UNIVERSITÀ DEGLI STUDI DI BERGAMO
DIPARTIMENTO DI INGEGNERIA

QUADERNI DEL DIPARTIMENTO

Department of Engineering

Working Paper

Series "*Mathematics and Statistics*"

n. 02/MS – 2014

MOMENT CONVERGENCE OF Z-ESTIMATORS

by

Ilia Negri and Yoichi Nishiyama

[§] L'accesso alle *Series* è approvato dal Comitato di Redazione. I *Working Papers* della Collana dei Quaderni del Dipartimento di Ingegneria dell'Informazione e Metodi Matematici costituiscono un servizio atto a fornire la tempestiva divulgazione dei risultati dell'attività di ricerca, siano essi in forma provvisoria o definitiva.

Moment convergence of Z-estimators

Ilia Negri and Yoichi Nishiyama*

Department of Engineering, University of Bergamo
Viale Marconi, 5, 24044, Dalmine (BG), Italy
ilia.negri@unibg.it
and
The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
nisiyama@ism.ac.jp

January 2, 2014

Abstract

The problem to establish not only the asymptotic distribution results for statistical estimators but also the moment convergence of the estimators has been recognized as an important issue in advanced theories of statistics. There is an authorised theory dealing with this problem for some M-estimators by Ibragimov and Has'minskii (1981). A large deviation inequality, which was a crucial point of Ibragimov and Has'minskii's (1981) theory, has been proved with a good generality by Yoshida (2011). The purpose of this paper is to present an alternative, simple theory to derive the moment convergence of Z-estimators; any large deviation type inequalities do not appear in our approach. Moreover, a merit of our approach is that the cases of parameters with different rate of convergence can be treated easily and smoothly. Applications to some diffusion process models and Cox's regression model are discussed.

1 Introduction

This paper is devoted to the convergence of moments for "Z-estimators", in other words, estimators that are the solutions to estimating equations. Let us first give a review on the moment convergence problem, and next we shall list up some examples to which our results can be applied.

*Corresponding author.

For an illustration, let us consider the simplest case of i.i.d. data. Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space, and let us be given a parametric family of probability densities $f(\cdot;\theta)$ with respect to μ, where $\theta \in \Theta \subset \mathbb{R}^d$. Let $X_1, X_2, \ldots$ be an independent sequence of $\mathcal{X}$-valued random variables from this parametric model. There are at least two ways to define the "maximum likelihood estimator (MLE)" in statistics. One way is to define it as the maximum point of the random function

$$\theta \mapsto \mathbb{M}_n(\theta) = \frac{1}{n}\sum_{k=1}^{n} \log f(X_k;\theta),$$

while the other is to do it as the solution to the estimating equation

$$\mathbb{Z}_n(\theta) = 0, \quad \text{or, in another notation,} \quad \dot{\mathbb{M}}_n(\theta) = 0,$$

where $\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta)$ is the gradient vector of $\mathbb{M}_n(\theta)$. The former is a special case of "M-estimators", and the latter is that of "Z-estimators"; see van der Vaart and Wellner (1996) for these terminologies.

It is well known that the MLE $\widehat{\theta}_n$ has the asymptotic normality: it holds for any *bounded* continuous function $f : \mathbb{R}^d \to \mathbb{R}$ that

$$\lim_{n\to\infty} E[f(\sqrt{n}(\widehat{\theta}_n - \theta_0))] = E[f(I(\theta_0)^{-1/2}Z)],$$

where $I(\theta_0)$ is the Fisher information matrix and Z is a standard Gaussian random vector. Furthermore, it is important for some advanced theories in statistics, including asymptotic expansions and model selections, to extend this kind of results for *bounded* continuous functions f to that for any continuous function f with *polynomial growth*, that is, any continuous function f for which there exist some constants $C = C_f > 0$ and $q = q_f > 0$ such that

$$|f(x)| \le C(1 + ||x||)^q, \quad \forall x \in \mathbb{R}^d. \tag{1}$$

See the discussion in Yoshida (2011) for the importance of this problem.

Notice here that, when we have an asymptotic distribution result of an estimator, namely $R_n(\widehat{\theta}_n - \theta_0) \to^d L(\theta_0)$ where R_n is a (possibly, random) diagonal matrix and the limit random vector $L(\theta_0)$ is not necessarily Gaussian, it is sufficient for the generalisation to the case where ψ is a continuous function satisfying (1) to check that $||R_n(\widehat{\theta}_n - \theta_0)||$ is *asymptotically L_p-bounded* for some $p > q$, that is,

$$\limsup_{n\to\infty} E[||R_n(\widehat{\theta}_n - \theta_0)||^p] < \infty.$$

The study to provide some methods to obtain the moment convergence with polynomial order goes back to Ibragimov and Has'minskii (1981) who considered the MLEs and the Bayes estimators (as some special cases of M-estimators) in the general framework of the locally asymptotically normal models. It should be emphasised that one of the important merits of Ibragimov and Has'minskii's program is perhaps that the theory, based on the likelihood, automatically yields

also the asymptotic efficiency. In their main theorems, it was assumed that an *exponential type large deviation inequality* holds for the rescaled log-likelihood ratio random field. However, checking the assumption in terms of the large deviation inequality was not always easy. Although there are some successful works of Yury Kutoyants, including his books published in 1984, 1994 and 2004, who applied the theory of Ibragimov and Has'minskii (1981) to some stochastic process models, developing a general theory to establish the large deviation inequality was an open problem for many years. Several years ago from now, N. Yoshida solved this problem, and his theory has been published in Yoshida (2011). The paper starts from pointing out that a *polynomial type large deviation inequality* is sufficient for the core part of Igragimov and Has'minskii's (1981) program, and the main contribution is to have proved the (polynomial type) large deviation inequality with a good generality. Uchida and Yoshida (2012) applied Yoshida's (2011) theory to establish the moment convergence of some M-estimators in ergodic diffusion process models with Kessler's (1997) adjustment. We also mention that Nishiyama (2010) pointed out that the moment convergence problem for M-estimators can be solved by using a maximal inequality instead of the large deviation inequalities, and that Kato (2011) took this type of approach to deal with some bootstrap M-estimators.

