Institute for Economic Studies, Keio University

Keio-IES Discussion Paper Series

無限次元凸線形モデルの局所漸近正規性

高梨耕作

Keio University

1858

CALAMVS GLADIO FORTIOR

無限次元凸線形モデルの局所漸近正規性
高梨 耕作
IES Keio DP 2017-012
2017 年 4 月 10 日

【要旨】

無限次元のパラメタ空間において，局所漸近正規性を証明する．
推定関数は凸関数に限るが，一様収束が成り立たない状況において，新たな位相を導入し，法則収束を示す．副産物として，尤度比検定統計量の漸近分布が求められる．

高梨耕作
慶應義塾大学経済学部
〒108-8345
東京都港区三田2-15-45
takanasi@econ.keio.ac.jp

Local Asymptotic Normality of Infinite-Dimensional Concave Extended Linear Models

Kōsaku Takanashi

Faculty of Economics, Keio University

Abstract

We study local asymptotic normality of M-estimates of convex minimization in an infinite dimensional parameter space. The objective function of M-estimates is not necessary differentiable and is possibly subject to convex constraints. In the above circumstance, narrow convergence with respect to uniform convergence fails to hold, because of the strength of it's topology. A new approach we propose to the lack-of-uniform-convergence is based on Mosco-convergence that is weaker topology than uniform convergence. By applying narrow convergence with respect to Mosco topology, we develop an infinite-dimensional version of the convexity argument and provide a proof of a local asymptotic normality. Our new technique also provides a proof of an asymptotic distribution of the likelihood ratio test statistic defined on real separable Hilbert spaces.

1 Introduction

We develop an infinite-dimensional version of local asymptotic normality and convexity arguments with non-differentiable objective functions in M-estimation of concave extended linear models. A new approach we propose is based on Mosco convergence that is weaker than uniform convergence in the topological sense. Because of the strength of uniform convergence, it does not fold in infinite dimensional circumstances. In this paper, we give proofs of local asymptotic normality on a real separable Hilbert space.

The basic set-up of the estimation problem we investigate is as follows. Let $\mathscr{H}$ be a real separable Hilbert space with the identical dual $\mathscr{H}^* = \mathscr{H}$. We denote the inner product and the associated norm in $\mathscr{H}$ by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ respectively. Let θ be a parameter vector in a parameter set Θ such that $\Theta \subseteq \mathscr{H}$. Suppose we have n observations $Z_1, \ldots, Z_n$ that are realizations of a random

vector Z on a arbitrary set E, and consider an M-estimator of the unknown parameter vector θ such that

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \rho\left(\theta, Z_i\right), \tag{1}$$

where $\rho : \Theta \times E \to (-\infty, \infty]$ is a criterion function. Define the empirical objective function in (1) as

$$F_n\left(\theta\right) \triangleq \frac{1}{n} \sum_{i=1}^{n} \rho\left(\theta, Z_i\right), \tag{2}$$

and its population counterpart as

$$F_0\left(\theta\right) \triangleq \mathbb{E}_Z\left[\rho\left(\theta, Z\right)\right]. \tag{3}$$

We further suppose the minimization problem in (1) is corresponding to a "concave extended linear model", that is,

1. ρ is a lower semi-continuous (l.s.c.) convex function with respect to θ (it is not necessarily smooth, though),

2. $F_0\left(\theta\right)$ is strictly convex in θ (see, e.g. Huang (2001)) and is uniquely minimized at a (pseudo-) true parameter $\theta_0 \in \Theta$.

Compared to the rate of convergence of the the M-estimator for the concave extended linear model, only a few studies have explored its asymptotic distribution and most of them is on the least squares regression case (e.g. Newey (1997), Huang (2003), Belloni et al. (2015)). Recently, Shang and Cheng (2013) obtain a general result on point-wise asymptotic normality of the M-estimator (1) based on functional Bahadur representation. In proving the asymptotic normality, however, they impose the smoothness condition on ρ so that it should be three times continuously differentiable with respect to θ. On the other hand, our new approach does not require the smoothness of ρ. The following is our motivating example.

Example (L_1 regression.)**.** Consider a nonparametric regression model with additive errors:

$$y = \langle x, \theta \rangle + \varepsilon, \tag{4}$$

where the regressor $x \in \mathscr{H}$ and the error term ε are mutually independent random variables; ε is assumed to be homoskedastic; and the conditional median

of ϵ given x is zero, i.e., $\inf\left\{q:\ P_{\varepsilon}\left(q\mid x\right)\geq\frac{1}{2}\right\}=0$ where $P_{\varepsilon}\left(\cdot\mid x\right)$ is the distribution function of ε conditional on x. We are interested in estimating $\theta\in\mathscr{H}$. Suppose we have observations $Z_i=(y_i,x_i):y_i\in\mathbb{R},x_i\in\mathscr{H},i=1,\ldots,n$ independently drawn from the regression model (4). With them, we may estimate θ via L_1 minimization with roughness penalty(see, for example Koenker et al. (1994)):

$$\hat{\theta}_n=\arg\min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}|y_i-\langle x_i,\theta\rangle|+\frac{\lambda}{2}\|\theta\|,\tag{5}$$

where λ is the smoothing parameter that converges to zero as $n\to\infty$. Obviously, this example gives the case in which the criterion function $\rho=|\cdot|$ is not continuously differentiable.

Uniform convergence of the objective function in (1) to its population counterpart:

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\rho\left(\theta,Z_i\right)-\mathbb{E}_Z\left[\rho\left(\theta,Z\right)\right]\right|\xrightarrow{p}0,$$

guarantees both consistency of of $\hat{\theta}_n$ and convergence of the optimal value of the objective function. In order to make the objective function satisfy the uniform convergence, we have to impose some compactness of the parameter space. These assumptions are rather restrictive for fully nonparametric settings. It is because of the theorem by Bakhvalov (Theorem 12.1.1. of Dudley (1999)) . When $\rho=|\cdot|$ and θ is in an infinite-dimensional space, we have

$$\sup_{\theta}\left|\frac{1}{n}\sum_{i=1}^{n}\rho\left(\theta,Z_i\right)-\mathbb{E}_Z\left[\rho\left(\theta,Z\right)\right]\right|\geq\gamma n^{-1/\infty}.$$

The left-hand side of the inequality does not converge uniformly.

Since ρ is convex, it seems that we may use the convexity lemma (e.g., Pollard (1991) and Kato (2009)) to ensure that point-wise convergence of convex functions implies uniform convergence. In the infinite-dimensional case, however, this argument for uniform convergence may fail. Let $\pi_n,\ n=1,2,\cdots$ be the sequence of projection operators on $\mathscr{H}$ onto $E_n\subset\mathscr{H}$ where $E_n\subsetneq E_{m>n}$. Consider a quadratic form $\langle\pi_n\theta,\theta\rangle$ for $\forall\theta\in\mathscr{H}$ that is considered as a convex function of θ. Then, as $n\to\infty$, $\langle\pi_n\theta,\theta\rangle$ converges point-wise to $\langle\theta,\theta\rangle$ but not uniformly.

To solve the aforementioned lack-of-uniform-convergence issue, we shall propose to apply an alternative mode of convergence, *Mosco convergence*, which is weaker than uniform convergence but still strong enough to enable statistical

applications. Mosco convergence of the objective function ensures the convergence of its minimizer (Attouch (1984)). We develop narrow convergence theory with respect to the Mosco metric, see also Geyer (1994), Dupacava and Wets (1988), Molchanov (2005), Knight (2003) and Bucher et al. (2014). There exist alternative forms of convergence that is equivalent to Mosco convergence but more easily verifiable. They include graph convergence (G-convergence) of subdifferential operators and strong convergence of resolvent. We shall explain these key concepts in Section 2. Using these equivalences, we can establish the consistency and narrow convergence of an M-estimator in an infinite-dimensional parameter space. Furthermore, Mosco convergence also ensures the invertibility of the "Hessian" operator.

If the parameter space is weakly compact, Mosco convergence of the convex objective function $F_n(\theta)$ in M-estimation ensures that both empirical minimizer $\hat{\theta}_n$ and empirical optimal value function $F_n(\hat{\theta}_n)$ will converges to the true parameter θ_0 and the true optimal value function $F_0(\theta_0)$ respectively. This property makes it possible to derive the asymptotic distribution of the optimal value function $F_n(\hat{\theta}_n)$. Namely,

1. The convex objective function $F_n(\theta)$ is locally asymptotically normal at θ_0 in Mosco topology:

$$n\left[F_n\left(\theta_0 + \frac{1}{\sqrt{n}}t\right) - F_n(\theta_0)\right] \rightsquigarrow \langle t, W\rangle + \frac{1}{2}\langle Vt, t\rangle,$$

where W is a normal random vector in a Hilbert space and $V = \nabla_\theta^2 F_0$ is the "Hessian" operator that is almost surely invertible.

