

# WEAK CONVERGENCE OF THE REGULARIZATION PATH IN PENALIZED M-ESTIMATION

JEAN-FRANÇOIS GERMAIN AND FRANÇOIS ROUEFF

*RENAULT DREAM-DTAA and Institut TELECOM, TELECOM ParisTech, LTCI CNRS*

ABSTRACT. We consider an estimator  $\hat{\beta}_n(\mathbf{t})$  defined as the element  $\phi \in \Phi$  minimizing a contrast process  $\Lambda_n(\phi, \mathbf{t})$  for each  $\mathbf{t}$ . We give some general results for deriving the weak convergence of  $\sqrt{n}(\hat{\beta}_n - \beta)$  in the space of bounded functions, where, for each  $\mathbf{t}$ ,  $\beta(\mathbf{t})$  is the  $\phi \in \Phi$  minimizing the limit of  $\Lambda_n(\phi, \mathbf{t})$  as  $n \rightarrow \infty$ . These results are applied in the context of penalized M-estimation, that is, when  $\Lambda_n(\phi, \mathbf{t}) = M_n(\phi) + \mathbf{t}J_n(\phi)$ , where  $M_n$  is a usual contrast process and  $J_n$  a penalty such as the  $\ell^1$  norm or the squared  $\ell^2$  norm. The function  $\hat{\beta}_n$  is then called a *regularization path*. For instance we show that the central limit theorem established for the lasso estimator in Knight and Fu [2000] continues to hold in a functional sense for the regularization path. Other examples include various possible contrast processes for  $M_n$  such as those considered in Pollard [1985].

## 1. INTRODUCTION

Let us consider a real-valued contrast process  $\{M_n(\phi), \phi \in \Phi\}$  based on an observed sample of size  $n$  and a contrast function  $M$  defined on the same parameter set  $\Phi$  and minimized at the point  $\beta$ . A penalized estimator with penalty weight  $\mathbf{t} \geq 0$  is defined as the minimizer of the contrast process

$$\Lambda_n(\phi, \mathbf{t}) = M_n(\phi) + \mathbf{t} J_n(\phi), \quad \phi \in \Phi, \quad (1)$$

where  $J_n$  is a non-negative function defined on  $\Phi$ , not depending on the observations but possibly on  $n$ , mainly to allow some appropriate normalization.

The use of penalties is popular for ill-posed problems and model selection, among which the ridge regression (see Hoerl and Kennard [1970]) and the lasso (see Tibshirani [1996]) are emblematic examples. In these two examples the contrast process  $M_n$  is the least-square criterion and the penalty function  $J_n$  is the squared  $\ell^2$  norm and the  $\ell^1$  norm, respectively. Consistency and central limit theorems are established in Knight and Fu [2000] precisely in the case where  $M_n$  is the least-square criterion and  $J_n$  is in a family of penalties including both the squared  $\ell^2$  norm and the  $\ell^1$  norm. They show that, when the penalty is properly normalized, the penalized mean square estimator is no longer asymptotically normal. Instead, its asymptotic distribution is given by the minimizer of a penalized quadratic form depending on a Gaussian vector (see *e.g.* [Knight and Fu, 2000, Theorem 2]). Their

*Date:* June 10, 2009 (first revision).

*1991 Mathematics Subject Classification.* Primary 62J07, 62F12, 60F17 Secondary: 62J05, 60F05, 62E20.

*Key words and phrases.* lasso, Akaike information criterion (AIC), penalized M-estimation, regularization path, weak convergence, pathwise argmin theorem.

*Corresponding author:* F. Roueff, Institut TELECOM, TELECOM ParisTech, LTCI CNRS.

asymptotic results hold as the number  $n$  of observations tends to infinity and for a fixed finite-dimensional model. Quite different results have been established when the dimension of the model increases with  $n$ , see Greenshtein and Ritov [2004], Zhao and Yu [2006], Bunea et al. [2007], Bickel et al. [2008] and the references therein. These results provide interesting properties of the lasso for model selection or prediction purposes in the context of sparse models. Although specific normalizations of the penalty (different from those required in Knight and Fu [2000]) are prescribed in these theoretical results, there exist numerous heuristic ways for choosing the penalty weight  $\mathbf{t}$  in practice. The first step is to minimize  $\Lambda_n(\phi, \mathbf{t})$  in (1) on  $\phi \in \Phi$  for a collection of non-negative weights  $\mathbf{t}$ , resulting in a collection of estimators  $\hat{\beta}_n(\mathbf{t})$ , which is called the *regularization path* (or the *solution path*). The Least Angle Regression (LAR) technique introduced by Efron et al. in Efron et al. [2004] provides, in most cases, the entire path, computed with the complexity of a linear regression. In a second step, some criterion is used to select  $\mathbf{t}$ , see e.g. Zou et al. [2007] where AIC and BIC procedures are proposed for the lasso. Because the whole path is used by the practitioner, we think that it is crucial to examine whether the convergence of  $\sqrt{n}(\hat{\beta}_n(\mathbf{t}) - \beta)$ , established in Knight and Fu [2000] for one fixed  $\mathbf{t}$ , continues to hold in a functional sense and, if it is the case, to determine the limit distribution. The goal of this paper is twofold. First we show that, under the same assumptions as in Knight and Fu [2000], the convergence holds in the space of locally bounded functions. Second we extend this result to more general contrast processes  $M_n$  such as generalized linear models (GLM) or least amplitude deviation (LAD). A key result is a *pathwise argmin theorem* which establishes the functional weak convergence of a path defined as the minimizer a collection of contrast processes, see Theorem 3.

For the moment let us give the asymptotic behavior of the lasso regularization path, which is the most simple application of our results and which naturally extends Knight and Fu [2000]. Consider the linear model

$$y_k = \mathbf{x}_k^T \beta + \varepsilon_k, \quad k = 1, 2, \dots \quad (2)$$

where  $\beta \in \mathbb{R}^p$  is an unknown parameter,  $(y_k)$  is a sequence of real-valued observations,  $(\mathbf{x}_k)$  is the sequence of regression vectors and  $(\varepsilon_k)$  is a strong white noise with variance  $\sigma^2$ . For any  $\mathbf{t} \geq 0$ , the lasso estimator  $\hat{\beta}_n(\mathbf{t})$  minimizes the penalized contrast process  $\Lambda_n(\phi, \mathbf{t})$  on  $\phi \in \mathbb{R}^p$ , where

$$\Lambda_n(\phi, \mathbf{t}) = \frac{1}{n} \sum_{k=1}^n (y_k - \mathbf{x}_k^T \phi)^2 + \mathbf{t} \lambda_n \sum_{i=1}^p |\phi_i|, \quad (3)$$

which is a specific form of (1). Denote  $\mathbf{X}_n = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ . We consider the following assumptions, for consistency and central limit theorem, respectively. The assumptions are the same as in Knight and Fu [2000].

**Assumption 1.**

- (i)  $C_n = n^{-1} \mathbf{X}_n^T \mathbf{X}_n \rightarrow C$ , where  $C$  is a positive-definite matrix;
- (ii)  $\lambda_n \rightarrow 0$ .

**Assumption 2.**

- (i) Assumption 1-(i) holds;
- (ii)  $\max_{1 \leq k \leq n} \|\mathbf{x}_k\|^2 = o(n)$ ;

(iii)  $\lambda_n = n^{-1/2}$ .

Assumptions 1-(i) and 2-(ii) are the classical assumptions for the asymptotic behavior of least squares estimators. The other assumptions provide the appropriate way of normalizing the  $\ell^1$  penalty.

**Theorem 1.** *Under Assumption 1,  $\widehat{\beta}_n(\mathbf{t})$  converges in probability to  $\beta$  locally uniformly in  $\mathbf{t} \in \mathbb{R}_+$ , that is*

$$\widehat{\beta}_n \xrightarrow{P} \beta \text{ in } \ell_o^\infty(\mathbb{R}_+, \mathbb{R}^p), \quad (4)$$

where  $\ell_o^\infty(\mathbb{R}_+, \mathbb{R}^p)$  denotes the space of locally bounded  $\mathbb{R}_+ \rightarrow \mathbb{R}^p$  functions.

We now define the limit process of the lasso regularization path, appropriately centered and normalized. Let  $U \sim \mathcal{N}(0, \sigma^2 C)$ . For any  $\mathbf{t} \geq 0$ , we define  $\widehat{\mathbf{u}}(\mathbf{t})$  as the point  $\phi \in \mathbb{R}^p$  which minimizes

$$\mathbb{L}(\phi, \mathbf{t}) = -2U^T \phi + \phi^T C \phi + \mathbf{t} \left[ \sum_{j=1}^p \phi_j \operatorname{sgn}(\beta_j) \mathbb{1}_{\{\beta_j \neq 0\}} + |\phi_j| \mathbb{1}_{\{\beta_j = 0\}} \right]. \quad (5)$$

It is easy to show that this defines  $\widehat{\mathbf{u}}(\mathbf{t})$  uniquely for all  $\mathbf{t} \geq 0$  (see the proof of Theorem 2). The distribution of  $\widehat{\mathbf{u}}$  as a function is not explicit but is not more complicated than its marginal distributions already described in Knight and Fu [2000], since the whole path is described as a deterministic function of the random variable (r.v.)  $U$ . An interesting property of  $\widehat{\mathbf{u}}(\mathbf{t})$  is that, with probability 1, the set of its components that vanish for  $\mathbf{t}$  large enough is given by the set of zero components of the true parameter  $\beta$ .

**Theorem 2.** *Under Assumption 2,*

$$\sqrt{n}(\widehat{\beta}_n - \beta) \rightsquigarrow \widehat{\mathbf{u}} \text{ in } \ell_o^\infty(\mathbb{R}_+, \mathbb{R}^p), \quad (6)$$

where  $\rightsquigarrow$  denotes the weak convergence.

*Remark 1.* The convergence in  $\ell_o^\infty(\mathbb{R}_+, \mathbb{R}^p)$  is equivalent to the uniform convergence on every compact subset of  $\mathbb{R}_+$ . In fact the convergences (4) and (6) cannot be improved in the sense that they do not hold uniformly on  $\mathbb{R}_+$ . To see why, observe that, by the definition of  $\widehat{\mathbf{u}}$ , its coordinates corresponding to non-vanishing  $\beta_j$  are unbounded as  $\mathbf{t} \rightarrow \infty$ . In contrast, the left-hand side of (6) is bounded since, for any  $n$ , there is a large enough  $\mathbf{t}$  for which  $\widehat{\beta}_n(\mathbf{t}) = 0$ . Note that this also implies that  $\sup_{\mathbf{t} \in \mathbb{R}_+} \|\widehat{\beta}_n(\mathbf{t}) - \beta\| \geq \|\beta\|$ , and thus that the consistency (4) does not hold if  $\ell_o^\infty(\mathbb{R}_+, \mathbb{R}^p)$  is replaced by the set of bounded  $\mathbb{R}_+ \rightarrow \mathbb{R}^p$  functions  $\ell^\infty(\mathbb{R}_+, \mathbb{R}^p)$  endowed with the sup norm.