In this paper, we shall consider the problem to prove the moment convergence of not M-estimators but Z-estimators. Since we have to assume that the random filed something like the log-likelihood is differentiable, our framework is more restrictive than that for M-estimators. Instead, the proof becomes simpler. Actually, any large deviation type inequalities do not appear in our approach, and our proof is just a combination of simple arguments based only on the usual Hölder's and Minkowskii's inequalities.

Another difference between Yoshida's (2011) and our theories is that we can easily treat also the cases where the rates of convergence are different over the components of θ. This is due to the fact that in our theory of Z-estimators we can multiply the gradient vector $\dot{\gamma}_n(\theta)$ of a contrast function $\gamma_n(\theta)$, where $\gamma_n(\theta)$ is typically the log-likelihood function, by a *matrix* R_n^{-2} to get a kind of law of large numbers, namely,

$$\dot{\mathbb{M}}_n(\theta) = R_n^{-2}\dot{\gamma}_n(\theta).$$

Typically, $R_n = \sqrt{n}I_d$ where I_d is the identity matrix, although a merit of our approach is that the diagonal components of R_n may be different in our framework. In contrast, in the framework of M-estimation theory the (scalar valued) contrast function $\gamma_n(\theta)$ with no assumption of differentiability has to be multiplied by a *scalar*. Yoshida (2011) overcame this difficulty by introducing some nuisance parameters in order to handle the components of different rates step by step.

In the rest of this section, we shall list up some examples which fit in our theories. In what follows, the parameter space Θ is a bounded, open, convex subset of $\mathbb{R}^d$, where d is a fixed, positive integer. The word "vector" always means "d-dimensional real column vector", and the word "matrix" does "$d \times d$ real matrix". The Euclidean norm is denoted by $||v|| := \sqrt{\sum_{i=1}^d |v^{(i)}|^2}$ for a vector v where $v^{(i)}$

denotes the i-th component of v, and by $||A|| := \sqrt{\sum_{i,j=1}^{d} |A^{(i,j)}|^2}$ for a matrix A where $A^{(i,j)}$ denotes the (i,j)-component of A. Note that $||Av|| \leq ||A|| \cdot ||v||$ and $||AB|| \leq ||A|| \cdot ||B||$ for vector v and matrices A, B. The notations $v^\top$ and $A^\top$ denote the transpose. We use also the notation $A \circ B$ defined by $(A \circ B)^{(i,j)} := A^{(i,j)} B^{(i,j)}$ for two matrices A, B (the Hadamard product). We denote by I_d the identity matrix. The notations $\to^p$ and $\to^d$ mean the convergence in probability and the convergence in distribution, as $n \to \infty$, respectively.

Example A: Moment estimators

Let $X_1, X_2, \ldots$ be an i.i.d. sample from a distribution P on $(\mathcal{X}, \mathcal{A})$. Let $\psi_1, \ldots, \psi_d$ be measurable functions on $\mathcal{X}$. Define

$$\mathbb{Z}_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} (\psi_1(X_k) - \theta_1, \ldots, \psi_d(X_k) - \theta_d)^\top.$$

Our result can be applied to these estimating functions whose derivative matrix $\dot{\mathbb{Z}}_n(\theta) = \{\dot{\mathbb{Z}}_n^{(i,j)}(\theta)\}_{(i,j) \in \{1,\ldots,d\}^2}$, where $\dot{\mathbb{Z}}_n^{(i,j)}(\theta) = \frac{\partial}{\partial \theta_j} \mathbb{Z}_n^{(i)}(\theta)$, is $-I_d$.

Example B: Ergodic diffusion process

Let $I = (l, r)$, where $-\infty \leq l < r \leq \infty$, be given. Let us consider an I-valued diffusion process $t \rightsquigarrow X_t$ which is the unique strong solution to the stochastic differential equation (SDE)

$$X_t = X_0 + \int_0^t S(X_s; \alpha) ds + \int_0^t \sigma(X_s; \beta) dW_s,$$

where $s \rightsquigarrow W_s$ is a standard Wiener process. The parameters come from $\alpha \in \Theta_A \subset \mathbb{R}^{d_A}$ and $\beta \in \Theta_B \subset \mathbb{R}^{d_A}$, and we denote $\theta = (\alpha^\top, \beta^\top)^\top$. We are supposed to be able to observe the process X at discrete time grids $0 = t_0^n < t_1^n < \cdots < t_n^n$, and we shall consider the asymptotic scheme $n\Delta_n^2 \to 0$ and $t_n^n \to \infty$ as $n \to \infty$, where

$$\Delta_n = \max_{1 \leq k \leq n} |t_k^n - t_{k-1}^n|,$$

and

$$\sum_{k=1}^{n} \left| \frac{|t_k^n - t_{k-1}^n|}{t_n^n} - \frac{1}{n} \right| \to 0, \quad \text{as } n \to \infty. \tag{2}$$

We will consider the following

$$\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta) = R_n^{-2} \dot{\gamma}_n(\theta),$$

where

$$\begin{aligned} &\gamma_n(\theta) \\ &= -\sum_{k: t_{k-1}^n \leq t_n^n} \left\{ \log \sigma(X_{t_{k-1}^n}; \beta) + \frac{|X_{t_k^n} - X_{t_{k-1}^n} - S(X_{t_{k-1}^n}; \alpha)|t_k^n - t_{k-1}^n||^2}{2\sigma(X_{t_{k-1}^n}; \beta)^2 |t_k^n - t_{k-1}^n|} \right\} \end{aligned}$$

and R_n is the diagonal matrix such that $R_n^{(i,i)}$ is $\sqrt{t_n^n}$ for $i = 1, \ldots, d_A$ and $\sqrt{n}$ for $i = d_A + 1, \ldots, d$ with $d = d_A + d_B$.

