Josai Mathematical Monographs
Vol. 11 (2018) pp. 3–14

Asymptotic behavior of regularized estimator under multiple and mixed-rates asymptotics

Yusuke Shimizu

Abstract. Masuda and Shimizu (2017) consider the uniform tail-probability estimate of a class of scaled regularized estimators under multiple and mixed-rates asymptotics in the sense of Radchenko (2008), where the associated statistical random fields may be non-differentiable and may fail to be partially locally asymptotically quadratic so that the conventional approach through the polynomial type large deviation inequality (PLDI) developed by Yoshida (2011) does not work directly. In this paper, we generalize the form of regularization terms considered in Masuda and Shimizu (2017), and derive the asymptotic behaviors including the moment convergence of estimator. Our setting includes sparsely regularized M-estimation such that sparse-bridge, the smoothly clipped absolute deviation and Seamless-L_0 regularization.

1. Introduction

Suppose that we observe data $\mathbf{X}_n$, the distribution of which is indexed by a finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^p$. In order to estimate θ based on $\mathbf{X}_n$, we usually introduce an appropriate (quasi-)likelihood or contrast function $\mathbb{H}_n : \Omega \times \Theta \to \mathbb{R}$, and estimate an optimal parameter value θ_0 by any point $\hat{\theta}_n \in \operatorname{argmin} \mathbb{H}_n$. For assessing asymptotic performance of $\hat{\theta}_n$ quantitatively, we look at the statistical random fields

$$\mathbb{M}_n(u;\theta_0) = \mathbb{H}_n(\theta_0 + A_n(\theta_0)u) - \mathbb{H}_n(\theta_0), \tag{1}$$

where $A_n(\theta_0)$ denotes the rate matrix such that $|A_n(\theta_0)| \to 0$ as $n \to \infty$ and the components may decrease at different rates; estimation with multiple-rates of convergence has appeared in the literature of, for example, econometrics [3]. Throughout this paper, we use the notation $|A|^2 = \operatorname{tr}(AA^\top)$ for a matrix A with $\top$ denoting the transpose. As is well-known, the weak convergence of $\mathbb{M}_n$ to some $\mathbb{M}_0$ over compact sets, the identifiability condition on $\mathbb{M}_0$, and the tightness of the

2010 Mathematics Subject Classification. Primary 62E20; Secondary 62F12.
Key Words and Phrases. regularized estimation; moment convergence; large deviation inequality; sparse estimation; stochastic differential equation; mixed-rates asymptotics.

scaled estimator $\hat{u}_n := A_n(\theta_0)^{-1}(\hat{\theta}_n - \theta_0)$ make the "argmin" functional continuous for $\mathbb{M}_n$: $\hat{u}_n \in \operatorname{argmin} \mathbb{M}_n \xrightarrow{\mathcal{L}} \operatorname{argmin} \mathbb{M}_0$. See e.g., [21, Section 5]. Further, when concerned with moments of $\hat{u}_n$-dependent statistics, such as the mean squared error, more than the weak convergence is required. Then the polynomial type large deviation inequality (PLDI) of [22], which estimates the tail of $\mathcal{L}(\hat{u}_n)$ in such a way that

$$\sup_{r>0} \sup_{n>0} r^L P(|\hat{u}_n| \geq r) < \infty \tag{2}$$

for a given $L > 0$, plays an important role: we set $\hat{u}_0 \in \operatorname{argmin} \mathbb{M}_0$ for a random variable $\hat{u}_0$. The moment convergence

$$E[|\hat{u}_n|^q] \to E[|\hat{u}_0|^q] \tag{3}$$

for some $q > 0$ holds if there exists a $q' > q$ such that $\sup_{n>0} E[|\hat{u}_n|^{q'}] < \infty$. Let us assume that the PLDI (2) holds for some $L > q'$. Then we obtain

$$\sup_{n>0} E[|\hat{u}_n|^{q'}] = \sup_{n>0} \int_0^\infty P(|\hat{u}_n|^{q'} > s) ds < \infty.$$

