Journal of Statistics: Advances in Theory and Applications
Volume 9, Number 2, 2013, Pages 105-133

CONVOLUTION REPRESENTATION IN PRACTICE

AO YUAN and QIZHAI LI

Howard University
Washington D. C. 20059
USA
e-mail: yuanao@hotmail.com

Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190
P. R. China

Abstract

The convolution theorem (Hájek [8]) characterizes the weak limit of any regular estimator as a convolution of two independent components. One is an optimal achievable part and another is a noise. Therefore, the optimal estimator is one without the noise part in its weak limit, which is a deeper characterization than the Cramer-Rao bound. However, this result is derived under the assumption that the specified model is the true one generating the data. In practice, any subjectively specified model is more or less deviated from the true one. The convolution representation (and the Cramer-Rao bound) should be modified to reflect this fact. Here, we study such modifications for the estimation of parameters under several cases: Euclidean parameter, Euclidean parameter with side information; Euclidean parameter with infinite-dimensional nuisance parameter; and the case of infinite-dimensional parameter. In each case, we decompose the weak limit of a regular estimator into three independent components, with one achievable optimal part, and two noise parts. When the specified model is indeed the true one, it reduces to existing convolution representation of two components.

2010 Mathematics Subject Classification: Primary 62G05; Secondary 62G20.

Keywords and phrases: convolution theorem, Cramer-Rao bound, efficient estimator, Euclidean parameter, infinite-dimensional parameter, nuisance parameter, optimal weak limit, side information.

Received June 18, 2013

1. Introduction

Let $f_0(\cdot)$ be the true density function generating the observed data, it may not necessarily be a member from some parametric family. Define the parameter $\theta_0 = G(f_0)$ for some known functional $G(\cdot)$. Since $f_0(\cdot)$ is unknown, in practice, a parametric model $f(\cdot|\theta)$ from some known parametric family is often specified as an approximate model to analyze the data.

If it happens that the model $f(\cdot|\theta_0)$ coincides with the true model $f_0(\cdot)$ at the parameter θ_0, then it is well-known that the maximum likelihood estimate (MLE) $\hat{\theta}_n$ of θ_0 will almost surely (a.s.) converge to

$$\arg\sup_{\theta\in\Theta}\int f_0(x)\log f(x|\theta)dx = \arg\sup_{\theta\in\Theta}\int f(x|\theta_0)\log f(x|\theta)dx,$$

which is achieved by θ_0, the true parameter.

On the other hand, if the model $f(\cdot|\theta)$ does not coincide with $f_0(\cdot)$ for some θ, it is known (Huber [10]; Pfanzagl [22]) that the MLE from the parametric model will a.s. converge to the pseudo-true parameter set Θ_1,

$$\Theta_1 = \arg\sup_{\theta\in\Theta}\int f_0(x)\log f(x|\theta)dx.$$

The points in Θ_1 may not necessarily correspond to the "true" parameter(s) generating the data. Similarly, in the Bayesian setting, if the wrong likelihood model is specified, the posterior will asymptotically concentrate on Θ_1 (Berk [4]).

However, estimators of θ_0 based on $f(\cdot|\theta)$ can still be consistent even if $f(\cdot|\theta_0) \neq f_0(\cdot)$, for example, if $(\cdot|\theta)$ and $f_0(\cdot)$ have the same mode. We are only interested in this case, as comparing inconsistent estimators is meaningless. Let $I(\theta_0)$ be the Fisher information (matrix) and $\xrightarrow{D}$ stand for convergence in distribution. Assume an estimate θ_n be consistent and

asymptotically normal, i.e., $\sqrt{n}(\theta_n - \theta_0) \xrightarrow{D} N(\mathbf{0}, \Omega)$. The Cramer-Rao theorem asserts that $\Omega \geq I^{-1}(\theta_0)$, here "$\geq$" is in the semi-definite matrix sense, and any estimator that achieve this lower bound is an efficient estimator. Under general conditions, the MLE and Bayes estimate are efficient estimators. The convolution theorem (Hájek [8]), based on the assumption that $f(\cdot|\theta)$ is the correct model for the observed data, states that for any regular estimator T_n with weak limit W, there are random variables Z and V such that

$$W = Z \oplus V, \quad Z \sim N(0, I^{-1}(\theta_0)),$$

where $Z \oplus V$ means independent summation of Z and V. Here, we see that Z is the optimal weak limit and V is undesirable noise. Inagaki [12] discovered a similar result as the above, in the same year, under considerably stronger conditions.

The Cramer-Rao theorem gives the achievable lower bound $I^{-1}(\theta_0)$ of the asymptotic variance of any asymptotically unbiased estimators. The convolution theorem further characterizes the achievable optimal weak limit of a regular estimator: It is the normal random variable Z with mean zero and variance $I^{-1}(\theta_0)$. An estimator is efficient iff its weak limit $W = Z$ or equivalently $V = 0$. The convolution theorem has had profound impact and generated considerable interest in the statistical field, and different versions of it (van der Vaart [32]; Pfanzagl [23]) and generalizations to infinite-dimensional parameters have been proposed (for example, Millar [21]; Schick and Susarla [27]; LeCam [19]; Beran [3]; Janssen and Ostrovski [13]), and in the Bayesian framework (van den Heuvel and Klassen [31]; Sen [28]). But to our knowledge, all these results are derived under the assumption that the model $f(\cdot|\theta_0)$ is the true one generating the data. This assumption is unlikely to be the case in practice, as any subjectively specified parametric model is more or less biased from the true one. Thus, these classical results should to be modified to reflect the model uncertainty, which is the motivation of this study.

We consider several cases, first the inference of Euclidean parameter(s), without and with side information on the specified model, then with nuisance parameter(s), and lastly the case of infinite-dimensional parameter. We show that generally in each case, the convolution representation has three independent components, which reduce to two components only if the assumed model is the true one generating the data. Thus, any parameter estimates based on the postulated model has bigger variation than that based on the existing result, and the modified Cramer-Rao lower bound is no smaller than the inverse Fisher information; it equals the latter only if the model happens to be the correct one.

2. Results

We first consider the case of estimation of Euclidean parameter(s) in the specified model, then the case with side information, with nuisance parameter, and the case of infinite-dimensional parameter. Let $X_1, \ldots, X_n, X$ be i.i.d. with $f_0(\cdot)$, which is unknown. In practice, the investigator often subjectively specifies it by a parametric model $f(\cdot|\theta)$ as a member from some known parametric family, with $\theta = (\theta_1, \ldots, \theta_d)'$ a d-dimensional parameter. Let $l_f(x|\theta) = \log f(x|\theta)$, $\dot{l}_f(x|\theta) = (\partial/\partial\theta) l_f(x|\theta)$, $\ddot{l}_f(x|\theta) = (\partial^2/\partial\theta\partial\theta') l_f(x|\theta)$, $L_n(\theta) = \sum_{i=1}^n l_f(X_i|\theta)$, $\dot{L}_n(\theta) = \sum_{i=1}^n \dot{l}_f(X_i|\theta)$, and $\ddot{L}_n(\theta) = \sum_{i=1}^n \ddot{l}_f(X_i|\theta)$. Let

$$\mathcal{F} = \{f_j(\cdot|\theta) : f_j(\cdot|\theta) \text{ be a density for each } \theta, \text{ and } f_j(\cdot|\theta_0) = f_0(\cdot),\ j \in J\}$$

be the class of all parametric densities, which pass through $f_0(\cdot)$ at θ_0. It is the class of all possible 'true' models with parameter θ for the observed data. Note that $f(\cdot|\theta)$ is a member of $\mathcal{F}$ only if it is a correctly specified parametric model of the data. Let $I_f(\theta) = -E_{f_0}[\ddot{l}_f(X|\theta)]$ and

$$f_*(\cdot|\theta) = \arg \max_{f_j(\cdot|\theta)\in\mathcal{F}} I_{f_j}(\theta_0)$$

be the 'most favorable' true model for the observed data, which is unknown. The max here is in the sense of matrix positive definiteness.

In our study, we assume the following conditions:

(C1) $l_f(\cdot|\theta)$ is twice differentiable with respect to θ.

(C2) $\int f_0(x)\dot{l}_f(x|\theta_0)dx = 0$.

(C3) $I_f(\theta) < \infty$ is non-singular in a neighbourhood of θ_0.

For the parametric model $f(\cdot|\theta)$, we will see that $I_f^{-1}(\theta_0)$ is the *effective information bound*, instead of the inverse Fisher information $I^{-1}(\theta_0)$, the classical information bound, where $I(\theta) = -E_{f(\cdot|\theta)}[\ddot{l}_f(X|\theta)]$. It is known that $I_f^{-1}(\theta_0) \geq I^{-1}(\theta_0)$ with "=" iff $f(\cdot|\theta_0) = f_0(\cdot)$ (Serfling [29], p.257).

