mathcal{L}_{P_0}(Z) \), then

\[ \mathcal{L}_{P_0}(Z) = \mathcal{L}_{P_0}(Z_{0h} + W) \]

where

\[ Z_{0h} \sim N(0, \langle \hat{\nu}(P_0), h \rangle_{P_0}^2 \langle h, h \rangle_{P_0}^{-1}) \]

and \( W \) is independent of \( Z_{0h} \). Let \( S_h \) denote the subspace spanned by the function \( h \). Since the variance of \( Z_{0h} \) is just \( \| \Pi(\hat{\nu}(P_0)|S_h)\|^2 \) it is easy to see that this quantity is maximized over \( \overline{T_m(P_0)} \) when \( h = h_0 \equiv \Pi(\hat{\nu}(P_0)|\overline{T_m(P_0)}) \). In this sense \( \mathcal{P}_{th_0} \) is the least-favorable sub-model in \( \mathcal{P}_{T,a} \) for the problem of the estimation of \( \nu(P_0) \). If \( \hat{\nu}_n \) is a "regular estimator with respect to \( \overline{T_m(P_0)} \)" of the parameter \( \nu \) at \( P_0 \) in the model \( \mathcal{P}_0 \) with limit law \( \mathcal{L}_{P_0}(Z) \), it must also be a "regular estimator with respect to \( S_{h_0} \)" of the parameter \( \nu \) at \( P_0 \) in the least-favorable model \( \mathcal{P}_{th_0} \).

By the Hajek-Le Cam Convolution Theorem we must have

\[ \mathcal{L}_{P_0}(Z) = \mathcal{L}_{P_0}(Z_{0h_0} + W) \]

where

\[ Z_{0h_0} \sim N(0, \| \Pi(\hat{\nu}(P_0)|\overline{T_m(P_0)})\|^2) \]
and \( W \) is independent of \( Z_{0h_0} \). We can in fact reach the same conclusion without assuming that \( T_m(P_0) \) is itself a tangent space.

In this section, we consider estimation of parameters in \( L_\infty(C) \). The ideas are essentially the same as those for the real line but the technicalities are more complicated. The reader is referred to Van der Vaart (1988) for a thorough exposition of these ideas.

While this theory is exactly what is needed for the constrained estimation problem where we observe \( X_1, \ldots, X_n \) i.i.d \( P_0 \in \mathcal{P}_{T,a} \), the supplementary sample moments problem does not fit into this framework since there the data is not i.i.d.. However the basic idea is the same. We begin by showing local asymptotic normality of the experiment in certain one-dimensional sub-models. Having done this it is an easy matter to get a convolution theorem for regular estimators in any one of these sub-models. From this we can then derive a convolution theorem for regular estimators in the whole model.

\( \S 4.2 \) Definitions. As a consequence of the fact that there is no "natural" choice for either the space where \( \nu_C \) takes its values, or, having chosen that space, for a \( \sigma \)-field defined on that space, there is also no "natural" definition of what constitutes an estimator sequence for \( \nu_C \). We present one possible definition for an estimator sequence here. In this definition we fix on a particular choice of range space for \( \nu_C \) and a particular choice of \( \sigma \)-field for that space. Not surprisingly, we make this choice to coincide with the choice made in Theorems 2 and 3 of section 3. We will then state a convolution theorem for sequences of "estimators" of \( \nu_C \) having certain (to be defined) regularity properties.

**Note.** We assume throughout the remainder of section 4 that \( C \) is a pointwise separable VC class of sets in \( B^d \) and \( E_{P_0} ||T||_b^2 < \infty \).

**Definition 4.** Suppose we observe \( X_n \in (A_n, \Omega_n) \). A sequence \( \{\nu_n(X_n)\} \equiv \{\nu_n\} \subset L_\infty(C) \) is said to be an \( (L_\infty(C), \mathcal{B}_C, P_0) \)-estimator sequence of \( \nu_C(P_0) \) if \( \nu_n \) is \( (A_n, \Omega_n) \mid (L_\infty(C), \mathcal{B}_C, P_0) \) measurable.
**Definition 5.** An \((L_\infty(C), \mathcal{B}_C, P_0)\)-estimator sequence, \(\{\nu_n\}\), of \(\nu_C(P_0)\) in the model \(\mathcal{P}_0\) is regular if for some maximal tangent space, \(T_m(P_0)\), to the model \(\mathcal{P}_0\) at \(P_0\), corresponding to each \(h \in T_m(P_0)\) there exists a sequence \(\{P_n\} \subset \mathcal{P}_{T,a}\) so that

\[(i) \quad \int \left[ \sqrt{n}((dP_n)^{1/2} - (dP_0)^{1/2}) - \frac{1}{2} h(dP_0)^{1/2} \right]^2 \to 0\]

\[(ii) \quad \left\| \sqrt{n}(\nu_C(P_n) - \nu_C(P_0)) - (1(\cdot), h) \right\|_{L_\infty(C)} \to 0\]

and

\[(iii) \quad \sqrt{n}(\nu_n - \nu_C(P_n)) \xrightarrow{P} \mathbb{Z}\]

as random elements of \((L_\infty(C), \mathcal{B}_C, P_0)\) where \(\mathbb{Z}\) is a tight random element of \(L_\infty(C)\) with sample paths in \(C(C, P_0)\).

In what follows, it will sometimes be necessary to specify the tangent space \(T_m(P_0)\) in Definition 5. We will do this by saying that \(\{\nu_n\}\) is regular with respect to the tangent space \(T_m(P_0)\).

**§4.3 Convolution Theorems for Estimators in the Constrained Estimation Problem**

We are now ready to state convolutions theorems for regular estimator sequences of \(\nu_C(P_0)\) in the model \(\mathcal{P}_{T,a}\).

**Theorem 4.** Suppose we observe \(X_1, \ldots, X_n\) i.i.d. \(P_0 \in \mathcal{P}_{T,a}\). Let \(\{\nu_n\}\) be an \((L_\infty(C), \mathcal{B}_C, P_0)\)-estimator sequence of \(\nu_C(P_0)\) in the model \(\mathcal{P}_{T,a}\). Suppose \(\{\nu_n\}\) is regular with limit law \(\mathcal{L}(\mathbb{Z})\). Then \(\mathbb{Z} \equiv \mathbb{Z}_{02} + \mathcal{W}\) where \(\mathbb{Z}_{02}\) is as defined in the statement of Corollary 1' and \(\mathcal{W}\) is a \(C(C, P_0)\)-valued tight random element of \(L_\infty(C)\), independent of \(\mathbb{Z}_{02}\).

The above theorems lead to the following definition for an efficient estimator of \(\nu_C(P_0)\) in the model \(\mathcal{P}_{T,a}\).

**Definition 6.** Suppose we observe \(X_1, \ldots, X_n\) i.i.d. \(P_0 \in \mathcal{P}_{T,a}\). Let \(\{\nu_n\}\) be an \((L_\infty(C), \mathcal{B}_C, P_0)\)-estimator sequence of \(\nu_C(P_0)\) in the model \(\mathcal{P}_{T,a}\). Suppose \(\{\nu_n\}\) is regular with limit law \(\mathcal{L}(\mathbb{Z})\). If

\[\mathbb{Z} \equiv \mathbb{Z}_{02}\]
where $Z_{02}$ is as defined in Corollary 1', we then say that $\{\nu_n\}$ is an $(L_\infty(C), B_{C, P_0})$-efficient estimator of $\nu_C(P_0)$.

**Corollary 3.** Let $C$ be a pointwise separable, VC class of sets in $B^d$. Suppose $0 \in \text{int } \Theta(P_0)$. Then $\nu(Q_n(a))$ is an $(L_\infty(C), B_{C, P_0})$-efficient estimator of $\nu_C(P_0)$ in the model $\mathcal{P}_{T, a}$.

Lemma 3 below is needed for the proof of Theorem 4.

**Lemma 3.** Define

$$T^*(P_0) = \{h \in L_2(P_0) : |h| < K, \text{some } K > 0, E_{P_0} h = 0, E_{P_0} h T = 0\}.$$  

Then $T^*(P_0)$ is a maximal tangent space at $P_0 \in \mathcal{P}_{T, a}$.