It has been known that the PLDI can be proved under modest conditions when $\mathbb{M}_n$ admit a locally asymptotically quadratic (LAQ) structure, which is satisfied for many situations including asymptotically mixed-normal type models under multi-scaling. Here, in the multi-scaling case where the random vector $\hat{\theta}_n$ converges at different rates, the LAQ structure at "first" step takes the form*

$$\mathbb{M}_n(u, \tau; \theta_0) = \Delta_n(\tau; \theta_0)[u] + \frac{1}{2}\Gamma_0(\tau; \theta_0)[u, u] + r_n(u, \tau; \theta_0), \tag{4}$$

where we are required to verify, among others, the following conditions which are to hold uniformly in "the second- and the subsequent-step" parameter τ, which is regarded as a nuisance parameter in the first step: sufficient integrability of the random linear form $\Delta_n(\tau; \theta_0)$; the non-degeneracy of the possibly random bi-linear form $\Gamma_0(\tau; \theta_0)$; and a kind of "non-explosiveness" of the scaled remainder term $(1 + |u|^2)^{-1} r_n(u, \tau; \theta_0)$, where the u-pointwise limit of $r_n(u, \tau; \theta_0)$, whenever exists, typically equals zero. For notational convenience, here and in the sequel we write $A[b_1, \ldots, b_m] = \sum_{i_1, \ldots, i_m} A_{i_1 \ldots i_m} b_{1 i_1} \ldots b_{m i_m}$ for multilinear forms $A = \{A_{i_1 \ldots i_m}\}_{i_1, \ldots, i_m}$ and $b_j = \{b_{j i_k}\}_{i_k}$; sometimes b_j themselves may be tensors, hence

*The sign in front of the quadratic term $(1/2)\Gamma_0(\tau; \theta_0)[u, u]$ is different from the original LAQ of [22] since we consider minimization of (1).

the resulting form is also a multilinear form, e.g. $A[B,C] = \{\sum_{i,j} A_{ij}B_{ik}C_{jl}\}_{k,l}$ for $A = \{A_{ij}\}$, $B = \{B_{ik}\}$, and $C = \{C_{jl}\}$. See [22, Section 5] for the detailed account of the above-mentioned multistep procedures. In many standard statistical models, the form (1) is enough to find the asymptotic distribution of all the components of $\hat{\theta}_n$.

In principle, any M-estimation procedure, typically producing an asymptotically mixed-normally distributed estimator, may have its "regularized" counterpart; we refer to [4] for some general backgrounds of statistical regularization. We are concerned here with extending the random-field structure to deal with possibly dependent data and a broader class of regularized M-estimation under the "mixed-rates" asymptotics. In particular, we will show how the PLDI of [22] can carry over to the mixed-rates M-estimation where the target statistical random fields may have components converging at different rates; we refer to [12] and [17] for details in case of linear regression with general regularization term. We will adopt the very general theoretical framework developed by [15, Sections 2 and 3]. It will be shown that the PLDI criterion of [22] can apply to some mixed-rates cases while it may require some modification when the key LAQ structure of the original statistical random field fails to hold; it may even happen that $r_n(u,\tau;\theta_0)$ diverges in probability. Indeed, most of the existing sparse estimation procedures may fall into this type of asymptotics. Consequently, with a true parameter being fixed, our moment-convergence result provides yet another theoretical insight about the regularized estimation, the well-established methodology especially in variable and/or model selection.[†] The logic of the sparse and more generally shrinkage estimation would be best and most clearly described by the context of multiple linear regression, with many deep theoretical interpretation such as geometrical (projection) characterization, variable selection, stabilized prediction performance, etc. See e.g. [10, Chapter 3].

There exist a lot of previous works on moment convergence of estimators. It serves as a fundamental tool when analyzing asymptotic behavior of the expectations of statistics depending on the estimator such as asymptotic bias and mean squared prediction error; to mention just a few, we refer to [5], [8], [11], [13], [16], [17], [18], [19], as well as [22]. Also, the convergence of moments of regularized sparse maximum-likelihood estimator of generalized linear model was deduced in [20] to verify the AIC type variable-selection. Further, [1] recently discussed

[†]It should be noted that the sparse estimation has received mixed reception from a kind of estimation singularity similar to that of the classical Hodge's super efficient estimator. The unpleasant feature of the sparse-type estimator essentially stems from non-uniformity in weak convergence with respect to the true value of parameters, see [13] for details.

optimal selection of random and k-fold cross-validation estimators, the theoretical backbone of which involves some moment bounds of the estimators used; the related paper [2] studied the uniform integrability of the ordinary least-squares estimator in the linear regression setting.

The goal of this paper is to generalize the form of regularization terms considered in [12], which derives the uniform tail-probability estimate of a class of scaled regularized estimators under multiple and mixed-rates asymptotics in the sense of [15], where the associated statistical random fields may be non-differentiable and may fail to be partially LAQ structure so that the conventional approach through the PLDI developed by [22] does not work directly.

Section 2 describes our model setup which includes sparsely regularized M-estimation such that the Seamless-L_0 regularization [6], the smoothly clipped absolute deviation [7] and the sparse-bridge [14]. Under the model descridbed in Section 2, we give a series of basic asymptotic statements in Section 3, where, in particular, the polynomial type large deviation estimate of the underlying statistical random fields will play a crucial role for the uniform tail-probability estimate concerning the scaled M-estimator; although the asymptotics is classical, in the literature there seems to exist no unified tools that can handle general M-estimation of multiple-rates and possibly mixed-rates type, and importantly, of possibly non-differentiable and non-convex type. The claims in Section 3 are the same as in [12, Section 3], although, note again that, the setting described in Section 2 generalizes that of [12].