The proofs of Theorem 1 and Theorem 2 are applications of some general results on the consistency of convex penalized M-estimators and on the weak convergence of Argmin's depending on an tuning parameter  $\mathbf{t}$  (the so called pathwise argmin theorem in the following). More general penalized contrasts will also be considered. Such extensions are of interest since the lasso regularization path has been extended to the case where  $M_n$  is different from the least-square criterion. In Park and Hastie [2007], a fast numerical algorithm is proposed for determining the regularization path when  $M_n$  is a regression function based on a negated log-likelihood of the canonical exponential family. In Germain [2007], a fast algorithm based on a dichotomy is proposed to explore the range of  $\mathbf{t}$ 's in the specific case of logistic regression penalized by the  $\ell^1$  norm.

The paper is organized as follows. In Section 2, we provide a pathwise argmin theorem (Theorem 3). Section 3 is concerned with the asymptotic behavior of the regularization path of a penalized contrast. Very mild conditions on the contrast and on the penalty are provided for obtaining the uniform consistency and the central limit of the path and a particular attention is given to the case where both the contrast and the penalty are convex. Except for the convex case, such results can actually be seen as special cases of the more general study of pathwise M-estimators, which is treated in Section 4. Finally we provide several examples of applications of these results in Section 5, including the  $\ell^1$ -penalized general linear model (GLM) introduced in Park and Hastie [2007], the penalized least absolute deviation (LAD) and the Akaike information criterion. Concluding remarks are provided in Section 6. The detailed proofs are deferred to the appendix for convenience.

## 2. A PATHWISE ARGMIN THEOREM

To obtain a CLT for the regularization path, we rely on a pathwise argmin theorem, which is of independent interest, and can be seen as an extension of [Kim and Pollard, 1990, Theorem 2.7] (see also [Van der Vaart and Wellner, 1996, Theorem 3.2.2]) to fit the context of a path defined as the minimizer of a collection of contrast processes.

Let us recall some of the terminology and notation used in Van der Vaart and Wellner [1996]. For a metric space  $\mathcal{D}$ , we say that a sequence of  $\mathcal{D}$ -valued maps  $(X_n)$  defined on  $\Omega$  converges weakly to a  $\mathcal{D}$ -valued map  $X$  defined on  $(\Omega, \mathcal{F})$ , and denote  $X_n \rightsquigarrow X$ , if  $X$  is a Borel map and, for any real-valued bounded continuous function  $f$  defined on  $\mathcal{D}$ ,

$$E^*[f(X_n)] \rightarrow E[f(X)] ,$$

where  $E$  denotes the expectation with respect to  $P$  and  $E^*$  denotes the *outer expectation*, defined for every real-valued map  $Z$  defined on  $\Omega$  by  $E^*[Z] = \inf\{E[U] : U \geq Z\}$ , where, in this sup, the r.v.  $U$  is taken measurable. The *inner expectation* and *inner probability* are respectively defined by  $E_*[Z] = -E^*[-Z]$  and  $P_*(A) = 1 - P^*(A^c)$ , where  $A^c$  denotes the complementary set of  $A$  in  $\Omega$ .

For any positive integer  $p$  and any set  $T$  we further denote by  $\ell^\infty(T, \mathbb{R}^p)$  the normed space of bounded functions  $f = (f_1, \dots, f_p)$  taking values in  $\mathbb{R}^p$  and defined on  $T$  endowed with the sup norm on  $T$ , denoted by

$$\|f\|_T = \sup_{t \in T, i \in \{1, \dots, p\}} |f_i(t)| .$$

We will simply denote  $\ell^\infty(T, \mathbb{R}^p)$  by  $\ell^\infty(T)$  for  $p = 1$ .

**Theorem 3.** *Let  $\Phi$  be a metric space endowed with a metric  $d$  and  $T$  be an arbitrary set. We suppose that we are in one of the two following cases*

- (C-1)  $T$  is a finite set. In this case, we set  $\mathcal{D} = \Phi^T$  endowed with the product topology;
- (C-2)  $\Phi = \mathbb{R}^p$  with  $p \geq 1$ ,  $d$  being the Euclidean metric. In this case, we set  $\mathcal{D} = \ell^\infty(T, \mathbb{R}^p)$ .

Let  $\{\mathbb{L}_n(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in T\}$  be a sequence of real-valued processes,  $\{\mathbb{L}(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in T\}$  be a real-valued process,  $\{\hat{\mathbf{u}}(\mathbf{t}), \mathbf{t} \in T\}$  be a  $\Phi$ -valued process, and  $\{\hat{\mathbf{u}}_n(\mathbf{t}), \mathbf{t} \in T\}$  be a sequence of  $\Phi$ -valued processes. Assume that

- (i) for any compact set  $K \subset \Phi$ ,  $\mathbb{L}_n \rightsquigarrow \mathbb{L}$  in  $\ell^\infty(K \times T)$  and  $\mathbb{L}$  is a tight Borel map taking values in  $\ell^\infty(K \times T)$ ;

(ii) for any  $\eta > 0$  and compact  $K \subset \Phi$ , we have almost surely that

$$\inf_{\mathbf{t} \in \mathbb{T}} [\inf\{\mathbb{L}(\phi, \mathbf{t}) : \phi \in K, d(\phi, \hat{\mathbf{u}}(\mathbf{t})) \geq \eta\} - \mathbb{L}(\hat{\mathbf{u}}(\mathbf{t}), \mathbf{t})] > 0; \quad (7)$$

(iii) for any  $\epsilon > 0$ , there exists a compact  $K \subset \Phi$  such that

$$P(\hat{\mathbf{u}}(\mathbf{t}) \in K \text{ for all } \mathbf{t} \in \mathbb{T}) \geq 1 - \epsilon; \quad (8)$$

(iv) for any  $\epsilon > 0$ , there exists a compact  $K \subset \Phi$  such that

$$\liminf P_*(\hat{\mathbf{u}}_n(\mathbf{t}) \in K \text{ for all } \mathbf{t} \in \mathbb{T}) \geq 1 - \epsilon; \quad (9)$$

(v)  $\hat{\mathbf{u}}_n$  is approximately minimizing  $\mathbb{L}_n$ ,

$$\sup_{\mathbf{t} \in \mathbb{T}} \left\{ \mathbb{L}_n(\hat{\mathbf{u}}_n(\mathbf{t}), \mathbf{t}) - \inf_{\phi \in \Phi} \mathbb{L}_n(\phi, \mathbf{t}) \right\}_+ = o_{P^*}(1). \quad (10)$$

Then there is a version of  $\hat{\mathbf{u}}$  in  $\mathcal{D}$  and  $\hat{\mathbf{u}}_n \rightsquigarrow \hat{\mathbf{u}}$ .

*Proof of Theorem 3 in the case (C-1).* In the case (C-1), where  $\mathbb{T}$  is finite, for any compact  $K \subset \Phi$ ,  $K^{\mathbb{T}}$  is a compact subset of  $\Phi^{\mathbb{T}}$  endowed with the metric

$$d_{\mathbb{T}}(\mathbf{u}, \mathbf{v}) = \sup_{\mathbf{t} \in \mathbb{T}} d(\mathbf{u}(\mathbf{t}), \mathbf{v}(\mathbf{t})).$$

Hence, in this case, Conditions (iii) and (iv) respectively say that  $\hat{\mathbf{u}}$  is tight and  $(\hat{\mathbf{u}}_n)$  is uniformly tight in  $\Phi^{\mathbb{T}}$ . In this case, the conclusion of Theorem 3 follows almost directly from Theorem 3.2.2 in Van der Vaart and Wellner [1996]. To see why, let us introduce the following contrast process

$$\mathcal{L}_n(\mathbf{v}) = \sup_{\mathbf{t} \in \mathbb{T}} \left\{ \mathbb{L}_n(\mathbf{v}(\mathbf{t}), \mathbf{t}) - \inf_{\phi \in \Phi} \mathbb{L}_n(\phi, \mathbf{t}) \right\}, \quad \mathbf{v} \in \Phi^{\mathbb{T}}. \quad (11)$$

Observe that defining  $\hat{\mathbf{u}}_n(\mathbf{t})$  as a minimizer of  $\mathbb{L}_n(\cdot, \mathbf{t})$  for all  $\mathbf{t} \in \mathbb{T}$  is equivalent to defining  $\hat{\mathbf{u}}_n$  directly as a minimizer of  $\mathcal{L}_n$ . In particular Condition (v) implies that

$$\mathcal{L}_n(\hat{\mathbf{u}}_n) \leq \inf_{\mathbf{v} \in \Phi^{\mathbb{T}}} \mathcal{L}_n(\mathbf{v}) + o_{P^*}(1),$$

that is,  $\hat{\mathbf{u}}_n$  is a near minimizer of  $\mathcal{L}_n$ . Condition (i), in turn, by the continuous mapping theorem, implies that  $\mathcal{L}_n \rightsquigarrow \mathcal{L}$  in  $\ell^\infty(\Phi^{\mathbb{T}})$ , where, for any  $\mathbf{v} \in \Phi^{\mathbb{T}}$ ,

$$\mathcal{L}(\mathbf{v}) = \sup_{\mathbf{t} \in \mathbb{T}} \left\{ \mathbb{L}(\mathbf{v}(\mathbf{t}), \mathbf{t}) - \inf_{\phi \in \Phi} \mathbb{L}(\phi, \mathbf{t}) \right\}.$$

Finally it is not too difficult to show that Condition (ii) implies that, almost surely, for all compact  $K \subset \Phi$  and  $\eta > 0$ ,

$$\inf \{ \mathcal{L}(\mathbf{v}) : \mathbf{v} \in K^{\mathbb{T}}, d_{\mathbb{T}}(\mathbf{v}, \hat{\mathbf{u}}) \geq \eta \} > 0 = \mathcal{L}(\hat{\mathbf{u}}).$$

This condition corresponds to the semicontinuity and argmax uniqueness conditions appearing in Theorem 3.2.2 in Van der Vaart and Wellner [1996]. Hence this theorem applies and yields  $\hat{\mathbf{u}}_n \rightsquigarrow \hat{\mathbf{u}}$  in the case (C-1).  $\square$

The proof in the case (C-2) is postponed to the appendix. The main originality of Theorem 3 lies in the case (C-2). In this case, Theorem 3.2.2 in Van der Vaart and Wellner [1996] cannot be directly applied because Condition (iv) is no longer a uniform tightness condition ( $K^{\mathbb{T}}$  is not a compact subset of  $\ell^\infty(\mathbb{T}, \mathbb{R}^p)$ ). The key idea, detailed in the appendix,

is to show that, under Conditions (ii)–(v), this asymptotic tightness of  $\widehat{\mathbf{u}}_n$  in  $\ell^\infty(\mathbb{T}, \mathbb{R}^p)$  is inherited from that of  $\mathbb{L}_n$  assumed in Condition (i).