When the estimator $\hat{\theta}_n$ is the MLE of θ_0 based on $f(\cdot|\theta)$,

$$\hat{\theta}_n - \theta_0 = -[\ddot{L}_n(\theta_n)]^{-1}\dot{L}_n(\theta_0),$$

where θ_n is an intermediate point between $\hat{\theta}_n$ and θ_0. So under (C1) and (C3) (and some further conditions on $\ddot{L}_n(\cdot)$),

$$n^{-1}\dot{L}_n(\theta_0) \xrightarrow{\text{a.s.}} \int f_0(x)\dot{l}_f(x|\theta_0)dx, \quad -n^{-1}\ddot{L}_n(\theta_n) \xrightarrow{\text{a.s.}} I_f(\theta_0),$$

thus $\hat{\theta}_n$ is asymptotically unbiased if and only if (C2) holds. Thus, (C2) seems necessary for many estimators based on $f(\cdot|\theta)$ to be asymptotically unbiased. Note in this case,

$$I_f(\theta) = -\int f_0(x)\frac{\ddot{f}(x|\theta)}{f(x|\theta)}dx + \int f_0(x)\dot{l}_f(x|\theta)\dot{l}_f'(x|\theta)dx$$

$$\neq E_{f_0}\left(\dot{l}_f(X|\theta)\dot{l}_f'(X|\theta)\right),$$

which differs from the classical result $I(\theta) = -E_{f(\cdot|\theta)}(\ddot{l}_f(X|\theta)) = E_{f(\cdot|\theta)}\left(\dot{l}_f(X|\theta)\dot{l}_f'(X|\theta)\right)$.

From now on, let $\theta_n = \theta_0 + n^{-1/2}b$ for some $b \in C$, the complex plane. A rate $n^{1/2}$ consistent estimator $T_n = T_n(X_1, \ldots, X_n)$ is said to be *regular*, if under $f(\cdot|\theta_n)$, $W_n := \sqrt{n}(T_n - \theta_n) \xrightarrow{D} W$ for some random variable W, and the result does not depend on the sequence $\{\theta_n\}$. Let $Z \oplus V$ denote the summation of two independent random variables Z and V; and $I(\theta)$ be the Fisher information for $f(\cdot|\theta)$ at θ. The convolution theorem (Hájek [8]), based on the assumption that $f(\cdot|\theta)$ is the correct model for the observed data, states that for any regular estimator T_n with weak limit W, there is a random variable V such that

$$W = Z \oplus V, \quad Z \sim N(0, I^{-1}(\theta_0)).$$

The Cramer-Rao theorem gives the lower bound of the asymptotic variance of any asymptotically unbiased estimators. The convolution theorem further characterizes the weak limit of an asymptotically optimal estimator: It is a normal random variable with mean zero and variance $I^{-1}(\theta_0)$. An estimator is efficient iff $V = 0$. Since any subjectively specified model is more or less deviated from the true one, below we modify this convolution result under the possibly wrong model $f(\cdot|\theta)$ specification. In some cases, the convergence rate of Euclidean or infinite-dimensional parameters can be different from $\sqrt{n}$. For example, for distributions with singularity of order α, the convergence rate of Euclidean parameter in the model is $r_n = n^{1/(1+\alpha)}$, $-1 < \alpha < 1 (\alpha \neq 0)$.

In this case, the local parameter is defined as $\theta_n = \theta_0 + r_n^{-1}b$, and the local likelihood ratio is often asymptotically non-normal, see Ibragimov and Has'minskii [11]. For Euclidean parameter takes only finite number of possible values, the convergence rate r_n is exponential (for example, Hammersley [9]; Robson [26]). Convergence rate r_n of infinite-dimensional parameters is often slower than $\sqrt{n}$. In these cases, the weak limit of $r_n(\theta_n - \theta_0)$ is often non-Gaussian, and the how to find specific form of the optimal weak limit in convolution representation, if exists, seems still open.

We first give a modification of the existing convolution theorem under the specified model (not necessarily the true one), for the case of Euclidean parameter.

Theorem 1. *Assume* (C1)-(C3). *Then for any rate* $\sqrt{n}$ *consistent regular estimator* $T_n(X_1, \ldots, X_n)$ *based on the model* $f(\cdot|\theta)$, *with weak limit* $W = \lim_n \sqrt{n}(T_n - \theta_n)$, *we have*

$$W = Z \oplus V \oplus U, \quad Z \sim N(0, I_{f_*}^{-1}(\theta_0)), \quad V \sim N(0, \Omega^{-1}(\theta_0)),$$

$$\Omega^{-1}(\theta_0) = I_{f_*}^{-1}(\theta_0)[(I_{f_*}(\theta_0) - I_f(\theta_0))^{-1} - I_{f_*}^{-1}(\theta_0)]^{-1} I_{f_*}^{-1}(\theta_0).$$

Remark 1. The noise U is due to the fact that the estimator T_n is not optimal based on the model $f(\cdot|\theta)$, thus for a regular estimator of θ_0 based on the given model, its optimal achievable weak limit is $Z \oplus V$, and the corresponding modified Cramer-Rao lower bound on asymptotic variance of any asymptotically unbiased estimator is $I_{f_*}^{-1}(\theta_0) + \Omega^{-1}(\theta_0)$ $= I_f^{-1}(\theta_0) \geq I^{-1}(\theta_0)$, with "=" iff $f(\cdot|\theta_0) = f_0(\cdot)$. Super-efficiency may happen, under some conditions, at some θ, in that there are some estimator, whose asymptotic variance can be smaller than $I^{-1}(\theta)$ at these θ, but all such points at most constitute a Lebesque null set

(LeCam [15]). The noise U is due to the deviation of model $f(\cdot|\theta_0)$ to the optimal true data generating model $f_*(\cdot)$. $V = 0$ iff $f(\cdot|\theta) = f_*(\cdot|\theta)$, then the optimal weak limit is Z, and we get the original convolution theorem.

Remark 2. When $\sqrt{n}(T_n - \theta_n)$ is asymptotically linear, the noise U can be further characterized. In this case, by the general central limit theorem (see Araujo and Giné [1]), the weak limit of $\sqrt{n}(T_n - \theta_n)$ must be of the form $N(a, \sigma^2) \oplus \delta \oplus Pios(\mu)$, where δ is some point mass, and $Pois(\mu)$ is a generalized Poisson distribution corresponding to a Lévy measure μ. Thus, we must have $\sigma^2 \geq I_{f_*}^{-1}(\theta_0) + \Omega^{-1}(\theta_0)$, and let $\sigma_0^2 = \sigma^2 - I_{f_*}^{-1}(\theta_0) - \Omega^{-1}(\theta_0)$, we have $U = N(0, \sigma_0^2) \oplus \delta \oplus Pios(\mu)$.

Droste and Wefelmeyer [7] derived Hájek's convolution representation with technically weaker conditions than regularity. A further almost everywhere version of this representation without the regularity condition by the works of a number of authors, was stated in Beran ([3], Theorem 2.3). We conjecture that Theorem 1 and the results below are still valid under these weaker conditions, but we will not persue them here for succinctness. LeCam [17] and Janssen and Ostrovski [13] generalized the convolution theorem to the case in which the optimal weak limit need not be Gaussian, and in the infinite-dimensional parameter space. Jeganathan [14] studied the case the optimal weak limit is mixed normal.

Let Θ_1 be as given in the Introduction. In the case $\Theta_1 = \{\theta_*\}$ has a single point, White [34] showed that

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{d} N(\mathbf{0}, C(\theta_*)),$$

where $C(\theta) = A^{-1}(\theta)B(\theta)A^{-1}(\theta)$, $A(\theta) = E_{f_0}(\ddot{l}_f(X|\theta))$, $B(\theta) = E_{f_0}(\dot{l}_f(X|\theta)\dot{l}_f'(X|\theta))$. Thus, when $f(\cdot|\theta) \notin \mathcal{F}$, the MLE based on it may not be efficient in the sense of $U = 0$ in Theorem 1.

With side information. In some cases, there is an additional information about the parameter, often be summarized by $E_f[g(X, \theta_0)] = 0$ for some known function g, or vector of functions. Incorporating such information often leads to improved accuracy of inference (for example, Qin and Lawless [25]; Xu and Wang [35]). Let $f(\cdot|\theta, g)$ be the density with the side information incorporated, and $l(\cdot|\theta, g)$ be the corresponding log-likelihood, though generally the forms may not be known. We need the following conditions:

(C4) $\Lambda := E_f[g(X, \theta_0)g'(X, \theta_0)] < \infty$ is invertible, and $A(\theta_0, g) := E_f[\dot{l}(X|\theta_0)g(X, \theta_0)] < \infty$.

(C5) $E_{P_0}[f^{-1}(X|\theta_0, g)\ddot{f}(X|\theta_0, g)] = 0$.