2. Setup

Let us begin with description of the basic model setup for Section 3. Throughout we are given an underlying probability space $(\Omega, \mathcal{F}, P)$. For the purpose of accelerating estimation performance, we consider M-estimation of an additive regularization type. We will focus on the case of two-scaling, where the target statistical parameter $\theta \in \Theta$ is divided into two parts, say $\theta = (\alpha, \beta)$; an extension to cases of more-than-two scaling is a trivial matter while making notation messy. We set $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^q$, and $\Theta = \Theta_\alpha \times \Theta_\beta$ to be a bounded convex domain in $\mathbb{R}^{p+q}$.

We are given a function $M_n : \Omega \times \Theta \to \mathbb{R}$, and regularization (possibly random) functions $\overline{R}_n^a(\alpha)$ and $\overline{R}_n^b(\beta)$. We then consider contrast functions $\mathbb{H}_n : \Omega \times \Theta \to \mathbb{R}$ of the form

$$\mathbb{H}_n(\theta) = \mathbb{H}_n(\alpha, \beta) = M_n(\alpha, \beta) + \overline{R}_n^a(\alpha) + \overline{R}_n^b(\beta). \tag{5}$$

The associated regularized M-estimator is defined to be any element (for brevity, implicitly assumed to exist)

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \overline{\Theta}} \mathbb{H}_n(\theta).$$

We quantitatively distinguish zero parameters from non-zero ones. We denote by $\theta_0 = (\alpha_0, \beta_0)$ the value we want to estimate (typically the true value of θ) and assume that it takes the form $\alpha_0 = (\alpha_0^\circ, \alpha_0^*) = ((\alpha_{0,k'}^\circ)_{k'}, (\alpha_{0,k''}^*)_{k''})$ and $\beta_0 = (\beta_0^\circ, \beta_0^*) = ((\beta_{0,l'}^\circ)_{l'}, (\beta_{0,l''}^*)_{l''})$ with

$$\alpha_{0,k'}^\circ = 0, \quad \beta_{0,l'}^\circ = 0, \quad \alpha_{0,k''}^* \neq 0, \quad \beta_{0,l''}^* \neq 0.$$

We set $\alpha_0^\circ \in \mathbb{R}^{p^\circ}$, $\beta_0^\circ \in \mathbb{R}^{q^\circ}$, $\alpha_0^* \in \mathbb{R}^{p^*}$ and $\beta_0^* \in \mathbb{R}^{q^*}$ with $p^\circ, q^\circ, p^*, q^* \in \mathbb{N}$; then, $p = p^\circ + p^*$ and $q = q^\circ + q^*$. Correspondingly, we write $\theta = (\theta^\circ, \theta^*)$ with $\theta^\circ = (\alpha^\circ, \beta^\circ)$ and $\theta^* = (\alpha^*, \beta^*)$ in the obvious manner. We also write $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n) = (\hat{\alpha}_n^\circ, \hat{\alpha}_n^*, \hat{\beta}_n^\circ, \hat{\beta}_n^*)$ with $\hat{\theta}_n^\circ = (\hat{\alpha}_n^\circ, \hat{\beta}_n^\circ)$ and $\hat{\theta}_n^* = (\hat{\alpha}_n^*, \hat{\beta}_n^*)$. In this paper, we focus on the following regularization terms:

$$\overline{R}_n^a(\alpha) = \sum_{k=1}^{p} \mathfrak{p}_{n,k}^a(\alpha_k), \qquad \overline{R}_n^b(\beta) = \sum_{l=1}^{q} \mathfrak{p}_{n,l}^b(\beta_l), \tag{6}$$

where $\mathfrak{p}_{n,k}^a(\cdot)$ and $\mathfrak{p}_{n,l}^b(\cdot)$ are random and non-negative functions such that $\mathfrak{p}_{n,k}^a(0) = \mathfrak{p}_{n,l}^b(0) = 0$ a.s. for any $k \in \{1, \ldots, p\}$ and $l \in \{1, \ldots, q\}$. Note that this type of regularization terms generalize that of [12], and subsume many of the existing types, e.g., [6], [7], [9], [14] and [23] for linear regression model. For convenience of reference in the regularity conditions given later, we write