The problem to establish the moment convergence for M-estimators in this model, where X is a multi-dimensional diffusion process, was considered by Yoshida (2011). Uchida and Yoshida (2012) relaxed the assumption $n\Delta_n^2 \to 0$ up to $n\Delta_n^a \to 0$, where $a \geq 2$ is a constant depending on the smoothness of the model, by using Kessler's (1997) method. However, their arguments consist of plural steps in order to handle the parameters α and β, whose rates of convergence are different, separately. In contrast, our theory makes it possible to treat both parameters simultaneously. Although we consider only the one-dimensional diffusion process X under the sampling scheme $n\Delta_n^2 \to 0$ in order to explain our core idea clearly within a reasonable number of pages, some extension to the case that Uchida and Yoshida (2012) considered would be possible. We leave this problem for readers.

Example C: Volatility of diffusion process

Let $I = (l, r)$, where $-\infty \leq l < r \leq \infty$, be given. Let us consider an I-valued diffusion process $t \rightsquigarrow X_t$ which is the unique strong solution to the SDE

$$X_t = X_0 + \int_0^t S(X_s)ds + \int_0^t \sigma(X_s; \theta)dW_s,$$

where $s \rightsquigarrow W_s$ is a standard Wiener process. Here, the drift coefficient $S(\cdot)$ is treated as an unknown nuisance function. We are supposed to be able to observe the process X at discrete time grids $0 = t_0^n < t_1^n < \cdots < t_n^n = T < \infty$, and we shall consider the asymptotic scheme (2).

We introduce

$$\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta) = \frac{1}{n}\dot{\gamma}_n(\theta),$$

where

$$\gamma_n(\theta) = -\sum_{k: t_{k-1}^n \leq t_n^n} \left\{ \log \sigma(X_{t_{k-1}^n}; \theta) + \frac{|X_{t_k^n} - X_{t_{k-1}^n}|^2}{2\sigma(X_{t_{k-1}^n}; \theta)^2 |t_k^n - t_{k-1}^n|} \right\}.$$

The rate matrix is given by $R_n = \sqrt{n} I_d$.

Example D: Cox's regression model

Let a sequence of counting processes $t \rightsquigarrow N_t^k$, $k = 1, 2, \ldots$, which do not have simultaneous jumps, be observed the time interval $[0, T]$. Suppose that $t \rightsquigarrow N_t^k$ has the intensity

$$\lambda_t^k(\theta) = \alpha(t) e^{\theta^\top Z_t^k} Y_t^k,$$

where the baseline hazard function α which is common for all k's is non-negative and satisfies that $\int_0^T \alpha(t)dt < \infty$, the random process $t \rightsquigarrow Z_t^k$ is a vector valued covariate for the individual k, and the random process $t \rightsquigarrow Y_t^k$ is given by

$$Y_t^k = \begin{cases} 1, & \text{if the individual } k \text{ is observed at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

This model was introduced by Cox (1972), and its asymptotic theory was developed by Andersen and Gill (1982).

We introduce

$$\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta) = \frac{1}{n}\dot{\gamma}_n(\theta),$$

where

$$\gamma_n(\theta) = \sum_{k=1}^{n} \int_0^T (\theta^\top Z_t^k - \log S_t^{n,0}(\theta)) dN_t^k$$

with

$$S_t^{n,0}(\theta) = \sum_{k=1}^{n} e^{\theta^\top Z_t^k} Y_t^k.$$

The rate matrix is $R_n = \sqrt{n} I_d$.

Some detailed discussions about moment convergence of Z-estimators for Examples B, C and D will be given in Sections 3.1, 3.2 and 3.3, respectively, while that for Example A is left for readers.

2 Moment convergence of Z-estimators

Let Θ be a bounded, open, convex subset of $\mathbb{R}^d$. Let an $\mathbb{R}^d$-valued random function $\mathbb{Z}_n(\theta)$ of $\theta \in \Theta$ which is continuously differentiable with the gradient vector $\dot{\mathbb{Z}}_n(\theta)$, defined on a probability space $(\Omega, \mathcal{F}, P)$ that is common for all $n \in \mathbb{N}$. (However, it will be clear from our proofs that if the limit matrices $V(\theta_0)$ and $\dot{Z}(\theta)$ appearing below are non-random then the underlying probability spaces need not be common for all $n \in \mathbb{N}$.)

As an important special case is that $\mathbb{Z}_n(\theta)$ is given as the gradient vector $\dot{\mathbb{M}}_n(\theta)$ of a rescaled contrast function $\mathbb{M}_n(\theta) = R_n^{-2}\gamma_n(\theta)$ of $\theta \in \Theta$ which is twice continuously differentiable with the gradient vector $\dot{\mathbb{M}}_n(\theta)$ and the Hessian matrix $\ddot{\mathbb{M}}_n(\theta)$, where R_n be a (possibly, random) diagonal matrix whose diagonal components are positive; that is, defining Q_n by $Q_n^{(i,j)} = (R_n^{(i,i)} R_n^{(j,j)})^{-1}$, put

$$\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta) = R_n^{-2}\dot{\gamma}_n(\theta) \quad \text{and} \quad \dot{\mathbb{Z}}_n(\theta) = \ddot{\mathbb{M}}_n(\theta) = Q_n \circ \ddot{\gamma}_n(\theta). \tag{3}$$

(In the typical cases, $R_n = \sqrt{n} I_d$ and $Q_n = n^{-1}\mathbf{1}$, where $\mathbf{1}$ denotes the matrix whose all components are 1.)