### 3. PENALIZED M-ESTIMATION

**3.1. Uniform consistency.** Standard results on the consistency of M-estimators (see *e.g.* [Van der Vaart, 1998, Theorem 5.7]) roughly say that if  $\widehat{\beta}_n$  is a sequence of minimizers of  $M_n$  on  $\Phi$ ,  $M_n$  tends to  $M$  with some uniformity and  $\beta$  is an isolated minimum of  $M$  on  $\Phi$ , then  $\widehat{\beta}_n$  converges to  $\beta$  in probability. We will use the following set of conditions which are slightly weaker than the classical ones.

**Assumption 3.** There exists  $\beta \in \Phi$  such that

- (i)  $\sup_{\phi \in \Phi} \{M(\phi) - M_n(\phi)\}_+ \xrightarrow{P} 0$ , where  $a_+ = \max(0, a)$  for any  $a \in \mathbb{R}$ ;
- (ii)  $M_n(\beta) \xrightarrow{P} M(\beta)$ ;
- (iii) for all  $\epsilon > 0$ ,  $\inf\{M(\phi) : \phi \in \Phi, d(\phi, \beta) \geq \epsilon\} > M(\beta)$ ,

where  $d$  is a metric endowing the metric space  $\Phi$ .

Let us briefly comment these assumptions. Conditions (i) and (ii) are generally replaced by the stronger uniform convergence condition  $\sup_{\phi \in \Phi} |M(\phi) - M_n(\phi)| \xrightarrow{P} 0$ . These weaker conditions are for instance useful when  $\Phi$  is non-compact since it is then sufficient to show the uniform convergence on a compact subset and provide a lower bound of  $M_n$  out of this compact. Condition (iii) is the standard condition which defines  $\beta$  as the (unique) isolated minimum of the limit contrast function.

We will show that, under Assumption 3, provided that  $J_n(\beta)$  tends to 0, the minimizer  $\widehat{\beta}_n(\mathbf{t})$  of  $\Lambda_n(\phi, \mathbf{t})$  converges to  $\beta(\mathbf{t})$ , *locally uniformly* in  $\mathbf{t}$ . To avoid making measurability assumptions on the path  $\mathbf{t} \mapsto \widehat{\beta}_n(\mathbf{t})$ , we need to work with outer probability to extend the probability to possibly non-measurable sets. Given a probability space  $(\Omega, \mathcal{F}, P)$ , we denote by  $P^*$  the *outer probability* defined on the subsets of  $\Omega$  by

$$P^*(A) = \inf\{P(B) : B \in \mathcal{F} \text{ with } A \subset B\}, \quad A \subseteq \Omega.$$

We say that a sequence  $(Y_n)$  of real-valued maps defined on  $\Omega$  converges in  $P^*$ -probability to 0 and denote  $Y_n \xrightarrow{P^*} 0$  if, for any  $\epsilon > 0$ ,  $P^* (\{|Y_n| \geq \epsilon\}) \rightarrow 0$ . Here  $\{|Y_n| \geq \epsilon\}$  is the usual short-hand notation for the subset  $\{\omega \in \Omega : |Y_n(\omega)| \geq \epsilon\}$ . When  $Y_n$  is measurable as a map taking values in  $\mathbb{R}$  endowed with the Borel  $\sigma$ -field, this is equivalent to the usual convergence in probability.

**Theorem 4.** *Suppose that Assumption 3 holds for some  $\beta \in \Phi$ ,  $M$  defined on  $\Phi$  and  $\{M_n(\phi), \phi \in \Phi\}$ , a sequence of real-valued processes. Let  $(J_n)$  be a sequence of non-negative functions defined on  $\Phi$  such that  $J_n(\beta) \rightarrow 0$ . Let  $\mathbb{T}$  be a compact subset of  $[0, \infty)$  and suppose that we have a  $\Phi$ -valued process  $\{\widehat{\beta}_n(\mathbf{t}), \mathbf{t} \geq 0\}$  such that*

$$\sup_{\mathbf{t} \in \mathbb{T}} \left\{ \Lambda_n(\widehat{\beta}_n(\mathbf{t}), \mathbf{t}) - \Lambda_n(\beta, \mathbf{t}) \right\}_+ \xrightarrow{P^*} 0, \quad (12)$$

where  $\Lambda_n$  is defined by (1). Then  $\widehat{\beta}_n(\mathbf{t})$  converges to  $\beta$  uniformly in  $\mathbf{t} \in \mathbb{T}$ , in  $P^*$ -probability, that is,

$$\sup_{\mathbf{t} \in \mathbb{T}} d(\widehat{\beta}_n(\mathbf{t}), \beta) \xrightarrow{P^*} 0. \quad (13)$$

*Remark 2.* In statistical applications the contrast function  $M$  in Assumption 3 depends on the unknown distribution of the contrast process  $M_n$  and thus  $\beta$  is an unknown point of  $\Phi$ . In particular, the convergence condition  $J_n(\beta) \rightarrow 0$  has to be verified for any  $\beta \in \Phi$  (but not uniformly in  $\beta$ ) and it simply amounts to correctly normalize the penalty  $J_n$  as  $n \rightarrow \infty$ .

*Remark 3.* The same result holds if the convergence in  $P$ -probability in Assumption 3-(i) is replaced by a convergence in  $P^*$ -probability. However, in applications, the smoothness properties of  $\phi \mapsto M_n(\Phi)$  and  $\phi \mapsto M(\Phi)$  usually imply that  $\sup_{\phi \in \Phi} \{M(\phi) - M_n(\phi)\}_+$  is a measurable function.

*Remark 4.* The fact that the outer probability  $P^*$  appears in (12) does not bring real difficulties in applications. Indeed Condition (12) follows from the definition of  $\hat{\beta}_n(\mathbf{t})$  as a near minimizer of  $\Lambda_n(\cdot, \mathbf{t})$ , that is, if  $\hat{\beta}_n(\mathbf{t})$  satisfies

$$\Lambda_n(\hat{\beta}_n(\mathbf{t}), \mathbf{t}) \leq \inf_{\phi \in \Phi} \Lambda_n(\phi; \mathbf{t}) + u_n,$$

with  $u_n = o_P(1)$  not depending on  $\mathbf{t}$ , e.g.  $u_n = 0$  (perfect minimizer) or  $u_n = n^{-1}$  (near minimizer). The numerical computation of a near minimizer is a difficult task in general, in particular in the presence of several local minima. We will focus on convexity assumptions in Section 3.2, which cover many cases of interest and which usually allow tractable numerical computation of  $\hat{\beta}_n(\mathbf{t})$  for any  $\mathbf{t}$ .

*Remark 5.* Although  $\hat{\beta}_n(\mathbf{t})$  is an r.v. for any  $\mathbf{t}$ , the map  $\sup_{\mathbf{t} \in T} \|\hat{\beta}_n(\mathbf{t}) - \beta\|$  defined on  $\Omega$  may not be measurable (it is in some particular cases, for instance if the map  $\mathbf{t} \mapsto \hat{\beta}_n(\mathbf{t})$  is continuous). This is where the outer probability is useful. Nevertheless, for any  $\mathbf{t} \geq 0$ , the event  $\{d(\hat{\beta}_n(\mathbf{t}), \beta) \geq \epsilon\}$  is measurable, and its probability is less than the left-hand side of Eq. (13); hence, for any  $\mathbf{t} \geq 0$ ,  $\hat{\beta}_n(\mathbf{t}) \xrightarrow{P} \beta(\mathbf{t})$ .

*Remark 6.* For  $L = 0$  in (13), we get a standard result on the consistency of M-estimators (without penalty). It is important to notice that the consistency of penalized M-estimators is obtained for free, in the sense that no additional assumption on  $M_n$  or  $M$  is required and the only assumption on  $J_n$  is  $J_n(\beta) \rightarrow 0$ .

**3.2. Uniform consistency in the convex case.** In this section, we consider the following assumption.

**Assumption 4** (convexity assumption).  $\Phi$  is a convex subset of an Euclidean space endowed with the norm  $\|\cdot\|$  and  $M_n$  is a convex real-valued function on  $\Phi$  almost surely. Let  $V \subseteq \Phi$  be a neighborhood of the point  $\beta$  and  $\Delta$  be a strictly convex real-valued function defined on  $V$  such that

- (i) for any  $\phi \in V$ ,  $M_n(\phi) \xrightarrow{P} \Delta(\phi)$ ;
- (ii)  $\Delta(\phi) \geq \Delta(\beta)$  for all  $\phi \in V$ .

Convex M-estimation is considered in Haberman [1989] and somewhat simplified in Niemiro [1992]. In the following result the convexity assumption is twofold. First it implies Assumption 3. Second, if the contrast the penalty  $J_n$  is strictly convex, then the minimization of (1) has a unique solution and this solution path is continuous, which allows to replace the outer probability in (13) by a standard probability. Convexity is also useful in practice since  $\hat{\beta}_n(\mathbf{t})$  can be computed using efficient numerical procedure for convex optimization (see Boyd and Vandenberghe [2004]).

**Theorem 5.** *Suppose that Assumption 4 holds. Let  $(J_n)$  be a sequence of non-negative functions defined on  $\Phi$  such that  $J_n(\beta) \rightarrow 0$  and define  $\Lambda_n$  as in (1). Then the 3 following assertions hold.*

(a) *For any  $L \geq 0$ , if we have a  $\Phi$ -valued process  $\{\hat{\beta}_n(\mathbf{t}), \mathbf{t} \geq 0\}$  satisfying (12),  $\hat{\beta}_n(\mathbf{t})$  converges to  $\beta$  uniformly in  $\mathbf{t} \in [0, L]$ , in  $P^*$ -probability, that is, (13) holds.*

(b) *If  $J_n$  is strictly convex on  $\Phi$ , then it is always possible to define a deterministic non-negative sequence  $(L_n)$  with  $L_n \rightarrow \infty$ , a sequence  $(A_n)$  of events in  $\mathcal{F}$  with  $P(A_n) \rightarrow 1$ , and, for each  $n$ , a collection  $\{\hat{\beta}_n(\mathbf{t}), \mathbf{t} \geq 0\}$  of r.v.'s satisfying the two following properties.*

(b1) *For all  $\mathbf{t} \in [0, L_n]$  and  $\omega \in A_n$ ,  $\Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}), \mathbf{t})$  is a minimum of  $\Lambda_n(\phi, \mathbf{t})$  on  $\phi \in \Phi$  and this minimum is unique for  $\mathbf{t} > 0$ .*

(b2) *For all  $\omega \in \Omega$ ,  $\hat{\beta}_n(\omega, \cdot)$  is a continuous function on  $(0, L_n]$  and on  $(L_n, \infty)$ .*

*As consequences, (12) holds for any  $L > 0$  and the uniform convergence (13) holds in  $P$ -probability, that is,*

$$\sup_{\mathbf{t} \in [0, L]} \|\hat{\beta}_n(\mathbf{t}) - \beta\| \xrightarrow{P} 0. \quad (14)$$

(c) *If  $M_n$  is strictly convex on  $\Phi$  for all  $n$ , then the conclusions of (b) hold with Properties (b1) and (b2) strengthened as follows.*

(c1) *For all  $\mathbf{t} \in [0, L_n]$  and  $\omega \in A_n$ ,  $\Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}), \mathbf{t})$  is the unique minimum of  $\Lambda_n(\phi, \mathbf{t})$  on  $\phi \in \Phi$ .*

(c2) *For all  $\omega \in \Omega$ ,  $\hat{\beta}_n(\omega, \cdot)$  is a continuous function on  $[0, L_n]$  and on  $(L_n, \infty)$ .*

*Remark 7.* The proof of Assertion (c) is somewhat simpler than Assertion (b). However, in some cases, the first purpose of the penalty  $J_n$  is precisely to solve an ill-posed problem such as in the ridge regression (see Hoerl and Kennard [1970]) where  $M_n(\phi) = \sum_k (y_k - \mathbf{x}_k^T \phi)^2$ ,  $J_n(\phi) \propto \|\phi\|^2$  and the regression matrix  $\mathbf{X}_n = [\mathbf{x}_1 \ \dots \ \mathbf{x}_n]^T$  is not full rank. Thus  $J_n$  is strictly convex and  $M_n$  is not, in which case Assertion (b) can be useful.