$$\overline{R}_n^a(\alpha) = \overline{R}_n^{a\circ}(\alpha^\circ) + \overline{R}_n^{a*}(\alpha^*) = \sum_{k'=1}^{p^\circ} \mathfrak{p}_{n,k'}^{a\circ}(\alpha_{k'}^\circ) + \sum_{k''=1}^{p^*} \mathfrak{p}_{n,k''}^{a*}(\alpha_{k''}^*), \tag{7}$$

$$\overline{R}_n^b(\beta) = \overline{R}_n^{b\circ}(\beta^\circ) + \overline{R}_n^{b*}(\beta^*) = \sum_{l'=1}^{q^\circ} \mathfrak{p}_{n.l'}^{b\circ}(\beta_{l'}^\circ) + \sum_{l''=1}^{q^*} \mathfrak{p}_{n,l''}^{b*}(\beta_{l''}^*). \tag{8}$$

Conditions on the ingredient of M_n, $\mathfrak{p}_{n,\cdot}^a$ and $\mathfrak{p}_{n,\cdot}^b$ will be imposed later on.

We will deal with a situation where the non-zero part of the first component α can be estimated faster than that of the second component β; more specifically, we will suppose that the sequence

$$\left(s_n^{-1}(\hat{\alpha}_n^* - \alpha_0^*),\, t_n^{-1}(\hat{\beta}_n^* - \beta_0^*)\right)$$

has a non-trivial asymptotic distribution for some possibly different positive sequence (s_n) and (t_n), both tending to zero and satisfying that $s_n = o(t_n)$. Although not explicitly mentioned, we presuppose that the "principal" part $M_n(\theta)$ reasonably makes sense even without regularization terms $\overline{R}^a_n(\alpha) + \overline{R}^b_n(\beta)$; most typically, the un-regularized case, where $\mathbb{H}_n(\theta) = M_n(\theta)$, corresponds to a negative of a (quasi) log-likelihood.

3. Asymptotics

Under the setting described in Section 2, we state results analogous to Theorems 3.4, 3.8, 3.12, 3.15, and 3.21 in [12]. Note that the type of regularization terms (6) generalize that of [12] (see [12, section 2] for details). First, we cite an assumption from [12]:

ASSUMPTION 3.1 (ASSUMPTION 3.1 IN [12]).

1. *(s_n) and (t_n) are positive nonrandom sequences such that* $\max(s_n, t_n) \to 0$ *and that* $s_n = o(t_n)$.
2. *There exist continuous random functions* $\overline{M}^a_0 : \Omega \times \Theta_\alpha \to \mathbb{R}$ *and* $\overline{M}^b_0 : \Omega \times \Theta \to \mathbb{R}$ *such that:*
   (a) $\sup_\alpha \left| s_n^2 \left\{ M_n(\alpha, \beta_0) - M_n(\alpha_0, \beta_0) \right\} - \overline{M}^a_0(\alpha) \right|$
   $+ \sup_\theta \left| t_n^2 \left\{ M_n(\alpha, \beta) - M_n(\alpha, \beta_0) \right\} - \overline{M}^b_0(\theta) \right| \xrightarrow{p} 0$;
   (b) $\operatorname{argmin}_\alpha \overline{M}^a_0(\alpha) = \{\alpha_0\}$ *a.s. and* $\operatorname{argmin}_\beta \overline{M}^b_0(\alpha_0, \beta) = \{\beta_0\}$ *a.s.*
3. $\sup_\alpha \left| s_n^2 \overline{R}^a_n(\alpha) \right| + \sup_\beta \left| t_n^2 \overline{R}^b_n(\beta) \right| \xrightarrow{p} 0$.

Under the Assumption 3.1, [12, Theorem 3.4], which describes the consistency of $\hat{\theta}_n = (\hat{\theta}^\circ_n, \hat{\theta}^*_n)$:

$$\hat{\theta}_n \xrightarrow{p} \theta_0,$$

holds as it is (see [12] for details).

For [12, Theorem 3.8], which derives $(\hat{u}_n, \hat{v}_n) = O_p(1)$ where

$$\hat{u}_n := s_n^{-1}(\hat{\alpha}_n - \alpha_0), \qquad \hat{v}_n := t_n^{-1}(\hat{\beta}_n - \beta_0), \tag{9}$$

we set the following additional assumption:

ASSUMPTION 3.2 (MODIFIED ASSUMPTION 3.6 IN [12]).