Theorem 2. *Under* (C1)-(C5). *Then for any rate* $\sqrt{n}$ *consistent regular estimator* $T_n(X_1, \ldots, X_n)$ *based on the model* $f(\cdot|\theta)$ *and side information* $E_f[g(X, \theta_0)] = 0$, *with weak limit* $W = \lim_n \sqrt{n}(T_n - \theta_n)$, *we have*

$$W = Z \oplus V \oplus U, \quad Z \sim N(0, I_{f_*}^{-1}(\theta_0|g)), \quad V \sim N(0, \Omega^{-1}(\theta_0|g)),$$

where $\Omega^{-1}(\theta_0|g) = I_{f_*}^{-1}(\theta_0|g)[(I_{f_*}(\theta_0|g) - I_f(\theta_0|g))^{-1} - I_{f_*}^{-1}(\theta_0|g)]^{-1} I_{f_*|g}^{-1}(\theta_0)$; $I_f(\theta_0|g) = E_{P_0}[e_f(X|\theta_0, g)e'(X|\theta_0, g)]$, $e_f(x|\theta_0, g) = I_f^{-1}(\theta_0)\dot{l}(X|\theta_0) - A'(\theta_0, g)\Lambda^{-1}g(x, \theta_0)$.

The noise U is due to the fact that the estimator T_n is not optimal based on the model $f(\cdot|\theta)$, thus for a regular estimator of θ_0 based on the given model, its optimal achievable weak limit is $Z \oplus V$, and the corresponding modified Cramer-Rao lower bound on asymptotic variance of any asymptotically unbiased estimator is $I_{f_*}^{-1}(\theta_0) + \Omega^{-1}(\theta_0) = I_f^{-1}(\theta_0) \geq I^{-1}(\theta_0)$, with "=" iff $f(\cdot|\theta_0) = f_0(\cdot)$. The noise V is due to the deviation of model $f(\cdot|\theta_0)$ from the optimal true data generating model $f_*(\cdot)$. $V = 0$ iff $f(\cdot|\theta) = f_*(\cdot|\theta)$, in which case the optimal weak limit is Z.

With nuisance parameters. Now, we consider the case there is a nuisance parameter g in the model $f(\cdot|\theta, g)$. We assume g is in a Banach space **B**, which include the Euclidean parameter as special case. To simplify presentation, we assume g has one component, the result for multi-components case is parallel. We need the score in this case, let $l_f(\theta, g) = \log f(x|\theta, g)$. There are several commonly used derivatives including Gâteaux, Hadamard (compactly), pathwise, Hellinger, and Fréchet differentials. Often, Gâteaux differentiability is too weak (even discontinuous functional can be Gâteaux differentiable), Fréchet differential is too strong (many commonly used statistical functionals do not have Fréchet differentiability), and Hadamard differential is stronger than Gâteaux and weaker than Fréchet and is considered appropriate to use in most statistical problems. The Hellinger differential is a special form of the Fréchet differential, pathwise differential is a special form of Hadamard differential and is often used for semi-and-nonparametric models. When all these differentials exist, they are all equal (except that the Hellinger differential and the score only differ by a factor of $f^{1/2}/2$), so their differences are only the existences of these differentials. The Hellinger differential on $f^{1/2}$ is used in many articles, instead of Hadamard differential on $l(x|\theta, g)$, to deal with differentials for **B**-valued parameters, it has the advantage that the referred quantities are automatically in $L^2(P_0)$ with norm 1. But for higher order differentials, this advantage is not obvious. In this article, we use the Hadamard differential for **B**-valued parameters, and assume the existence for all referred quantities. Let P_0 be the probability distribution for the 'true' model f_0; $L_2(P_0)$ be the Hilbert space of all functions h with $\|h\|_{P_0}^2 = \int h^2(t)P_0(dt) < \infty$ and define the inner product $< h, g >_{P_0} := \int h(t)g(t)P_0(dt) (h, g \in L_2(P_0))$. Let $< h, g > = \int h(t)g(t)dt$ and $\|h\|^2 = < h, h >, \forall h, g \in \mathbf{B}$. For fixed x and θ and g, let $l_f^{(\cdot)}(x|\theta, g)$ be the ordinary partial derivative of $l_f(x|\theta, g)$ with respect to its first

component θ, and we adopt the following version of definition of the Hadamard differential (Bickel et al. [5], p.454) $l_f^{(\cdot,\cdot)}(\cdot|\theta, g; h) : \mathbf{B} \to L^2(P_0)$ of $l_f(x|\theta, g)$ with respect to g in the direction $h \in \mathbf{B}$: for all compact subset S of $\mathbf{B}$,

$$\frac{l_f(x|\theta, g+\epsilon h) - l_f(x|\theta, g) - l_f^{(\cdot,\cdot)}(x|\theta, g; \epsilon h)}{\epsilon} \to 0, \text{ as } \epsilon \to 0 \text{ uniformly in } h \in S.$$

We can define the second order differentials $l_f^{(\cdot,\cdot)}(x|\theta, g)$, $l_f^{(\cdot,\cdot)}(x|\theta, g; h)$ as the Hadamard differential on $l_f^{(\cdot,\cdot)}(x|\theta, g)$ with respect to g, and $l_f^{(\cdot,\cdot)}(x|\theta, g; h_1, h_2)$ be that of $l_f^{(\cdot,\cdot)}(x|\theta, g, h_1)$ with respect to g in the direction h_2, the latter is a bi-linear operator: $\mathbf{B} \times \mathbf{B} \to L^2(P_0)$. When these second order differentials exist, we say $l_f(\cdot|\theta, g)$ is twice differentiable.

For fixed θ and g, $l_f^{(\cdot,\cdot)}(\cdot|\theta, g, h) : \mathbf{B} \to L^2(P_0)$ is a linear operator. Now, let $f_n(\cdot) = f(\cdot|\theta_n, g_n)$ be the local model, with $\theta_n = \theta_0 + n^{-1/2}b$ and $g_n(\cdot) = g(\cdot) + n^{-1/2}h(\cdot)$, for some $g, h \in \mathbf{H}$. Let

$$\tilde{\mathcal{F}} = \{f_j(\cdot|\theta, g) : f_j(\cdot|\theta, g) \text{ be a density for each } \theta \in R, g \in \mathbf{H},$$

and $f_j(\cdot|\theta_0, g) = f_0(\cdot)$, for some $g, j \in J\}$ be the class of all parametric densities pass through $f_0(\cdot)$ at (θ_0, g). To simplify notation, let $\rho_f(x) = l_f^{(\cdot,\cdot)}(x|\theta_0, g)$, $A_f(x, h) = l_f^{(\cdot,\cdot)}(x|\theta_0, g, h)$, and $\alpha_f(x; b, g, h) = b\rho_f(x) + A_f(x, h)$. Note for fixed x, $A_f(x, \cdot) : \mathbf{B} \to L^2(P_0)$ is a linear operator. By the projection theorem in Luenberger ([20], p.59) and the assumption $L^2(P_0) \subset \mathbf{B}$, there is an $h_f^* \in L^2(P_0)$ such that

$$\rho_f(\cdot) - A_f(\cdot, h_f^*) \perp_{P_0} A_f(\cdot, h), \quad \forall h \in L^2(P_0).$$

Let $(A_f)^* : L^2(P_0) \to \mathbf{B}$ be the adjoint of A_f, it is determined by $< h, (A_f)^* g > = < Ah, g >_{P_0}$, $\forall h \in \mathbf{B}$, $\forall g \in L^2(P_0)$. When $(A_f)^* A_f$ is invertible, $h_f^* = ((A_f)^* A_f)^{-1} (A_f)^* \rho_f$. Let

$$\beta_f(x; \theta, b, g, h) = b^2 l_f^{(\cdot,\cdot)}(x|\theta, g) + 2b l_f^{(\cdot,\cdot)}(x|\theta, g; h) + l_f^{(\cdot,\cdot)}(x|\theta, g; h, h),$$

$$I_f(\theta_0, g, h_f^*) = - E_{P_0} \beta(X; \theta_0, 1, g, - h_f^*),$$

and

$$f_*(\cdot|\theta, g) = \arg \max_{f_j(\cdot|\theta, g, h_{f_j}^*) \in \widetilde{\mathcal{F}}} I_{f_j}(\theta_0, g, h_{f_j}^*).$$

Assume the following conditions:

(C6) $l(\cdot|\theta, g)$ is twice differentiable with respect to (θ, g).

(C7) $\int f_0(x) l_f^{(\cdot,)}(x|\theta_0, g) dx = \int f_0(x) l_f^{(\cdot,\cdot)}(x|\theta_0, g; h) dx = 0$, $\forall h \in \mathbf{B}$.

(C8) $I_f(\theta, g, h_f^*) < \infty$ is non-singular in a neighbourhood of (θ_0, g).