Note that (ii) of Definition 5 implies that for any $C \in C$, $\nu_C(P_0)(C)$ is differentiable with respect to some maximal tangent space, $T_m(P_0)$, and $\dot{\nu}_C(P_0)(C) = 1_C$. Since $T^*(P_0)$ is maximal we must have

$$T_m(P_0) = \{h \in L_2(P_0) : E_{P_0} h = 0, E_{P_0} h T = 0\}$$

for any maximal tangent space $T_m(P_0)$. A simple calculation shows that for any $A, B \in C$, any $a \equiv (a_1, a_2) \in \mathbb{R}^2$

$$a' \Gamma_{P_0}(A, B)a = \|\Pi(a_1 \dot{\nu}_C(P_0)(A) + a_2 \dot{\nu}_C(P_0)(B)|T_m(P_0))\|^2$$

with $\Gamma_{P_0}(\cdot, \cdot)$ as defined in (3.2) and $\dot{\nu}_C(P_0)(\cdot)$ as defined above. With this observation Theorem 4 follows from Theorem 5.3.1 of Bickel et al. (1989).

### §4.4 Convolution Theorems for Estimators in the Supplementary Sample Moments Experiment.

The results stated here are clumsier than those stated in §4.3 for the constrained estimation problem. Specifically, in order to get convolution theorems, we must make some hypotheses about the distribution of the random variable

$$\frac{1}{\sqrt{m}} \sum_{i=1}^m T(Y_i) \equiv \sqrt{m} T_m.$$  

We need to do this in order to get a Local Asymptotic Normality (LAN) result in one-dimensional sub-models. In Theorem 5 below we
hypothesize such an LAN result. In Lemma 4 we specify certain conditions on the pdf of $P_0 T^{-1}$ under which such an LAN result is true. Corollary 4 is a restatement of Theorem 5 with these conditions inserted in place of the LAN hypothesis. These conditions are, no doubt, much stronger than what are needed to get such results, see Le Cam and Yang(1988) for weaker (though harder to interpret) hypotheses.

As before, let $\mathcal{P}$ be the collection of all probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$.

**Theorem 5.** Suppose we observe $X_1, \ldots, X_n \in \mathbb{R}^d$, $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} T(Y_i) \in \mathbb{R}^b$ where $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are i.i.d. $P_0 \in \mathcal{P}$, $\frac{n}{m} \to \lambda$ as $n \wedge m \to \infty$, $T$ is some Borel measurable function. Let $T_m(P_0)$ be some maximal tangent space to the model $\mathcal{P}$ at $P_0$ and $\{\nu_n\}$ be some $(L_{\infty}(C), B_{C,P_0})$-estimator sequence of $\nu_C(P_0)$ with limit law $\mathcal{L}(\mathcal{Z})$ concentrated on $C(C,P_0)$. Suppose that for every $h \in T_m(P_0)$ there exists a sequence $\{P_{th}\} \subset \mathcal{P}$ for which

(i) (1) and (2) hold

(ii) $\{\nu_n\}$ is a regular $(L_{\infty}(C), B_{C,P_0})$-estimator sequence in the model $\{P_{th}\}$, and

(iii) if $f_{nm}^h$ denotes the pdf of the random variable $(X_1, \ldots, X_n, \frac{1}{\sqrt{m}} \sum_{i=1}^{m} T(Y_i))$

where $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are i.i.d. $P_{\frac{1}{\sqrt{m}} h}$ and $f_{nm}^0$ denotes the pdf of the same random variable where $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are i.i.d. $P_0$,

$$L_n\left(\frac{\theta}{\sqrt{n}}\right) - L_n(0)$$

$$= \log f_{nm}^h(X_1, \ldots, X_n, \sqrt{m} T_m) - \log f_{nm}^0(X_1, \ldots, X_n, \sqrt{m} T_m)$$

$$= \frac{\theta}{\sqrt{\lambda}} \sqrt{m} T_m'(\text{Var}_{P_0} T) - \text{Cov}_{P_0}(T, h) - \frac{1}{2} \frac{\theta^2}{\lambda} \text{Cov}_{P_0}(T, h)'(\text{Var}_{P_0} T)^{-1} \text{Cov}_{P_0}(T, h)$$

$$+ \frac{\theta}{\sqrt{n}} \sum_{i=1}^{n} h(X_i) - \frac{1}{2} \theta^2 E_{P_0} h^2 + R_n$$

where $R_n \to 0$ in $P_0$-probability.

Then

$$\mathcal{Z} \overset{d}{=} \mathcal{Z}_{\theta, \lambda} + W$$
where $\mathbb{Z}_{03}$ is a mean-zero, tight, Gaussian process with sample paths in $C(C, P_0)$ and covariance function $\Gamma^X_{P_0}(\cdot, \cdot)$ as defined in (3.3) and $W$ is a tight $C(C, P_0)$-valued random element of $(L_\infty(C), B_C, P_0)$, independent of $\mathbb{Z}_{03}$.

Before we state Corollary 4 we need to introduce some notation. Let

$$T^*_m(P_0) = \{h \in L_2(P_0) : |h| < K, \text{ some } K > 0 \text{ and } E_{P_0} h = 0\}.$$

It is an easy matter to show that $T^*_m(P_0)$ is a maximal tangent space for the model $\mathcal{P}$ (see, for example, the proof of Lemma 3). For any $h \in T^*_m(P_0)$ define

$$\mathcal{P}_h = \{P_{\theta h} \in \mathcal{P} : \frac{dP_{\theta h}}{dP_0} = 1 + \theta h, \text{ for all } \theta \text{ such that } |\theta h| < 1\}.$$

**Corollary 4.** Suppose we observe $X_1, \ldots, X_n \in \mathbb{R}^d$, $\frac{1}{\sqrt{m}} \sum_{i=1}^m T(Y_i) \in \mathbb{R}^b$ where $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are i.i.d. $P_0 \in \mathcal{P}$, $\frac{n}{m} \to \lambda$ as $n \wedge m \to \infty$ and $T$ is some Borel measurable function. Suppose that $P_0 T^{-1}$ has a bounded density with respect to Lebesgue measure and $E_{P_0} ||T||_b^3 < \infty$. Suppose that $\{\nu_n\}$ is an $(L_\infty(C), B_C, P_0)$-regular estimator sequence of $\nu_C(P_0)$ in the model $\mathcal{P}_h$ with limit law $L(\mathbb{Z})$ concentrated on $C(C, P_0)$, for each $h \in T^*_m(P_0)$. Then

$$\mathbb{Z} \overset{d}{=} \mathbb{Z}_{03} + W$$

where $\mathbb{Z}_{03}$ is a mean-zero, tight, Gaussian process with sample paths in $C(C, P_0)$ and covariance function $\Gamma^X_{P_0}(\cdot, \cdot)$ as defined in (3.3) and $W$ is a tight random element of $(L_\infty(C), B_C, P_0)$, independent of $\mathbb{Z}_{03}$.

**Definition 7.** Suppose we observe $X_1, \ldots, X_n \in \mathbb{R}^d$, $\frac{1}{\sqrt{m}} \sum_{i=1}^m T(Y_i) \in \mathbb{R}^b$ where $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are i.i.d. $P_0 \in \mathcal{P}$, $\frac{n}{m} \to \lambda$ as $n \wedge m \to \infty$, $T$ is some Borel measurable function. Let $\overline{T}_m(P_0)$ be some maximal tangent space to the model $\mathcal{P}$ at $P_0$. If $\{\nu_n\}$ is an $(L_\infty(C), B_C, P_0)$-estimator sequence for which corresponding to each $h \in \overline{T}_m(P_0)$, there exists a sequence $\{P_n\} \subset \mathcal{P}$ for which (i), (ii) and (iii) of Definition 5 hold and also (iii) of Theorem 5 holds with $\theta = 1$ and $P_{\frac{1}{\sqrt{n}}} \equiv P_n$. If
\{\nu_n\} has limit law \(L(\mathbb{Z})\) concentrated on \(C(C, P_0)\) we say that \(\{\nu_n\}\) is an efficient estimator if

\[\mathbb{Z} \overset{d}{=} \mathbb{Z}_{03}.\]

**Corollary 5.** Suppose we observe \(X_1, \ldots, X_n \in \mathbb{R}^d, \frac{1}{\sqrt{m}} \sum_{i=1}^m T(Y_i) \in \mathbb{R}^b\) where \(X_1, \ldots, X_n, Y_1, \ldots, Y_m\) are i.i.d. \(P_0 \in \mathcal{P}, \frac{n}{m} \to \lambda\) as \(n \wedge m \to \infty\) and \(T\) is some Borel measurable function. Suppose that \(P_0T^{-1}\) has a bounded density with respect to Lebesgue measure and \(0 \in \text{int } \Theta(P_0)\). Then \(\nu_c(Q_n(T_N))\) is an efficient estimator of \(\nu_c(P_0)\).