Turning back to the general setup, we shall state a theorem to give an asymptotic representation for Z-estimators. Although this result is not really novel, we will give a full (and short) proof for references.

Theorem 2.1 *Consider the setting described in the first paragraph of this section. Suppose there exists a sequence of matrices $V_n(\theta_0)$ which are regular almost surely such that for any sequence of Θ-valued random vectors $\widetilde{\theta}_n$ converging in probability to θ_0,*

$$\dot{\mathbb{Z}}_n(\widetilde{\theta}_n) - (-V_n(\theta_0)) \to^p 0.$$

Suppose also that

$$(R_n\mathbb{Z}_n(\theta_0), V_n(\theta_0)) \to^d (L(\theta_0), V(\theta_0)),$$

where R_n be a (possibly, random) diagonal matrix whose diagonal components are positive, $L(\theta_0)$ is a random vector, and $V(\theta_0)$ is a random matrix which is regular almost surely (we do not assume that $V(\theta_0)$ and $L(\theta_0)$ are independent).

Then, for any sequence of Θ-valued random vectors $\widehat{\theta}_n$ which converges in probability to θ_0 and satisfies that $||R_n\mathbb{Z}_n(\widehat{\theta}_n)|| = o_P(1)$, it holds that

$$\begin{aligned} R_n(\widehat{\theta}_n - \theta_0) &= V_n(\theta_0)^{-1} R_n \mathbb{Z}_n(\theta_0) + o_P(1) \\ &\to^d V(\theta_0)^{-1} L(\theta_0). \end{aligned}$$

In this theorem, the consistency of the sequence of Z-estimators $\widehat{\theta}_n$ has been assumed. A method to show this property will be given in Lemma 2.2 below, whose proof is omitted because it can be proved exactly in the same way as Theorems 5.7 and 5.9 of van der Vaart (1998).

Lemma 2.2 *Suppose that for some $\theta_0 \in \Theta$, it holds that*

$$\sup_{\theta \in \Theta} ||\mathbb{Z}_n(\theta) - Z_{\theta_0}(\theta)|| \to^p 0,$$

where the random field $\theta \rightsquigarrow Z_{\theta_0}(\theta)$ of the limit satisfies that

$$\inf_{\theta : ||\theta - \theta_0|| > \varepsilon} ||Z_{\theta_0}(\theta)|| > 0 = ||Z_{\theta_0}(\theta_0)||, \quad \textit{almost surely,} \quad \forall \varepsilon > 0.$$

Then, for any sequence of Θ-valued random vectors $\widehat{\theta}_n$ such that $||\mathbb{Z}_n(\widehat{\theta}_n)|| = o_P(1)$, it holds that $\widehat{\theta}_n \to^p \theta_0$.

Now, we give a theorem to establish the moment convergence of Z-estimators, which is the main result in this section.

Theorem 2.3 *Consider the setting described in the first paragraph of this section. Let some constants $p \geq 1$ and $a, b > 1$ such that $\frac{1}{a} + \frac{1}{b} = 1$ be given; see a remark at the end of the theorem for the case where we may set $a = 1$.*

Suppose that for some $\theta_0 \in \Theta$,

$$||R_n\mathbb{Z}_n(\theta_0)|| \quad \textit{is asymptotically } L_{pa}\textit{-bounded.} \tag{4}$$

Suppose also that there exist a constant $\gamma \in (0, 1]$ and some random matrices $\dot{Z}_{\theta_0}(\theta)$ indexed by $\theta \in \Theta$ such that

$$\lim_{n \to \infty} E\left[\sup_{\theta \in \Theta} ||R_n^{\gamma}(\dot{\mathbb{Z}}_n(\theta) - \dot{Z}_{\theta_0}(\theta))||^{pa/\gamma}\right] = 0. \tag{5}$$

Suppose further that either of the following [M1] or [M2] is satisfied:

[M1] There exists a random matrix $J(\theta_0)$ which is positive definite almost surely such that $\dot{Z}(\theta) \leq -J(\theta_0)$ for all $\theta \in \Theta$, almost surely, and that $E[||J(\theta_0)^{-1}||^{pb/\gamma}] < \infty$;

[M2] $E[\sup_{\theta\in\Theta} ||\dot{Z}_{\theta_0}(\theta)^{-1}||^{pb/\gamma}] < \infty$, where the random matrices $\dot{Z}_{\theta_0}(\theta)$'s are assumed to be regular almost surely.

Then, for any sequence of Θ-valued random vectors $\widehat{\theta}_n$ such that $||R_n\mathbb{Z}_n(\widehat{\theta}_n)||$ is asymptotically L_{pa}-bounded, it holds that $||R_n(\widehat{\theta}_n - \theta_0)||$ is asymptotically L_p-bounded. Therefore, in this situation, whenever we also have that $R_n(\widehat{\theta}_n - \theta_0) \to^d G(\theta_0)$ where $G(\theta_0)$ is a random vector, it holds for any continuous function f satisfying (1) for $q \in (0, p)$ that

$$\lim_{n\to\infty} E[f(R_n(\widehat{\theta}_n - \theta_0))] = E[f(G(\theta_0))],$$

where the limit is also finite.

When the last condition in [M1] is satisfied with $||J(\theta_0)||^{-1}$ which is bounded or the first condition in [M2] is satisfied with $\sup_{\theta\in\Theta} ||\dot{Z}_{\theta_0}(\theta)^{-1}||$ which is bounded, the constant a appearing in the above claim may be replaced by 1.