**3.3. Functional central limit theorem.** Some general conditions for proving  $\sqrt{n}$  asymptotic normality for M-estimators rely on the so called stochastic differentiability condition introduced in Pollard [1985]. They exploit the idea introduced in Huber [1967] of using strong differentiability conditions on the limit contrast function rather than on the contrast process. Moreover it is explained in Pollard [1985] how the empirical process theory can be used to prove the stochastic differentiability condition. Extensions of these ideas can be found in Van der Vaart and Wellner [1996].

In Pollard [1985], Pollard proves the asymptotic normality of M-estimators based on a contrast process of the form

$$M_n(\phi) = n^{-1} \sum_{k=1}^n g(\xi_k, \phi) = P_n g(\cdot, \phi), \quad (15)$$

where  $(\xi_k)$  is a sequence of  $\mathcal{X}$ -valued random variables and  $g$  is a  $\mathcal{X} \times \mathbb{R}^p$  function satisfying the following Taylor expansion around a given point  $\beta \in \mathbb{R}^p$ ,

$$g(x, \phi) = g(x, \beta) + (\phi - \beta)^T \Delta(x) + \|\phi - \beta\| r(x, \phi). \quad (16)$$

We will show that if the  $\sqrt{n}$  asymptotic normality conditions in Pollard [1985] are verified and if the penalty satisfies mild asymptotic conditions then the penalized version of the

M-estimator satisfies a CLT similar to the CLT in Knight and Fu [2000] for the mean square criterion. Moreover this CLT applies to the regularization path in a functional sense.

Let us recall Pollard's conditions that we will use on the contrast process  $M_n$  defined by (15) and (16).

- (P-1)  $(\xi_k)$  is a sequence of i.i.d. random variables with distribution  $P$ ;
- (P-2) the function  $M(\phi) = Pg(\cdot, \phi)$  has a nonsingular second derivative  $\Gamma$  at  $\beta \in \mathbb{R}^p$ ;
- (P-3)  $P\|\Delta\|^2 < \infty$  and  $P\Delta = 0$ ;
- (P-4) the stochastic differentiability condition holds on  $r$ , that is, for any sequence of positive r.v.  $(r_n)$  such that  $r_n \xrightarrow{P} 0$ ,

$$\sup_{\|\phi - \beta\| \leq r_n} \frac{|\nu_n r(\cdot, \phi)|}{1 + \sqrt{n}\|\phi - \beta\|} \xrightarrow{P} 0. \quad (17)$$

Here we used the notations, standard in the empirical process literature,  $Pf$ ,  $P_n f$  and  $\nu_n f$  for  $\int f dP$ ,  $n^{-1} \sum_{k=1}^n f(\xi_k)$  and  $\sqrt{n}(P_n f - Pf)$ , respectively. Theorem 6 below provides a central limit theorem for the regularization path defined on the penalized contrast (1) when  $M_n$  satisfies Pollard's conditions (P-1)–(P-4) with some mild conditions on the penalty  $J_n$ .

**Theorem 6.** *Let  $\Phi = \mathbb{R}^p$ ,  $p \geq 1$  and  $\mathbb{T}$  be a compact subset of  $[0, \infty)$ . Define  $\Lambda_n$  as in (1), where  $M_n$  is defined by (15) and satisfies Pollard's conditions (P-1)–(P-4) and  $J_n$  is a sequence of deterministic non-negative functions defined on  $\mathbb{R}^p$ . Further assume that there exists a positive constant  $C$  such that*

$$n |J_n(\phi) - J_n(\beta)| \leq C (1 + \sqrt{n}\|\phi - \beta\|) \quad \text{for } \|\phi - \beta\| \leq 1, \quad (18)$$

and, for any compact  $K \subset \mathbb{R}^p$ ,

$$\sup_{\phi \in K} \left| n J_n(\beta + n^{-1/2}\phi) - n J_n(\beta) - J_\infty(\phi) \right| \rightarrow 0, \quad (19)$$

where  $J_\infty$  is a real-valued function on  $\Phi$ . Let  $\{\hat{\beta}_n, \mathbf{t} \in \mathbb{T}\}$  be a sequence of  $\Phi$ -valued processes satisfying (12) and such that the uniform  $P^*$ -consistency (13) holds. Let  $W$  be a centered Gaussian  $p$ -dimensional vector with covariance  $P(\Delta\Delta^T)$  and define

$$\mathbb{L}(\phi, \mathbf{t}) = W^T \phi + \phi^T \Gamma \phi + \mathbf{t} J_\infty(\phi). \quad (20)$$

Finally assume that there exists a  $\Phi$ -valued process  $\{\hat{\mathbf{u}}(\mathbf{t}), \mathbf{t} \in \mathbb{T}\}$  such that Conditions (ii) and (iii) in Theorem 3 hold. Then there is a version of  $\hat{\mathbf{u}}$  in  $\ell^\infty(\mathbb{T}, \mathbb{R}^p)$  and

$$\sqrt{n}(\hat{\beta}_n - \beta) \rightsquigarrow \hat{\mathbf{u}}. \quad (21)$$

The following lemma shows that the penalties considered in Knight and Fu [2000] satisfy Conditions (18) and (19).

**Lemma 1.** *Let  $\gamma > 0$  and define, for all  $\phi = (\phi_1, \dots, \phi_p) \in \mathbb{R}^p$ ,*

$$J_n^{(\gamma)}(\phi) = n^{(1 \wedge \gamma)/2 - 1} \sum_{k=1}^p |\phi_k|^\gamma. \quad (22)$$

Then for any  $\beta \in \mathbb{R}^p$ , there exists  $C > 0$  such that, for all  $\phi \in \mathbb{R}^p$ ,

$$n \left| J_n^{(\gamma)}(\phi) - J_n^{(\gamma)}(\beta) \right| \leq C (1 + \sqrt{n}\|\phi - \beta\| + \sqrt{n}\|\phi - \beta\|^{1 \vee \gamma}), \quad (23)$$

and, for any compact  $K \subset \mathbb{R}^p$ ,

$$\sup_{\phi \in K} \left| nJ_n^{(\gamma)}(\beta + n^{-1/2}\phi) - nJ_n^{(\gamma)}(\beta) - J_\infty^{(\gamma)}(\phi) \right| \rightarrow 0, \quad (24)$$

where

$$J_\infty^{(\gamma)}(\phi) = \begin{cases} \sum_{j=1}^p |\phi_j|^\gamma \mathbb{1}_{\{\beta_j=0\}} & \text{if } \gamma < 1 \\ \sum_{j=1}^p \left\{ \phi_j \operatorname{sgn}(\beta_j) \mathbb{1}_{\{\beta_j \neq 0\}} + |\phi_j| \mathbb{1}_{\{\beta_j=0\}} \right\} & \text{if } \gamma = 1 \\ \gamma \sum_{j=1}^p \phi_j \operatorname{sgn}(\beta_j) |\beta_j|^{\gamma-1} \mathbb{1}_{\{\beta_j \neq 0\}} & \text{if } \gamma > 1. \end{cases} \quad (25)$$

*Remark 8.* The limit penalties in (25) correspond to those in Theorems 2 and 3 in Knight and Fu [2000], except for the multiplicative constant  $\gamma$  in the case  $\gamma > 1$ , which seems to have been forgotten in Knight and Fu [2000].

#### 4. PATHWISE M-ESTIMATION

It turns out that the specific form of the contrast  $\Lambda_n$  in (1) is not fundamental for the basic arguments yielding the consistency and the CLT in Theorems 4 and 6, respectively. Here we provide results formulated in the more general form where  $\widehat{\beta}_n(\mathbf{t})$  is a near minimizer of  $\Lambda_n(\cdot, \mathbf{t})$  for all  $\mathbf{t} \in \mathbb{T}$ . Moreover the true parameter  $\beta$  itself is defined as a map on  $\mathbb{T}$ , with  $\beta(\mathbb{T})$  defined as the minimizer of  $\mathbb{L}(\cdot, \mathbf{t})$  for all  $\mathbf{t} \in \mathbb{T}$ . We refer this general situation as *pathwise M-estimation*.

**4.1. Uniform consistency.** Theorem 4 is obtained by applying the following general result on pathwise M-estimators.

**Proposition 1.** *Let  $\Phi$  be a subset of a metric space endowed with the metric  $d$  and  $\mathbb{T}$  be any set. Let  $\Lambda$  be a real-valued function defined on  $\Phi \times \mathbb{T}$ ,  $\{\Lambda_n(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in \mathbb{T}\}$  be a sequence of real-valued processes,  $\beta$  be a  $\mathbb{T} \rightarrow \Phi$  map and  $\{\widehat{\beta}_n(\mathbf{t}), \mathbf{t} \in \mathbb{T}\}$  be a sequence of  $\Phi$ -valued processes such that*

- (i)  $\sup_{\phi \in \Phi} \sup_{\mathbf{t} \in \mathbb{T}} \{\Lambda(\phi; \mathbf{t}) - \Lambda_n(\phi, \mathbf{t})\}_+ \xrightarrow{P^*} 0;$
- (ii)  $\sup_{\mathbf{t} \in \mathbb{T}} |\Lambda_n(\beta(\mathbf{t}), \mathbf{t}) - \Lambda(\beta(\mathbf{t}), \mathbf{t})| \xrightarrow{P^*} 0;$
- (iii) For all  $\epsilon > 0$ ,

$$\inf_{\mathbf{t} \in \mathbb{T}} [\inf\{\Lambda(\phi; \mathbf{t}) : \phi \in \Phi, d(\phi, \beta(\mathbf{t})) \geq \epsilon\} - \Lambda(\beta(\mathbf{t}), \mathbf{t})] > 0;$$

- (iv)  $\sup_{\mathbf{t} \in \mathbb{T}} \left\{ \Lambda_n(\widehat{\beta}_n(\mathbf{t}), \mathbf{t}) - \Lambda_n(\beta(\mathbf{t}), \mathbf{t}) \right\}_+ \xrightarrow{P^*} 0.$

Then,  $\widehat{\beta}_n(\mathbf{t})$  converges to  $\beta(\mathbf{t})$  uniformly in  $\mathbf{t} \in \mathbb{T}$ , in  $P^*$ -probability, that is,

$$\sup_{\mathbf{t} \in \mathbb{T}} d(\widehat{\beta}_n(\mathbf{t}), \beta(\mathbf{t})) \xrightarrow{P^*} 0. \quad (26)$$

**4.2. Functional central limit theorem.** We now extend the setting of Pollard [1985] to pathwise M-estimation. First we obtain the  $\sqrt{n}$ -rate of convergence in the sup norm; second we apply Theorem 3 to obtain a functional CLT for pathwise M-estimators. Theorem 6 is a direct application of this result in the context of penalized M-estimation.