1. $M_n \in \mathcal{C}^3(\Theta)$ *a.s., and it holds that:*
   (a) $\sup_{\beta} |s_n \partial_\alpha M_n(\alpha_0, \beta)| + |t_n \partial_\beta M_n(\theta_0)| = O_p(1)$;
   (b) $\sup_{\alpha} |s_n t_n \partial_\alpha \partial_\beta M_n(\alpha, \beta_0)| = O_p(1)$;
   (c) $\sup_{\theta} \left|s_n^2 \partial_\zeta \partial_\alpha^2 M_n(\theta)\right| + \sup_{\theta} \left|t_n^2 \partial_\zeta \partial_\beta^2 M_n(\theta)\right| = O_p(1)$ *for* $\zeta = \alpha, \beta$;
   (d) *There exist symmetric random functions* $\Gamma_0^\alpha : \Omega \times \Theta_\alpha \to \mathbb{R}^p \otimes \mathbb{R}^p$ *and* $\Gamma_0^\beta : \Omega \times \Theta \to \mathbb{R}^q \otimes \mathbb{R}^q$ *such that*

$$\left|s_n^2 \partial_\alpha^2 M_n(\theta_0) - \Gamma_0^\alpha(\alpha_0)\right| + \left|t_n^2 \partial_\beta^2 M_n(\theta_0) - \Gamma_0^\beta(\theta_0)\right| \xrightarrow{p} 0,$$

   with $\lambda_{\min}\big(\Gamma_0^\alpha(\alpha_0)\big) \wedge \lambda_{\min}\big(\Gamma_0^\beta(\theta_0)\big) > 0$ *a.s.*

2. *For all* $a_0,\ b_0 \neq 0$ *and* $m > 0$,

$$\sup_{k'',l''} \sup_{(a',b'):|a'|\vee|b'|\leq m} \frac{s_n \left|\mathfrak{p}_{n,k''}^{a*}(a') - \mathfrak{p}_{n,k''}^{a*}(a_0)\right| + t_n \left|\mathfrak{p}_{n,l''}^{b*}(b') - \mathfrak{p}_{n,l''}^{b*}(b_0)\right|}{|a' - a_0| + |b' - b_0|} = O_p(1).$$

Then, the following corollary is derived from [12, Theorem 3.8].

COROLLARY 3.3. *We have* $(\hat{u}_n, \hat{v}_n) = O_p(1)$ *under Assumptions 3.1 and 3.2.*

For the sparse consistency $P(\hat{\theta}_n^\circ = 0) \to 1$, we add the following assumption:

ASSUMPTION 3.4 (MODIFIED ASSUMPTION 3.11 IN [12]). *There exist constants* $\underline{a}_{k'}, \underline{b}_{l'} \in (0, 1/2)$ *such that*

$$P\Bigg(s_n^2 \partial_{\alpha_{k'}^\circ} M_n(0, \ldots, 0, \hat{\alpha}_{n,k'}^\circ, \ldots, \hat{\alpha}_{n,p^\circ}^\circ, \hat{\alpha}_n^*, \hat{\beta}_n)\hat{\alpha}_{n,k'}^\circ + s_n^2 \mathfrak{p}_{n,k'}^{a\circ}(\hat{\alpha}_{n,k'}^\circ) \geq -\underline{a}_{k'} \lambda_{\min}\big(\Gamma_0^{\alpha^\circ}(\alpha_0)\big)|\hat{\alpha}_{n,k'}^\circ|^2\Bigg) \to 1,$$

$$P\Bigg(t_n^2 \partial_{\beta_{l'}^\circ} M_n(\hat{\alpha}_n, 0, \ldots, 0, \hat{\beta}_{n,l'}^\circ, \ldots, \hat{\beta}_{n,q^\circ}^\circ, \hat{\beta}_n^*)\hat{\beta}_{n,l'}^\circ + t_n^2 \mathfrak{p}_{n,l'}^{b\circ}(\hat{\beta}_{n,l'}^\circ) \geq -\underline{b}_{l'} \lambda_{\min}\big(\Gamma_0^{\beta^\circ}(\theta_0)\big)|\hat{\beta}_{n,l'}^\circ|^2\Bigg) \to 1$$

for each $k' \in \{1, \ldots, p^\circ\}$ *and* $l' \in \{1, \ldots, q^\circ\}$.

Then, it is straightforward to prove the following corollary by making use of [12, Theorem 3.12].

COROLLARY 3.5. *We have* $P(\hat{\theta}_n^\circ = 0) \to 1$ *under Assumptions 3.1, 3.2 and 3.4.*