2. The asymptotic distribution of the optimal value function $F_n(\hat{\theta}_n)$ is

$$n\left[F_n(\hat{\theta}_n) - F_n(\theta_0)\right] \rightsquigarrow \langle \hat{t}, W\rangle + \frac{1}{2}\langle V\hat{t}, \hat{t}\rangle,$$

where $\hat{t} = \sqrt{n}(\hat{\theta}_n - \theta_0)$.

As a by-product, the asymptotic distribution of the likelihood ratio statistic can be derived. These results are established in a fully nonparametric setting.

The rest of this paper is organized as follows. In Section 2, we describe the Mosco convergence and introduce the narrow convergence in the Mosco topology. In Section 3, we derive local asymptotic normality of an convex objective function in an infinite-dimensional Hilbert space. We also provide the asymptotic distribution of the likelihood ratio statistic by using the local asymptotic normality. Appendixes give some technical lemmas.

Notations

Let $\rightsquigarrow$ denote narrow convergence and $\xrightarrow{P}$ denote convergence in probability. We use empirical process notation: $\mathbb{G}_n \rho = \frac{1}{\sqrt{n}} \sum_{i=1}^n \rho(\theta, Z_i) - \mathbb{E}[\rho(\theta, Z_i)]$. We denote $\|\theta\|$ as l_2-norm or L_2-norm of an element of Hilbert space $\theta \in \mathscr{H}$. Let $\theta_n \xrightarrow{s} \theta_0$ denote convergence in strong topology, e,g,$\|\theta_n - \theta_0\| \to 0$ and $\theta_n \xrightarrow{w} \theta_0$ denote convergence in weak topology, e,g, $\langle \theta_n, \theta^* \rangle \to \langle \theta_0, \theta^* \rangle$ for all identical dual $\theta^* \in \mathscr{H}^* (= \mathscr{H})$. We denote the limit in weak topology as w-$\lim_{n \to \infty} \theta_n$.

2 Mosco Convergence

First, we introduce a mode of convergence, *Mosco convergence*, for proper lower semi-continuous (l.s.c.) convex functions on a real separable Hilbert space. For l.s.c. convex functions on a finite dimensional Euclidean space, point-wise convergence is equivalent to locally uniform convergence. For functions defined on an infinite-dimensional space, however, this is not the case. Mosco convergence, on the other hand, still ensures $\arg\min$ convergence of l.s.c. convex functions on an infinite-dimensional space, though it is weaker than locally uniform convergence. In this section, we also provide preliminary results related to Mosco convergence for later use.

Mosco convergence and similar concepts in a non-stochastic environment are considered in Mosco (1969), Attouch (1984) and Beer (1993). Mosco convergence is particularly useful in the context of functional optimization, making it well suited to M-estimation.

Definition 1. [Mosco Convergence]
Let $f_n : \mathscr{H} \to (-\infty, \infty]$, $n = 1, 2, \ldots$ be a sequence of proper l.s.c. convex functions. f_n is said to be Mosco-convergent to the l.s.c. convex function $f : \mathscr{H} \to (-\infty, \infty]$ if and only if the following two conditions hold.
(M1) For each $\theta \in \mathscr{H}$, there exist a convergent sequence $\theta_n \xrightarrow{s} \theta$ such that $\limsup_n f_n(\theta_n) \leq f(\theta)$.
(M2) $\liminf_n f_n(\theta_n) \geq f(\theta)$ whenever $\theta_n \xrightarrow{w} \theta$.
In this paper, we let "$f_n \xrightarrow{M} f$" denote "f_n Mosco-converges to f."

The variational properties of Mosco convergence are given by the following theorem (Theorem 1.10 in Attouch (1984)), which ensures the convergence of both empirical minimizer and empirical minimum value of the objective function to the true ones. Suppose $\arg\min f_n \neq \emptyset$, and existence of $\arg\min f_n$ and $\inf f_n$ are proved in Appendix A.3.

Theorem 2. *We assume the same definitions for $f_1, f_2, \cdots$ and f. If $f_n \overset{M}{\to} f$, then*

$$\limsup_{n \to \infty} (\arg \min f_n) \subset \arg \min f,$$

in the weak topology, e.g.,

$$\langle \arg \min f_n, h \rangle \to \langle \arg \min f, h \rangle \quad (\forall h \in \mathscr{H}^*),$$

where the $\limsup$ is defined as

$$\limsup_{n \to \infty} F_n \triangleq \left\{ w\text{-}\lim_{n \to \infty} y_{n_k} \ : \ y_{n_k} \in F_{n_k} \text{ for some } n_k \to \infty \right\}.$$

If there is a weakly compact set $K \subset \mathscr{H}$ such that $\arg \min f_n \subset K$ for all n, then $\lim_{n \to \infty} (\inf f_n) = \inf f$.

It is difficult to prove Mosco convergence directly in general settings. Fortunately, several equivalence conditions for Mosco convergence are known in the literature. One of the most convenient conditions for Mosco convergence is point-wise convergence of subdifferentials of functions.

To deal with this mode of convergence, we introduce several basic tools in convex analysis: subdifferential and resolvent. For more details and proofs on these subjects, see Aubin and Frankowska (1990). For fixed $Z \in E$, we can define a set-valued mapping $\partial \rho (\theta, Z) : \Theta \times E \to \mathscr{H}$ by

$$\partial \rho (\theta, Z) = \left\{ \theta \in \mathscr{H} : {}^{\forall}\zeta \in \mathscr{H}, \ \rho (\zeta, Z) \geq \rho (\theta, Z) + \langle \zeta - \theta, \ \theta \rangle \right\}.$$

Such $\partial \rho (\theta, \cdot)$ is said to be the *subdifferential* of ρ at θ. For each fixed θ, $\partial \rho (\theta, Z)$ is considered as a possibly set-valued function of Z. We may regard $\partial \rho (\theta, Z)$ as a generalized derivative of ρ at θ, for each fixed Z. If ρ is Gâteaux differentiable at θ and has a continuous Gâteaux derivative $\nabla \rho (\theta)$, then $\partial \rho (\theta, Z) = \nabla \rho (\theta, Z)$.

Example (L_1 regression(continued))**.** The criterion function $\rho (\theta, Z) = |y - \langle x, \theta \rangle|$ is a proper l.s.c. convex function and has the subdifferential such that

$$\partial \rho (\theta, Z) = \begin{cases} \text{sgn} (y - \langle x, \theta \rangle) x, & \text{if } y - \langle x, \theta \rangle \neq 0; \\ [-1, 1] x, & \text{if } y - \langle x, \theta \rangle = 0, \end{cases}$$

where $\text{sgn} (y - \langle x, \theta \rangle) = \begin{cases} 1, & \text{if } (y - \langle x, \theta \rangle) > 0 \\ -1, & \text{if } (y - \langle x, \theta \rangle) < 0. \end{cases}$

Proof. Proof is given in Appendix A.1. $\qquad \square$

Lemma 3 ("Optimization Theory" Indicator Function)**.** *The indicator function* Ψ_A *is defined by*

$$\Psi_A(\theta) = \begin{cases} 0 & (\theta \in A) \\ \infty & (\theta \notin A) \end{cases}$$

where the set A is a convex subset of Θ. The normal cone $N_A(a)$ is defined by

$$N_A(a) = \{\theta^\star \in \mathscr{H} : \langle \theta - a, \theta^\star \rangle \leqq 0, \ \forall \theta \in A\}.$$

Then, $N_A(a) = \partial \Psi_A(a)$, where $N_A(a)$ is such that $0 \in N_A(a)$.