**Proposition 2.** *Let  $\Phi$  be a subset of a metric space endowed with the metric  $d$  and  $\mathbb{T}$  be any set. Let  $\{\Lambda_n(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in \mathbb{T}\}$  be a sequence of real-valued processes,  $\beta$  be a  $\mathbb{T} \rightarrow \Phi$  map and  $\{\hat{\beta}_n(\mathbf{t}), \mathbf{t} \in \mathbb{T}\}$  be a sequence of  $\Phi$ -valued processes such that*

$$\sup_{\mathbf{t} \in \mathbb{T}} \left\{ \Lambda_n(\hat{\beta}_n(\mathbf{t}), \mathbf{t}) - \Lambda_n(\beta(\mathbf{t}), \mathbf{t}) \right\}_+ = O_{P^*}(n^{-1}), \quad (27)$$

and the uniform  $P^*$ -consistency (26) holds. Assume that we have the following decomposition of the contrast process,

$$\Lambda_n(\phi, \mathbf{t}) - \Lambda_n(\beta(\mathbf{t}), \mathbf{t}) = G_n(\phi, \mathbf{t}) + H(\phi, \mathbf{t}) + d(\phi, \beta(\mathbf{t})) R_n(\phi, \mathbf{t}), \quad (28)$$

where  $G_n$ ,  $H$  and  $R_n$  satisfy

(i)  $\{G_n(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in \mathbb{T}\}$  is a sequence of real-valued processes such that

$$\sup_{\phi \in \Phi} \sup_{\mathbf{t} \in \mathbb{T}} \frac{n |G_n(\phi, \mathbf{t})|}{1 + \sqrt{n} d(\phi, \beta(\mathbf{t}))} = O_{P^*}(1); \quad (29)$$

(ii)  $H$  is a real-valued function defined on  $\Phi \times \mathbb{T}$  such that there exists  $\epsilon > 0$  for which

$$\inf_{\mathbf{t} \in \mathbb{T}} \inf \left\{ \frac{H(\phi, \mathbf{t})}{d^2(\phi, \beta(\mathbf{t}))} : \phi \in \Phi, d(\phi, \beta(\mathbf{t})) \leq \epsilon \right\} > 0; \quad (30)$$

(iii)  $\{R_n(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in \mathbb{T}\}$  is a sequence of real-valued processes such that, for any positive random sequence  $(r_n)$  converging to 0 in  $P^*$ -probability,

$$\sup_{\mathbf{t} \in \mathbb{T}} \sup \{|R_n(\phi, \mathbf{t})|; \phi \in \Phi, d(\phi, \beta(\mathbf{t})) \leq r_n\} = o_{P^*}(r_n) + O_{P^*}(n^{-1/2}). \quad (31)$$

Then,  $\hat{\beta}_n(\mathbf{t})$  converges to  $\beta(\mathbf{t})$  uniformly in  $\mathbf{t} \in \mathbb{T}$ , in  $P^*$ -probability, with rate at least  $\sqrt{n}$ , that is,

$$\sqrt{n} \sup_{\mathbf{t} \in \mathbb{T}} d(\hat{\beta}_n(\mathbf{t}), \beta(\mathbf{t})) = O_{P^*}(1). \quad (32)$$

Applying Proposition 2 and Theorem 3, we get the following result.

**Theorem 7.** *Let  $\Phi = \mathbb{R}^p$ ,  $p \geq 1$ , and  $\mathbb{T}$  be any set. Let  $\{\Lambda_n(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in \mathbb{T}\}$  be a sequence of real-valued processes,  $\beta$  be a  $\mathbb{T} \rightarrow \Phi$  map and  $\{\hat{\beta}_n(\mathbf{t}), \mathbf{t} \in \mathbb{T}\}$  be a sequence of  $\Phi$ -valued processes such that*

$$\sup_{\mathbf{t} \in \mathbb{T}} \left\{ \Lambda_n(\hat{\beta}_n(\mathbf{t}), \mathbf{t}) - \Lambda_n(\beta(\mathbf{t}), \mathbf{t}) \right\}_+ = o_{P^*}(n^{-1}), \quad (33)$$

and the uniform  $P^*$ -consistency (26) holds. Assume that the decomposition (28) of the contrast process holds where  $G_n$ ,  $H$  and  $R_n$  satisfy:

(i)  $\{G_n(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in \mathbb{T}\}$  is a sequence of real-valued processes satisfying (29);

(ii)  $H$  is a real-valued function defined on  $\Phi \times \mathbb{T}$  and there exists a function  $\Gamma$  defined on  $\mathbb{T}$  and taking values in the set of non-negative symmetric  $p \times p$  matrices such that, denoting by  $\lambda_{\min}(\Gamma(\mathbf{t}))$  and  $\lambda_{\max}(\Gamma(\mathbf{t}))$  the smallest and largest eigenvalues of  $\Gamma(\mathbf{t})$ ,

$$0 < \inf\{\lambda_{\min}(\Gamma(\mathbf{t})), \mathbf{t} \in \mathbb{T}\} < \sup\{\lambda_{\max}(\Gamma(\mathbf{t})), \mathbf{t} \in \mathbb{T}\} < \infty, \quad (34)$$

and, as  $\phi \rightarrow \beta$  in  $\ell^\infty(\mathbb{T}, \mathbb{R}^p)$ ,

$$\|H(\phi(\cdot), \cdot) - (\phi - \beta)^T \Gamma(\phi - \beta)\|_{\mathbb{T}} = o(\|\phi - \beta\|_{\mathbb{T}}^2) ; \quad (35)$$

(iii)  $\{R_n(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in \mathbb{T}\}$  is a sequence of real-valued processes such that, for any positive random sequence  $(r_n)$  converging to 0 in  $P^*$ -probability,

$$\sup_{\mathbf{t} \in \mathbb{T}} \sup \{|R_n(\phi, \mathbf{t})| ; \phi \in \Phi, d(\phi, \beta(\mathbf{t})) \leq r_n\} = o_{P^*}(r_n) + o_{P^*}(n^{-1/2}) . \quad (36)$$

Let us further define

$$\widehat{G}_n(\phi, \mathbf{t}) = nG_n(\beta(\mathbf{t}) + n^{-1/2}\phi, \mathbf{t}) , \quad (37)$$

and assume that there exists a real-valued process  $\{G(\phi, \mathbf{t}), \phi \in \Phi, \mathbf{t} \in \mathbb{T}\}$  such that, for any compact  $K \subset \Phi$ ,  $G$  is tight in  $\ell^\infty(K \times \mathbb{T}, \mathbb{R}^p)$  and  $\widehat{G}_n \rightsquigarrow G$  in  $\ell^\infty(K \times \mathbb{T}, \mathbb{R}^p)$ . Define

$$\mathbb{L}(\phi, \mathbf{t}) = G(\phi, \mathbf{t}) + \phi^T \Gamma(\mathbf{t}) \phi , \quad (38)$$

and assume that there exists a  $\Phi$ -valued process  $\{\widehat{\mathbf{u}}(\mathbf{t}), \mathbf{t} \in \mathbb{T}\}$  such that Conditions (ii) and (iii) in Theorem 3 hold. Then there is a version of  $\widehat{\mathbf{u}}$  in  $\ell^\infty(\mathbb{T}, \mathbb{R}^p)$  and

$$\sqrt{n}(\widehat{\beta}_n - \beta) \rightsquigarrow \widehat{\mathbf{u}} . \quad (39)$$

*Remark 9.* Observe that Eq. (33) is a strengthened version of (32) and that (34) and (35) imply (30). Hence Conditions (i)–(iii) in Theorem 7 imply Conditions (i)–(iii) in Proposition 2.

## 5. EXAMPLES

The uniform consistency and a functional central limit theorem for the lasso regularization path are respectively given in Theorems 1 and 2. Theorems 5 and 6 allow many extensions, some examples of which are given in this section. In Pollard [1985], a wide variety of models and functions  $g$  are shown to satisfy Conditions (P-1)–(P-4). These conditions apply for the general linear model (GLM) as this model satisfies the pointwise assumptions of [Pollard, 1985, Section 4] (provided some moment conditions). They also apply for the least absolute deviation (LAD) criterion, see Example 8 in [Pollard, 1985, Section 6] (provided again some moment conditions on the model). We briefly write the corresponding results in these two cases as examples of applications of Theorem 6. Uniform consistencies for both examples are obtained as applications of Theorem 5, since in these cases  $M_n$  is convex. For these two examples, we consider the  $\ell^1$  and  $\ell^2$  penalties. They fit the conditions of Theorem 6 as they satisfy (18) and (19) by Lemma 1. Observe however that the function  $J_\infty$  in Lemma 1 depends on the chosen penalty and thus so does the limit  $\widehat{\mathbf{u}}$  in (21). We conclude this section with a discussion on the Akaike information criterion (AIC), which corresponds to a  $\ell^0$  penalty.