Theorem 3. *Assume* (C6)-(C8) *and that* $L^2(P_0) \subset \mathbf{B}$. *Then for any rate* $\sqrt{n}$ *consistent regular estimator* $T_n(X_1, \ldots, X_n)$ *of* θ_0 *based on the model* $f(\cdot|\theta, g)$ ($g \in \mathbf{B}$ *is a nuisance parameter*), *with weak limit* $W = \lim_n \sqrt{n}(T_n - \theta_n)$, *we have*

$$W = Z \oplus V \oplus U, \quad Z \sim N(0, I_{f_*}^{-1}(\theta_0, g, h_f^*)), \quad V \sim N(0, \Omega^{-1}(\theta_0, g)),$$

$$\Omega^{-1}(\theta_0, g) = I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*)[(I_{f_*}(\theta_0, g, h_{f_*}^*) - I_f(\theta_0, g, h_f^*))^{-1} - I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*)]^{-1}$$

$$\times I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*).$$

As in the cases before, the noise U is due to the fact that the estimator T_n is not optimal based on the model $f(\cdot|\theta, g)$, thus for a regular

estimator of θ_0 based on the given model, its achievable optimal weak limit is $Z \oplus V$, and the corresponding modified Cramer-Rao lower bound on asymptotic variance of any asymptotically unbiased estimator, in the presence of nuisance parameter $g \in \mathbf{H}$, is $I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*) + \Omega^{-1}(\theta_0, g)$ $= I_f^{-1}(\theta_0, g, h_f^*) \geq I^{-1}(\theta_0, g, h^*) := E_{f(\cdot|\theta_0, g)}\alpha_f^2(X; 1, g, -h_f^*)$, with "=" iff $f(\cdot|\theta_0, g) = f_0(\cdot)$. The noise V is due to the deviation of model $f(\cdot|\theta_0, g)$ to the optimal true data generating model $f_*(\cdot|\theta, g)$. $V = 0$ iff $f(\cdot|\theta, g) = f_*(\cdot|\theta, g)$, and then we get the result in Begun et al. [2].

Infinite dimensional parameter. Now, we consider estimation of g in the model $f(\cdot|g)$, $g \in \mathbf{B}$. For fixed g, let $\dot{l}_f(\cdot|g; h) : \mathbf{B} \to L^2(P_0)$ be the Hadamard differential of $l_f(\cdot|g)$ in the direction $h \in \mathbf{B}$. For conciseness, we concentrate on the case g has only one component. If $g = (g_1, \ldots, g_k)'$ with each $g_j \in \mathbf{B}$, $\dot{l}_f(x|g, h)$ will be $\mathbf{B}^k \to \mathbf{B}^k$, and the results will be parallel, but the presentation and notations will be more involved. In that case $h = (h_1, \ldots, h_k)'$, and $\dot{l}_f(x|g, h) = (\dot{l}_f(x|g_1, h_1), \ldots, \dot{l}_f(x|g_k, h_k))'$.

Let $\mathbf{B}^*$ be the dual space of $\mathbf{B}$, and $\forall b^* \in \mathbf{B}^*$ and $\forall b \in \mathbf{B}$, denote $b^*(b)$ the value of b^* at b. When there is an inner product $< \cdot, \cdot >_{\mathbf{B}}$ on $\mathbf{B}$, by the Rize representation, corresponds to this inner product, there is a unique element $b_* \in \mathbf{B}$ such that $b^*(b) =< b, b_* >_{\mathbf{B}}$ for all $b \in \mathbf{B}$. In fact, let $< a, c >= ac$ be the inner product in Euclidean space, then $b^*(b) =$ $< b^*(b), 1 >=< b, b^{*\top}1 >_{\mathbf{B}}$, $Ab \in \mathbf{B}$, thus, $b_* = b^{*\top}1$, the adjoint of b^* evaluate at 1. Note in some texts, such as in Bickel et al. ([5], Chapter 5), b^* and b_* are denote by the same notation for simplicity.

For fixed g_0, let $\dot{l}_f^{\top}(\cdot|g_0, b) : \mathbf{B} \to L^2(P_0)$ be the adjoint operator of $\dot{l}_f(x|g_0, \cdot)$, which is determined by

$$< \dot{l}_f(\cdot|g_0, h), b >_{\mathbf{B}} = < h, \dot{l}_f^{\top}(\cdot|g_0, b) >_{P_0}, \quad \forall h \in L^2(P_0), \quad \forall b \in \mathbf{B}.$$

For fixed x and (g_0, h), the second order Hadamard differential $\ddot{l}_f(x|g_0, h, h_1)$ of $l_f(x|g_0)$ in the direction h_1 is defined analogously. It is the Hadamard differential of $\dot{l}_f(x|g_0; h)$ with respect to g_0 in the direction h_1.

For $h \in \mathbf{B}$, $I_f(g_0|h, h) := E_{P_0}[\ddot{l}_f(X|g_0; h, h)]$ is the information of g_0 in the model $f(\cdot|g_0)$ at direction h, and $I_f(g|\cdot, \cdot) : \mathbf{B} \times \mathbf{B} \to R$ is a *covariance functional*. Just as a covariance (matrix) uniquely determines a zero mean Gaussian random variable in Euclidean space, the covariance functional uniquely determines a zero mean (in the Pettis sense) tight Gaussian random element in $\mathbf{B}$ (see, for example, Vakhania et al. [30]). As there is no Lebesque measure in general Banach space, and so no density function with respect to such measure, the distribution of random element Z in $\mathbf{B}$ is often characterized by the real random variable b^*Z for each $b^* \in \mathbf{B}^*$. For Gaussian random element Z with mean zero and covariance functional $C(h, h)$, the distribution of b^*Z is the Gaussian random variable with mean zero and variance $C(b^*, b^*) := C(b_*1, b_*1)$, with b_* the Rize representer of b^* with respect to the inner product $< \cdot, \cdot >_{P_0}$.

Unlike the Euclidean case, many parameters in $\mathbf{B}$ are not rate-$\sqrt{n}$ estimable, but still some of them are. Let $\xrightarrow{D}$ stands for weak convergence in R^k, and $\overset{D}{\Rightarrow}$ for that in $l^{\infty}(T) = \{g(\cdot) : \|g\|_T := \sup_{t \in T}|g(t)| < \infty\}$, with respect to the metric $\|\cdot\|_T$. Note weak limit of random

elements in **B** is characterized via that of any linear functional of the elements. For $g \in \mathbf{B}$ is not rate-$n^{1/2}$ estimable, often its rate is slower than $n^{1/2}$, the weak limit is often non-Gaussian, and convolution representation for this case have not been seen. Let $g_n = g + n^{-1/2}h$ for $h \in \mathbf{B}$. Since there is no Borel measure on **B**, distributions of random element V in **B** is characterized by that of the real random variables b^*V for $b^* \in \mathbf{B}^*$. We define an estimator $\hat{g}_n$ of g to be *regular*, if under $f(\cdot|g_n)$,

$$\sqrt{n}(\hat{g}_n - g_n) \stackrel{D}{\Rightarrow} V,$$

for some tight random element $V \in \mathbf{B}$, which does not depend on the sequence $\{g_n\}$; and $\hat{g}_n$ to be *weakly regular* (Bickel et al. [5], p.181), if $\forall b^* \in \mathbf{B}^*$,

$$\sqrt{n}b^*(\hat{g}_n - g_n) \stackrel{D}{\rightarrow} b^*V.$$

Let

$\mathcal{F}_0 = \{f_j(\cdot|g): f_j(\cdot|g)$ be a density for each $g \in \mathbf{H}$, and $f_j(\cdot|g_0) = f_0(\cdot), j \in J\}$ be the class of all parametric densities pass through $f_0(\cdot)$ at g_0.

Assume the following conditions:

(C9) $l_f(\cdot|g)$ is twice Hadamard differentiable with respect to g.

(C10) $\int f_0(x)\dot{l}_f(x|g_0;\, h)dx = 0,\ \forall h \in \mathbf{B}$.

(C11) $I_f(g_0, h, h)$ is positive definite.

For infinite-dimensional parameters, the convergence rates of their estimators are often slower than $\sqrt{n}$ and the weak limits are often non-Gaussian, a natural (and difficult) question is whether and under what conditions the convoluiton theorem holds for the general case? When the

convergence rate is not $\sqrt{n}$, the problem is much harder, as the LAN property no longer holds with such rates. However, a number of papers have tackled this question, such as in Millar [21] and LeCam [19]. These authors considered very general parameter spaces, established convolution results for estimators regardless of their convergence rates or forms of their weak limits. But these results are mostly of the existence type, not the specific type. Also, it is unclear whether one of the two components in their convolution representation is optimally achievable. For example, given an infinite-dimensional parameter and/or the corresponding likelihood model, although the optimal convergence rate for estimators of this parameter can be determined in principle (LeCam [18]; Birgé [6]), but it is still unknown, if there is an optimal weak limit of its estimators, and what is its specific form, if it exists. Pötzelberger et al. [24] gave examples in which the infinite-dimensional version of the convolution theorem does not hold in general abstract space, but does hold under some regularity conditions, the results are of existence type. Janssen and Ostrovski [13] gave more detailed account of the optimal weak limit. Their Theorem 2.3 gives a convolution result for arbitrary convergence rate a_n, for linear functions of infinite-dimensional parameter and their estimate, with the assumption that the two involed estimators are asymptotically joint Gaussian. They gave the optimal weak limit as the minimal variance random element defined in their condition (a), p.7, but how to find this random element is still not clear. Also, the joint asymptotic Gaussian assumption can only be satisfied for a few parameters in the infinite-dimensional spaces, and in these cases often $a_n = \sqrt{n}$. Their Theorems 3.1 and 4.1 established convolution results for infinite-dimensional parameters in abstract spaces, but again the results are of existence type.