The Asymptotic non-degenerate distribution of $(\hat{u}_n^*, \hat{v}_n^*) = (s_n^{-1}(\hat{\alpha}_n^* - \alpha_0^*), t_n^{-1}(\hat{\beta}_n^* - \beta_0^*))$ can be also derived with modifications. We define some notations:

$$\Delta_n(\theta_0) := D_n \partial_{\theta^*} M_n(\theta_0), \qquad \Gamma_n(\theta_0) := D_n \partial^2_{\theta^*} M_n(\theta_0) D_n,$$
$$\Delta \overline{R}_n^{a*}(u^*) := \overline{R}_n^{a*}(\alpha_0^* + s_n u^*) - \overline{R}_n^{a*}(\alpha_0^*), \qquad \Delta \overline{R}_n^{b*}(v^*) := \overline{R}_n^{b*}(\beta_0^* + t_n v^*) - \overline{R}_n^{b*}(\beta_0^*),$$

where $D_n := \mathrm{diag}(s_n I_{p^*}, t_n I_{q^*})$ and, I_{p^*} and I_{q^*} are $p^* \times p^*$ and $q^* \times q^*$ identity matrix, respectively. Then, we cite an assumption from [12]:

ASSUMPTION 3.6 (ASSUMPTION 3.14 IN [12]). *There exist random variables* Δ_0 *and* Γ_0*, and random functions* $\Delta \overline{R}_0^{a*}(u^*)$ *and* $\Delta \overline{R}_0^{b*}(v^*)$ *such that*

$$\Big(\Delta_n(\theta_0),\, \Gamma_n(\theta_0),\, \Delta \overline{R}_n^{a*}(\cdot),\, \Delta \overline{R}_n^{b*}(\cdot)\Big) \xrightarrow{\mathcal{L}} \Big(\Delta_0,\, \Gamma_0,\, \Delta \overline{R}_0^{a*}(\cdot),\, \Delta \overline{R}_0^{b*}(\cdot)\Big)$$

in $\mathcal{C}(K_0 \times K_1)$ *for every compact* $K_0 \times K_1 \subset \mathbb{R}^{p^*} \times \mathbb{R}^{q^*}$*, and that the random function*

$$\mathbb{M}_0(u^*, v^*) := \Delta_0[w^*] + \frac{1}{2}\Gamma_0[w^*, w^*] + \Delta \overline{R}_0^{a*}(u^*) + \Delta \overline{R}_0^{b*}(v^*)$$

has an a.s. unique minimum at $(u^*, v^*) = (\hat{u}_0^*, \hat{v}_0^*)$.

As a variant of [12, Theorem 3.15], we get the following corollary.

COROLLARY 3.7. *We have* $(\hat{u}_n^*, \hat{v}_n^*) \xrightarrow{\mathcal{L}} (\hat{u}_0^*, \hat{v}_0^*)$ *under Assumptions 3.1, 3.2, 3.4 and 3.6.*

Remark 3.8. *We can derive the asymptotically mixed normality of* $\hat{\theta}_n^*$ *with slight modifications (see* [12, Corollary 3.17] *for details).*

Finally as for [12, Theorem 3.20]:

$$\sup_{r>0} \sup_{n>0} r^L P(|(\hat{u}_n, \hat{v}_n)| \geq r) < \infty, \tag{10}$$

which gives us the tail probability estimates of $(\hat{u}_n, \hat{v}_n)$, we set the following as-

sumptions:

ASSUMPTION 3.9 (ASSUMPTION 3.18 IN [12]).

1. *There exist nonrandom functions* $\widetilde{M}_0^a : \Theta_\alpha \to \mathbb{R}$ *and* $\widetilde{M}_0^b : \Theta_\beta \to \mathbb{R}$, *and positive constants* δ_1^a, δ_1^b, χ^a *and* χ^b *such that for all* $K > 0$,
   - $\sup_{n>0} E\left[\sup_{\theta\in\Theta}\left|s_n^{-2\delta_1^a}\left\{s_n^2\big(M_n(\alpha,\beta)-M_n(\alpha_0,\beta)\big)-\widetilde{M}_0^a(\alpha)\right\}\right|^K\right] < \infty.$
   - $\sup_{n>0} E\left[\sup_{\beta\in\Theta_\beta}\left|t_n^{-2\delta_1^b}\left\{t_n^2\big(M_n(\alpha_0,\beta)-M_n(\alpha_0,\beta_0)\big)-\widetilde{M}_0^b(\beta)\right\}\right|^K\right] < \infty.$
   - $\widetilde{M}_0^a(\alpha) \geq \chi^a|\alpha-\alpha_0|^2, \quad \widetilde{M}_0^b(\beta) \geq \chi^b|\beta-\beta_0|^2.$
2. *There exist nonrandom matrices* $C_0(\beta)$ *and* $C_0 > 0$, *and constants* δ_2^a, $\delta_2^b \in (0, 1/2]$ *such that for all* $K > 0$,
   - $\inf_\beta \lambda_{\min}(C_0(\beta)) > 0.$
   - $\sup_{n>0} E\left[\sup_{\beta\in\Theta_\beta}\left(s_n^{-2\delta_2^a}\left|s_n^2\partial_\alpha^2 M_n(\alpha_0,\beta)-C_0(\beta)\right|\right)^K\right] < \infty.$
   - $\sup_{n>0} E\left[\left(t_n^{-2\delta_2^b}\left|t_n^2\partial_\beta^2 M_n(\alpha_0,\beta_0)-C_0\right|\right)^K\right] < \infty.$
   - $\sup_{n>0} E\left[\sup_{\alpha\in\Theta_\alpha}\left|s_n t_n^{2(1-\delta_2^b)}\partial_\alpha\partial_\beta^2 M_n(\alpha,\beta_0)\right|^K\right] < \infty.$
3. *For all* $K > 0$,
   - $\sup_{n>0} E\left[\sup_{\beta\in\Theta_\beta}|s_n\partial_\alpha M_n(\alpha_0,\beta)|^K\right] < \infty,\ \sup_{n>0} E\left[|t_n\partial_\beta M_n(\alpha_0,\beta_0)|^K\right] < \infty.$
   - $\sup_{n>0} E\left[\sup_{\theta\in\Theta}\left|s_n^2\partial_\alpha^3 M_n(\theta)\right|^K\right] < \infty,\ \sup_{n>0} E\left[\sup_{\theta\in\Theta}\left|t_n^2\partial_\beta^3 M_n(\theta)\right|^K\right] < \infty.$
   - $\sup_{n>0} E\left[\sup_{\alpha\in\Theta_\alpha}|s_n t_n\partial_\alpha\partial_\beta M_n(\alpha,\beta_0)|^K\right] < \infty.$