To prove Corollary 5 it is enough to show that the conditions of Definition 7 hold with \(T_m(P_0) = T^*_m(P_0)\) and \(\{P_n\} = \{P_{\sqrt{n}, h}\}\) where \(P_{\sqrt{n}, h}\) was defined in (4). With this choice of \(T_m(P_0)\) and \(\{P_n\}\), showing that (i) and (ii) of Definition 5 hold is easy. To show that (iii) of Definition 5 holds, see example 1 of Sheehy and Wellner (1988). Lemma 4 implies that (iii) of Theorem 5 holds and this will complete the proof.

Corollary 4 follows from Theorem 5 and Lemma 4 in which a Local Asymptotic Normality (LAN) result for the supplementary sample moments experiment is stated. Lemma 5—in which we get a convolution theorem for sequences of estimators of real-valued parameters in one-dimensional sub-models of \(\mathcal{P}\)—is needed in the proof of Theorem 5.

**Lemma 4.** Suppose \(P_0T^{-1}\) has a bounded density with respect to Lebesgue measure and \(E_{P_0}||T||_b^2 < \infty\). For a fixed \(h \in T^*_m(P_0)\) and some \(\theta \in \mathbb{R}^1\) let \(P_n = P_{\sqrt{n}, h}\) where \(P_{\sqrt{n}, h}\) was defined in (4). Let \(f_{nm}\) denote the pdf of the random variable \((X_1, \ldots, X_n, \frac{1}{\sqrt{m}} \sum_{i=1}^m T(Y_i))\) where \(X_1, \ldots, X_n, Y_1, \ldots, Y_m\) are i.i.d. \(P_0\), and \(f_{nm}^0\) denote the pdf of the same random variable where \(X_1, \ldots, X_n, Y_1, \ldots, Y_m\) are i.i.d. \(P_0\). Then,

\[L_n\left(\frac{\theta}{\sqrt{n}}\right) - L_n(0)\]
\(\log f_{nm}(X_1, \ldots, X_n, \sqrt{m}T_m) - \log f_{nm}^0(X_1, \ldots, X_n, \sqrt{m}T_m)\)
\[= \frac{\theta}{\sqrt{\lambda}} \sqrt{m}T_m' (\text{Var}_P T)^{-1} \text{Cov}_P(T, h) - \frac{1}{2} \frac{\theta^2}{\lambda} \text{Cov}_P(T, h)'(\text{Var}_P T)^{-1} \text{Cov}_P(T, h) + \frac{\theta}{\sqrt{n}} \sum_{i=1}^{n} h(X_i) - \frac{1}{2} \theta^2 E_P h^2 + R_n\]

where \(R_n \to 0\) in \(P_0\)-probability.

**Lemma 5.** Let \(\{\nu_n\}\) be a Borel-measurable estimator sequence of \(\nu(P_0) \in \mathbb{R}^1\) with limit law \(L(Z)\). Suppose \(h \in T(P_0)\), where \(T(P_0)\) is some tangent space to the model \(\mathcal{P}_0\) at \(P_0\). Suppose also that for this \(h\) there exists a sequence \(\{P_n\} \subset \mathcal{P}_{T,a}\) so that

(i)
\[
\int \left[ \sqrt{n}((dP_n)^{1/2} - (dP_0)^{1/2}) - \frac{1}{2} h(dP_0)^{1/2} \right]^2 \to 0
\]

(ii)
\[
|\sqrt{n}(\nu(P_n) - \nu(P_0)) - \langle \dot{\nu}(P_0), h \rangle| \to 0
\]

(iii)
\[
\sqrt{n}(\nu_n - \nu_c(P_n)) \xrightarrow{P} Z
\]

and

(iv) (iii) of Theorem 5 holds with \(\theta \equiv 1, P_{\frac{1}{\sqrt{n}}, h} \equiv P_n\) hold.

Then,
\[
Z \overset{d}{=} Z_{0h} + W_h
\]

where
\[
Z_{0h} \sim N(0, (\text{Cov}_P(\dot{\nu}(P_0), h))^2[E_P h^2 + \frac{1}{\lambda} \text{Cov}_P(T, h)'(\text{Var}_P T)^{-1} \text{Cov}_P(T, h)]^{-1}),
\]

and \(W_h\) is independent of \(Z_{0h}\).
§5 Proofs.

§5.1 Proofs of Results in Section 2.

Proof of Lemma 1. The proof given here is an adaptation of the proof of theorem 1 of Haberman (1984). Write \( K \equiv \text{cosupp} P_0 T^{-1} \) and \( K_n \equiv \text{cosupp} \mathbb{P}_n T^{-1} \). The result given in (C1) above implies that to prove existence of \( Q_n(a_n) \) with \( \mathbb{P}_n \)-density as given in (2.3), it is enough to show

\[(a) \quad a_n \in \text{relint} K_n, \text{with probability one, for all } n \text{ sufficiently large} \]

and

\[(b) \quad P^n_0 [u' T(X_i) = u' E P_0 T, \text{for all } u \in H, i = 1, \ldots, n] = 1. \]

where \( P^n_0 \) denotes the \( n \)-fold product measure.

We first show (a) holds. Since \( a \in \text{relint} K \), for some \( \epsilon > 0 \), \( B(a, \epsilon) \cap K \subseteq \text{relint} K \), where \( B(a, \epsilon) \), denotes the ball centered at \( a \) of radius \( \epsilon \) in \( \mathbb{R}^d \). This in turn is true if and only if, for some points \( t_j, 1 \leq j \leq c \), for any \( a_0 \in B(a, \epsilon) \cap K \), \( a_0 = \sum_{j=1}^{c} \alpha_j t_j \), where \( \sum_{j=1}^{c} \alpha_j = 1 \), \( \alpha_j > 0 \), \( 1 \leq j \leq c \), and for every open neighborhood \( U \) of a \( t_j, 1 \leq j \leq c \), \( P_0 T^{-1}(U) > 0 \). For some open \( U_j \), \( 1 \leq j \leq c \), one thus has \( P_0 T^{-1}(\bar{U}_j) > 0 \), where \( \bar{U}_j = U_j \cap K \) with \( a_0 \) in the relative interior of the convex hull of \( \{s_j : j = 1, \ldots, c\} \) for any \( \{s_j : j = 1, \ldots, c\} \), with \( s_j \in \bar{U}_j \), \( 1 \leq j \leq c \) and every \( a_0 \in B(a, \epsilon) \cap K \). Now use the fact that for any \( j = 1, \ldots, c \), the probability that no \( T(X_i) \in \bar{U}_j, i = 1, \ldots, n \), infinitely often, is zero and the fact that \( a_n \to_{a.s.} a \) to conclude that (a) is true.

We now show that (b) holds. Define

\[ A_u = \{x : u' T(x) = u' E P_0 T\}. \]

By definition of \( H \), if \( u \in H \) then \( P_0(A_u) = 1 \). Let \( \{u_n\}_{n \geq 1} \) be a dense subset of \( H \) and define \( A = \bigcap_{n=1}^{\infty} A_{u_n} \). If \( x \in A \), then \( u' T(x) = u' E P_0 T \) for all \( u \in H \). Clearly,
\( P_0(A) = 1. \) To show that when \((a)\) is true, we can assume with probability one \(\theta_n(a_n) \in H^{-1},\) it is then enough to show that

\[
(c) \quad P^n_0(\bigcap_{i=1}^n [X_i \in A]) = 1.
\]

However \((c)\) is clearly true and we are done. \[ \square \]

**Proof of Lemma 2.** This lemma can be proved using an argument similar to that used in the proof of Theorem 3 of Haberman(1984). \[ \square \]