Convolution theorem and information bound for rate $\sqrt{n}$ -estimable parameters of the form $\nu(P): \mathbf{P} \to \mathbf{B}$, for some known functional $\nu(\cdot)$ and a family $\mathbf{P}$ of distributions, can be found in BKRW [5] and van der Vaart and Wellner [33]. There the results are in terms of the Hadamard

differential of ν. Below, we only consider rate $\sqrt{n}$ -estimable parameter $g \in \mathbf{B}$, which is implemented in the likelihood model, not in the form $g = \nu(P)$ for some known $\nu(\cdot)$, and we do not assume the model to be true.

Theorem 4. *Assume* (C9)-(C11), *and that* $\hat{g}_n$ *is a regular estimator of* g *with weak limit* W.

(i) *If* $f(\cdot|g) \notin \mathcal{F}_0$, *then*

$$\sqrt{n}(\hat{g}_n - g_0) \stackrel{D}{\Rightarrow} W = Z \oplus U \oplus V,$$

where Z *is the Gaussian element with zero mean and covariance functional* $I^{-1}_{f^o}(g_0, h, h) = < h, G^{-1}_{f^o} h >_{P_0})$, $f^o \in \mathcal{F}_0$, $G^{-1}_{f^o}$ *is the inverse of the linear operator* $G_{f^o} : \mathbf{B} \to \mathbf{B}^*$, *which is determined by* $I_{f^o}(h_1, h_2) = < h_2, G_{f^o} h_2 >_{P_0}$, $\forall h_1, h_2 \in \mathbf{B}$, *and* $U, V \in \mathbf{B}$ *are random elements independent of* Z.

(ii) *If* $f(\cdot|g) \in \mathcal{F}_0$, *let* $\dot{l}^{\tau}_f(\cdot|g_0, h)$ *be the adjoint of* $\dot{l}_f(\cdot|g_0, h)$. *If* $f(\cdot|g) \in \mathcal{F}_0$ *and the range* $R(\dot{l}^{\tau}_f(x|g_0, \cdot)) = \mathbf{B} = L^2(P_0)$, *then*

$$\sqrt{n}(\hat{g}_n - g_0) \stackrel{D}{\Rightarrow} W = Z \oplus U,$$

where Z *is the Gaussian element with zero mean and covariance functional* $I^{-1}_f(g_0, \cdot, \cdot)$, $I^{-1}_f(g_0, b^*_1, b^*_2) = E_{P_0}[\widetilde{I}(X|g_0, b^*_1)\widetilde{I}(X|g_0, b^*_2)]$, *and for fixed* x, $\widetilde{I}(x|g_0, b^*) := \widetilde{I}(x|g_0, b_*)$ *is the solution of the equation*

$$\dot{l}^{\tau}_f(x|g_0, \widetilde{I}(x|g_0, b)) = b, \quad \forall b \in \mathbf{B}.$$

Let $W_n(b^*) = \sqrt{n} b^*(\hat{g}_n - g_0)$, it is a random process indexed by $b^* \in \mathbf{B}^*$. Let $\|b^*\|_{\mathbf{B}^*} = \sup_{\|h\|=1, h\in\mathbf{B}} |b^* h|$, and the distance $d(\cdot,\cdot)$ in $\mathbf{B}^*$ as

$d(b_1^*, b_2^*) = \|b_1^* - b_2^*\|_{\mathbf{B}^*}$. Using Theorem 1 (ii), under suitable compactness conditions (see, for example, van der Vaart and Wellner [33]) on $\mathbf{B}^*$, we can have $W_n(\cdot) \overset{D}{\Rightarrow} W(\cdot) = Z(\cdot) \oplus U(\cdot)$ in $l^\infty(\mathbf{B}^*)$, with $Z(\cdot)$ the Gaussian random element indexed on $\mathbf{B}^*$, i.e., $\mathbf{B}^*$ is a Gaussian Donsker class for $W_n(\cdot)$.

Acknowledgement

This work is supported in part by the National Center for Research Resources at NIH grant 2G12RR003048 (Yuan), and the National Nature Science Foundation of China, No. 61134013 (Li).

References

[1] A. Araujo and E. Giné, The Central Limit Theorems for Real and Banach Valued Random Variables, Wiley, New York, 1980.

[2] J. M. Begun, W. J. Hall, W. Huang and J. A. Wellner, Information and asymptotic efficiency in parametric-nonparametric models, Annals of Statistics 11 (1983), 432-452.

[3] R. Beran, The role of Hájek's convolution theorem in statistical theory, Kybernetika 31 (1995), 221-237.

[4] R. Berk, Limiting behaviour of posterior distributions when the model is incorrect, Annals of Mathematical Statistics 37 (1966), 51-58.

[5] P. J. Bickel, C. A. Klaassen, Y. Ritov and J. A. Wellner, Efficient and Adaptive Estimation for Semiparametric Models, Johns Hopkins University Press, Baltimore, Maryland, 1993.

[6] L. Birgé, Approximation dans les espaces métriques et théorie de l'estimation, Z. Wahrscheinlichkeitstheorie Verw. Gebiete 65 (1983), 181-237.

[7] W. Droste and W. Wefelmeyer, On Hájek's convolution theorem, Statistical Decisions 2 (1984), 131-144.

[8] J. Hájek, A characterization of limiting distributions of regular estimates, Z. Wahrsch. und Verw. Gebiete 14 (1970), 323-330.

[9] J. M. Hammersley, On estimating restricted parameters, Journal of the Royal Statistical Society, Series B 12 (1950), 192-240.

[10] P. J. Huber, The behaviour of maximum likelihood estimates under non-standard conditions, Proc. Fifth Berkeley Symp. Math. Statist. Probab. 1 (1967), 221-233.

[11] I. A. Ibragimov and R. Z. Has'minskii, Statistical Estimation Asymptotic Theory, Springer-Verlag, New York, 1981.

[12] N. Inagaki, On the limiting distribution of a sequence of estimators with uniformity property, Annals of the Institute of Statistical Mathematics 22 (1970), 1-13.

[13] A. Janssen and V. Ostrovski, The convolution theorem of Hájek and LeCam-revisited, Statistics and Decision 1 (2005), submitted.

[14] P. Jeganathan, On the asymptotic theory of estimation when the limit of the log-likelihood ratio is mixed normal, Sankhá 44 (1982), 173-212.

[15] L. LeCam, On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, University of California Publ. Statist. 1 (1953), 277-330.

[16] L. LeCam, Locally asymptotically normal families of distributions, Univ. California Publ. Statist. 3 (1960), 37-98.

[17] L. LeCam, Limits of experiments, In: Proc. Sixth Berkeley Symp. Math. Statist. Prob. (L. LeCam, J. Neymen and E. Scott, eds.), Univ. California Press, Berkeley 1 (1972), 245-261.

[18] L. LeCam, Convergence of estimates under dimensionality restrictions, Annals of Statistics 1 (1973), 38-53.

[19] L. LeCam, An infinite dimensional convolution theorem, Proc. 5, Purdue Int. Symp. Stat. Decis. Theory and Reatedl. Topics, Purdue Univ./USA, Springer, 5 (1994), 401-411.

[20] D. Luenberger, Optimization by Vector Space Methods, John Wiley & Sons, Inc., New York, London, Sydney, Toronto, 1969.

[21] P. W. Millar, Non-parametric applications of an infinite-dimensional convolution theorem, Z. Wahrscheinlichkeitstheorie Verw. Geb. 68 (1985), 545-556.

[22] J. Pfanzagl, On the measurability and consistency of minimum contrast estimators, Metrika 14 (1969), 249-272.

[23] J. Pfanzagl, Parametric Statistical Theories, Water de Gruyter, 1994.

[24] K. Pötzelberger, W. Schachermayer and H. Strasser, On the convolution theorem for infinite-dimensional parameter spaces, Manuscript, (2000).

[25] J. Qin and J. L. Lawless, Empirical likelihood and general estimating equations, Annals of Statistics 22 (1994), 300-325.

[26] D. S. Robson, Admissible and minimax integer-valued estimators of an integer valued parameters, Annals of Mathematical Statistics 29 (1958), 801-812.

[27] A. Schick and V. Susarla, An infinite dimensional convolution theorem with applications to random censoring and missing data models, Journal of Statistical Planning and Inference 24 (1990), 13-23.

[28] P. K. Sen, The Hájek convolution theorem and empirical Bayes estimation: Parameters, semiparametrics and nonparametrics, Journal of Statistical Planning and Inference 91 (2000), 541-556.