Remark. When the last condition in [M1] is satisfied with $||J(\theta_0)||^{-1}$ which is bounded or the first condition in [M2] is satisfied with $\sup_{\theta\in\Theta} ||\dot{Z}_{\theta_0}(\theta)^{-1}||$ which is bounded, the constant a appearing in the above claim may be replaced by 1.
Remark. The crucial point in the course of applying this theorem is to check the condition (5) together with [M1] or [M2]. This is clearly satisfied for moment estimators described in Example A.
Remark. The condition [M1] above is corresponding to the case $\rho = 2$ of the conditions [A3] and [A5] in Yoshida (2011), which are, in the notations of our style,

$$M_{\theta_0}(\theta) - M_{\theta_0}(\theta_0) \leq -\chi(\theta_0)||\theta - \theta_0||^\rho, \quad \forall\theta \in \Theta,$$

where $M_{\theta_0}(\theta)$ denotes the "limit" of $\mathbb{M}_n(\theta)$, and high order moment conditions on the positive random variable $\chi(\theta_0)^{-1}$.

Proof of Theorem 2.1. Recalling (3), it follows from the Taylor expansion that

$$(R_n\mathbb{Z}_n(\widehat{\theta}_n))^{(i)} = (R_n\mathbb{Z}_n(\theta_0))^{(i)} + (\dot{\mathbb{Z}}_n(\widetilde{\theta}_n)R_n(\widehat{\theta}_n - \theta_0))^{(i)}, \quad i = 1, \ldots, d. \tag{6}$$

So we have

$$R_n(\widehat{\theta}_n - \theta_0) = A_n + B_nR_n(\widehat{\theta}_n - \theta_0), \tag{7}$$

where

$$\begin{aligned} A_n &= V_n(\theta_0)^{-1}R_n(\mathbb{Z}_n(\theta_0) - \mathbb{Z}_n(\widehat{\theta}_n)), \\ B_n &= V_n(\theta_0)^{-1}(\dot{\mathbb{Z}}_n(\widetilde{\theta}_n) + V_n(\theta_0)), \end{aligned}$$

and $\widetilde{\theta}_n$ is a random vector on the segment connecting θ_0 and $\widehat{\theta}_n$. It follows from the extended continuous mapping theorem (e.g., Theorem 1.11.1 of van der Vaart

and Wellner (1996)) that $V_n(\theta_0)^{-1} \to^p V(\theta_0)^{-1}$, thus we have $||A_n|| = O_P(1)$ and $||B_n|| = o_P(1)$. It therefore holds that

$$||R_n(\widehat{\theta}_n - \theta_0)|| \leq O_P(1) + o_P(1) \cdot ||R_n(\widehat{\theta}_n - \theta_0)||,$$

which implies that $||R_n(\widehat{\theta}_n - \theta_0)|| = O_P(1)$. Hence, going back to (7) we obtain

$$R_n(\widehat{\theta}_n - \theta_0) = R_n A_n + o_P(1) = V_n(\theta_0)^{-1} R_n \mathbb{Z}_n(\theta_0) + o_P(1).$$

The last claim is also a consequence of the extended continuous mapping theorem. The proof is finished. □

Proof of Theorem 2.3. We will give a proof for the case where [M1] is assumed. The proof for the case where [M2] is assumed is similar (and simpler), so it is omitted.

Due to (6) again, we have

$$R_n(\widehat{\theta}_n - \theta_0) = C_n + (D_n^{(1)} + D_n^{(2)}) R_n(\widehat{\theta}_n - \theta_0),$$

where

$$\begin{aligned} C_n &= J(\theta_0)^{-1} R_n(\mathbb{Z}_n(\theta_0) - \mathbb{Z}_n(\widehat{\theta}_n)), \\ D_n^{(1)} &= J(\theta_0)^{-1}(\dot{\mathbb{Z}}_n(\widetilde{\theta}_n) - \dot{Z}_{\theta_0}(\widetilde{\theta}_n)), \\ D_n^{(2)} &= J(\theta_0)^{-1}(\dot{Z}_{\theta_0}(\widetilde{\theta}_n) + J(\theta_0)), \end{aligned}$$

where $\widetilde{\theta}_n$ is a random vector on the segment connecting θ_0 and $\widehat{\theta}_n$.

From now on, we consider the case $\gamma \in (0,1)$; the proof for the case $\gamma = 1$ is easier, and it is omitted. Since $-D_n^{(2)}$ is non-negative definite almost surely, it follows from Minkowskii's and Hölder's inequalities that

$$\begin{aligned} &(E[||R_n(\widehat{\theta}_n - \theta_0)||^p])^{1/p} \\ &\quad \leq (E[||(I_d - D_n^{(2)}) R_n(\widehat{\theta}_n - \theta_0)||^p])^{1/p} \\ &\quad \leq (E[||C_n||^p])^{1/p} + (E[||R_n^\gamma D_n^{(1)}||^{p/\gamma}])^{\gamma/p} (E[||R_n^{1-\gamma}(\widehat{\theta}_n - \theta_0)||^{p/(1-\gamma)}])^{(1-\gamma)/p} \\ &\quad \leq O(1) + o(1) \cdot (E[||R_n^{1-\gamma}(\widehat{\theta}_n - \theta_0)||^{p/(1-\gamma)}])^{(1-\gamma)/p}, \end{aligned}$$

where we have used Hölder's inequality again to get

$$E[||C_n||^p] \leq (E[||J(\theta_0)^{-1}||^{pb}])^{1/b} (E[||R_n(\mathbb{Z}_n(\theta_0) - \mathbb{Z}_n(\widehat{\theta}_n))||^{pa})^{1/a}$$

and

$$E[||R_n^\gamma D_n^{(1)}||^{p/\gamma}] \leq (E[||J(\theta_0)^{-1}||^{pb/\gamma}])^{1/b} (E[||R^\gamma(\dot{\mathbb{Z}}_n(\widetilde{\theta}_n) - \dot{Z}_{\theta_0}(\widetilde{\theta}_n))||^{pa/\gamma})^{1/a};$$

if $||J(\theta_0)||^{-1}$ is bounded, we can get this kind of bound with $a = 1$.