Proof. a

$$\theta^\star \in \partial \Psi_A(a) \Leftrightarrow \Psi_A(a) + \langle \theta - a, \theta^\star \rangle \leqq \Psi_A(\theta) \ (\forall \theta \in A)$$
$$\Leftrightarrow \langle \theta - a, \theta^\star \rangle \leqq \Psi_A(\theta) \ (\forall \theta \in A)$$
$$\Leftrightarrow \langle \theta - a, \theta^\star \rangle \leqq 0 \ (\forall \theta \in A)$$
$$\Leftrightarrow \theta^\star \in N_A(a)$$

Then, $N_A(a) = \partial \Psi_A(a)$. $\qquad \square$

Subdifferential operator for proper l.s.c. convex funtions holds distributive law:

$$\partial(f_1 + f_2) = \partial f_1 + \partial f_2$$

where f_1 and f_2 are proper l.s.c. convex functions on $\mathscr{H}$ (see Theorem 3.16. in Phelps (1992)). When $\mathscr{H}$ is real separable, subdifferential operator is exchangeable with respect to integral (Clarke (1983) page 76.):

$$\partial f(\theta) = \partial \int_E f(\theta, Z) \, \mathbb{P}_Z(dZ) = \int_E \partial f(\theta, Z) \, \mathbb{P}_Z(dZ).$$

Example (L_1 regression (continued).)**.** The limit criterion $\mathbb{E}[|y - \langle x, \theta \rangle|]$ is convex function and has the subdifferential

$$\partial \mathbb{E}[|y - \langle x, \theta \rangle|] = \mathbb{E}[\partial |y - \langle x, \theta \rangle|],$$

and

$$\mathbb{E}[\partial |y - \langle x, \theta \rangle|] = \mathbb{E}[x \cdot \text{sgn}(y - \langle x, \theta \rangle)]$$
$$= \mathbb{E}[\mathbb{E}[x\{1 - 2\mathbb{I}(y - \langle x, \theta \rangle \leq 0)\} | x]] \qquad (6)$$
$$= \mathbb{E}[x\{1 - 2P_\varepsilon(q - \langle x, \theta \rangle | x)\}].$$

where $P_\varepsilon \left(\cdot \mid x \right)$ is the distribution function of ε conditional on x.

In this paper, we assume that the subdifferential $\partial \rho$ is selected and measurable in Z. In general, because $\partial \rho$ is a set-valued mapping, the selection is not unique. Nonetheless, we can show that not only such measurable selections exists but also the set of all measurable selector $S_{\partial \rho}$ is identical to $\partial \rho$.

Proposition 4. *There exists a measurable selector of the subdifferential ∂f, i.e., $S_{\partial f} \neq \emptyset$. Moreover, $S_{\partial f} = \partial f$.*

Proof. Proof is given in Appendix A.2. □

Consider a map

$$J_\lambda^{\partial f} \theta = \left\{ z \in \mathscr{H} \ : \ z + \lambda \partial f \left(z \right) \ni \theta \right\}.$$

Such a map should be single-valued (on Proposition 3.5.3 in Aubin and Frankowska (1990)). Such $J_\lambda^{\partial f}, \lambda > 0$ are called *resolvents* of ∂f and denoted by

$$^\forall \lambda > 0, \quad J_\lambda^{\partial f} = \left(I + \lambda \partial f \right)^{-1}.$$

The following theorem states the equivalence between Mosco convergence and strong convergence of resolvents and G-convergence of subdifferential operators. The proofs are given in Theorem 3.26. and Theorem 3.66. of Attouch (1984).

Theorem 5. *Let $\mathscr{H}$ be a real separable Hilbert space. Let $(f_n)_{n \in \mathbb{N}}$, $f_n : \mathscr{H} \to (-\infty, \infty]$, $^\forall n \in \mathbb{N}$ be a proper l.s.c. convex function. The following statements are equivalent.*
(1) $f_n \xrightarrow{M} f_0$.
(2) $^\forall \lambda > 0$, $^\forall \theta \in \mathscr{H}$, $J_\lambda^{\partial f_n} \theta \to J_\lambda^{\partial f} \theta$ strongly in $\mathscr{H}$ as n goes to ∞.
(3) $\begin{cases} \partial f_n \xrightarrow{G} \partial f_0, \\ ^\exists (\theta_0, \eta_0) \in \partial f_0 \ ^\exists (\theta_n, \eta_n) \in \partial f_n \text{ such that } \theta_n \xrightarrow{s} \theta_0, \ \eta_n \xrightarrow{s} \eta_0, \ f_n (\theta_n) \to f_0 (\theta_0), \end{cases}$
where $\partial f_n \xrightarrow{G} \partial f_0$ means that, for every $(\theta_0, \eta_0) \in \partial f_0$, there exists a sequence $(\theta_n, \eta_n) \in \partial f_n$ such that $\theta_n \to \theta_0$ strongly in $\mathscr{H}$, $\eta_n \to \eta_0$ strongly in $\mathscr{H}^* \left(= \mathscr{H} \right)$.

Statement (3) in Theorem 5 is called G-convergence of monotone operators. This states that point-wise convergence of all measurable selectors of subdifferential operators is equivalent to Mosco convergence of functionals. When the subdifferential is calculable, point-wise convergence of measurable selectors are easy to verify.

Example (L_1 regression(continued)). From the foregoing theorems, it will be seen that the law of large numbers(LLN) of subdifferential $\partial\rho(\theta)$ implies the Mosco convergence. From Lemma 10 and the LLN in Banach spaces for each sequence of mesurable selectors of $\partial\rho(\theta)$, we have the LLN of subdirrential $\partial\rho(\theta)$:

$$\frac{1}{n}\sum_{i=1}^{n}\partial\rho(\theta, Z_i) \xrightarrow{P} \mathbb{E}[\partial\rho(\theta, Z)]$$
$$= \partial\mathbb{E}[\rho(\theta, Z)].$$

Thus this fact establish the consistency of local functional estimation.

(2) in the above theorem give a metric that induces the Mosco convergence. Based on resolvet, Attouch (1984) (p. 365) gives a metric that induces graph convergence on the space of subdifferential operators:

$$d_G(\partial f, \partial g) \triangleq \sum_{k\in\mathbb{N}}\frac{1}{2^k}\inf\left\{1, \left\|J_{\lambda_0}^{\partial f}\theta_k - J_{\lambda_0}^{\partial g}\theta_k\right\|\right\},$$

for any subdifferential operators ∂f and ∂g where λ_0 is taken strictly positive and $\{\theta_k; k\in\mathbb{N}\}$ is a dense subset of $\mathscr{H}$. This metric d_G induces the Mosco convergence topology and is complete. Convergence in d_G are equivalent to the convergence results in (1)$\sim$(3) in Theorem 5.

Hoffman-Jørgensen weak convergence theory performs in a metric space. Generally, epi-convergence does not usually work with a metric but a semimetric. Even if functions f, g are different each other, it is possible f epiconverge to g (see, Section 3 in Bucher et al. (2014)). Fortunately in the case where the functional space is constituted by convex functions, we can obtain a metric space as described above. We shall define a weak convergence in the following way.

Definition 6. [Mosco Convergence in Distribution]

A sequence of random elements f_n in the space of proper l.s.c. convex functions $\mathscr{H} \to (-\infty, \infty]$ is said to be Mosco converges in distribution to the random element f_0 in the space of proper l.s.c. convex functions if $f_n \rightsquigarrow f_0$ with metric d_G. We use the notation $f_n \overset{M}{\rightsquigarrow} f_0$.

3   Local Asymptotic Normality

First, we show that the reparametrized objective function admits a certain quadratic expansion. A common starting point in developing an asymptotic distribution theory for an M-estimator is to define a centered stochastic process

based on the objective function. Recall that $F_n(\theta) = \frac{1}{n} \sum_i \rho(\theta, Z_i)$ is the objective function for the M-estimator (1). We may define such a centered stochastic process as

$$H_n(\theta, t) \triangleq n \left[F_n \left(\theta + \frac{1}{\sqrt{n}} t \right) - F_n(\theta) \right], \tag{7}$$

$$Q_0(t) \triangleq \langle t, W \rangle + \frac{1}{2} \langle Vt, t \rangle, \tag{8}$$

where W is an $N(\mathbf{0}, A)$ random vector in a Hilbert space and V is a "Hessian" operator. $H_n(\theta_0, t)$ is interpreted as the log likelihood ratio for hypothesis testing against the local alternative, i.e., $\mathcal{H}_0: \ \theta = \theta_0$; $\mathcal{H}_1: \ \theta = \theta_0 + \frac{1}{\sqrt{n}} t$. Define the locally asymptotically quadratic (LAQ) as follows.

Definition 7 (LAQ LeCum and Yang (2000)(p. 120)))**.** The convex objective function $F_n(\theta)$ is said to be locally asymptotically quadratic at θ if there exists a random matrix $V_{n,\theta}$ and a random vector $\Delta_{n,\theta}$ such that

$$H_n(\theta, t) = \langle t, \Delta_{n,\theta} \rangle + \frac{1}{2} \langle V_{n,\theta} t, t \rangle + o_{p_{n,\theta}}(1),$$

and the matrix $V_{n,\theta}$ and their limit $(V_{n,\theta} \rightsquigarrow) V_\theta$ are almost surely invertible.