[29] R. Serfling, Approximation Theorems of Mathematical Statistics, Wiley, 1980.

[30] N. N. Vakhania, V. I. Tarieladze and S. A. Chobanyan, Probability Distributions on Banach Spaces, D. Reidel Publishing Company, Dordrecht/Boston/Lancaster/Tokyo, 1987.

[31] E. R. van den Heuvel and C. A. J. Klassen, Bayes convolution, International Statistical Review 67 (1999), 287-299.

[32] A. W. van der Vaart, An asymptotic representation theorem, International Statistical Review 59 (1991), 97-121.

[33] A. W. van der Vaart and J. A. Wellner, Weak Convergence and Empirical Processes: With Application to Statistics, Springer-Verlag, New York, 1996.

[34] H. White, Maximum likelihood estimation of misspecified models, Economitrica 50 (1982), 1-25.

[35] J. Xu and J. Wang, Maximum likelihood estimation of linear models for longitudinal data with inequality constraint, Communications in Statistics : Theory and Methods 37(6) (2008), 931-946.

Appendix

Proof of Theorem 1. Let $\phi_Y(t) = E_{f(\cdot|\theta)}[\exp\{itY\}]$ be the characteristic function of a random variable Y under model $f(\cdot|\theta)$. We are to show $\lim_n \phi_{W_n}(t) = \phi_U(t)\phi_V(t)\phi_Z(t)$ with $V \sim N(0, \Omega^{-1}(\theta_0))$ and for some U. In fact, by (C2) and (C3), we have the following modified version of locally asymptotic normality (LeCam [16]) of the likelihood ratio:

$$\lambda_n := L_n(\theta_n) - L_n(\theta_0) = bS_n - b^2 I_f(\theta_0)/2 + o_P(1),$$

where $S_n = n^{-1/2}\sum_{i=1}^n \dot{l}_f(X_i|\theta_0) \xrightarrow{D} S \sim N(0, I_{f,1}(\theta_0))$, with $I_{f,1}(\theta_0) = E_{f_0}[\dot{l}(X|\theta_0)\dot{l}'(X|\theta_0)]$. When $f(\cdot|\theta) = f_*(\cdot|\theta)$, $I_{f,1}(\theta_0) = I_f(\theta_0) = I_{f_*}(\theta_0)$.

Note

$$I_f(\theta_0) = I_{f_*}(\theta_0) - [I_{f_*}(\theta_0) - I_f(\theta_0)] := I_{f_*}(\theta_0) - I_\delta(\theta_0).$$

In the above, when $f(\cdot|\theta) \notin \mathcal{F}$, it is known that $I_{f_*}(\theta_0) \geq I_f(\theta_0)$ (Serfling [29], p.257); and by definition of $f_*(\cdot|\theta)$, when $f(\cdot|\theta) \in \mathcal{F}$, we still have $I_{f_*}(\theta_0) \geq I_f(\theta_0)$, thus $I_\delta(\theta_0) \geq 0$.

By the formula $(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$, we have $I_f^{-1}(\theta_0) = I_{f_*}^{-1}(\theta_0) + I_{f_*}^{-1}(\theta_0)(I_\delta^{-1}(\theta_0) - I_{f_*}^{-1}(\theta_0))^{-1}I_{f_*}^{-1}(\theta_0) := I_{f_*}^{-1}(\theta_0) + \Omega^{-1}(\theta_0)$. Here $\Omega(\theta_0)$ is positive definite, hence a covariance matrix.

By assumption of regularity,

$$\phi_{W_n}(t) = E_{f(\cdot|\theta_0)}[\exp\{itW_n\}] = E_{f(\cdot|\theta_n)}[\exp\{it(W_n - b)\}]$$

$$= E_{f(\cdot|\theta_0)}[\exp\{it(W_n - b) + \lambda_n\}] \to E[\exp\{it(W_n - b) + bS - b^2 I_f(\theta_0)/2\}],$$

where the last step above is by the same argument as in Begun et al. [2]. Below for simple of exposition, we assume θ is 1-dimensional, the proof is similar for multivariate θ. Since $b \in C$ is arbitrary, take $b = -itI_f^{-1}(\theta_0)$, we get

$$it(W - b) + bS - b^2 I_f(\theta_0)/2 = it(W - I_f^{-1}(\theta_0)S) - I_f^{-1}(\theta_0)t^2/2,$$

thus,

$$\lim_n \phi_{W_n}(t) = E[\exp\{it(W - I_f^{-1}(\theta_0)S\}]\exp\{-I_f^{-1}(\theta_0)t^2/2\}$$

$$= \exp\{-I_{f_*}^{-1}(\theta_0)t^2/2\}\exp\{-\Omega^{-1}(\theta_0)t^2/2\}E[\exp\{it(W - I_f^{-1}(\theta_0)S\}]$$

$$= \phi_Z(t)\phi_V(t)\phi_{W - I_f^{-1}(\theta_0)S}(t).$$

Now take $U = W - I_f^{-1}(\theta_0)S$, and note that when $f(\cdot|\theta) = f_*(\cdot|\theta)$, $I_f(\theta_0) = I_{f_*}(\theta_0)$, and $V = 0$, the proof is complete.

Proof of Theorem 2. Let $f(\cdot|\theta, g)$ be the density with the side information in g incorporated, $l(x|\theta, g)$ be its log-likelihood, and $I_f(\theta_0|g) = -E_{P_0}\ddot{l}(X|\theta_0, g)$ be the corresponding information for θ_0. By (C6), $I_f(\theta_0|g) = E_{P_0}[\dot{l}_f(X|\theta_0, g)\dot{l}_f'(X|\theta_0, g)]$. Define $L_n(\theta|g)$ accordingly. Now, the local likelihood ratio has the asymptotic form

$$\lambda_n := L_n(\theta_n|g) - L_n(\theta_0|g) = bS_n - b^2 I_f(\theta_0|g)/2 + o_P(1),$$

where $S_n = n^{-1/2}\sum_{i=1}^n \dot{l}_f(X_i|\theta_0, g) \xrightarrow{D} S \sim N(0, I_f(\theta_0|g))$. Here $I_f(\theta_0|g)$ generally has an unknown form as is $f(\cdot|\theta_0, g)$, we need to evaluate it in terms of $I_f(\theta_0)$ and functional(s) of g.

For this, let $\gamma(g) = E_f[g(X, \theta)]$ for the side information constraint, and $\dot{\gamma}(g)(x, \theta)$ be the adjoint (evaluated at 1) of its pathwise derivative (for definition, see, for example, Bickel et al. [5]). By Proposition A.5.2 in Bickel et al. [5], $\dot{\gamma}(g)(x, \theta) = g(x, \theta)$. Define the inner product (matrix) $< s_1, s_2 >= E_f[s_1(X)s_2'(X)] = \int s_1(x)s_2'(x)f(x)dx$, the norm (matrix) $\|s_1\|^2 =< s_1, s_1 >$ and $\|s_1\|^{-2} := (\|s_1\|^2)^{-1}$ when $\|s_1\|^2$ is non-degenerate.

Without side information, the efficient influence function for estimating θ under model $f(\cdot|\theta)$ is $e_f(x|\theta_0) = I_f^{-1}(\theta_0)\dot{l}_f(x|\theta_0)$, and the optimal asymptotic covariance for any regular estimator of θ_0 based on the model $f(\cdot|\theta)$ is $E_f[e_f(X|\theta_0)e_f'(X|\theta_0)] = I_f^{-1}(\theta_0)$. Let $\prod(v|v_1)$ be the projection of v onto $[v_1]$, the linear span of v_1 with respect to f, and $v_1^\perp$ be the orthogonal complement of $[v_1]$ with respect to f. With side information given by $E_f[g(X, \theta)] = 0$, by Example 3.2.3 in Bickel et al. [5], the efficient influence function is

$$\begin{aligned} e_f(x|\theta_0, g) &= \prod(e_f(x|\theta_0)|\dot{\gamma}(g)^\perp) \\ &= e_f(x|\theta_0) - \prod(e_f(x|\theta_0)|\dot{\gamma}(g)) \end{aligned}$$

$$= e_f(x|\theta_0) - < e_f(\cdot|\theta_0),\ \dot{\gamma}(g) > \|\dot{\gamma}(g)\|^{-2}\dot{\gamma}(g)(x)$$

$$= e_f(x|\theta_0) - A'(\theta_0,\ g)\Lambda^{-1}g(x,\ \theta_0).$$

Thus, $I_f(\theta_0|g) = \|e_f(X|\theta_0,\ g)\|_{P_0}^{-2}$.