**Proof of Theorem 1.** For ease of notation write \(Q_n \equiv Q_n(a_n), \ Q \equiv Q(a), \ \theta_n \equiv \theta_n(a_n)\) and \(\theta \equiv \theta(a).\) For any \(C \in \mathcal{C},\)

\[
\begin{align*}
(a) \quad (Q_n - Q)(C) \\
&= \frac{\int_C \exp(\theta'_n T) d\mathbb{P}_n}{\int \exp(\theta'_n T) d\mathbb{P}_n} - \frac{\int_C \exp(\theta' T) dP_0}{\int \exp(\theta' T) dP_0} \\
&= \frac{1}{\int \exp(\theta'_n T) d\mathbb{P}_n} \left[ \int [1 + \frac{\int_C \exp(\theta' T) dP_0}{\int \exp(\theta' T) d\mathbb{P}_n} - \exp(\theta' T) \right] d\mathbb{P}_n \\
&\quad + \frac{1}{\int \exp(\theta'_n T) d\mathbb{P}_n} \left[ \int (1 - \frac{\int \exp(\theta'_n T) d\mathbb{P}_n}{\int \exp(\theta' T) d\mathbb{P}_n}) \exp(\theta' T) d(\mathbb{P}_n - P_0) \right].
\end{align*}
\]

Lemma 7 (in section 6) shows that if

\[
(b) \quad \mathcal{F}_\delta = \{1_C e^{\theta' T}, e^{\theta' T} : \theta \in B_\delta(0), C \in \mathcal{C} \}
\]

then

\[
(c) \quad \mathcal{F}_\delta \text{ is a VC graph class of functions.}
\]

(See Definition 10 in section 6.) Clearly, if \( \mathcal{C} \) is pointwise separable, then \( \mathcal{F}_\delta \) is also. We use \((c)\) and Theorems II.24 and II.25 of Pollard(1984) and the fact that under the conditions of the theorem \(\theta_n \to \theta\) a.s. as \(n \to \infty\) to deduce that

\[
\sup_{C \in \mathbb{R}^d} |\int_C \exp(\theta'_n T) d\mathbb{P}_n - \int_C \exp(\theta' T) dP_0| = o(1) \quad \text{a.s.}
\]
and

\[ \sup_{C \subset \mathbb{R}^d} \left| \int_1^C \exp(\theta'T) d\mathbb{P}_n - \int_1^C \exp(\theta'T) d\mathbb{P}_0 \right| = o(1) \text{ a.s.} \]

We can then rewrite (a) in the following way,

\[
\begin{align*}
(Q_n - Q)(C) &= \frac{1}{\int \exp(\theta'T) d\mathbb{P}_0 + o(1)} \int (1 - Q(C) + o(1))(\exp(\theta'T_n) - \exp(\theta'T)) d\mathbb{P}_n \\
&\quad + \frac{1}{\int \exp(\theta'T) d\mathbb{P}_0 + o(1)} \int (1 - Q(C) + o(1)) \exp(\theta'T) d(\mathbb{P}_n - \mathbb{P}_0) \\
&= o(1) \text{ a.s. uniformly in } C \in \mathcal{C}. \quad \Box
\end{align*}
\]

§5.2 Proofs of Results in Section 3.

We begin by finding the asymptotic distribution of \( \theta_n(a) \) and \( \theta_n(T_N) \)

**Proposition 3.** Suppose \( a \in \text{relintcosupp}_0 T^{-1} \) and \( \Theta(P_0) \) is open. Suppose also that

\[ E_{P_0} \|T\|^2 \exp(2\theta(a)'T) < \infty. \]

Then

(i) \( \sqrt{n}(\theta_n(a) - \theta(a)) = -\int \text{Var}_{Q(a)} T^{-1} (T - a) \frac{dQ(a)}{dP_0} dX_n + o_P(1) - b_n \)

where \( b_n \in H \).

**Proposition 3'.** Suppose \( E_{P_0} T = a \) and \( 0 \in \text{int} \Theta(P_0) \). Then

(i) \( \sqrt{n}(\theta_n(a) = -\int \text{Var}_{P_0} T^{-1} (T - a) dX_n + o_P(1) - b_n \)

where \( b_n \in H \).

**Proposition 4.** Suppose \( 0 \in \text{int} \Theta(P_0), \frac{n}{m} \to \lambda \geq 0 \) and \( m \wedge n \to \infty \). Then

\[
\sqrt{n}(\theta_n(T_N) = (\text{Var}_{P_0} T)^{-1} \left\{ \int \frac{\sqrt{\lambda}}{1 + \lambda} (T - a) dX_n^2 - \frac{1}{1 + \lambda} \int (T - a) dX_n \right\} + o_P(1) - b_n
\]
where \( b_n \in H \).

**Proof of Proposition 3.** For simplicity of notation we write

\[
\theta_n(a) \equiv \theta_n, \quad \theta(a) \equiv \theta, \quad Q_n(a) \equiv Q_n, \quad Q(a) \equiv Q.
\]

By definition of \( \theta_n \) and \( \theta \), on a set of probability one, for \( n \) sufficiently large

(a)

\[
0 = \sqrt{n} \left( \frac{\int T \exp(\theta_n' T) d\Pi_n}{\int \exp(\theta_n' T) d\Pi_n} - \frac{\int T \exp(\theta' T) dP_0}{\int \exp(\theta' T) dP_0} \right)
\]

\[
= \frac{1}{\int \exp(\theta_n' T) d\Pi_n} \left[ \int \sqrt{n}(\exp(\theta_n' T) - \exp(\theta' T))^2 \left( T - \frac{\int T \exp(\theta' T) dP_0}{\int \exp(\theta' T) dP_0} \right) d\Pi_n \right]
\]

\[
+ \frac{1}{\int \exp(\theta_n' T) d\Pi_n} \left[ \int \left( T - a \frac{\int \exp(\theta_n' T) d\Pi_n}{\int \exp(\theta' T) dP_0} \right) \exp(\theta' T) dX_n \right].
\]

Use a Taylor series expansion to write

(b)

\[
\sqrt{n}(\exp(\theta_n' T) - \exp(\theta' T)) = \sqrt{n}(\theta_n - \theta)' T \exp(\theta' T) + \frac{1}{2} \sqrt{n}(\theta_n - \theta)' TT'(\theta_n - \theta) \exp(\theta' T)
\]

where \( ||\theta^* - \theta|| \leq ||\theta_n - \theta|| \). Use the fact that \( \theta \in \text{int} \Theta(P_0) \) and \( \theta_n \to_{a.s.} \theta \) as \( n \to \infty \) to get

(c)

\[
\int \exp(\theta_n' T) d\Pi_n = \int \exp(\theta' T) dP_0 + o_P(1)
\]

and

(d)

\[
\int \exp(\theta' T) d\Pi_n = \int \exp(\theta' T) dP_0 + o_P(1).
\]

Putting (a), (b), (c) and (d) together we get

(e)

\[
0 = \left( \frac{1}{\int \exp(\theta' T) dP_0} + o_P(1) \right) \left[ \int (T - a + o_P(1))' \exp(\theta' T) dP_0 + o_P(1) \right] \sqrt{n}(\theta_n - \theta)
\]

\[
+ \left( \frac{1}{\int \exp(\theta' T) dP_0} + o_P(1) \right) \sqrt{n}(\theta_n - \theta) o_P(1)
\]

\[
+ \left[ \int (T - a) \frac{dQ}{dP_0} dX_n \right] + o_P(1)
\]

\[
= [\text{Var}_Q T + o_P(1)] \sqrt{n}(\theta_n - \theta) + \sqrt{n}(\theta_n - \theta) o_P(1)
\]

\[
+ \left[ \int (T - a) \frac{dQ}{dP_0} dX_n \right] + o_P(1).
\]
We see from (e) that it must be true that

\[(f) \quad \| \sqrt{n}(\theta_n - \theta) \| = O_P(1).\]

and so

\[
\text{Var}_Q T \sqrt{n}(\theta_n - \theta) = - \int (T - a) \frac{dQ}{dP_0} d\mathbb{X}_n + o_P(1)
\]

or

\[(g) \quad \sqrt{n}(\theta_n - \theta) = - \int (\text{Var}_Q T)^{-1} (T - a) \frac{dQ}{dP_0} d\mathbb{X}_n + o_P(1) - b_n\]

where the \(b_n\) term is the projection of the first two terms on the right hand side of (g) onto \(H\). We include it to ensure that the quantity \(\sqrt{n}(\theta_n - \theta_0) \in H^\perp\).

Proof of Proposition 3'. The proof is very similar to that of Proposition 1 and so will be omitted.