Notice that

$$\begin{aligned} & ||R_n^{1-\gamma}(\widehat{\theta}_n - \theta_0)||^{1/(1-\gamma)} \\ \leq & \sqrt{d^{(1/(1-\gamma))-1} \sum_{i=1}^{d} |R_n^{(i,i)}|^2 |\widehat{\theta}_n^{(i)} - \theta_0^{(i)}|^{2/(1-\gamma)}} \\ \leq & ||R_n(\widehat{\theta}_n - \theta_0)|| \cdot d^{1/(2-2\gamma)} \cdot |\mathcal{D}(\Theta)|^{\gamma/(1-\gamma)}, \end{aligned}$$

where $\mathcal{D}(\Theta)$ denotes the diameter of Θ. So we obtain

$$\begin{aligned} & (E[||R_n(\widehat{\theta}_n - \theta_0)||^p])^{1/p} \\ \leq & O(1) + o(1) \cdot (E[||R_n(\widehat{\theta}_n - \theta_0)||^p])^{(1-\gamma)/p} \\ \leq & O(1) + o(1) \cdot (E[||R_n(\widehat{\theta}_n - \theta_0)||^p] \vee 1)^{1/p}, \end{aligned}$$

which yields that

$$E[||R_n(\widehat{\theta}_n - \theta_0)||^p] \leq O(1) + o(1) \cdot E[||R_n(\widehat{\theta}_n - \theta_0)||^p].$$

Therefore, $||R_n(\widehat{\theta}_n - \theta_0)||$ is asymptotically L_p-bounded. □

3 Examples

In this section we give some detailed discussions about moment convergence of Z-estimators for Examples B, C and D, respectively.

3.1 Example B: Ergodic diffusion process

Recall the description of Example A in Section 1, where the first d_A-components α of the parameter $\theta = (\alpha^\top, \beta^\top)^\top$ is involved in the drift coefficient, and the latter d_B-components β is in the diffusion coefficient. Recalling also the definition of the rate matrix R_n there, let us consider the $(d_A + d_B)$-dimensional random vectors $\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta)$ and the $(d_A + d_B) \times (d_A + d_B)$-random matrices $\dot{\mathbb{Z}}_n(\theta) = \ddot{\mathbb{M}}_n(\theta)$ given by the trivial notations as follows:

$$\begin{aligned} \dot{\mathbb{M}}_n(\theta) &= (\dot{\mathbb{M}}_n^A(\theta)^\top, \dot{\mathbb{M}}_n^B(\theta)^\top)^\top, \\ \ddot{\mathbb{M}}_n(\theta) &= \begin{pmatrix} \ddot{\mathbb{M}}_n^A(\theta) & \ddot{\mathbb{M}}_n^C(\theta) \\ \ddot{\mathbb{M}}_n^C(\theta)^\top & \ddot{\mathbb{M}}_n^B(\theta) \end{pmatrix}. \end{aligned}$$

Below, we will use the following notation: for a given constant $p \geq 1$ and a given sequence of positive constants r_n,

$$\xi_n = o_{M(p)}(r_n^{-1}) \quad \Longleftrightarrow \quad r_n E[||\xi_n||^p] \to 0. \tag{8}$$

Notice that $\xi_n = o_{M(1)}(r_n^{-1})$ implies that $\xi_n = o_P(r_n^{-1})$.

Under some regularity conditions which are usually assumed in the asymptotic theory for ergodic diffusion process models, it is standard to show the following facts (see e.g. the appendix of Kessler (1997) and Nishiyama (2011) for the detailed proofs of the techniques that are omitted in Kessler's (1997) appendix):

$$\left|\left|\dot{\mathbb{M}}_n^A(\theta_0) - \frac{1}{t_n^n}\sum_{k:t_{k-1}^n \le t_n^n} \frac{\dot{S}(X_{t_{k-1}^n};\alpha_0)}{\sigma(X_{t_{k-1}^n};\beta_0)}(W_{t_k^n} - W_{t_{k-1}^n})\right|\right| = o_{M(p)}((t_n^n)^{-1/2}),$$

$$\left|\left|\dot{\mathbb{M}}_n^B(\theta_0) - \frac{1}{n}\sum_{k:t_{k-1}^n \le t_n^n} \frac{\dot{\sigma}(X_{t_{k-1}^n};\beta_0)}{\sigma(X_{t_{k-1}^n};\beta_0)}\left\{\frac{|W_{t_k^n} - W_{t_{k-1}^n}|^2}{|t_k^n - t_{k-1}^n|} - 1\right\}\right|\right| = o_{M(p)}(n^{-1/2}),$$

$$\sup_{\theta\in\Theta}\left|\left|\ddot{\mathbb{M}}_n^A(\theta) - \frac{1}{t_n^n}\sum_{k:t_{k-1}^n \le t_n^n} H^A(X_{t_{k-1}^n};\theta_0,\theta)|t_k^n - t_{k-1}^n|\right|\right| = o_{M(p)}((t_n^n)^{-1/2}),$$

$$\sup_{\theta\in\Theta}\left|\left|\ddot{\mathbb{M}}_n^B(\theta) - \frac{1}{n}\sum_{k:t_{k-1}^n \le t_n^n} H^B(X_{t_{k-1}^n};\theta_0,\theta)\right|\right| = o_{M(p)}(n^{-1/2}),$$

$$\sup_{u\in[0,1]}\sup_{\theta\in\Theta}||\ddot{\mathbb{M}}_n^C(u,\theta)|| = o_{M(p)}(n^{-1/4}),$$