Remark. Recall that locally asymptotically mixed normality (LAMN) is equivalent to LAQ with a restriction: $\Delta_{n,\theta}, V_{n,\theta}$ converge to normal distributions. Locally asymptotically normality (LAN) is equivalent to LAMN with the limiting matrix V_θ is deterministic.

3.1 Second Order Differentiability

In typical situations, we assume that the function F_0 has a quadratic expansion at θ_0 and their Hessian is often supposed to be continuously invertible (Theorem 3.3.1. of van der Vaart and Wellner (1996)). In an infinite-dimensional case, the assumption that the Hessian operator is continuously invertible is harder to ascertain. However, if the convex function F_0 has a *generalized second order differentiability* (defined later), its "generalized Hessian" is continuously invertible.

Define the Young-Fenchel conjugate f^* of convex function f as

$$f^*(\eta) \triangleq \sup_\theta (\langle \eta, \theta \rangle - f(\theta)).$$

The conjugate f^* has a strong link between a convex function f in the second order differentiability. Recall the case of a convex function defined on finite

dimensional parameters. A convex function f defined on the Euclid space $\mathbb{R}^d$ is second order differentiable and the Hessian $\nabla^2 f(\theta)$ of f at θ is nondegererate. Then the conjugate function f^* is second order differentiable at $y = \nabla f(\theta)$, and its Hessian $\nabla^2 f^*(\eta)$ at y is the inverse of $\nabla^2 f(\theta)$, i.e.,

$$\nabla^2 f(\theta) = \left(\nabla^2 f^*(\eta)\right)^{-1}.$$

In order to maintain a duality-type of this relation in an infinite-dimensional space, we shall define the second order differential concepts based on Mosco convergence. Mosco convergence ensures the continuity of this type of conjugation (Kato (1989) and Borwein and Noll (1994)).

Define *second difference quotient* of f at $\theta \in \mathscr{H}$ relative to $\eta^* \in \partial f(\theta)$ as

$$\Delta_{f,\theta,\eta,t}(h) \triangleq \frac{f(\theta + th) - f(\theta) - t\langle \eta^\star, h\rangle}{t^2}$$

and define a *purely quadratic* continuous convex function as

$$q(h) \triangleq \frac{1}{2}\langle Vh, h\rangle,$$

where V is a closed symmetric positive linear operator. f is said to have *generalized second order differentiability* at θ relative to $\eta^\star \in \partial f(\theta)$ if there exists a purely quadratic function q such that the second order difference quotient $\Delta_{f,\theta,\eta,t}(\cdot)$ converges to $q(\cdot)$ in the Mosco sense, i.e.,

$$\Delta_{f,\theta,\eta,t}(h) \xrightarrow[t\downarrow 0]{M} q(h).$$

The closed symmetric positive linear operator V is called the *generalized Hessian* of f at θ relative to $\eta \in \partial f(\theta)$.

Mosco convergence is invariant under Young-Fenchel conjugation, so that Mosco convergence of $\Delta_{f,\theta,\eta,t}(h)$ is equivalent to Mosco convergence of $(\Delta_{f,\theta,\eta,t}(h))^* = \Delta_{f*,\eta,\theta,t}(h)$. And generalized Hessian of f^* at η relative to $\theta \in \partial f^*(\eta)$ is V^{-1}.

Next, we derive sufficient conditions under which the objective function of M-estimation has generalized second order differentiability. ∂f is called *weak* Gâteaux differentiable* at θ if there exists a bounded linear operator $T : \mathscr{H} \to \mathscr{H}^*$ such that

$$\lim_{t\to 0} \frac{1}{t}(\eta_t^* - \eta^*) = Vh,$$

in the weak* sense for any fixed $h \in \mathscr{H}$ and all $\eta_t^* \in \partial f(\theta + th)$, $\eta^* \in \partial f(\theta)$ where $\partial f(\theta)$ must consist of a single element η^*. We use the notation $T = \nabla \partial f(\theta)$ for the operator T. For the generalized differentiability, we quote the

following result of Borwein and Noll (1994).

Theorem 8. *(a variant of Propotion 6.4. of Borwein and Noll (1994))*
Let $(Z, \mathcal{Z}, \mathbb{P}_Z)$ be a probability space and $\Theta \subseteq \mathcal{H}$ be a separable Hilbert space. Suppose $\rho : \Theta \times Z \to (-\infty, \infty]$ is measurable on $(Z, \mathcal{Z}, \mathbb{P}_Z)$ and convex at any $\theta \in \Theta$ and define a closed convex integral functional f on $\Theta \subset \mathcal{H}$ as

$$f(\theta) = \int_Z \rho(\theta, z) \, d\mathbb{P}_Z(z).$$

Then f is generalized second order differentiable at θ if and only if $\partial \rho$ is weak* Gâteaux differentiable and

$$\operatorname*{ess\,sup}_{z \in Z} |\nabla \partial \rho(\theta, z)| < \infty.$$

Example *(L_1 regression(continued))*. Let $Z = (Y, X)$ be a random vector, where Y is real-valued while X is the covariate and $X \in \mathcal{H}$. Note that objective function of L_1 regression is

$$\begin{aligned} F(\theta) &= \mathbb{E}\left[|Y - \langle x, \theta \rangle|\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[|Y - \langle x, \theta \rangle| \, | \, X\right]\right]. \end{aligned}$$

Then, L_1 regression objective function $F(\theta)$ is generalized second order differentialbe at θ if and only if $\partial \mathbb{E}\left[|Y - \langle x, \theta \rangle| \, | \, X\right]$ is weak* Ĝateaux differentiable and

$$\operatorname*{ess\,sup}_{x \in X} |\nabla \partial \mathbb{E}\left[|Y - \langle x, \theta \rangle| \, | \, X\right]| < \infty.$$

From (6), weak* Gâteaux differentiability of $\partial \mathbb{E}\left[|Y - \langle x, \theta \rangle| \, | \, X\right]$ at θ is equivalent to the Gâteaux differentiability of the distribution function $F_e(q - \langle x, \theta \rangle | x)$ at θ. If the distribution function $F_e(q - \langle x, \theta \rangle | x)$ is Gâteaux differentiable at θ, essential boundedness of $\operatorname{ess\,sup}_{x \in X} |\nabla \partial \mathbb{E}\left[|Y - \langle x, \theta \rangle| \, | \, X\right]| < \infty$ will be automatically satisfied.

Therefore, in order to obtain invertiblity of "generalized Hessian", we impose the following assumption on ρ:

Assumption. *A*
$\partial \rho(\cdot)$ *is weak* Gâteaux differentiable at θ_0 and*

$$\operatorname*{ess\,sup}_{z \in E} |\nabla \partial \rho(\theta_0, z)| < \infty.$$

This assumption is a "low-level" condition which are sufficient for locally

asymptotically quadratic at θ_0 than that of Geyer (1994). Of course, this result is attributed to the convexity of the objective function.

3.2 LAN

Define auxiliary stochastic process as

$$G_n\left(t\right) \triangleq n\left\langle \frac{1}{\sqrt{n}}t, \partial F_n\left(\theta_0\right)\right\rangle + n\left[F_0\left(\theta_0 + \frac{1}{\sqrt{n}}t\right) - F_0\left(\theta_0\right)\right],$$
$$G_n'\left(t\right) \triangleq n\left\langle \frac{1}{\sqrt{n}}t, \partial F_n\left(\theta_0\right)\right\rangle + \frac{1}{2}\left\langle Vt, t\right\rangle.$$

We also impose the following assumption. Considering Proposition 4 : the set of all mesurable selectors of a subdifferential coincides with its own subdifferential, we denote any measurable selector of $\partial\rho\left(\cdot\right)$ as itself.

Assumption. B

Every mesurable selector in $\partial\rho\left(\theta, Z\right)$ has a bounded variance: $\forall \theta \in \Theta$, $\mathbb{E}\left[\left\|\partial\rho\left(\theta, Z\right)\right\|^2\right] < \infty$, and there is a sequence of mesurable selectors satisfying a central limit theorem in the Hilbert space:

$$\mathbb{G}_n\partial\rho\left(\theta_0, Z\right) \rightsquigarrow N\left(0, A\right),$$

for some trace class covariance operator A.

Proposition 9. *LAN*

1. $H_n\left(t\right)$ Mosco-converges to $G_n'\left(t\right)$ in probability.

2. $G_n'\left(t\right)$ converges in law to $Q_0\left(t\right)$. Then, $H_n\left(t\right)$ Mosco-converge in law to $Q_0\left(t\right)$.