Proof of Proposition 4. The proof is very similar to that of Proposition 1. We give an outline here. For simplicity of notation we write

\[
\theta_n(T_N) = \theta_n, \quad Q_n(T_N) = Q_n.
\]

By definition of \(\theta_n\), on a set of probability one, for \(n\) sufficiently large

\[(a) \quad \sqrt{n}(T_N - a) = \sqrt{n} \left( \frac{\int T \exp(\theta'_n T) dP_n}{\int \exp(\theta'_n T) dP_n} - \int T dP_0 \right) = [\text{Var}_P \hat{T} + o_P(1)] \sqrt{n} \theta_n + o_P(1) \sqrt{n} \theta_n + \int (T - a) d\mathbb{X}_n + o_P(1).
\]

The left hand side of (a) can be written as

\[
\sqrt{n}(T_N - a) = \frac{n}{n + m} \int (T - a) d\mathbb{X}_n + \frac{m}{n + m} \sqrt{\frac{n}{m}} \int (T - a) d\mathbb{X}_m
\]

\[(b) \quad = \frac{\lambda}{1 + \lambda} \int (T - a) d\mathbb{X}_n + \frac{\sqrt{\lambda}}{1 + \lambda} \int (T - a) d\mathbb{X}_m + o_P(1).
\]

\[= O_P(1).\]
From (a) and (b) we see that

\[ \sqrt{n}||\theta_n|| = o_P(1) \]

and so

\[ \frac{\lambda}{1+\lambda} \int (T-a)dXX_n + \frac{\sqrt{\lambda}}{1+\lambda} \int (T-a)dXX_m = \sqrt{n}\theta_n \text{Var}_P T + \int (T-a)dXX_n + o_P(1). \]

So we can write

\( c \)

\[ \sqrt{n}\theta_n = (\text{Var}_P T)^{-} \left\{ \frac{\sqrt{\lambda}}{1+\lambda} \int (T-a)dXX_m - \frac{1}{1+\lambda} \int (T-a)dXX_n \right\} + o_P(1) - b_n \]

where \( b_n \) is the projection onto \( H \) of the first two terms on the left hand side of (c). We include this term to ensure \( \sqrt{n}\theta_n \in H^\perp. \)

**Proof of Theorem 2.** For simplicity of notation we write \( Q_n(a) \equiv Q_n, \ Q(a) \equiv Q, \ \theta_n(a) \equiv \theta_n \) and \( \theta(a) \equiv \theta. \) Let \( C \in \mathcal{C}. \) Then on a set of probability one, for \( n \) sufficiently large,

\( a \)

\[ \sqrt{n}(Q_n(C) - Q(C)) = \frac{1}{\int \exp(\theta_n T) d\Pi_n} \left[ \int \frac{1}{\int \exp(\theta T) d\Pi_n} \left( \frac{1}{\exp(\theta_n T) - \exp(\theta T)} \sqrt{n}(\exp(\theta_n T) - \exp(\theta T)) d\Pi_n \right) \right] \]

\[ + \frac{1}{\int \exp(\theta_n T) d\Pi_n} \left[ \int (1 - Q(C)) \frac{\exp(\theta_n T) d\Pi_n}{\exp(\theta T) d\Pi_n} \exp(\theta T) dXX_n \right]. \]

Arguing as in the proof of Proposition 1 yields

\( a \)

\( b \)

\( \frac{dQ}{dP_0} \)

\[ (1C - Q(C)) \frac{dQ}{dP_0} dXX_n + o_P(1) \int (1C - (Q(C)) \frac{dQ}{dP_0} dXX_n + o_P(1). \]

where the \( o_P(1) \) terms are uniformly small in \( \mathcal{C}. \) We now show that the third term on the right hand side of (b) is \( O_P(1). \) Define

\[ \mathcal{F} = \{(1C - Q(C)) \frac{dQ}{dP_0} : C \in \mathcal{C}\}. \]
Theorem 10 of Pollard (1982) shows that $\mathcal{F}$ is a sparse class of functions (see Definition 11 in §6). Corollary 13 of Pollard (1982) implies then that $\mathcal{F}$ is totally bounded in $L_2(P_0)$. It is easy to see that $\mathcal{C}$ being pointwise separable implies $\mathcal{F}$ is pointwise separable. In the proof of Theorem VII.21 of Pollard (1984) it is shown that for any $\epsilon > 0$, we can find a finite collection of open balls in $L_\infty(\mathcal{F})$, each with finite radius, centered at points in $C(\mathcal{F}, P_0)$, such that the probability that $X_n$ is in the union of those balls, call it $B_\epsilon$, is greater than $1 - \epsilon$. Since $\mathcal{F}$ is totally bounded, each element of $C(\mathcal{F}, P_0)$ must be bounded. So for some $M > 0$ we must have that $\varphi \in B_\epsilon$ implies $||\varphi||_{L_\infty(\mathcal{F})} < M$. From this we may conclude that

$$o_P(1) \int (1 - Q(C)) \frac{dQ}{dP_0} dX_n = o_P(1).$$

Thus, from Proposition 3 and (b), using the fact that $b'_n T = b'_n E_{P_0} T$, a constant, with probability one, we get

$$\sqrt{n}(Q_n(C) - Q(C)) = \int [(1 - Q(C)) - (\text{Cov}_Q(1_C, T)'(\text{Var}_Q T)^{-}(T - a)] \frac{dQ}{dP_0} dX_n + o_P(1)$$

where the $o_P(1)$ term is uniformly small over $\mathcal{C}$.

Proof of Theorem 2′. The proof is exactly analogous to that of Theorem 1 and so is omitted.

Proof of Theorem 3. For simplicity of notation we write $Q_n(T_N) \equiv Q_n$, $\theta_n(T_N) = \theta_n$. As in the proof of Theorem 2 we can show that on a set of probability one, for $n$ sufficiently large, for any $C \in \mathcal{C}$,

$$\sqrt{n}(Q_n(C) - P_0(C)) = \text{Cov}_{P_0}(1_C, T)'\sqrt{n}\theta_n + \int (1 - Q(C)) dX_n + o_P(1)$$

where the $o_P(1)$ term is uniformly small in $\mathcal{C}$. Using the fact that $b'_n T = b'_n a$, a
constant a.s. \( P_0 \), we get from Proposition 4,

\[
\sqrt{n}(Q_n(C) - P_0(C)) = \int \frac{\sqrt{\lambda}}{1 + \lambda} \text{Cov}_{P_0}(1_C, T)(\text{Var}_{P_0}T)^{-1}(T - a)dX^2_m
\]

\[
+ \int [(1_C - P_0(C)) - \frac{1}{1 + \lambda} \text{Cov}_{P_0}(1_C, T)(\text{Var}_{P_0}T)^{-1}(T - a)dX_n
+ o_P(1)
\]

where the \( o_P(1) \) term is uniformly small in \( C \).

**Proof of Corollary 1.** Write \( Q(a) \equiv Q \). Define

\[
\mathcal{F} = \{(1_C - Q(a)(C)) - \text{Cov}_Q(1_C, T)(\text{Var}_Q T)^{-1}(T - a)\frac{dQ}{dP_0} : C \in C\}
\]

and

\[
f(\cdot, C) = 1_C - Q(C)) - \text{Cov}_Q(1_C, T)(\text{Var}_Q T)^{-1}(T - a)\frac{dQ}{dP}.
\]

Theorem 10 of Pollard (1982) implies \( \mathcal{F} \) is a sparse collection of functions under the hypotheses of the Corollary. It is easy then to see that \( \mathcal{F}, C \) and \( f(\cdot, C) \) satisfy the conditions of Lemma 6 (in §6) and so we can conclude that

\[
\mathbb{Z}_n \Rightarrow \mathbb{Z}_0
\]

as random elements of \((L_\infty(C), B_C, P_0)\), where \( \mathbb{Z}(\cdot) \) is as described in the statement.

The proofs of Corollaries 1' and 2 are very similar to that of Corollary 1 and so are not included here.

§5.3 Proofs of Results in Section 4.