where

$$\begin{aligned} H^A(x;\theta_0,\theta) &= \frac{\ddot{S}(x;\alpha)(S(x;\alpha_0) - S(x;\alpha)) - \dot{S}(x;\alpha)\dot{S}(x;\alpha)^\top}{\sigma(x;\beta)^2}, \\ H^B(x;\theta_0,\theta) &= \left\{\frac{\ddot{\sigma}(x;\beta)}{\sigma(x;\beta)^3} - 3\frac{\dot{\sigma}(x;\beta)\dot{\sigma}(x;\beta)^\top}{\sigma(x;\beta)^4}\right\}(\sigma(x;\beta_0)^2 - \sigma(x;\beta)^2) \\ &\quad -2\frac{\dot{\sigma}(x;\beta)\dot{\sigma}(x;\beta)^\top}{\sigma(x;\beta)^2}. \end{aligned}$$

The regularity conditions for the above claims depend on the constant $p \ge 1$ appearing in "$o_{M(p)}(r_n^{-1})$" which we need to have.

The assumption (4) can be checked by applying Burkholder-Davis-Gundy's inequality to the main part of $R_n\dot{\mathbb{M}}_n(\theta_0)$. On the other hand, noting also $t_n^n \le n$, we can apply Remark 1 (ii) of Uchida and Yoshida (2012) to show that the assumption (5) for $\ddot{\mathbb{M}}_n(\theta)$ is satisfied with the limits

$$\ddot{M}_{\theta_0}(\theta) = \begin{pmatrix} \ddot{M}_{\theta_0}^A(\theta) & 0 \\ 0 & \ddot{M}_{\theta_0}^B(\theta) \end{pmatrix},$$

where

$$\ddot{M}_{\theta_0}^A(\theta) = \int_I H^A(x;\theta_0,\theta)\mu_{\theta_0}(dx) \quad \text{and} \quad \ddot{M}_{\theta_0}^B(\theta) = \int_I H^B(x;\theta_0,\theta)\mu_{\theta_0}(dx),$$

and μ_{θ_0} denotes the invariant distribution of X when the true value is θ_0. In order to make the assumption [M1] or [M2] fulfilled, we have to introduce the parametric

model for the drift and diffusion coefficients nicely. An example for which the assumption [M1] can be easily checked is $S(\cdot;\alpha) = \alpha^\top a(\cdot)$ and $\sigma(\cdot;\beta) = e^{\beta^\top b(\cdot)}$, where $a(\cdot)$ and $b(\cdot)$ are some vectors of known functions, assuming that $b(\cdot)$ is bounded. The assumption [M2] would be satisfied in more general parametric models, because $\ddot{\mathbb{M}}(\theta)$'s are non-random in this example.

3.2 Example C: Volatility of diffusion process

Recall the description of Example C in Section 1. An interesting point of this example is that the limit of $-\ddot{\mathbb{M}}_n(\widetilde{\theta}_n)$ is random.

Let a constant $p \geq 1$ be given, and recall the notation (8). Under some regularity conditions, it holds that

$$\left|\left|\dot{\mathbb{M}}_n(\theta_0) - \frac{1}{n}\sum_{k:t^n_{k-1}\leq t^n_n} \frac{\dot{\sigma}(X_{t^n_{k-1}};\theta_0)}{\sigma(X_{t^n_{k-1}};\theta_0)}\left\{\frac{|W_{t^n_k} - W_{t^n_{k-1}}|^2}{|t^n_k - t^n_{k-1}|} - 1\right\}\right|\right| = o_{M(p)}(n^{-1/2}),$$

$$\sup_{\theta\in\Theta}\left|\left|\ddot{\mathbb{M}}_n(\theta) - \frac{1}{n}\sum_{k:t^n_{k-1}\leq t^n_n} H(X_{t^n_{k-1}};\theta_0,\theta)\right|\right| = o_{M(p)}(n^{-1/2}),$$

where

$$\begin{aligned} H(x;\theta_0,\theta) &= \left\{\frac{\ddot{\sigma}(x;\theta)}{\sigma(x;\theta)^3} - 3\frac{\dot{\sigma}(x;\theta)\dot{\sigma}(x;\theta)^\top}{\sigma(x;\theta)^4}\right\}(\sigma(x;\theta_0)^2 - \sigma(x;\theta)^2) \\ &\quad -2\frac{\dot{\sigma}(x;\theta)\dot{\sigma}(x;\theta)^\top}{\sigma(x;\theta)^2}. \end{aligned}$$

The regularity conditions for the above claims depend on the constant $p \geq 1$ which we need to have. Moreover, under some standard conditions, it holds that for any sequence of random vectors $\widetilde{\theta}_n$ such that $||\widetilde{\theta}_n - \theta_0|| \to^p 0$,

$$||\ddot{\mathbb{M}}_n(\widetilde{\theta}_n) + V_n(\theta_0)|| \to^p 0,$$

where

$$V_n(\theta_0) = \frac{2}{n}\sum_{k:t^n_{k-1}\leq t^n_n} \frac{\dot{\sigma}(X_{t^n_{k-1}};\theta_0)\dot{\sigma}(X_{t^n_{k-1}};\theta_0)^\top}{\sigma(X_{t^n_{k-1}};\theta_0)^2}.$$

Also, it follows that

$$(\sqrt{n}\dot{\mathbb{M}}_n(\theta_0), V_n(\theta_0)) \to^d (V(\theta_0)^{1/2}Z, V(\theta_0))$$

where Z is a standard Gaussian random vector which is independent of the random matrix $V(\theta_0)$ given by

$$V(\theta_0) = 2\int_0^T \frac{\dot{\sigma}(X_s;\theta_0)\dot{\sigma}(X_s;\theta_0)^\top}{\sigma(X_s;\theta_0)^2} ds.$$