**5.1.  $\ell^1$ -penalized GLM.** Consider a canonical exponential family of density

$$p(y|\theta) = h(y) \exp\{y\theta - b(\theta)\} ,$$

with respect to a dominating measure  $\mu$ . The function  $b$ , sometimes called the log-repartition function, is given by

$$b(\theta) = \log \int h(y) \exp\{y\theta\} \mu(dy) ,$$

and thus is strictly convex and infinitely differentiable. In a GLM, one observes a sequence of i.i.d.  $\mathbb{R} \times \mathbb{R}^p$ -valued r.v.'s  $(y_k, \mathbf{x}_k)$ ,  $k = 1, \dots, n$ , where  $y_k$  have conditional density  $p(\cdot | \mathbf{x}_k^T \boldsymbol{\beta})$ , given  $\mathbf{x}_k$ , with  $\boldsymbol{\beta} \in \mathbb{R}^p$  denoting the unknown parameter of interest. In this context, the non-penalized contrast process is given by the negated log-likelihood

$$M_n(\boldsymbol{\phi}) = n^{-1} \sum_{k=1}^n g((\mathbf{x}_k, y_k), \boldsymbol{\phi}),$$

where  $g((\mathbf{x}, y), \boldsymbol{\phi}) = -y\mathbf{x}^T \boldsymbol{\phi} + b(\mathbf{x}^T \boldsymbol{\phi})$ . Using that  $g$  is convex and smooth, and assuming some appropriate moment conditions on  $\mathbf{x}_1$  for obtaining Pollard's conditions (P-1)–(P-4), we get the uniform consistency and a functional CLT on the regularization path  $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$  defined as the minimizer of (1) with  $J_n(\boldsymbol{\phi}) = n^{-1/2} \sum_{i=1}^p |\phi_i|$  (this is the  $\ell^1$  penalty  $J_n^{(1)}$  defined in (22)). In particular, for any  $L > 0$ ,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightsquigarrow \widehat{\mathbf{u}} \text{ in } \ell^\infty([0, L], \mathbb{R}^p),$$

where the limit  $\widehat{\mathbf{u}}$  is defined as in the lasso case as the minimizer of (5) with  $C = \mathbb{E}[b''(\mathbf{x}_1^T \boldsymbol{\beta}) \mathbf{x}_1 \mathbf{x}_1^T]$  (assumed positive-definite) and  $U \sim \mathcal{N}(0, C)$ . The numerical computation of  $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$  can be processed as proposed in Park and Hastie [2007].

**5.2.  $\ell^1$  and  $\ell^2$ -penalized LAD.** Given a sequence of  $\mathbb{R} \times \mathbb{R}^p$ -valued r.v.'s  $(y_k, \mathbf{x}_k)$ ,  $k = 1, \dots, n$ , the LAD criterion is defined as

$$M_n(\boldsymbol{\phi}) = n^{-1} \sum_{k=1}^n |y_k - \mathbf{x}_k^T \boldsymbol{\phi}|.$$

It can be used to estimate the parameter  $\boldsymbol{\beta} \in \mathbb{R}^p$  of a linear regression model  $y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k$ , with  $(\varepsilon_k)$  and  $(\mathbf{x}_k)$  two independent sequence of i.i.d. r.v.'s. This contrast process is an alternative to the mean square criterion, resulting in an estimator less sensitive to the presence of outliers (for  $\mathbf{x}_k = 1$ , the minimizer of  $M_n$  is the sample median). In contrast to the previous case, the contrast is not smooth, since the first derivative is discontinuous. However, as shown *e.g.* in Pollard [1985], the minimizer of this contrast is asymptotically normal, provided some moment conditions and that

$$G(\boldsymbol{\phi}) = \mathbb{E} [|\varepsilon_1 + \mathbf{x}_1^T (\boldsymbol{\beta} - \boldsymbol{\phi})|]$$

has a non-singular second derivative at  $\boldsymbol{\phi} = \boldsymbol{\beta}$ . Observe that

$$G(\boldsymbol{\phi}) = \mathbb{E} \left[ \mathbf{x}_1^T (\boldsymbol{\beta} - \boldsymbol{\phi}) + 2 \int_0^{\mathbf{x}_1^T (\boldsymbol{\phi} - \boldsymbol{\beta})} F(s) ds \right],$$

where  $F$  denotes the cumulative distribution function of  $\varepsilon_1$ . Thus, if  $\varepsilon_1$  is distributed from a continuous density  $f$ , the second derivative of  $G$  at  $\boldsymbol{\beta}$  is  $\Gamma = 2f(0)\mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^T]$ . Because the LAD criterion uses the  $\ell^1$  error function, the  $\ell^2$  penalty  $J_n(\boldsymbol{\phi}) = n^{-1/2} \sum_{i=1}^p \phi_i^2$  could seem more reasonable. On the contrary Theorem 6 suggests that using an  $\ell^1$  error function contrast does not modify the asymptotic distribution of the regularization path, only the choice of the penalty does. In other words, the regularization path of the  $\ell^1$  and  $\ell^2$ -penalized LAD has similar asymptotic distributions as the lasso and the ridge regression, respectively. Let us now precise the limit distribution of the regularization path  $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$  defined as the minimizer of (1) with  $J_n(\boldsymbol{\phi}) = n^{-1/2} \sum_{i=1}^p |\phi_i|$  and  $J_n(\boldsymbol{\phi}) = n^{-1/2} \sum_{i=1}^p \phi_i^2$  respectively

(these are the  $\ell^1$  and  $\ell^2$  penalty  $J_n^{(1)}$  and  $J_n^{(2)}$  defined in (22)). Under appropriate moment conditions on  $(\varepsilon_1, \mathbf{x}_1)$  implying Pollard's conditions (P-1)–(P-4) (in particular  $\mathbb{E}[\text{sgn}(\varepsilon_1)] = 0$ ,  $\mathbb{E}[\|\mathbf{x}_1\|^2] < \infty$  so that  $\mathbb{E}[\Delta] = 0$ ,  $\mathbb{E}[\|\Delta\|^2] < \infty$  and  $G$  is minimized at  $\phi = \beta$ ), one has, for any  $L > 0$ ,

$$\sqrt{n}(\widehat{\beta}_n - \beta) \rightsquigarrow \widehat{\mathbf{u}} \text{ in } \ell^\infty([0, L], \mathbb{R}^p),$$

where the limit  $\widehat{\mathbf{u}}$  is defined as the minimizer of (20) where  $\Gamma$  is the (non-singular) second derivative of  $G$  at  $\phi = \beta$ ,  $W \sim \mathcal{N}(0, \mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^T])$  and  $J_\infty$  depends on the penalty. Namely, for the  $\ell^1$  penalty, one has  $J_\infty = J_\infty^{(1)}$  and for the  $\ell^2$  penalty, one has  $J_\infty = J_\infty^{(2)}$ , where  $J_\infty^{(\gamma)}$  is defined by (25).

**5.3. Akaike information criterion and the  $\ell^0$  penalty.** Consider a parametric family of densities  $\{p_\phi, \phi \in \Phi\}$  defined on  $\mathcal{X}^n$  for modelling the distribution of the observations  $\xi_1, \dots, \xi_n$ . The Akaike information criterion (AIC) was proposed in Akaike [1973] as the negated log-likelihood criterion penalized by the dimension of the parameter. It can be defined (up to a multiplicative factor which does not change its minimizer) as

$$\text{AIC}(\phi) = \Lambda_n(\phi, 1),$$

where  $\Lambda_n$  is defined by (1) with  $M_n(\phi) = -n^{-1} \log p_\phi(\xi_1, \dots, \xi_n)$  and

$$J_n^{(0)}(\phi) = n^{-1} \# \{k : \phi_k \neq 0\},$$

where  $\#A$  denotes the cardinality of the set  $A$ . We note that it corresponds to a  $\ell^0$  penalty, that is, to  $\gamma = 0$  in (22) although this case is not considered in Knight and Fu [2000]. It is not usually assumed that  $\Phi$  is finite-dimensional in the presentation of the AIC. However, in practice, the minimization of  $\text{AIC}(\phi)$  requires numerically minimizing  $M_n(\phi)$  for each possible submodel, which corresponds to a given value of the sequence  $(\mathbb{1}(\phi_k \neq 0))_{k \geq 1}$ . This makes sense only in a finite-dimensional setting,  $\Phi \subseteq \mathbb{R}^p$ , with  $p$  not too large (say  $p \leq 15$ ) since  $2^p$  numerical minimizations of  $M_n$  are then necessary.

Observe that, for any fixed  $\phi \in \mathbb{R}^p$  we have  $nJ_n^{(0)}(\phi) \leq p$  and, for any  $\beta \in \mathbb{R}^p$  and any  $r > 0$ , we have, for  $n$  large enough,

$$nJ_n^{(0)}(\beta + n^{-1/2}\phi) - nJ_n^{(0)}(\beta) = J_\infty^{(0)}(\phi) \quad \text{for all } \|\phi\| \leq r,$$

where

$$J_\infty^{(0)}(\phi) = \sum_{k=1}^p \mathbb{1}(\beta_k = 0 \text{ and } \phi_k \neq 0).$$

It follows that the contrast  $J_n^{(0)}$  satisfies the assumptions (18) and (19) in Theorem 6 and thus we may apply this result to obtain the limit behavior of the minimizer of the AIC in the i.i.d. case, that is, when

$$M_n(\phi) = P_n g(\cdot, \phi) \quad \text{with } g(x, \phi) = -\log p_\phi(x).$$

Here,  $p_\phi$  denotes the density of one observation in the parametric family  $\{p_\phi, \phi \in \Phi\}$ . Suppose that this model satisfy Assumption 3 and the Pollard's conditions (P-1)–(P-4) with  $\beta$  denoting the true parameter and with  $\Gamma = P(\Delta \Delta^T)$  equal to the Fisher information matrix at parameter  $\beta$ . We may thus apply Theorem 4 and Theorem 6 successively to the minimizing sequence

$$\widehat{\beta}_n = \underset{\phi \in \Phi}{\text{Argmin}} \text{AIC}(\phi).$$

We obtain  $\widehat{\beta}_n \rightsquigarrow \beta$  and  $\sqrt{n}(\widehat{\beta}_n - \beta) \rightsquigarrow \widehat{u}$ , where  $\widehat{u}$  is defined as the minimizer of (20) with  $J_\infty = J_\infty^{(0)}$  and  $\mathbf{t} = 1$ . Observe that, in the limit penalty  $J_\infty^{(0)}$ , only the vanishing coordinates of the true parameter  $\beta$  are penalized. In other words, for a coordinate  $k$  such that  $\beta_k = 0$  and only for such a coordinate, we have  $\widehat{u}_k = 0$  with positive probability. This property highlights the (well known) ability of the AIC criterion to correctly select the correct model.

Finally we note that the AIC can easily be extended to a collection of contrast  $\Lambda_n(\phi, \mathbf{t})$ , where  $\mathbf{t}$  is a positive penalty weight (the case  $\mathbf{t} = 1$  corresponding to the standard AIC). The solution path  $\widehat{\beta}_n(\mathbf{t})$  which minimizes  $\Lambda_n(\phi, \mathbf{t})$  for all  $\mathbf{t} > 0$  is not more difficult to compute than  $\widehat{\beta}_n(1)$  once one has minimized  $M_n(\phi)$  for the  $2^p$  possible submodels. One easily sees that the solution path is piece-wise constant with multiple solutions at the discontinuities. Multiple solutions for a finite set of penalty weight  $\mathbf{t}$  are also present in the limit contrast (20) with  $J_\infty = J_\infty^{(0)}$ . One can show that there exists almost surely a unique minimizer  $\widehat{u}(\mathbf{t})$  of the limit contrast  $\mathbb{L}(\cdot, \mathbf{t})$  for all  $\mathbf{t} \in \mathbb{T}$  if and only if the closure of the set  $\mathbb{T}$  has zero Lebesgue measure. In the latter case, one also has that  $\widehat{u}$  satisfies Condition (ii) in Theorem 3. This non-uniqueness problem of the minimizer of the limit contrast did not appear in the previous examples because for both  $\ell^1$  and  $\ell^2$  penalties, the limit contrast was strictly convex. This is non-longer true for the  $\ell^0$  penalty so that the convergence  $\sqrt{n}(\widehat{\beta}_n - \beta) \rightsquigarrow \widehat{u}$  cannot hold in a functional sense in this case. Nevertheless, the convergence continues to hold for the  $\ell^0$  penalty in the sense of the finite-dimensional convergence because, for a given finite number of penalty weights  $\mathbf{t}$ , there is a unique minimizer  $\widehat{u}(\mathbf{t})$  of  $\mathbb{L}(\cdot, \mathbf{t})$  almost surely.