Now, the rest proof is the same as that in Theorem 1, by taking $b = -itI_f^{-1}(\theta_0|g) = -it\|e_f(X|\theta_0,\ g)\|_{P_0}^2$. Similarly, $I_{f^*}(\theta_0|g) \geq I_f(\theta_0|g)$. We get

$$\lim_n \phi_{W_n}(t) = E[\exp\{it(W - I_f^{-1}(\theta_0|g)R\}]\exp\{-I_f^{-1}(\theta_0|g)t^2/2\}$$

$$= \exp\{-I_{f_*}^{-1}(\theta_0|g)t^2/2\}\exp\{-\Omega^{-1}(\theta_0|g)t^2/2\}E[\exp\{it(W - I_f^{-1}(\theta_0|g)S\}]$$

$$= \phi_Z(t)\phi_V(t)\phi_{W-I_f^{-1}(\theta_0|g)S}(t).$$

Now take $U = W - I_f^{-1}(\theta_0|g)S$, and note that when $f(\cdot|\theta) = f_*(\cdot|\theta)$, $I_f(\theta_0|g) = I_{f_*}(\theta_0|g)$, and $V = 0$, the proof is complete.

Proof of Theorem 3. We use the same method as in the proof of Theorem 1. Let $\alpha_f(x;\ b,\ g,\ h) = bl_f^{(\cdot)}(x|\theta_0,\ g) + l_f^{(\cdot,\cdot)}(x|\theta_0,\ g,\ h)$. We used the index f to denote its dependence on the specified model $f(\cdot|\theta,\ g)$. In this case, as in Begun et al. [2], we have, under P_0,

$$\lambda_n = L_n(\theta_n,\ g_n) - L_n(\theta_0,\ g)$$

$$= n^{-1/2}\sum_{i=1}^n \alpha_f(X_i;\ b,\ g,\ h) - \frac{1}{2}I_f(\theta_0,\ g,\ b,\ -h) + o_P(1),$$

where $I_f(\theta_0,\ g,\ b,\ h) = -E_{P_0}\beta_f(X;\ \theta_0,\ g,\ b,\ -h)$. Note When $f(\cdot|\theta_0,\ g) = f_0(\cdot)$, we have $I_f(\theta_0,\ g,\ b,\ h) = \|\alpha_f(X;\ \theta_0,\ g,\ b,\ h)\|_{P_0}^2$ as in Begun et al. [2], in which they used Hellinger differential, so the notation α there differs the α_f used here by a factor $f^{-1/2}$, and their $\sigma^2 = 4\|\alpha_f^2\|_\mu$ with

μ the Lebesque measure, here we used Hadamard differential, and the $I_f(\theta_0, g, b, h)$ here will be their σ^2 when $f(\cdot|\theta_0, g) = f_0(\cdot)$. Note by (6), $E_{f_0}\alpha_f(X; b, g, h) = 0$, and so for fixed (b, g, h),

$$n^{-1/2}\sum_{i=1}^{n}\alpha_f(X_i; b, g, h) \xrightarrow{D} S(b, g, h) \sim N(0, \sigma_f^2(\theta_0, g, b, h)),$$

where $\sigma_f^2(\theta_0, g, b, h) = \|\alpha_f(X; \theta_0, g, b, h)\|_{P_0}^2$. Thus, as in the proof of Theorem 1, we have

$$\phi_{W_n}(t) \to \phi_W(t) = E[\exp\{it(W - b) + S(b, g, h) - I_f(\theta_0, g, b, -h)/2\}].$$

Since $b \in C^1$ and $h \in \mathbf{H}$ are arbitrary, we first choose $h = h_f^*$ to minimize $\sigma_f^2(\theta_0, g, b, h)$. By definition of h_f^*, we have $\alpha_f(x; b, g, h) = b(\rho_f(x) - A_f(x, h_f^*)) + A_f(x, bh_f^* + h)$, and $\|\alpha_f(X; b, g, h)\|_{P_0}^2 = b^2\|\rho_f(X) - A_f(X, h_f^*)\|_{P_0}^2 + \|A_f(x, bh_f^* + h)^2\|_{P_0}$, $\forall h \in L^2(P_0)$, with "=" iff $h = -bh_f^*$. Now take $h = -bh_f^*$, we get

$$\begin{aligned}\phi_W(t) &= E[\exp\{it(W - b) + S(b, g, -bh_f^*) - I_f(\theta_0, g, b, bh_f^*)/2\}\\ &= E[\exp\{it(W - b) + bS(1, g, -h_f^*) - b^2 I_f(\theta_0, g, h_f^*)/2\}.\end{aligned}$$

Now, similarly as before, take $b = -itI_f^{-1}(\theta_0, g, h_f^*)$, we have

$$\phi_W(t) = E[\exp\{it[W - I_f^{-1}(\theta_0, g, h_f^*)S(1, g, -h_f^*)]\}]\exp\{-I_f^{-1}(\theta_0, g, h_f^*)t^2/2\}.$$

Similarly as in the proof of Theorem 1, we have $I_{f_*}(\theta_0, g, h_{f_*}^*) \geq I_f(\theta_0, g, h_f^*)$ with "=" iff $f(\cdot|\theta, g) = f_*(\cdot|\theta, g)$. Let $\Omega^{-1}(\theta_0, g) = I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*)$ $[(I_{f_*}(\theta_0, g, h_{f_*}^*) - I_f(\theta_0, g, h_f^*))^{-1} - I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*)]^{-1} I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*)$, take $U = W - I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*)S(1, g, -h_f^*)$, then

$$\phi_W(t) = E[\exp\{it(W - I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*)S(1, g, -h_f^*))\}]\exp\{-t^2 I_{f_*}^{-1}(\theta_0, g, h_{f_*}^*)/2\}$$

$$\times \exp\{-t^2\Omega^{-1}(\theta_0, g)/2\} = \phi_U(t)\phi_Z(t)\phi_V(t).$$

This complete the proof.

Proof of Theorem 4. We first generalizes the LAN condition to **B**-vaued parameter case. By the Taylor expansion using Hadamard differentials, we have

$$\lambda_n = L_n(g_n) - L_n(g_0)$$

$$= n^{-1/2}\sum_{i=1}^{n} \dot{l}_f(X_i|g_0; h)$$

$$+ \frac{1}{2} n^{-1}\sum_{i=1}^{n} \ddot{l}_f(X_i|g_0; h, h) + o_P(1),$$

the remainder $o_P(1)$ is by the definition of second order Hadamard differentiability.

By (C10), $E_{P_0}\dot{l}_f(X, g_0, h) = 0$, $\forall h \in \dot{\mathbf{P}}_0$, so by the central limit theorem, $S_n := n^{-1/2}\sum_{i=1}^{n} \dot{l}_f(X_i|g_0, h) \xrightarrow{D} S(h) \sim N(0, I_{f,1}(g_0; h, h))$, and by the strong law of large numbers, $n^{-1}\sum_{i=1}^{n} \ddot{l}_f(X_i|g_0, h, h) \xrightarrow{\text{a.s.}} E_{P_0}$ $\ddot{l}_f(X|g_0, h, h) = -I_f(g_0; h, h)$. Thus, we get

$$\lambda_{\dot{n}} = S(h) - \frac{1}{2} I_f(g_0; h, h) + o_P(1).$$

(i) Let $W_n = \sqrt{n}(\hat{g}_n - g_0)$, and $\phi_Y(b^*) = E[\exp\{ib^*(Y)\}] : \mathbf{B}^* \to C$ be the characteristic functional of a random element $Y \in \mathbf{B}$. The proof below is true for any inner product $< \cdot, \cdot >_{\mathbf{B}}$ on **B**, although $< \cdot, \cdot >_{P_0}$ is convenient. Let $b_* \in \mathbf{B}$ be the Rize representer of b^* for the inner

product $< \cdot, \cdot >_{\mathbf{B}}$. Then $\phi_Y(b^*) = \phi_Y(b_*) = E[\exp\{i < Y, b_* >_{\mathbf{B}}\}]$. Like characteristic function for random variables, characteristic functional and distribution of random element uniquely determine each other in $\mathbf{B}$, and two random elements are independent iff the characteristic functional of their summation is the product of those of each other. For random element, weak convergence implies convergence of corresponding characteristic functionals, but the convergence of characteristic functionals of a sequence of random elements to some limit one is not enough for the weak convergence of the sequence to the corresponding weak limit, it needs weakly relatively compactness of the sequence $\{W_n\}$ (see Theorem 4.3.1, p.224, Vakhania et al. [30]), just like weak convergence of random processes. However, the weak convergence of W_n in $l^\infty(T)$ is already given in its definition of regularity, so we only need to show $\lim_n \phi_{W_n}(s) = \phi_Z(s)\phi_U(s)\phi_V(s)$ for all $s \in \mathbf{B}$.