Due to the above facts Theorem 2.1 yields that for any consistent estimator $\widehat{\theta}_n$ for θ_0 satisfying $||\dot{\mathbb{M}}_n(\widehat{\theta}_n)|| = o_P(n^{-1/2})$ we have $\sqrt{n}(\widehat{\theta}_n - \theta_0) \to^d V(\theta_0)^{-1/2} Z$,

Next let us apply Theorem 2.3. The assumption (4) for $\sqrt{n}\dot{\mathbb{M}}_n(\theta_0)$ can be checked by using Burkholder-Davis-Gundy's inequality. In the case of this example, checking that the assumption (5) for $\ddot{\mathbb{M}}_n(\theta)$ is satisfied with

$$\ddot{M}_{\theta_0}(\theta) = \int_0^T H(X_t; \theta_0, \theta) dt$$

is easy. In order to make the assumption [M1] or [M2] fulfilled, we again have to introduce the parametric model for the diffusion coefficients nicely. An example for which the former assumption in [M1] can be easily checked is $\sigma(\cdot; \theta) = e^{\theta^\top g(\cdot)}$, where $g(\cdot)$ are some vectors of known, bounded functions. The latter assumption in [M1] is then reduced to

$$E\left[\left|\left|\left(\int_0^T g(X_t) g(X_t)^\top dt\right)^{-1}\right|\right|^{pb/\gamma}\right] < \infty,$$

for which we can give a clear sufficient condition for the function g at least in the one-dimensional case (for example, just assume $|g(\cdot)|^2 \geq c$ for a constant $c > 0$).

3.3 Example D: Cox's regression model

Recall the description of Example D in Section 1.

Introducing the notations

$$\begin{aligned}
S_t^{n,0}(\theta) &= \sum_{k=1}^n e^{\theta Z_t^k} Y_t^k, \\
S_t^{n,1}(\theta) &= \sum_{k=1}^n Z_t^k e^{\theta Z_t^k} Y_t^k, \\
S_t^{n,2}(\theta) &= \sum_{k=1}^n (Z_t^k)^\top Z_t^k e^{\theta Z_t^k} Y_t^k,
\end{aligned}$$

we suppose that

$$\sup_{\theta \in \Theta} \sup_{t \in [0,T]} \left|\left| \frac{1}{n} S_t^{n,l}(\theta) - \mathcal{S}_t^l(\theta) \right|\right| \to^p 0, \quad l = 0, 1, 2,$$

where the limits $t \rightsquigarrow \mathcal{S}_t^l$ are some stochastic processes (c.f. Andersen and Gill (1982) who assumed that $\mathcal{S}^l$'s are not random).

Then, some arguments similar to Section 3.1 are possible for

$$\begin{aligned}
\dot{M}_{\theta_0}(\theta) &= \int_0^T \left(\frac{\mathcal{S}_t^1(\theta_0)}{\mathcal{S}_t^0(\theta_0)} - \frac{\mathcal{S}_t^1(\theta)}{\mathcal{S}_t^0(\theta)} \right) \mathcal{S}_t^0(\theta_0)\alpha(t)dt, \\
V_n(\theta_0) &= \frac{1}{n}\int_0^{uT} \frac{S_t^{n,0}(\theta_0)S_t^{n,2}(\theta_0) - S_t^{n,1}(\theta_0)S_t^{n,1}(\theta_0)^\top}{S_t^{n,0}(\theta_0)}\alpha(t)dt, \\
V(\theta) &= \int_0^T \frac{\mathcal{S}_t^0(\theta)\mathcal{S}_t^2(\theta) - \mathcal{S}_t^1(\theta)\mathcal{S}_t^1(\theta)^\top}{\mathcal{S}_t^0(\theta)^2}\mathcal{S}_t^0(\theta_0)\alpha(t)dt
\end{aligned}$$

The details are omitted.

Acknowledgements. This work was supported by Italian MIUR, Grant 2009 (I.N.) and by Grant-in-Aid for Scientific Research (C), 24540152, from Japan Society for the Promotion of Science (Y.N.).

References

[1] Aït-Sahalia, Y. (1999). Transition densities for interest rate and other nonlinear diffusion. *J. Finance* **54**, 1361-1395.

[2] Andersen, P.K. and Gill, R.D. (1982). Cox's regression models for counting processes: A large sample study. *Ann. Statist.* **10**, 1100-1120.

[3] Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B* **34**, 187-220.

[4] Ibragimov, I.A. and Has'minskii, R.Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, New York.

[5] Kato, K. (2011). A note on moment convergence of bootstrap M-estimators. *Statist. Decision* **28**, 51-61.

[6] Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.* **24**, 211-229.

[7] Kutoyants, Yu.A. (1984). *Parameter Estimation for Stochastic Processes.* Heldermann, Berlin.

[8] Kutoyants, Yu.A. (1994). *Identification of Dynamical Systems with Small Noise.* Kluwer Academic Publishers, Dordrecht.

[9] Kutoyants, Yu.A. (2004). *Statistical Inference for Ergodic Diffusion Processes.* Springer-Verlag, London.

[10] Nishiyama, Y. (2010). Moment convergence of M-estimators. *Statist. Neerlandica* **64**, 505-507.

[11] Nishiyama, Y. (2011). *Statistical Analysis by the Theory of Martingales.* (In Japanese.) ISM Series **1**, Kindaikagakusha, Tokyo.

[12] Uchida, M. and Yoshida, N. (2012). Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Process. Appl.* **122**, 2885-2924.

[13] van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

[14] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer-Verlag, New York.

[15] Yoshida, N. (2011). Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Ann. Inst. Statist. Math.* **63**, 431-479.