## 6. CONCLUSION

We extended the works of Knight and Fu (2000) in several ways by showing that the asymptotic distribution that they exhibited for the penalized least squared continues to hold 1) for the solution path in a functional sense 2) for a wide variety of contrasts extending the least squares case. We provided several examples of interest. An interesting feature of penalized estimation is that the form of the limit distribution of the regularization path only depends on the penalty since for any *standard* contrast, it is given as the path minimizing (20) with  $J_\infty$  only depending on the penalty. The marginal limit distribution is discussed in Knight and Fu (2000) for  $\ell^\gamma$  penalties with  $\gamma > 0$ . As pointed out in this reference, a particular feature of  $\ell^1$  penalty is that the limit distribution is compatible with model selection properties but introduce an additional bias on the non-vanishing components. We have shown that the model selection property is preserved by the  $\ell^0$  penalty, without introducing an additional bias on the non-vanishing components. However, the  $\ell^0$  penalty is much less numerically tractable for a large dimension of the parameter space and the central limit theorem on the solution path only holds in a finite-dimensional sense. This latter result were derived in Section 5 for the AIC in the i.i.d. case. A similar analysis can clearly be carried out for the AIC applied to time series models or for Mallows's  $C_p$  criterion.

## ACKNOWLEDGEMENTS

We would like to thank the referees for their valuable comments.

## REFERENCES

- H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.
- P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 2008. To appear.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.
- F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007. ISSN 1935-7524.
- B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *Annals of Statistics*, 32:407–499, 2004.
- J.-F. Germain. A Two-steps Model Selection Procedure Based on the Regularization Path of a  $L_1$ -Penalized Logistic Likelihood. *Proceedings of SFdS*, June 2007.
- E. Greenshtein and Y. Ritov. Persistency in High Dimensional Linear Predictor-Selector and the Virtue of Over-Parametrization. *Bernoulli*, 10:971–988, 2004.
- S. J. Haberman. Concavity and estimation. *Ann. Statist.*, 17(4):1631–1661, 1989. ISSN 0090-5364.
- A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.
- P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 221–233. Univ. California Press, Berkeley, Calif., 1967.
- J. K. Kim and D. Pollard. Cube root asymptotics. *Ann. Statist.*, 18(1):191–219, 1990. ISSN 0090-5364.
- K. Knight and W. Fu. Asymptotics for LASSO-Type Estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.
- W. Niemi. Asymptotics for  $M$ -estimators defined by convex minimization. *Ann. Statist.*, 20(3):1514–1533, 1992. ISSN 0090-5364.
- M. Y. Park and T. Hastie.  $L_1$ -regularization path algorithm for generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(4):659–677, 2007. ISSN 1369-7412.
- D. Pollard. New Ways to Prove Central Limit Theorems. *Econometric Theory*, 1(3):295–313, December 1985.
- R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.
- R. Tibshirani. Regression Shrinkage and Selection via the LASSO. *J. Royal. Statist. Soc.*, B(58):229–243, 1996.
- A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
- A. W. Van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. With applications to statistics.
- P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, 2006. ISSN 1532-4435.
- H. Zou, T. Hastie, and R. Tibshirani. On the “degrees of freedom” of the lasso. *Ann. Statist.*, 35(5):2173–2192, 2007. ISSN 0090-5364.

**Appendix: detailed proofs.**

*Proof of Theorem 3 in the case (C-2).* Recall that, in the case (C-2), we set  $\Phi = \mathbb{R}^p$  and  $\mathcal{D} = \ell^\infty(\mathbb{T}, \mathbb{R}^p)$ . By Theorem 1.5.4 in Van der Vaart and Wellner [1996], to show that  $\hat{\mathbf{u}}$  admits a version in  $\mathcal{D}$  with  $\hat{\mathbf{u}}_n \rightsquigarrow \hat{\mathbf{u}}$  in  $\mathcal{D}$ , it is sufficient to show that the finite-dimensional distributions of  $\hat{\mathbf{u}}_n$  converge to those of  $\hat{\mathbf{u}}$  and that  $(\hat{\mathbf{u}}_n)$  is asymptotically tight. The convergence of the finite-dimensional distributions follows from the case (C-1) that we already proved. Hence to conclude the proof in the case (C-2), it only remains to show that  $(\hat{\mathbf{u}}_n)$  is asymptotically tight. In the following we show that this uniform tightness is inherited from that of  $(\mathbb{L}_n)$  in  $\ell^\infty(K \times \mathbb{T})$ . Asymptotic tightness follows from an equicontinuity criterion. The proof has now two steps. In Step 1, we construct a metric  $\tilde{\rho}$  on  $\mathbb{T}$  based on a metric  $\rho$  that makes  $\mathbb{L}_n$  asymptotically uniformly equicontinuous. In Step 2 we use the metric  $\tilde{\rho}$  to prove an equicontinuity criterion for  $\hat{\mathbf{u}}_n$ .

**Step 1.** By successively applying Lemma 1.3.8 and Theorem 1.5.7 in Van der Vaart and Wellner [1996], Condition (i) implies that, for any compact set  $K \subset \mathbb{R}^p$ ,  $\mathbb{L}_n$  is asymptotically tight in  $\ell^\infty(K \times \mathbb{T})$  and there exists a semi-metric  $\rho$  on  $K \times \mathbb{T}$  such that  $(K \times \mathbb{T}, \rho)$  is totally bounded and  $\mathbb{L}_n$  is asymptotically uniformly  $\rho$ -equicontinuous in probability. This means that, for any  $\epsilon, \alpha > 0$ , there exists  $\delta > 0$  such that

$$\limsup P^* \left( \sup_{(\mathbf{u}, \mathbf{u}') \in \mathcal{S}_\delta(K)} |\mathbb{L}_n(\mathbf{u}) - \mathbb{L}_n(\mathbf{u}')| > \alpha \right) \leq \epsilon, \quad (40)$$

where

$$\mathcal{S}_\delta(K) = \{((\phi, \mathbf{t}), (\phi', \mathbf{t}')) \in (K \times \mathbb{T})^2 : \rho((\phi, \mathbf{t}), (\phi', \mathbf{t}')) < \delta\}.$$

Clearly, the semi-metric  $\rho$  can be assumed to be bounded and not to depend on the compact set  $K$  without loss of generality; in other words, a bounded semi-metric  $\rho$  can be defined on  $\mathbb{R}^p \times \mathbb{T}$  so that  $(\mathbb{R}^p \times \mathbb{T}, \rho)$  is totally bounded and  $\mathbb{L}_n$  is asymptotically uniformly  $\rho$ -equicontinuous in probability on  $K \times \mathbb{T}$  for any compact set  $K$ . We shall use this semi-metric in the following to show that  $\hat{\mathbf{u}}_n$  is asymptotically uniformly  $\tilde{\rho}$ -equicontinuous in probability, where  $\tilde{\rho}$  is the semi-metric defined on  $\mathbb{T}$  by

$$\tilde{\rho}(\mathbf{t}, \mathbf{t}') = \sup_{\phi \in \mathbb{R}^p} \rho((\phi, \mathbf{t}), (\phi, \mathbf{t}')).$$

By [Van der Vaart and Wellner, 1996, Theorem 1.5.7], the asymptotic uniform  $\tilde{\rho}$ -equicontinuity in probability implies that  $(\hat{\mathbf{u}}_n)$  is asymptotically tight.

**Step 2.** It now remains to show that  $(\hat{\mathbf{u}}_n)$  is asymptotically uniformly  $\tilde{\rho}$ -equicontinuous in probability. Let  $\eta$  and  $\epsilon$  be two arbitrarily small positive numbers. By Conditions (iii) and (iv), we may choose a compact  $K \subset \mathbb{R}^p$  such that

$$P(B) \leq \epsilon \quad \text{and} \quad \limsup P^*(B_n) \leq \epsilon, \quad (41)$$

where

$$B = \{\hat{\mathbf{u}}(\mathbf{t}) \in K \text{ for all } \mathbf{t} \in \mathbb{T}\}^c \quad \text{and} \quad B_n = \{\hat{\mathbf{u}}_n(\mathbf{t}) \in K \text{ for all } \mathbf{t} \in \mathbb{T}\}^c.$$

Using Condition (ii), we may find  $\alpha > 0$  arbitrarily small such that

$$P \left( \inf_{\mathbf{t} \in \mathbb{T}} \left[ \inf \{ \mathbb{L}(\phi, \mathbf{t}) : \phi \in K, \|\phi - \hat{\mathbf{u}}(\mathbf{t})\| \geq \eta/2 \} - \mathbb{L}(\hat{\mathbf{u}}(\mathbf{t}), \mathbf{t}) \right] \leq 4\alpha \right) \leq \epsilon. \quad (42)$$

We further choose  $\delta > 0$  so that Inequality (40) holds, that is

$$\limsup P^*(E_n) \leq \epsilon, \quad (43)$$

where

$$E_n = \left\{ \sup_{(\mathbf{u}, \mathbf{u}') \in \mathcal{S}_\delta(K)} |\mathbb{L}_n(\mathbf{u}) - \mathbb{L}_n(\mathbf{u}')| > \alpha \right\}.$$

Finally, Condition (v) gives that

$$\limsup P^*(C_n) = 0, \quad (44)$$

where

$$C_n = \left\{ \sup_{\mathbf{t} \in \mathbb{T}} \left\{ \mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}), \mathbf{t}) - \inf_{\phi \in \Phi} \mathbb{L}_n(\phi, \mathbf{t}) \right\}_+ > \alpha \right\}.$$

On  $B_n^c$ , we notice that  $((\widehat{\mathbf{u}}_n(\mathbf{t}'), \mathbf{t}), (\widehat{\mathbf{u}}_n(\mathbf{t}'), \mathbf{t}')) \in \mathcal{S}_\delta(K)$  for every  $(\mathbf{t}, \mathbf{t}')$  such that  $\tilde{\rho}(\mathbf{t}, \mathbf{t}') < \delta$ . Hence, on  $B_n^c \cap E_n^c$ , we have

$$\tilde{\rho}(\mathbf{t}, \mathbf{t}') < \delta \Rightarrow \mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}'), \mathbf{t}) \leq \mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}'), \mathbf{t}') + \alpha. \quad (45)$$