ASSUMPTION 3.10 (ASSUMPTION 3.19.1 IN [12]). *There exist constants* ν^a, $\nu^b \in (0, 1/2)$ *such that for any* $K > 0$,

$$\sup_{n>0} E\left[\sup_{\alpha\in\Theta_\alpha}\left(s_n^{1+2\nu^a}\overline{R}_n^a(\alpha)\right)^K\right] < \infty, \quad \sup_{n>0} E\left[\sup_{\beta\in\Theta_\beta}\left(t_n^{1+2\nu^b}\overline{R}_n^b(\beta)\right)^K\right] < \infty.$$

ASSUMPTION 3.11 (MODIFIED ASSUMPTION 3.19.2. IN [12]). *There exist* $\kappa^a,\ \kappa^b \in (0,2)$, *and for any* $u \neq 0$ *there exist random variables* $z_u^a,\ z_u^b > 0$ *a.s. such that for any* $v \in \mathbb{R}$*:*

- $\limsup_{n\to\infty} \left|\mathfrak{p}_{n,k''}^{a*}(u+s_n v) - \mathfrak{p}_{n,k''}^{a*}(u)\right| \leq z_u^a |v|^{\kappa^a}$ *a.s. for all* $k'' \in \{1,\dots,p^*\}$
- $\limsup_{n\to\infty} \left|\mathfrak{p}_{n,l''}^{b*}(u+t_n v) - \mathfrak{p}_{n,l''}^{b*}(u)\right| \leq z_u^b |v|^{\kappa^b}$ *a.s. for all* $l'' \in \{1,\dots,q^*\}$
- $E[|z_u^a|^K] < \infty, \quad E[|z_u^b|^K] < \infty$ *for every* $K > 0$.

Assumption 3.9 is borrowed from [22, Theorem 3(c)], hence we should note that the assumed differentiability of M_n is not essential and could be relaxed; see also Remark 3.14 below. Then, we obtain the following claim.

COROLLARY 3.12. *For any* $L > 0$, (10) *holds under Assumptions 3.9–3.11. Additionally if we have the weak convergence* $(\hat{u}_n^\circ, \hat{u}_n^*, \hat{v}_n^\circ, \hat{v}_n^*) \xrightarrow{\mathcal{L}} (\hat{u}_0^\circ, \hat{u}_0^*, \hat{v}_0^\circ, \hat{v}_0^*)$ *for some random vector* $(\hat{u}_0^\circ, \hat{u}_0^*, \hat{v}_0^\circ, \hat{v}_0^*)$*, then the moment convergence*

$$E[f(\hat{u}_n, \hat{v}_n)] \to E[f(\hat{u}_0^\circ, \hat{u}_0^*, \hat{v}_0^\circ, \hat{v}_0^*)]$$

holds for all continuous $f : \mathbb{R}^{p+q} \to \mathbb{R}$ *of at most polynomial growth.*