**Proof of Lemma 3.** Obviously \( T^*(P_0) \) is a subspace of \( L_2(P_0) \). We first show that \( \overline{T^*(P_0)} \) contains every tangent space. That is, if \( T(P_0) \) is a tangent space then

\[
T(P_0) \subset \overline{T^*(P_0)} = \{ h \in L_2(P_0) : \int h dP_0 = 0, \int hT dP_0 = 0 \}.
\]
Suppose $h \in T(P_0)$. Then for some $\{P_t\} \subset \mathcal{P}_{T,a}$ we must have

$$(a) \quad \int [t^{-1}((dP_t)^{1/2} - (dP_0)^{1/2}) - \frac{1}{2} h(dP_0)^{1/2}]^2 \to 0 \text{ as } t \to 0.$$ 

We show that

$$\int h dP_0 = 0 \text{ and } \int Th dP_0 = 0.$$ 

Write

$$f_{tt} = \frac{dP_t}{d\mu_t} \text{ and } f_t = \frac{dP_0}{d\mu_t}$$

where $\mu_t$ is some measure dominating $P_t$ and $P_0$, and

$$A = \{x : f_t(x) > 0\}.$$

Then

$$| - \int h dP_0| = | \int [t^{-1}(f_{tt} - f_t) - hf_t] d\mu_t|$$

$$\leq | \int t^{-1}f_{tt} 1_A d\mu_t| + | \int f_t^{1/2} \left( t^{-1} \left( \frac{f_{tt}}{f_t^{1/2}} - f_t^{1/2} \right) - \frac{1}{2} hf_t^{1/2} \right) 1_A d\mu_t|$$

$$\leq | \int t^{-1}f_{tt} 1_A d\mu_t| + | \int f_t^{1/2} \left[ t^{-1} \left( \frac{f_{tt}}{f_t^{1/2}} - f_t^{1/2} \right) - \frac{1}{2} hf_t^{1/2} \right] 1_A d\mu_t|$$

$$+ | \int f_t^{1/2} (t^{-1}(f_{tt}^{1/2} - f_t^{1/2}) - \frac{1}{2} hf_t^{1/2}) 1_A d\mu_t|$$

$$\leq | \int t^{-1}f_{tt} 1_A d\mu_t| + | \int f_t^{1/2} [t^{-1}(f_{tt}^{1/2} - f_t^{1/2}) - \frac{1}{2} hf_t^{1/2}] 1_A d\mu_t|$$

$$+ | \int \frac{1}{2} hf_t^{1/2} (f_t^{1/2} - f_{tt}^{1/2}) d\mu_t|$$

$$+ | \int f_t^{1/2} (t^{-1}(f_{tt}^{1/2} - f_t^{1/2}) - \frac{1}{2} hf_t^{1/2}) 1_A d\mu_t|$$

$$= o(1)$$

by (1) and repeated application of the Cauchy-Schwarz Inequality.

We now show $\int Th dP_0 = 0$ must hold. Again choose $h \in T(P_0)$ and a sequence $\{P_t\} \subset \mathcal{P}_{T,a}$ as defined in (a). We show that we can find a sequence $\{P_t^*\} \subset \mathcal{P}_{T,a}$ for which (a) holds (with $\{P_t^*\}$ in place of $\{P_t\}$) and for which
\[ \int |T|^2 dP_t^* \leq 2 \int |T|^2 dP_0 \] for all \( t \geq 0 \). Then, using an argument similar to that used above to show \( \int h dP_0 = 0 \), it is easy to show that \( \int T h dP_0 = 0 \) and we will be done.

Define, for \( K > 0 \)

\[ A_K = \{ x : |h(x)| < K \}, \]

and

\[ h_K(x) = h(x) 1_{A_K/2}(x) - E_{P_0} h 1_{A_K/2}. \]

Notice,

\[ |h_K(x)| < K \text{ and } E_{P_0} h_K(X) = 0. \]

Define \( \{ P_t^* \} \) by

\[ dP_t^* = (1 + th_{1/t}) dP_0, \quad t > 0. \]

Then, for \( t > 0 \)

\[ \int \|T\|^2 dP_t^* \leq \int \|T\|^2 dP_0 + \int \|T\|^2 dP_0. \]

A straightforward argument can be used to show

\[ \int t^{-1}((dP_t^*)^{1/2} - (dP_0)^{1/2}) - \frac{1}{2} h(dP_0)^{1/2}]^2 \to 0 \text{ as } t \to 0. \]

To show that \( T^*(P_0) \) is itself a tangent space, we need to show that if \( h \in T^*(P_0) \), there exists \( \{ P_t \} \subset \mathcal{P}_{T,a} \) so that

\[ \int t^{-1}((dP_t)^{1/2} - (dP_0)^{1/2}) - \frac{1}{2} h(dP_0)^{1/2}]^2 \to 0 \text{ as } t \to 0. \]

It is easy to show that the sequence \( \{ P_t \} \) defined by \( \frac{dP_t}{dP_0} = 1 + th \) will do and we are done.

**Proof of Corollary 3.** We need to show regularity of the estimator \( \nu_C(Q_n(a)) \). The proof of this can be found in example 1 of Sheehy and Wellner(1988).

**Notation:** For any \( \{ C_1, \ldots, C_k \} \subset L_\infty(C) \), any \( \varphi \in L_\infty(C) \), define

\[ (\varphi(C_1), \ldots, \varphi(C_k)) \equiv \varphi(C_1, \ldots, C_k) \]
Proof of Theorem 5. Under the hypotheses of the theorem, Lemma 5 implies that for each \( \{C_1, \ldots, C_k\} \subset C\), \( k \geq 1 \) and any \( a \in \mathbb{R}^k \),

\[(a) \quad a'Z(C_1, \ldots, C_k) \overset{d}{=} a'Z_{0h}(C_1, \ldots, C_k) + a'W_h(C_1, \ldots, C_k)\]

where, if \( \Gamma_h(C_1, \ldots, C_k) \) is the \( k \times k \) matrix whose \( ij \)th element is given by

\[\gamma_{ij} = \text{Cov}_{P_0}(1_{C_i}, h)\text{Cov}_{P_0}(1_{C_j}, h)[E P_0 h^2 + \frac{1}{\lambda} \text{Cov}_{P_0}(T, h)'(\text{Var}_{P_0} T)\text{Cov}_{P_0}(T, h)]^{-1}\]

then

\[Z_{0h}(C_1, \ldots, C_k) \sim N(0, \Gamma_h(C_1, \ldots, C_k)),\]

and \( W_h(C_1, \ldots, C_k) \) is independent of \( Z_0(C_1, \ldots, C_k) \). To see this, in the statement of Lemma 5, write

\[\nu(P_0) = \sum_{i=1}^{k} a_i \nu(C_0)(C_i) \quad \text{and} \quad \dot{\nu}(P_0) = \sum_{i=1}^{k} a_i \dot{1}_{C_i}.\]

Define

\[h_a = \sum_{i=1}^{k} a_i(1_{C_i} - P_0(C_i) - \frac{1}{1+\lambda}\text{Cov}_{P_0}(1_{C_i}, T)(\text{Var}_{P_0} T)^{-1}(T - E P_0 T)).\]

Then \( h_a \in \overline{T_m(P_0)} \) and so there exists a sequence \( \{h_{na}\} \subset T_m(P_0) \) with \( \|h_{na} - h_a\|_{P_0} \to 0 \) as \( n \to \infty \). It can be shown that

\[(b) \quad a'Z_{0h_{na}}(C_1, \ldots, C_k) \Rightarrow a'Z_{03}(C_1, \ldots, C_k)\]

and

\[(c) \quad a'W_{h_{na}}(C_1, \ldots, C_k) \Rightarrow a'W(C_1, \ldots, C_k)\]

where \( Z_{03}(C_1, \ldots, C_k) \) and \( W(C_1, \ldots, C_k) \) are as defined in the statement of the theorem. From \( a) \), \( b) \) and \( c) \) we can thus conclude that

\[Z(C_1, \ldots, C_k) \overset{d}{=} Z_{03}(C_1, \ldots, C_k) + W(C_1, \ldots, C_k)\]
with $\mathbb{Z}_{03}$ and $W$ as defined in the statement. Since $\mathbb{Z}$ and $\mathbb{Z}_{03}$ exist as $C(C, P_0)$-valued tight random elements of $L_\infty(C)$, it remains only to identify $W(C_1, \ldots, C_k)$ as $(W(C_1), \ldots, W(C_k))$ where $W$ is a $C(C, P_0)$-valued tight random element of $L_\infty(C)$. Lemma 8 in section 6 makes such an identification.