Proof. We shall prove the first statement. In order that $H_n\left(t\right)$ converges in Mosco to $G_n\left(t\right)$, we will apply Theorem 5 to $H_n\left(t\right)$ and $G_n\left(t\right)$. All we have to do is to show the graph convergence of the subdifferential $\partial H_n\left(t\right)$ to $\partial G_n\left(t\right)$ in probability. Considering proposition 4, we denote any measurable selector of $\partial\rho\left(\cdot\right)$ as itself in the following proof below. Calculate subdifferential of H_n, G_n

with respect to t, we obtain

$$\partial H_n(t) = \sqrt{n}\partial F_n\left(\theta_0 + \frac{1}{\sqrt{n}}t\right)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n \partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t\right),$$

$$\partial G_n(t) = \sqrt{n}\partial F_n(\theta_0) + \sqrt{n}\partial F_0\left(\theta_0 + \frac{1}{\sqrt{n}}t\right)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n \partial\rho(\theta_0, Z_i) + \sqrt{n}\mathbb{E}\left[\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z\right)\right].$$

Recall $\partial f_n \overset{G}{\to} \partial f_0$ means that for every $(\theta_0, \eta_0) \in \partial f_0$, there exists a sequence $(\theta_n, \eta_n) \in \partial f_n$ such that $\theta_n \to \theta_0$ strongly in $\mathscr{H}$, $\eta_n \to \eta_0$ strongly in $\mathscr{H}^*\,(=\mathscr{H})$. $\partial H_n \overset{G}{\to} \partial G_n$ means that there exists a sequence of measurable selectors of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right)$ such that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right) \to \frac{1}{\sqrt{n}}\sum_{i=1}^n \partial\rho(\theta_0, Z_i) + \sqrt{n}\mathbb{E}\left[\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z\right)\right],$$

strongly in $\mathscr{H}$.

The random variable

$$\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right) - \partial\rho(\theta, Z_i),$$

converges monotonically to non-negative random variable. Because $F_0(\theta) = \mathbb{E}[\rho(\theta)]$ is second order differentiable in the generalized sense,

$$\mathbb{E}\left[\lim_{n\to\infty} \partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right) - \partial\rho(\theta_0, Z_i)\right] = 0,$$

so,

$$\lim_{n\to\infty} \partial\rho\left(x\theta + \frac{1}{\sqrt{n}}t, Z_i\right) - \partial\rho(\theta_0, Z_i) = 0 \text{ a.s.}.$$

Fix t and define a (selected) random variable ξ_{ni} by

$$\xi_{ni} = \frac{1}{\sqrt{n}}\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right) - \frac{1}{\sqrt{n}}\partial\rho(\theta_0, Z_i).$$

Note that

$$\mathbb{E}\left[\frac{1}{\sqrt{n}}\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z\right) - \frac{1}{\sqrt{n}}\partial\rho(\theta_0, Z)\right] = \mathbb{E}\left[\frac{1}{\sqrt{n}}\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z\right)\right],$$

14

where $\mathbb{E}\left[\frac{1}{\sqrt{n}}\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z\right)\right]$ is singleton. Therefore, for any selected ξ_{ni},

$$\sum_{i=1}^{n}\xi_{ni} = \partial H_n(t) - \partial G_n(t) + \sqrt{n}\mathbb{E}\left[\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z\right)\right],$$

and

$$\text{Var}\left[\sum_{i=1}^{n}\xi_{ni}\right] = \mathbb{E}\left[(\partial H_n(t) - \partial G_n(t))^2\right].$$

Since $\xi_{n1}, \ldots, \xi_{nn}$ are i.i.d., we have

$$\text{Var}\left[\sum_{i=1}^{n}\xi_{ni}\right] \leq \sum_{i=1}^{n}\mathbb{E}\left[\xi_{ni}^2\right],$$

for any selected ξ_{ni}. We have the equality

$$\sum_{i=1}^{n}\mathbb{E}\left[\xi_{ni}^2\right] = n\mathbb{E}\left[\left\{\frac{1}{\sqrt{n}}\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right) - \frac{1}{\sqrt{n}}\partial\rho(\theta_0, Z_i)\right\}^2\right]$$

$$= \mathbb{E}\left[\left\{\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right) - \partial\rho(\theta_0, Z_i)\right\}^2\right].$$

By weak* differentiability of $\mathbb{E}[\partial\rho]$ at θ_0, the limit of any measurable selector of $\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right) - \partial\rho(\theta_0, Z_i)$ has expectation zero. From the Assumption B : for every measurable selector $\mathbb{E}\left[\{\partial\rho(\theta, Z_i)\}^2\right] < \infty$ for each θ in the neighborhood of θ_0 and from Lebesgue dominated convergence theorem, we have

$$\mathbb{E}\left[\left\{\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right) - \partial\rho(\theta_0, Z_i)\right\}^2\right] \to 0, \qquad (n \to \infty).$$

Thus, $\text{Var}\left[\sum_{i=1}^{n}\xi_{ni}\right] \leq \sum_{i=1}^{n}\mathbb{E}\left[\xi_{ni}^2\right] \to 0$. By Chebyshev inequality, we have

$$\partial H_n(t) - \partial G_n(t) \xrightarrow{P} 0,$$

for fixed t. Then, $H_n(t)$ converges in Mosco to $G_n(t)$ in probability.

From Assumption A, F_0 is second order differentiable in generalized sense:

$$\frac{\left\{F_0\left(\theta_0 + \frac{1}{\sqrt{n}}t\right) - F_0(\theta_0) - \frac{1}{\sqrt{n}}\langle\partial F_0(\theta_0), t\rangle\right\}}{\left(\frac{1}{\sqrt{n}}\right)^2} \xrightarrow{M} \frac{1}{2}\langle Vt, t\rangle.$$

Therefore, combining aforementioned result, we obtain the result that $H_n(t)$

15

Mosco-converges to $G'_n(t)$ in probability.

The second statement of Proposition 9 is derived from Assumption B and a.s. representation theorem (Theorem 1.10.4. of van der Vaart and Wellner (1996)). We get

$$\xi_n \left(= \frac{1}{\sqrt{n}} \sum_{i=1}^n \partial\rho(\theta_0, X_i) \right) \rightsquigarrow \xi,$$

and an almost sure representation $\tilde{\xi}_n \to \tilde{\xi}$ a.s., where $\tilde{\xi}_n$ has the same law as ξ_n and $\tilde{\xi}$ the same law as ξ. This provide the Mosco convergence in distribution of G'_n to Q_0. $\qquad\square$

The aforementioned proposition achieves mosco convergence of H_n to its limit Q_0. Note that $t = \sqrt{n}(\theta - \theta_0)$ minimizes $H_n(t)$.

Next, we will also show convergence of the minimizer of H_n to that of Q_0, provided that the minimizer is almost surely unique. This follows from the following lemma.

Lemma 10. *The minimizer of the function $Q_0(t) = \langle t, W \rangle + \frac{1}{2} \langle Vt, t \rangle$ is single valued.*

Proof. Let $t_0 = \arg\min_t Q_0(t)$. Suppose there exists $t_1 (\neq t_0)$ such that

$$\langle t_1, W \rangle + \frac{1}{2} \langle Vt_1, t_1 \rangle = \langle t_0, W \rangle + \frac{1}{2} \langle Vt_0, t_0 \rangle = \alpha.$$

Then,

$$\left\langle \frac{t_1 + t_0}{2}, W \right\rangle + \frac{1}{2} \left\langle V\frac{t_1 + t_0}{2}, \frac{t_1 + t_0}{2} \right\rangle$$
$$< \frac{1}{2} \langle t_1, W \rangle + \frac{1}{2} \langle t_0, W \rangle + \frac{1}{2} \left(\frac{1}{2} \langle Vt_1, t_1 \rangle + \frac{1}{2} \langle Vt_0, t_0 \rangle \right)$$
$$= \frac{1}{2}\alpha + \frac{1}{2}\alpha = \alpha.$$

This means $Q_0\left(\frac{t_1+t_0}{2}\right) < \alpha$, which is contradiction. $\qquad\square$

We apply the previous results to consider the asymptotic distribution of $\sqrt{n}\left\langle \hat{\theta} - \theta_0, \theta^* \right\rangle$ in the weak topology.

Corollary 11. *Asymptotic Normality*
Let W be an $N(0, A)$ distribution. Under Assumption A and B, we obtain the asymptotic distribution of $\sqrt{n}\left\langle \hat{\theta}_n - \theta_0, \theta^* \right\rangle$ as following;

$$\sqrt{n}\left\langle \hat{\theta}_n - \theta_0, \theta^* \right\rangle \rightsquigarrow \left\langle V^{-1}W, \theta^* \right\rangle \qquad \forall \theta^* \in \Theta$$

where V^{-1} is generalized Hessian of Young-Fenchel conjugate of $F_0(\theta)$.