Remark 3.13. *The proof of Corollary 3.12 remains unchanged from that of* [12, Theorem 3.20] *except for*

$$\begin{aligned}
&\sup_{n>0} E\left[\sup_{u\in U_n^a(r)} \left(\frac{1}{1+|u|^2}\left|\overline{R}_n^{a*}(\alpha_0^* + s_n u^*) - \overline{R}_n^{a*}(\alpha_0^*)\right|\right)^d\right] \\
&\quad\lesssim \sup_{n>0} E\left[\sup_{u\in U_n^a(r)} \left(\frac{1}{1+|u|^2}\sum_{k''=1}^{p^*}\left|\mathfrak{p}_{n,k''}^{a*}(\alpha_0^* + s_n u^*) - \mathfrak{p}_{n,k''}^{a*}(\alpha_0^*)\right|\right)^d\right] \\
&\quad\lesssim E\left[\sup_{u\in U_n^a(r)} \left(\frac{|u^*|^{\kappa^a}}{1+|u|^2}\right)^d\right] \lesssim r^{-(\kappa^a-2)d}.
\end{aligned}$$

Here, we used Assumption 3.11. See [12] *for details.*

Remark 3.14. *Trivially, it is not essential for the discussions so far that the LAQ part* M_n *is twice continuously differentiable. All the assertions presented in this section can also go for possibly non-differential* M_n *as long as statistical random fields associated with* M_n *is of LAQ:* $M_n(\theta_0 + A_n u) - M_n(\theta_0) = \Delta_n[u] + \frac{1}{2}\Gamma_0[u,u] +$

$r_n(u)$ with (quasi-)score sequence Δ_n, asymptotic (quasi-)information matrix Γ_0, and remainder term $r_n(u) = o_p(1)$ (locally uniformly in u). The resulting set of conditions becomes somewhat less concise.

Acknowledgements. The author is grateful to a referee for careful reading and several valuable comments. He also thanks to Professor H. Masuda for his helpful comments.

References

[1] G. Afendras and M. Markatou. Optimality of training/test size and resampling effectiveness of cross-validation estimators of the generalization error. *arXiv:1511.02980*, 2015.
[2] G. Afendras and M. Markatou. Uniform integrability of the OLS estimators, and the convergence of their moments. *arXiv:1511.02962*, 2015.
[3] B. Antoine and E. Renault. Efficient minimum distance estimation with multiple rates of convergence. *J. Econometrics*, 170(2):350–367, 2012.
[4] P. J. Bickel and B. Li. Regularization in statistics. *Test*, 15(2):271–344, 2006. With comments and a rejoinder by the authors.
[5] N. H. Chan and C.-K. Ing. Uniform moment bounds of Fisher's information with applications to time series. *Ann. Statist.*, 39(3):1526–1550, 2011.
[6] L. Dicker, B. Huang, and X. Lin. Variable selection and estimation with the seamless-L_0 penalty. *Statist. Sinica*, 23(2):929–962, 2013.
[7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.
[8] D. F. Findley and C.-Z. Wei. AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. *J. Multivariate Anal.*, 83(2):415–450, 2002.
[9] L. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.
[10] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.
[11] C.-K. Ing and C.-Y. Yang. Predictor selection for positive autoregressive processes. *J. Amer. Statist. Assoc.*, 109(505):243–253, 2014.
[12] H. Masuda and Y. Shimizu. Moment convergence in regularized estimation under multiple and mixed-rates asymptotics. *Math. Methods Statist.*, 26(2):81–110, 2017.
[13] B. M. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, 100(9):2065–2082, 2009.
[14] P. Radchenko. Reweighting the lasso. *2005 Proceedings of the American Statistical Association [CD-ROM]*, 2005.
[15] P. Radchenko. Mixed-rates asymptotics. *Ann. Statist.*, 36(1):287–309, 2008.
[16] Y. Sakamoto and N. Yoshida. Asymptotic expansion formulas for functionals of ϵ-Markov processes with a mixing property. *Ann. Inst. Statist. Math.*, 56(3):545–597, 2004.
[17] Y. Shimizu. Moment convergence of regularized least-squares estimator for linear regression model. *Ann. Inst. Statist. Math.*, 69(5):1141–1154, 2017.
[18] M. Uchida and N. Yoshida. Information criteria in model selection for mixing processes.

Stat. Inference Stoch. Process., 4(1):73–98, 2001.

[19] M. Uchida and N. Yoshida. Asymptotic expansion and information criteria. *SUT J. Math.*, 42(1):31–58, 2006.

[20] Y. Umezu, Y. Shimizu, H. Masuda, and Y. Ninomiya. AIC for non-concave penalized likelihood method. *arXiv:1509.01688v2*, 2015.

[21] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[22] N. Yoshida. Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Ann. Inst. Statist. Math.*, 63(3):431–479, 2011.

[23] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

Yusuke SHIMIZU

Department of Mathematics, Josai University
Keyakidai 1-1, Sakado-shi, Saitama 350-0295, Japan
E-mail: yshimizu@josai.ac.jp