By regularity of $\hat{g}_n$ and expression of λ_n, we have

$$\begin{aligned}\phi_{W_n}(s) &= E_{f(\cdot|g_0)}[\exp\{i < s, W_n >_{\mathbf{B}}\}] = E_{f(\cdot|g_n)}[\exp\{i < s, W_n - h >_{\mathbf{B}}\}] \\ &= E_{f(\cdot|g_0)}[\exp\{i < s, W_n - h >_{\mathbf{B}} + L_n(g_n) - L_n(g_0)\}] \to \phi_W(s) \\ &= E[\exp\{i < s, W - h >_{\mathbf{B}} + S(h) - \frac{1}{2} I_f(g_0; h, h)\}]. \qquad \text{(A.1)}\end{aligned}$$

Since $I_f(g_0; \cdot, \cdot)$ is positive definite by (C10) and symmetric bi-linear by definition, so $I_f(g_0, \cdot, \cdot) : \mathbf{B} \times \mathbf{B} \to R$ is a positive definite bi-linear form, thus, there is a symmetric linear operator (covariance operator) $G_f : \mathbf{B} \to \mathbf{B}^*$, such that $I_f(g_0, h_1, h_2) = (G_f h_1)(h_2)\ \forall h_1, h_2 \in \mathbf{B}$ (see, for example, p.145, Vakhania et al. [30]) . Also, G_f has an inverse G_f^{-1} since $I_f(g_0 : \cdot, \cdot)$ is positive definite. Without confusion, denote G_f as the Rize representer of G_f for the inner product $< \cdot, \cdot >_{\mathbf{B}}$, then

$I_f(g_0, h_1, h_2) =< G_f h_1, h_2 >_{\mathbf{B}} =< h_1, G_f h_2 >_{\mathbf{B}}$. Also, G_f has the square-root decomposition $G_f = G_f^{1/2} \circ G_f^{1/2}$ for some linear operator $G_f^{1/2}$ (c.f. Lemma 1.1 and its proof, p.149, Vakhania et al. [30]). Thus $< h_1, G_f h_2 >_{\mathbf{B}} =< G_f^{1/2} h_1, G_f^{1/2} h_2 >_{\mathbf{B}}, \forall h_1, h_2 \in \mathbf{B}$.

Since the right hand side in (A.1) is independent of h by definition of regularity, we can choose h as we want. Take $h = -iG_f^{-1}s$. Then $S(h) = -iG_f^{-1}S(s) = -i < 1, G_f^{-1}S(s) >$, where $< \cdot, \cdot >$ is the inner product in Euclidean space, given by $< a, b >= ab, \forall a, b \in R$. Note $[G_f^{-1}S](\cdot)$ is linear, it has an adjoint $(G_f^{-1}S)^{\tau} : R \to \mathbf{B}$, so that $< 1, G_f^{-1}S(s) >= < s, (G_f^{-1}S)^{\tau}1 >_{\mathbf{B}}$. Also, $-i < s, h >_{\mathbf{B}} = - < s, G_f^{-1}s >_{\mathbf{B}}, - I_f(g_0, h, h) = - < h, G_f h >_{\mathbf{B}} =< G_f^{-1}s, G_f G_f^{-1}s >_{\mathbf{B}} =< s, G_f^{-1}G_f G_f^{-1}s >_{\mathbf{B}} =< s, G_f^{-1}s >_{\mathbf{B}}$, and so

$$i < s, W - h >_{\mathbf{B}} + S(h) - \frac{1}{2} I_f(g_0; h, h)$$

$$= i < s, W - (G_f^{-1}S)^{\tau}1 >_{\mathbf{B}} - \frac{1}{2} < s, G_f^{-1}s >_{\mathbf{B}}.$$

Now, we have

$$\phi_W(s) = E[\exp\{i < s, W - (G_f^{-1}S)^{\tau}1 >_{\mathbf{B}}\}] \exp\{-\frac{1}{2} < s, G_f^{-1}s >_{\mathbf{B}}\}.$$

When $f \notin \mathcal{F}_0, \forall f^{\circ} \in \mathcal{F}_0, I_{f^{\circ}}(g_0; h, h) > I_f(g_0; h, h), \forall h \in \mathbf{B}$. Write $I_f(g_0; h, h) = I_{f^{\circ}}(g_0; h, h) - [I_{f^{\circ}}(g_0; h, h) - I_f(g_0; h, h)] := I_{f^{\circ}}(g_0; h, h) - I_{\delta}(g_0; h, h)$. Let $G_{f^{\circ}}^{-1}$ and G_{δ}^{-1} be the counter parts of G_f^{-1}, corresponding to $I_{f^{\circ}}(g_0; h, h)$ and $I_{\delta}(g_0, h, h)$, then G_{δ}^{-1} is positive definite

and $< s, G_{f^\circ}^{-1} s >_{\mathbf{B}} =< s, G_f^{-1} s > + < s, G_\delta^{-1} s >_{\mathbf{B}}, \forall s \in \mathbf{B}$. So, we get

$$\phi_W(s) = E[\exp\{i < s, W - (G_f^{-1}S)^\tau 1 >_{\mathbf{B}}\}]\exp\{-\frac{1}{2} < s, G_{f^\circ}^{-1} s >_{\mathbf{B}}\}\exp\{-\frac{1}{2} < s, G_\delta^{-1} s >_{\mathbf{B}}\}.$$

On the right hand side above, the second factor is the characteristic functional $\phi_Z(s)$ of the Gaussian element Z with (Pettis) mean zero and covariance functional $I_f^{-1}(g_0;\, s,\, s) := < s, G_{f^\circ}^{-1} s >_{\mathbf{B}}$, while the first factor is the characteristic functional $\phi_U(s)$ of $U := W - (G_f^{-1}S)^\tau 1$, the third is that of the Gaussian element with mean zero and covariance functional $< s, G_\delta^{-1} s >_{\mathbf{B}}$, and the factorization implies U, V, and Z are independent.

(ii) In this case, $I_f(g_0;\, h,\, h) = I_{f_0}(g_0;\, h,\, h) = E_{f_0}[\dot{l}_f(X|g_0,\, h)\dot{l}_f(X|g_0,\, h)]$. We use van der Vaart's differentiability to find the expression for $I_{f_0}^{-1}(g_0,\, \cdot,\, \cdot\,)$. Let $\psi(g) = g$ be the parameter of interest, it has Hadamard differential, in the direction h, is $\dot{\psi}(g;\, h) = h : \mathbf{B} \to \mathbf{B} = L^2(P_0)$. Let $\dot{\psi}^\tau(g;\, \cdot) : \mathbf{B} \to \mathbf{B} = L^2(P_0)$ be its adjoint, then for fixed g, $< \dot{\psi}^\tau(g;\, h),\, b >_{L^2(P_0)} =< h,\, \dot{\psi}(g;\, b) >_{L^2(P_0)} =< h,\, b >_{L^2(P_0)}, \forall h,\, b \in \mathbf{B}$. Thus, $\dot{\psi}^\tau(g;\, b) = b, \forall b \in \mathbf{B}$. For each fixed $b^* \in \mathbf{B}^*$, we have $b^*[\dot{\psi}(g;\, h)] = < b^*\dot{\psi}(g;\, h),\, 1 >=< \dot{\psi}(g;\, h),\, b_* >_{L^2(P_0)} =< h,\, \dot{\psi}^\tau(g;\, b_*) >_{L^2(P_0)}, \forall h \in \mathbf{B}$, where b_* is the Rize representer of b^* with respect to the inner product $< \cdot, \cdot >_{L^2(P_0)}$. In this way, we also identify $\dot{\psi}^\tau(g;\, \cdot)$ as a linear operator: $\mathbf{B}^* \to \mathbf{B}$, by defining $\dot{\psi}^\tau(g;\, b^*) := \dot{\psi}^\tau(g;\, b_*)$. Let $\dot{l}_f^\tau(\cdot|g_0,\, h) : L^2(P_0) \to \mathbf{B}$ be the adjoint of $\dot{l}_f(\cdot|g_0,\, h)$, which is determined by $< \dot{l}_f(\cdot|g_0,\, h),\, b >_{P_0} =< h,\, \dot{l}_f^\tau(\cdot|g_0,\, b) >_{P_0}, \forall h,\, b \in \mathbf{B}$. Since by assumption, we have $\mathbf{B} = R(\dot{\psi}^\tau(g,\, \cdot)) = R(\dot{l}_f^\tau(x|g_0,\, \cdot\,))$, and note pathwise differential is a type

of Hadamard differential, by Theorem 5.4.1 (Bickel et al. [5], p.202), the efficient influence function $\widetilde{I}(x|g_0,\cdot): \mathbf{B}^* \to \mathbf{B}$ for estimating b^*g_0 is the solution of the equation

$$\dot{l}_f^{\tau}(x|g_0, \widetilde{I}(x|g_0, b^*)) = \dot{\psi}^{\tau}(g_0; b_*) = b_*, \quad \forall b^* \in \mathbf{B}^*,$$

and the inverse information covariance functional $\widetilde{I}_f^{-1}(g_0;\cdot\cdot): \mathbf{B}^* \times \mathbf{B}^* \to R$ for g is

$$I_f^{-1}(g_0, b_1^*, b_2^*) = E_{P_0}[\widetilde{I}(X|g_0, b_1^*)\widetilde{I}(X|g_0, b_2^*)].$$

Thus, the optimal achievable lower bound of the asymptotic variance for estimating b^*g_0 based on the model $f(\cdot|g)$ is $I_f^{-1}(g_0, b^*, b^*)$.

■