Proof. From Proposition 9, $H_n\left(\theta_0, \hat{t}\right)$ converges weakly to $Q_0(t)$ in Mosco topology. Applying a.s. representation theorem(Theorem1.10.4 in van der Vaart and Wellner (1996)) we get an almost sure representation $H_n \xrightarrow{M} Q_0$ *a.s..* By Theorem 5 we have

$$\lim_{n\to\infty} \left(\arg\min H_n\right) \to \arg\min Q_N \ a.s.$$

in the weak topology. This provide

$$\sqrt{n}\left\langle \hat{\theta}_n - \theta_0, \theta^*\right\rangle \rightsquigarrow \left\langle V^{-1}W, \theta^*\right\rangle \qquad \forall \theta^* \in \Theta.$$

$\square$

Example (L_1 regression (continued).)**.** Suppose the distribution function $F_e\left(q - \langle x, \theta\rangle | x\right)$ is Gâteaux differentiable at θ and denote their differential as operator V. Under Assumption A and B, for any $x_0 \in \mathscr{H}$,

$$\sqrt{n}\left\langle x_0, \hat{\theta}_n - \theta_0\right\rangle \rightsquigarrow N\left(0, V^{-1}A\right).$$

For the implement, we need a consistent estimators of the generalized Hessian. From the fact of the properties of the generalized differential, the natural candidates are

$$\lim_{h_n\to 0} \frac{1}{k_n}\left(\hat{\eta}_{k_n}^\star - \hat{\eta}^\star\right)$$

in the weak* sense for any fixed $h \in \mathscr{H}$ and all $\hat{\eta}_{k_n}^\star \in \partial f\left(\hat{\theta} + k_n h\right), \hat{\eta}^\star \in \partial f\left(\hat{\theta}\right)$
.

3.3 Likelihood Ratio Test Statistic

Using the previous LAN result, we derives the asymptotic distribution of the likelihood ratio statistic. Let $A_n = \sqrt{n}\left(\Theta - \theta_0\right)$ and $A_{n,0} = \sqrt{n}\left(\Theta_0 - \theta_0\right)$. The likelihood ratio statistic is written by the form

$$\Lambda_n = \inf_{t\in A_n} H_n\left(\theta_0, t\right) - \inf_{t\in A_{n,0}} H\left(\theta_0, t\right).$$

By the previous LAN result, for large n, the likelihood ratio process is similar to the same as in the normal experiment. And by the Mosco convergence argument in theorem 5, if the parameter space is weakly compact, the empirical optimal value of convex function achieve the true optimal.

17

Assumption. C

The parameter set Θ is weakly compact. In a Hilbert space setting $\Theta \subset \mathscr{H}$, weakly compactness is equal to boundedness: for all $\theta \in \Theta$, there exists constant C such that $\|\theta\| \leq C$.

Lemma 12. *Let W be an $N(0, A)$ distribution and repeat (7);*

$$H_n(\theta, t) = n\left[F_n\left(\theta + \frac{1}{\sqrt{n}}t\right) - F_n(\theta)\right].$$

Let $\hat{t} = \sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ denote this minimizer. Under Assumption A-C, the asymptotic distribution of the optimal value function

$$H_n(\theta_0, \hat{t}) = n\left[F_n\left(\hat{\theta}_n\right) - F_n(\theta_0)\right]$$

is the distribution of $Q_N(\hat{t})$.

Proof. From Proposition 9, $H_n(\theta_0, \hat{t})$ converges weakly to $Q_N(t)$ in Mosco topology. Applying a.s. representation theorem(Theorem1.10.4 in van der Vaart and Wellner (1996)) we get an almost sure representation $H_n \overset{Mosco}{\longrightarrow} Q_N$ a.s.. By Theorem 2 and Assumption C, we have

$$\lim_{n\to\infty}(\inf H_n) = \inf Q_N.$$

This provide the optimal value of function H_n converges weakly to Q_N. $\qquad\square$

Define an objective function with convex constraint $G(\theta)$ from $\mathscr{H}$ to $(-\infty, \infty]$ by

$$G_n(\theta) = F_n(\theta) + \Psi_A(\theta)$$

where Ψ_A is defined by

$$\Psi_A(\theta) = \begin{cases} 0 & (\theta \in A) \\ \infty & (\theta \notin A) \end{cases}$$

and A is convex. Because F_n and Ψ_A are convex function, $G_n(\theta)$ are also convex function with respect to θ for all n. Redefine (7), (8) as

$$H_n^{A_{n,0}}(\theta, t) \triangleq n\left[F_n\left(\theta + \frac{1}{\sqrt{n}}t\right) - F_n(\theta)\right] + \Psi_{A_{n,0}}(t)$$

$$Q_0^A(t) \triangleq \langle t, Z \rangle + \frac{1}{2}\langle Vt, t \rangle + \Psi_{T_{A_0}(\theta_0)}(t)$$

where $T_A(\theta)$ is tangent cone:

$$T_{A_0}(\theta) = \limsup_{\tau \downarrow 0} \frac{\Theta_0 - \theta_0}{\tau}.$$

From the result of lemma (3) and lemma (12), we obtain the asymptotic distribution of the optimal value function

$$H_n^A\left(\theta_0, \hat{t}\right) \rightsquigarrow Q_N\left(\hat{t}\right).$$

The above result yeilds the asymptotic distribution of the likelihood ratio statistics Λ_n. The proof strategy is based on van der Vaart (1998), Chapter 16, Theorem 16.7.

Proposition 13. *Assume the parameter spaces Θ and Θ_0 is convex. And assume Assumption A-C. If the sets A_n and $A_{n,0}$ converge to sets A and A_0, then the sequence of likelihood ratio statistics Λ_n converges under $\theta_0 + \frac{t}{\sqrt{n}}$ in distribution to*

$$\left\| V^{-\frac{1}{2}} W + V^{\frac{1}{2}} t \left(\in A_{n,0}\right) \right\|^2 - \left\| V^{-\frac{1}{2}} W + V^{\frac{1}{2}} t \left(\in A_n\right) \right\|^2$$

where W is an $N(\mathbf{0}, A)$ random vector.

Proof. By Lemma 12 and simple algebra

$$
\begin{aligned}
\Lambda_n &= \inf_{t \in A_n} H_n(\theta_0, t) - \inf_{t \in A_{n,0}} H(\theta_0, t) \\
&= 2 \inf_{t \in A_n} \left(n \left\langle \frac{1}{\sqrt{n}} t, \partial F_n(\theta_0) \right\rangle + \frac{1}{2} \langle Vt, t \rangle \right) \\
&\quad - 2 \inf_{t \in A_{n,0}} \left(n \left\langle \frac{1}{\sqrt{n}} t, \partial F_n(\theta_0) \right\rangle + \frac{1}{2} \langle Vt, t \rangle \right) + o_P(1) \\
&= \left\| V^{-\frac{1}{2}} \mathbb{G}_n \partial \rho(\theta_0) + V^{\frac{1}{2}} \hat{t} \left(\in A_{n,0}\right) \right\|^2 - \left\| V^{-\frac{1}{2}} \mathbb{G}_n \partial \rho(\theta_0) + V^{\frac{1}{2}} \hat{t} \left(\in A_n\right) \right\|^2 + o_P(1)
\end{aligned}
$$

the proposition follows by the continuous mapping theorem. $\qquad \square$

Example (L_1 regression(continued)). Consider a likelihood ratio statistics for testing the value of $\langle \theta_0, x_0 \rangle$ at any $x_0 \in E$. For some prespecified point (x_0, c), we consider the following hypothesis:

$$H_0 : \langle \theta_0, x_0 \rangle \leq 0 \quad \text{vs.} \quad H_1 : \langle \theta_0, x_0 \rangle > 0.$$

The objective function under the null constrained is defined as

$$F_n\left(\theta^{H_0}\right) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \langle x_i, \theta^{H_0}\rangle\right| + \frac{\lambda}{2}\left\|\theta^{H_0}\right\|$$

where $\theta^{H_0} \in H_0 = \{\theta \in \Theta : \langle\theta_0, x_0\rangle \leq 0\}$. Note that the set H_0 is convex. We define the generalized likelihood ratio test statistic as

$$\Lambda_n = F_n\left(\hat{\theta}^{H_0}\right) - F_n\left(\hat{\theta}_n\right),$$

where $\hat{\theta}^{H_0}$ is the M-estimator under convex constraint:

$$\hat{\theta}^{H_0} = \arg\min_{\theta^{H_0} \in H_0} F_n\left(\theta^{H_0}\right).$$

If the null the interior of the hypothesis H_0 contains the true parameter θ_0, the sequence of Λ_n converges to zero in distribution. This means that an error of the first kind converges to zero under that the null hypothesis is true. If the true parameter θ_0 belongs to the boundary: $\langle\theta_0, x_0\rangle = 0$, the sets $\sqrt{n}\left(\Theta_0 - \theta_0\right)$ converge to the $H_0 = \{\theta : \langle\theta, x_0\rangle \leq 0\}$. The sequence of Λ_n converges in distribution to the distribution of the square distance of a standard normal vector to the half-space $V^{\frac{1}{2}}H_0 = \left\{\theta : \left\langle\theta, V^{-\frac{1}{2}}x_0\right\rangle \leq 0\right\}$, that is the distribution of $(W \vee 0)^2$.