Proof of Lemma 4. We will only give an outline of the proof. To prove the lemma is true, we break the expression in $(i)$ into two parts, one involving the loglikelihood of the $X_i$'s and the other involving the loglikelihood of $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} T(Y_i)$. The first part is straightforward to deal with. For the second part we must show that the loglikelihood can be approximated by the log of the pdf of a multivariate normal distribution with mean $E_{P_n} T$ and covariance matrix $\text{Var}_{P_n} T$ evaluated at $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} T(Y_i)$. To show that this can be done we use the fact that under the hypotheses of the theorem the pdf of $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} T(Y_i)$, under $P_n$ can be uniformly approximated by the pdf of the multivariate normal distribution with mean $E_{P_n} T$ and covariance matrix $\text{Var}_{P_n} T$. (This can be proved using an argument similar to that used to prove Theorem XV.5.2, Feller, Vol.II). Since $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} T(Y_i)$ is $O_{P_0}(1)$ the result follows.

Proof of Lemma 5. The proof is similar to that of the Hajek-Le Cam Convolution Theorem, see Theorem 3.1 (Roussas(1972)).

§6 Technicalities.

Definition 8. Let $F$ be a collection of functions in $L_2(P_0)$. We say that $F$ is pointwise separable if there exists a countable subclass $F^*$ of $F$ such that for any $f \in F$ there exists a sequence $\{f_n\} \subset F^*$ with $f_n(x) \rightarrow f(x)$ for all $x \in \mathbb{R}^d$.

Definition 9. (Vapnik and Červonenkis(1971)) Call a class $C$ of subsets of $\mathbb{R}^d$ a VC class of sets of degree $v$ if each set $S$ of $v$ points in $\mathbb{R}^d$ has fewer than $2^v$ distinct subsets of the form $C \cap S$ picked out by members of $C$.

Definition 10. Call a class $F$ of real-valued functions on $\mathbb{R}^d$ a VC graph class if 

$$
C_F = \{ C_f : C_f = \{(x,t) : 0 \leq t \leq f(x)\}, f \in F \} 
$$
is a VC class of sets in $\mathbb{R}^{d+1}$.

**Definition 11.** (Pollard(1982)) Let $\mathcal{F}$ be a collection of functions with envelope function $F$ (i.e. $|f| < F$, for every $f \in \mathcal{F}$). For each finite subset $S$ of $\mathbb{R}^d$ and each positive $\delta$, take $N_F(\delta, S, \mathcal{F})$ to be the smallest value of $m$ for which there are functions $\varphi_1, \ldots, \varphi_m$ in $\mathcal{F}$ such that

$$\min_i \sum_{x \in S} [f(x) - \varphi_i(x)]^2 \leq \delta^2 \sum_{x \in S} F(x)^2$$

for every $f$ in $\mathcal{F}$. Define

$$H(\delta) = H_F(\delta, \mathcal{F}) = \sup_S \log N_F(\delta, S, \mathcal{F})$$

the supremum running over all finite subsets of $\mathbb{R}^d$. If

$$\sum_{j=1}^{\infty} 2^{-j} H_F(2^{-j}, \mathcal{F}) < \infty$$

we say that $\mathcal{F}$ is sparse.

**Lemma 6.** Let $\mathcal{C}$ be a separable (in $L_2(P_0)$) collection of sets in $\mathbb{R}^d$. Let $f : \mathbb{R}^d \times \mathcal{C} \to \mathbb{R}^1$ be a uniformly continuous function from $\mathcal{C}$ into $L_2(P_0)$, i.e. given any $\epsilon > 0$ there exists a $\delta > 0$ such that $\|1_{C_1} - 1_{C_2}\|_{L_2(P_0)} < \delta$ implies $\|f(\cdot, C_1) - f(\cdot, C_2)\|_{L_2(P_0)} < \epsilon$. Define

$$\mathcal{F} = \{f(\cdot, C) : C \in \mathcal{C}\}.$$

Suppose $\mathcal{F}$ is a sparse, pointwise separable collection of functions with envelope $F \in L_2(P_0)$. Define $\mathcal{Z}_n \in L_\infty(\mathcal{C})$ by

$$\mathcal{Z}_n(C) = \mathcal{X}_n \circ f(\cdot, C)$$

where $\mathcal{X}_n$ was defined in (3.1). Then,

$$\mathcal{Z}_n \Rightarrow \mathcal{Z}_0$$
as random elements of $(L_\infty(C), B_{C,P_0})$ where $\mathbb{Z}_0$ is a tight, Gaussian, random element of $L_\infty(C)$ whose sample paths all belong to $C(C, P_0)$.

**Proof.** In Theorem VII.21 of Pollard(1984) he shows that under the above hypotheses $\mathbb{X}_n \Rightarrow \mathbb{X}_0$ as random elements of $(L_\infty(\mathcal{F}), B_{\mathcal{F},P_0})$ where $\mathbb{X}_0$ is a tight, Gaussian, random element of $L_\infty(\mathcal{F})$ whose sample paths all belong to $C(\mathcal{F}, P_0)$. The proof proceeds in the following way:-

He shows that

(i) $\mathbb{X}_n$ is a random element of $(L_\infty(\mathcal{F}), B_{\mathcal{F},P_0})$ and

(ii) Given any $\varepsilon > 0$, there exists a compact set, $K$, of completely regular points such that for every open set, $G$, containing $K$, for which $[\mathbb{X}_n \in G]$ is measurable, $\liminf \Pr[\mathbb{X}_n \in G] > 1 - \varepsilon$. Specifically, $K$ is constructed as an intersection of sets $D_1, D_2, \ldots$ where $D_k$ is a finite union of closed balls of radius $\frac{1}{k}$ centered at points of $C(\mathcal{F}, P_0)$.

(Note that $\varphi \in L_\infty(\mathcal{F})$ is a completely regular point, with respect to a $\sigma$-field $B$, if all open balls centered at $\varphi$ are contained in $B$.)

It is not difficult to see that we can adapt the proof of this theorem to show $\mathbb{Z}_n \Rightarrow \mathbb{Z}_0$ as random elements of $(L_\infty(C), B_{C,P_0})$. In fact, we need to show that (i) and (ii) are true with $\mathbb{Z}_n$ in place of $\mathbb{X}_n$, $B_{C,P_0}$ in place of $B_{\mathcal{F},P_0}$ and $C(C, P_0)$ in place of $C(\mathcal{F}, P_0)$.

Showing (ii) is true is easy. Just notice that $f(\cdot, C)$ being uniformly continuous over $\mathcal{C}$ implies $C(\mathcal{F}, P_0) \subset C(C, P_0)$. We now show that $\mathbb{Z}_n$ is a random element of $(L_\infty(C), B_{C,P_0})$.

Because each

$$\mathbb{Z}_n(C) = \mathbb{X}_n \circ f(\cdot, C) = \frac{1}{n} \sum_{i=1}^{n} [f(X_i, C) - E_{P_0} f(X_i, C)]$$

is a real random variable, the finite dimensional projections create no difficulty. It remains then to show that for any $\varphi \in C(C, P_0)$
\( (b) \quad \sup_C |Z_n(C) - \varphi(C)| = \sup_C |X_n \circ f(\cdot, C) - \varphi(C)| \)

is measurable. However, the fact that \( \mathcal{F} \) is pointwise-separable with envelope function \( F \in L_2(P_0) \) and \( \mathcal{C} \) is separable (in \( L_2(P_0) \)) implies

\[ \sup_{\mathcal{C}} |Z_n(C) - \varphi(C)| = \sup_{\mathcal{C}^*} |Z_n(C) - \varphi(C)| \]

where \( \mathcal{C}^* \) is a countable collection of sets in \( \mathcal{C} \). Measurability of the quantity in (b) is thus clear and we are done.

The following lemma is used in the proof of Theorem 1.

**Lemma 7.** Suppose \( \mathcal{C} \) is a VC class of sets in \( \mathbb{R}^d \). Then

\[ \mathcal{F} = \{ e^{\theta^T A} : A \in \mathcal{C}, \theta \in B_\delta(\theta_0) \} \]

is a VC graph class.