A   Appendix

A.1   Proof of Subdifferential Calculus of $\rho = |y - \langle x, \theta\rangle|$

Here we show the subdifferential calculus of $\rho = |y - \langle x, \theta\rangle|$. We use the following lemma.

Lemma 14. *The subdifferential of $\|\theta\| = \langle\theta, \theta\rangle$ is $\partial\|\theta\| = \{\theta\}$, $\theta \in \mathscr{H}$.*

Proof. For $\theta \in \mathscr{H}$,

$$\langle\eta, \theta\rangle - \langle\theta, \theta\rangle = \langle\eta - \theta, \theta\rangle, \quad \eta \in \mathscr{H},$$

then $\partial\|\theta\| = \{\theta\}$. $\qquad\qquad\square$

Proposition (Subdifferential Calculus of $\rho = |y - \langle x, \theta\rangle|$)**.** *The criterion function $\rho(\theta, Z) = |y - \langle x, \theta\rangle|$ is a proper l.s.c. convex function and has the subdif-*

ferential such that

$$\partial\rho\left(\theta, Z\right) = \begin{cases} sgn\left(y - \langle x,\theta\rangle\right)x, & \text{if } y - \langle x,\theta\rangle \neq 0; \\ [-1,1]\,x, & \text{if } y - \langle x,\theta\rangle = 0. \end{cases}$$

Proof. Let $t \in [-1,1]$, $\theta = tx$. For all $\zeta \in \mathcal{H}$,

$$\langle tx, \zeta - \theta\rangle = t\langle x,\zeta\rangle - ty \leq t\left|\langle x,\zeta\rangle - y\right| \leq |t|\left|\langle x,\zeta\rangle - y\right| \leq \left|\langle x,\zeta\rangle - y\right|.$$

Then, $\theta = tx \in \partial\rho\left(y - \langle x,\theta\rangle = 0\right)$ and $[-1,1]\,x \subset \partial\rho\left(y - \langle x,\theta\rangle = 0\right)$.

Next, we shall show the inverse inclusion: $\partial\rho\left(y - \langle x,\theta\rangle = 0\right) \subset [-1,1]\,x$. Let $\theta \in \partial\rho\left(y - \langle x,\theta\rangle = 0\right)$ and assume $\theta \neq x$. From $\theta \in \partial\rho\left(y - \langle x,\theta\rangle = 0\right)$, we have

$$|y - \langle x,\zeta\rangle| \geq \langle\zeta - \theta,\ \theta\rangle, \qquad {}^{\forall}\zeta \in \mathcal{H}. \tag{9}$$

From now on, set $H = \{\eta \in \mathcal{H}: \langle x,\eta\rangle = y\}$ and $G = \{\eta \in \mathcal{H}: \langle\eta,\theta\rangle = \langle\theta,\theta\rangle\}$, we shall show that $H = G$. When $\dim\left(\mathcal{X}\right) = 1$, $H = G = \left\{\frac{y}{x^*}\right\}$. Assume $\dim\{\mathcal{H}\} > 2$. First $\eta \in H \Rightarrow \eta \in G$, pick $\eta \in H$: $\langle x,\eta\rangle = y$ we have $\eta = \theta$, so $\langle\eta,\theta\rangle = \langle\theta,\theta\rangle$. Then, $H \subset G$. We shall show the inverse inclusion $G \subset H$. Assume $\eta \in G$ and $\eta \notin H$. Because $\theta \neq x$, there exists $u \in \mathcal{H}$ such that $\langle\theta,u\rangle \neq y$. Put $p = \langle x,\eta\rangle u - \langle x,u\rangle\eta + \theta$, because u and η are linear independent, $p \neq \theta$. On the other hand

$$\begin{aligned} \langle x,p\rangle &= \langle x, \langle x,\eta\rangle u - \langle x,u\rangle\eta + \theta\rangle \\ &= \langle x,\eta\rangle\langle x,u\rangle - \langle x,u\rangle\langle x,\eta\rangle + y \\ &= y. \end{aligned}$$

This is contradiction, therefore $G \subset H$. Finally, we have $G = H$.

Now, set

$$x' \triangleq \zeta - \frac{y - \langle x,\zeta\rangle}{y - \langle x,v\rangle}\left(v - \theta\right), \qquad {}^{\forall}\zeta \in \mathcal{H},$$

Then, we have

$$\begin{aligned} \langle x,x'\rangle &= \langle x,\zeta\rangle - \frac{y - \langle x,\zeta\rangle}{y - \langle x,v\rangle}\langle x, v - \theta\rangle \\ &= \langle x,\zeta\rangle - \frac{y - \langle x,\zeta\rangle}{y - \langle x,v\rangle}\left(\langle x,v\rangle - y\right) \\ &= \langle x,\zeta\rangle + y - \langle x,\zeta\rangle \\ &= y. \end{aligned}$$

21

Furthermore $x' \in H \Rightarrow x' \in G$. Therefore,

$$
\begin{aligned}
\langle \theta, \theta \rangle &= \langle \theta, x' \rangle \\
&= \langle \theta, \zeta \rangle - \frac{y - \langle x, \zeta \rangle}{y - \langle x, v \rangle} \langle \theta, \ v - \theta \rangle \\
&= \langle \theta, \zeta \rangle - \frac{\langle \theta, \ v - \theta \rangle}{y - \langle x, v \rangle} (y - \langle x, \zeta \rangle) \\
&= \langle \theta, \zeta - \theta \rangle - \frac{\langle \theta, \ v - \theta \rangle}{y - \langle x, v \rangle} (y - \langle x, \zeta \rangle) \\
&= \langle \theta, \zeta - \theta \rangle - \frac{\langle \theta, \ v - \theta \rangle}{y - \langle x, v \rangle} (\langle x, \theta \rangle - \langle x, \zeta \rangle),
\end{aligned}
$$

and we get $\langle \theta, \zeta - \theta \rangle = t \langle x, \zeta - \theta \rangle$ where $t = \frac{\langle \theta, \ v - \theta \rangle}{y - \langle x, v \rangle} \neq 0$. Because of (9), $\langle \theta, \ v - \theta \rangle \leq |y - \langle x, v \rangle|$ and

$$
\begin{aligned}
- \langle \theta, \ v - \theta \rangle = \langle \theta, \ \theta - v \rangle &\leq | - \langle x, \theta - v \rangle | \\
&= | \langle x, v \rangle - y | \\
&= | y - \langle x, v \rangle |,
\end{aligned}
$$

Since $\langle \theta, \zeta - \theta \rangle \neq 0, \langle x, \zeta - \theta \rangle \neq 0$. We have $|\langle \theta, \ v - \theta \rangle| \leq |y - \langle x, v \rangle|$, $|t| \leq 1$. Therefor, $\partial \rho (y - \langle x, \theta \rangle = 0) \subset [-1, 1] x$. $\qquad \square$

A.2   Proof of Proposition 4

Set the following notation;

$(\Omega, \mathcal{F}, \mathbb{P})$: probability triple

$(\mathscr{H}, \mathcal{H})$: real separable Hilbert space with Borel σ-field

$2^{\mathscr{H}}$: the family of all nonempty subsets of $\mathscr{H}$

$F : \Omega \to 2^{\mathscr{H}}$: set-valued function.

The inverse image $F^{-1}(X)$ is defined by

$$
F^{-1}(X) = \{ \omega \in \Omega : \ F(\omega) \cap X \neq \varnothing \}.
$$

A set-valued function $F : \Omega \to 2^{\mathscr{X}}$ is called measurable if $F^{-1}(X)$ is measurable for every closed subset X of $\mathscr{X}$. For $1 \leq p \leq \infty$ define a selection of F by

$$
S_F^p = \{ f \in L_p [\Omega, \mathcal{F}, \mu] : \ f(\omega) \in F(\omega) \ a.e. (\mu) \}.
$$

The key notion of set-valued mesurable mapping is decomposability.