**Proof.** We need to show that \( \mathcal{F} \) is a VC graph class, i.e. we need to show

\[ \gamma = \{(x, t) : 0 \leq t \leq e^{\theta^T(x)} 1_A(x), A \in \mathcal{C}, \theta \in B_\delta(\theta_0) \} \]

is a VC class of sets. A typical set in \( \gamma \) is of the form

\[ \{(x, t) : t > 0, \log t \leq \theta^T(x), x \in A \} \cup \{(x, 0) : x \in A^c\} \]

\[ = C_{A, \theta} \cup A^c \times \{0\}, \quad A \in \mathcal{C}, \theta \in B_\delta(\theta_0). \]

Let

\[ \gamma_1 = \{ A^c \times \{0\} : A \in \mathcal{C} \} \]

and let

\[ \gamma_2 = \{ C_{A, \theta} : \theta \in B_\delta(\theta_0), A \in \mathcal{C} \}. \]
Theorem 9.2.6 and (9.1.8) of Dudley (1982) implies $\gamma_1$ is a VC class of sets. Define

$$g : \mathbb{R}^d \times \mathbb{R}^+ \rightarrow \mathbb{R}^b \times \mathbb{R}^1 \text{ by } g(x, t) = (T(x), \log t).$$

Then

$$C_{A, \theta} = \{(x, t) : t > 0, \log t \leq \theta' T(x)\} \cap A \otimes \mathbb{R}^+$$

$$= g^{-1}(\Delta_{\theta}) \cap A \otimes \mathbb{R}^+$$

where

$$\Delta_{\theta} = \{(r, s) \in \mathbb{R}^{d+1} : (r, s) \cdot (\theta, -1) > 0\}.$$

Theorem 9.2.2 of Dudley (1982) implies that

$$\gamma_3 = \{g^{-1}(\Delta_{\theta}) : \theta \in B_{\delta}(\theta_0)\}$$

is a VC class of sets, since $\{\Delta_{\theta} : \theta \in B_{\delta}(\theta_0)\}$ is one. Theorem 9.2.6 of Dudley (1982) implies that

$$\gamma_4 = \{A \otimes \mathbb{R}^+ : A \in \mathcal{C}\}$$

is a VC class of sets. We then conclude, again from Theorem 9.2.6 of Dudley (1982), that

$$\gamma = \{B_3 \cap B_4 \cup B_1 : B_1 \in \gamma_1, B_3 \in \gamma_3, B_4 \in \gamma_4\}$$

is a VC class of sets and the lemma is proved.

The proof of the following Lemma is contained in the proof of Theorem 5.3.1, Bickel et al. (1989). We include it here for completeness.

**Lemma 8.** Let $\mathcal{Z}$ and $\mathcal{Z}_0$ be tight random elements of $L_\infty(\mathcal{C})$ with sample paths in $C(\mathcal{C}, P_0)$, where $\mathcal{C}$ is a totally bounded subset in $L_2(P_0)$. Suppose for every $\{C_1, \ldots, C_k\} \subset \mathcal{C}$, $k \geq 1$,

(i) $$\mathcal{Z}(C_1, \ldots, C_k) \overset{d}{=} \mathcal{Z}_0(C_1, \ldots, C_k) + \mathcal{W}(C_1, \ldots, C_k)$$

where $\mathcal{W}(C_1, \ldots, C_k)$ is independent of $\mathcal{Z}_0(C_1, \ldots, C_k)$. Then

$$\mathcal{Z} \overset{d}{=} \mathcal{Z}_0 + \mathcal{W}$$
where \( W \) is a tight, random element of \( L_\infty(C) \) with sample paths in \( C(C, P_0) \) and \( W \) is independent of \( Z_0 \).

**Proof.** Since \( Z \) and \( Z_0 \) exist as \( C(C, P_0) \)-valued random elements of \( L_\infty(C) \), it remains only to identify \( W(C_1, \ldots, C_k) \) as \((W(C_1), \ldots, W(C_k))\) where \( W \) is a \( C(C, P_0) \)-valued random elements of \( L_\infty(C) \).

Let \( k \geq 1 \). Since \( C \) is totally bounded there exists a finite \( \frac{1}{2k} \) net for \( C \), say \( \{C_1, \ldots, C_m\} \) with \( m = m(k) \). Thus for each \( C \in C \) there is a \( C^* \in \{C_1, \ldots, C_m\} \) such that \( ||C - C^*||_{L_2(P_0)} < \frac{1}{2k} \). Define \( \Delta_i : C \to \mathbb{R}^1 \) by

\[
\Delta_i(C) = \frac{v_i(C)}{\sum_{j=1}^{m} v_j(C)} \quad i = 1, \ldots, m
\]

Each \( \Delta_i \) is a uniformly continuous function that vanishes off the ball of radius \( \frac{1}{k} \) about \( C_i \). For every \( C \in C \) there is at least one \( C_i \) so that \( v_i(C) > 1/2 \). Thus the \( \Delta_i \) are non-negative uniformly continuous functions that sum to 1 everywhere in \( C \).

Now define interpolation functions \( Z_m \) and \( Z_{0m} \) of \( Z \) and \( Z_0 \) by

\[
(b) \quad Z_m(C) \equiv \sum_{i=1}^{m} \Delta_i(C)Z(C_i), \quad C \in C
\]

\[
(c) \quad Z_{0m}(C) \equiv \sum_{i=1}^{m} \Delta_i(C)Z(C_i), \quad C \in C
\]

and use the random vector \( W(C_1, \ldots, C_m) \equiv (W_1, \ldots, W_m)' \) to define a process \( W_m \) by

\[
(d) \quad W_m(C) = \sum_{i=1}^{m} \Delta_i(C)W_i.
\]

Denote the laws of these processes on \((C(C, P_0), B_{C, P_0})\) by \( \tilde{P}_m \equiv \mathcal{L}(Z_m) \), \( \tilde{P}_{0m} \equiv \mathcal{L}(Z_{0m}) \) and \( \tilde{R}_m \equiv \mathcal{L}(W_m) \).
It follows from (i) and linearity of the interpolated processes that

\[ \mathcal{L}(\mathcal{Z}_m) = \mathcal{L}(\mathcal{Z}_{0m} + \mathcal{W}_m). \]

Now since \( \tilde{P}_m \Rightarrow \tilde{P} \equiv \mathcal{L}(\mathcal{Z}) \) and \( \tilde{P}_{0m} \Rightarrow \tilde{P}_0 \equiv \mathcal{L}(\mathcal{Z}_0) \), given any \( \epsilon > 0 \) there exist compact sets \( K_1 \equiv K_1(\epsilon) \) and \( K_2 \equiv K_2(\epsilon) \) such that \( \liminf_m \tilde{P}_m(K_1) \geq 1 - \epsilon \) and \( \liminf_m \tilde{P}_{0m}(K_2) \geq 1 - \epsilon \), respectively. Hence, for \( m \geq \) some \( m(\epsilon) \)

\[
1 - \epsilon \leq \tilde{P}_m(K_1) = \int \tilde{R}_m(K_1 - x) d\tilde{P}_{0m}(x) \leq \int_{K_2} \tilde{R}_m(K_1 - x) d\tilde{P}_{0m}(x) + \epsilon
\]
or

\[
1 - 2\epsilon \leq \int_{K_2} \tilde{R}_m(K_1 - x) d\tilde{P}_{0m}(x).
\]
Thus \( \tilde{R}_m(K_1 - x) \geq 1 - 2\epsilon \) for some sequence \( \{x_m\} \subset K_2 \). Hence the compact set

\[ K_1 - K_2 \equiv \{K_1 - x : x \in K_2\} \supset \{K_1 - \{x_m\}\} \]

has \( \tilde{R}_m(K_1 - K_2) \geq 1 - 2\epsilon \) for \( m \geq m(\epsilon) \) and it follows that \( \liminf_m \tilde{R}_m(K_1 - K_2) \geq 1 - 2\epsilon \). Hence \( \tilde{R}_m \) is tight and there exists a weakly convergent subsequence \( \{\tilde{R}_{m'}\} \), \( \tilde{R}_{m'} \Rightarrow \tilde{R} \) on \( (C(\mathcal{C}, \mathcal{P}_0), \mathcal{B}_{\mathcal{C}, \mathcal{P}_0}) \). Let \( \mathcal{L}(\mathcal{W}) = \tilde{R} \) and the conclusion of the Lemma follows.

Acknowledgments. Part of this research was carried out at the University of Washington in Seattle as part of the requirements for a Ph.D. degree in statistics. I would like to thank Professor Jon Wellner for his help and encouragement during the course of this research. Thanks are also due to Aad van der Waart with whom I had some discussions about the results in section 4.
References