Definition 15. Decomposability [Section3 in Hiai and Umegaki (1977)]

Let M be a set of measurable functions $f : \Omega \mapsto \mathscr{H}$. M is called decomposable

with respect to $\mathcal{F}$ if $f_1, f_2 \in M$ and $A \in \mathcal{F}$ implies

$$\mathbb{I}_A f_1 + \mathbb{I}_{\Omega \setminus A} f_2 \in M.$$

For proof of Proposition 4, we need lemmas from Hiai and Umegaki (1977).

Lemma 16. *[Lemma1.1. in Hiai and Umegaki (1977)]*
Let F be measurable set-valued function. If S_F^p is nonempty, then there exists a sequence $\{f_n\} \in S_F^p$ such that $F(\omega) = cl\{f_n(\omega)\}$ for all $\omega \in \Omega$.

Lemma 17. *[Lemma 2.1. in Hiai and Umegaki (1977)]*
Let $\phi : \Omega \times \mathscr{H}$ be $\mathcal{F} \otimes \mathcal{H}$-measurable. Assume $(\Omega, \mathcal{F}, \mathbb{P})$ is complete and $\phi(\omega, \theta)$ is l.s.c. in θ for every fixed ω. Then the function

$$\omega \mapsto \inf\{\phi(\omega, \theta) : \theta \in F(\omega)\},$$

is measurable.

Lemma 18. *[Theorem 3.1. in Hiai and Umegaki (1977)]*
$M = S_F$ *if and only if M is decomposable.*

For the set-valued random variables the following Theorem and definition were given by Hiai and Umegaki

Proposition. *4 There is a measurable selector of subdifferential ∂f i.e., $S_{\partial f} \neq \emptyset$. And the set of all measurable selector is identical to subdifferential ∂f: $S_{\partial f} = \partial f$.*

Proof. Let $h(\gamma, z)$ as

$$h(\gamma, z) = \inf_{|\beta - \alpha| \leq 1} \{f(\beta, z) - f(\alpha, z) - \langle \beta - \alpha, \gamma \rangle\}.$$

Fix α. γ is a subdifferential of $f(\cdot, z)$ at α iff $h(\gamma, z) \geq 0$. For every z, $h(\gamma, \cdot)$ is measurable. From Lemma 17 $\gamma(\cdot)$ is measurable.

Let $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ be measurable selector of subdifferential $\partial f(\alpha, \cdot)$ satisfying

$$f(\beta, \cdot) \geq f(\alpha, \cdot) - \langle \beta - \alpha, \gamma_1(\cdot) \rangle,$$
$$f(\beta, \cdot) \geq f(\alpha, \cdot) - \langle \beta - \alpha, \gamma_2(\cdot) \rangle.$$

From the following inequality

$$f(\beta, \cdot) \geq f(\alpha, \cdot) - \langle \beta - \alpha, \mathbb{I}_A(\cdot)\gamma_1(\cdot) + \mathbb{I}_{\Omega \setminus A}(\cdot)\gamma_2(\cdot) \rangle,$$

$\partial f (\alpha, \cdot)$ is decomposable. Therefore, from Lemma 18 and Lemma 16, $S_{\partial f} = \partial f$. $\qquad \square$

A.3 A Existence of Minimum

Proposition. *Existence of Minimum*

Suppose $f : \Theta \to (-\infty, \infty]$ is a lower semi-continuous convex (l.s.c.) functional and its domain Θ is bounded. Then there exists $\arg\min_\theta f(\omega, \theta)$ and $\inf_\theta f(\omega, \theta)$.

Proof. Let C be a convex subset of a Banach space. From the separation theorem, C is closed in norm topology if and only if C is closed in the weak topology(Correspondence of closedness). f is lsc on Θ in the norm topology if and only if f is lsc in the weak topology.

For each $a \in \mathbb{R}$ put

$$G_a = \{\theta \in \Theta : \ f(\theta) > a\}.$$

G_a is open in the weak topology and $\Theta = \bigcup_{a \in \mathbb{R}} G_a$. Since Θ is weakly compact, there is finite subcover such that

$$\Theta = \bigcup_{i=1}^{n} G_{a_i}.$$

Putting $a_0 = \min\{a_1, \cdots, a_n\}$, we have $f(\theta) > a_0$ for all $\theta \in \Theta$. There exists a real number $b = \inf\{f(\theta) : \ \theta \in \Theta\}$.

Suppose $f(\theta) > b$ for all $\theta \in \Theta$, then

$$\Theta = \bigcup_{n=1}^{\infty} \left\{\theta : \ f(\theta) > b + \frac{1}{n}\right\}.$$

Since Θ is weakly compact,

$$\Theta = \bigcup_{i=1}^{m} \left\{\theta : \ f(\theta) > b + \frac{1}{n_i}\right\}.$$

Put $b_0 = \min\left\{b + \frac{1}{n_1}, \cdots, b + \frac{1}{n_m}\right\}$, we have $f(\theta) > b_0$ for all θ. Therefore we have

$$b = \inf\{f(\theta) : \ \theta \in \Theta\} \geq b_0 > b.$$

This is a contradiction. $\qquad \square$

Acknowledgements

References

ATTOUCH, H. (1984): *Variational Convergence for Functions and Operators*, Pitman Publishing.

AUBIN, J. AND H. FRANKOWSKA (1990): *Set-Valued Analysis*, Birkhauser.

BEER, G. (1993): *Topologies on Closed and Closed Convex Sets*, Kluwer Academic Publishing.

BELLONI, A., V. CHERNOZHUKOV, D. CHETVERIKOV, AND K. KATO (2015): "Some New Asymptotic Theory for Least Squares Series:Pointwise and Uniform Results," *J. Econometrics*, 186, 345–366.

BORWEIN, J. AND D. NOLL (1994): "Second Order Differentiability of Convex Functions in Banach Spaces," *Trans. Amer. Math. Soc.*, 342, 43–81.

BUCHER, A., J. SEGERS, AND S. VOLGUSHEV (2014): "When Uniform Weak Convergence Fails: Empirical Processes for Dependence Functions and Residuals via Epi- and Hypographs," *Ann. Statist.*, 42, 1598–1634.

CLARKE, F. (1983): *Optimization and Nonsmooth Analysis*, New York: Wiley.

DUDLEY, R. (1999): *Uniform Central Limit Theorems*, Cambridge University Press.

DUPACAVA, J. AND R. WETS (1988): "Asymptotic Behavior of Statistical Estimators and of Optimal Solution of Stochastic Optimization Problems," *Ann. Statist.*, 16, 1517–1549.

GEYER, C. (1994): "On the Asymptotics of Constrained M-Estimation," *Ann. Statist.*, 22, 1993–2010.

HIAI, F. AND H. UMEGAKI (1977): "Integrals, Conditional Expectations, and Martingales of Multivalued Functions," *J. Multivariate Anal.*, 7, 149–182.

HUANG, J. (2001): "Concave Extended Linear Modeling: A Theoretical Synthesis," *Statist. Sinica*, 11, 173–197.

——— (2003): "Local Asymptotics for Polynomial Spline Regression," *Ann. Statist.*, 31, 1600–1635.

KATO, K. (2009): "Asymptotics for argmin processes: Convexity argumants," *J. Multivariate Anal.*, 100, 1816–1829.

KATO, N. (1989): "On the Second Derivatives of Convex Functions on Hilbert Spaces," *Proc. Amer. Math. Soc.*, 106, 697–705.

KNIGHT, K. (2003): "Epi-convergence in distribution and stochastic equi-semicontinuity," *Unpublished Manuscript.*

KOENKER, R., P. NG, AND S. PORTNOY (1994): "Quantile smoothing spline," *Biometrika*, 81, 673–680.

LECUM, L. AND G. YANG (2000): *Asymptotics in Statistics*, Springer.

MOLCHANOV, I. (2005): *Theory of Random Sets*, Springer.

MOSCO, U. (1969): "Convergence of Convex Sets and of Solutions of Variational Inequalities," *Adv. Math.*, 3, 510–585.

NEWEY, W. (1997): "Convergence rates and Asymptotic Normality for Series Estimators," *J. Econometrics*, 79, 147–168.

PHELPS, R. (1992): *Convex Functions, Monotone Operators and Differentiability*, Springer, 2nd ed.

POLLARD, D. (1991): "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186–199.

SHANG, Z. AND G. CHENG (2013): "Local and Global Asymptotic Inference in Smoothing Spline Models," *Ann. Statist.*, 41, 2608–2638.

VAN DER VAART, A. (1998): *Asymptotic statistics*, Cambridge University Press.

VAN DER VAART, A. AND J. WELLNER (1996): *Weak convergence and empirical processes*, Springer.