Suppose for a moment that we are on the set

$$D_n = \left\{ \sup_{\tilde{\rho}(\mathbf{t}, \mathbf{t}') < \delta} \|\widehat{\mathbf{u}}_n(\mathbf{t}) - \widehat{\mathbf{u}}_n(\mathbf{t}')\| > \eta \right\}.$$

Then we may find  $(\mathbf{t}, \mathbf{t}') \in \mathbb{T}^2$  such that  $\tilde{\rho}(\mathbf{t}, \mathbf{t}') < \delta$  and  $\|\widehat{\mathbf{u}}_n(\mathbf{t}) - \widehat{\mathbf{u}}_n(\mathbf{t}')\| > \eta$ . On  $C_n^c$ , we further have  $\mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}'), \mathbf{t}') \leq \inf_{\phi \in \Phi} \mathbb{L}_n(\phi, \mathbf{t}') + \alpha$ . Intersecting with  $B_n^c \cap E_n^c$  and applying (45), we obtain

$$\mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}'), \mathbf{t}) \leq \inf_{\phi \in \Phi} \mathbb{L}_n(\phi, \mathbf{t}') + 2\alpha \leq \mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}), \mathbf{t}') + 2\alpha \leq \mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}), \mathbf{t}) + 3\alpha,$$

where the last inequality is obtained by exchanging  $\mathbf{t}$  with  $\mathbf{t}'$  in (45). Applying again that we are on  $C_n^c$ , we have  $\mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}), \mathbf{t}) \leq \inf_{\phi \in \Phi} \mathbb{L}_n(\phi, \mathbf{t}) + \alpha$ , and thus, with the last display, we get

$$\max(\mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}), \mathbf{t}), \mathbb{L}_n(\widehat{\mathbf{u}}_n(\mathbf{t}'), \mathbf{t})) \leq \inf_{\phi \in \Phi} \mathbb{L}_n(\phi, \mathbf{t}) + 4\alpha \leq \inf_{\phi \in K} \mathbb{L}_n(\phi, \mathbf{t}) + 4\alpha.$$

Since  $\|\widehat{\mathbf{u}}_n(\mathbf{t}) - \widehat{\mathbf{u}}_n(\mathbf{t}')\| > \eta$  and  $\widehat{\mathbf{u}}_n(\mathbf{t})$  and  $\widehat{\mathbf{u}}_n(\mathbf{t}')$  belong to  $K$  on  $B_n^c$ , we just proved that  $D_n \cap C_n^c \cap B_n^c \cap E_n^c$  is included in

$$F_n = \left\{ \inf_{\mathbf{t} \in \mathbb{T}} \left[ \inf_{(\phi, \phi') \in \mathcal{B}_\eta(K)} \max(\mathbb{L}_n(\phi, \mathbf{t}), \mathbb{L}_n(\phi', \mathbf{t})) - \inf_{\phi \in K} \mathbb{L}_n(\phi, \mathbf{t}) \right] \leq 4\alpha \right\},$$

where

$$\mathcal{B}_\eta(K) = \{(\phi, \phi') \in K^2 : \|\phi - \phi'\| > \eta\}.$$

Using Condition (i) and the continuous mapping Theorem, we have  $\limsup P^*(F_n) \leq P(F)$ , where

$$F = \left\{ \inf_{\mathbf{t} \in \mathbb{T}} \left[ \inf_{(\phi, \phi') \in \mathcal{B}_\eta(K)} \max(\mathbb{L}(\phi, \mathbf{t}), \mathbb{L}(\phi', \mathbf{t})) - \inf_{\phi \in K} \mathbb{L}(\phi, \mathbf{t}) \right] \leq 4\alpha \right\}.$$

Since  $D_n \cap C_n^c \cap B_n^c \cap E_n^c \subset F_n$ , using (41), (43) and (44), we further obtain

$$\limsup P^*(D_n) \leq P(F) + 2\epsilon.$$

Observe that for all  $(\phi, \phi') \in \mathcal{B}_\eta(K)$  and  $\mathbf{t} \in \mathbb{T}$ , we have  $\|\phi - \widehat{\mathbf{u}}(\mathbf{t})\| > \eta/2$  or  $\|\phi' - \widehat{\mathbf{u}}(\mathbf{t})\| > \eta/2$ . Hence, for all  $\mathbf{t} \in \mathbb{T}$ ,

$$\inf_{(\phi, \phi') \in \mathcal{B}_\eta(K)} \max(\mathbb{L}_n(\phi, \mathbf{t}), \mathbb{L}_n(\phi', \mathbf{t})) \geq \inf \{\mathbb{L}(\phi'', \mathbf{t}) : \phi'' \in K, \|\phi'' - \widehat{\mathbf{u}}(\mathbf{t})\| \geq \eta/2\} .$$

Further, by definition of  $B$ , we have on  $B^c$  that for all  $\mathbf{t} \in \mathbb{T}$ ,  $\inf_{\phi \in K} \mathbb{L}(\phi, \mathbf{t}) \leq \mathbb{L}(\widehat{\mathbf{u}}(\mathbf{t}), \mathbf{t})$ . This and the last display show that  $F \cap B^c$  is included in

$$\left\{ \inf_{\mathbf{t} \in \mathbb{T}} \left[ \inf \{\mathbb{L}(\phi, \mathbf{t}) : \phi \in K, \|\phi - \widehat{\mathbf{u}}(\mathbf{t})\| \geq \eta/2\} - \mathbb{L}(\widehat{\mathbf{u}}(\mathbf{t}), \mathbf{t}) \right] \leq 4\alpha \right\} ,$$

which, by (42), has probability at most  $\epsilon$  for our choice of  $\alpha$ . Since  $K$  has been chosen so that  $P(B) \leq \epsilon$ , we finally get

$$\limsup P^*(D_n) \leq 4\epsilon .$$

This exactly says that  $(\widehat{\mathbf{u}}_n)$  is asymptotically uniformly  $\tilde{\rho}$ -equicontinuous in probability and the proof is achieved.  $\square$

*Proof of Theorem 5.* Let  $\epsilon > 0$  and denote by  $B' = \{\phi : \|\phi - \beta\| \leq 2\epsilon\}$  and  $B = \{\phi : \|\phi - \beta\| \leq \epsilon\}$  the balls centered at  $\beta$  with radii  $2\epsilon$  and  $\epsilon$ . We choose  $\epsilon$  small enough so that  $B' \subseteq V$ . We first show that Assumption 3 holds for  $M$  defined on  $\Phi$  by

$$M(\phi) = \begin{cases} \Delta(\phi) & \text{if } \phi \in B , \\ \Delta(\beta) + \alpha/2 & \text{otherwise ,} \end{cases} \quad (46)$$

where

$$\alpha = \inf_{\phi \in B' \setminus B} \Delta(\phi) - \Delta(\beta) > 0 . \quad (47)$$

The positiveness of  $\alpha$  follows from the strict convexity of  $\Delta$  and Assumption 4-(ii). Assumption 3-(ii) follows from Assumption 4-(i). Assumption 3-(iii) follows from the strict convexity of  $\Delta$ , Assumption 4-(ii) and the definition of  $M$  in (46). It only remains to prove that Assumption 3-(i) holds. By [Rockafellar, 1970, Theorem 10.8 ] and arguing as in the proof of Lemma 3 in Niemiro [1992] for getting the result in the sense of the convergence in probability, the pointwise convergence in Assumption 4-(i) implies the uniform convergence on the compact set  $B'$ , that is,

$$\sup_{\phi \in B'} |M_n(\phi) - \Delta(\phi)| \xrightarrow{P} 0 . \quad (48)$$

Let  $\Omega'$  be a probability 1 set on which  $M_n$  is convex and define

$$A_n = \left\{ \sup_{\phi \in B'} |M_n(\phi) - \Delta(\phi)| \leq \alpha/4 \right\} \cap \Omega' .$$

The set  $A_n$  is measurable since  $M_n$  and  $\Delta$  are convex on  $\Phi$  and thus the sup can be replaced by a sup on a countable dense subset of  $B'$  without changing the definition of  $A_n$ . Let  $\omega \in A_n$ . For all  $\phi \in B' \setminus B$  and  $\mathbf{t} \in [0, L]$ , we have  $M_n(\omega, \phi) \geq \Delta(\phi) - \alpha/4$ ,  $\Delta(\phi) \geq \Delta(\beta) + \alpha$ , and, since  $\beta \in B'$ ,  $\Delta(\beta) \geq M_n(\omega, \beta) - \alpha/4$ . Hence

$$\inf_{\phi \in B' \setminus B} M_n(\omega, \phi) \geq M_n(\omega, \beta) + \alpha/2 .$$

By convexity of the function  $M_n(\omega, \cdot)$  and of the set  $\Phi$ , the last display implies that

$$\inf_{\phi \in \Phi \setminus B} M_n(\omega, \phi) \geq M_n(\omega, \beta) + \alpha/2 .$$

For all  $\omega \in A_n$ , using the definition of  $M$  in (46), we thus have, for all  $\phi \in \Phi \setminus B$ ,

$$\{M(\phi) - M_n(\omega, \phi)\}_+ = \{\Delta(\beta) + \alpha/2 - M_n(\omega, \phi)\}_+ \leq |\Delta(\beta) + M_n(\omega, \beta)| .$$

Using this with (48) and  $P(A_n) \rightarrow 1$ , we get Assumption 3-(i). We conclude that Assumption 3 holds and we obtain Assertion (a) as an application of Theorem 4.

Next we show Assertion (b) and thus assume that  $J_n$  is strictly convex. The proof of Assertion (c) is similar and thus omitted. We set

$$L_n = \frac{\alpha}{4J_n(\beta)} ,$$

so that  $L_n \rightarrow \infty$  by assumption on  $J_n(\beta)$  and  $\mathbf{t}J_n(\beta) \leq \alpha/4$  for all  $\mathbf{t} \leq L_n$ . Let  $\omega \in A_n$ . Then, for all  $\phi \in B' \setminus B$  and  $\mathbf{t} \in [0, L_n]$ , using that  $\Lambda_n(\omega, \phi, \mathbf{t}) \geq M_n(\omega, \phi)$  and  $M_n(\omega, \beta) = \Lambda_n(\omega, \beta, \mathbf{t}) - \mathbf{t}J_n(\beta) \geq \Lambda_n(\omega, \beta) - \alpha/4$ , we obtain

$$\inf_{\mathbf{t} \in [0, L_n]} \inf_{\phi \in B' \setminus B} \Lambda_n(\omega, \phi, \mathbf{t}) \geq \Lambda_n(\omega, \beta, \mathbf{t}) + \alpha/4 .$$

Since  $J_n$  is strictly convex, so is the function  $\Lambda_n(\omega, \cdot, \mathbf{t})$  for  $\mathbf{t} > 0$ . By convexity of the set  $\Phi$ , the previous display implies that for all  $\mathbf{t} \in [0, L_n]$ , the minimum of  $\Lambda_n(\omega, \phi, \mathbf{t})$  on  $\phi \in \Phi$  is attained within  $B$ . By strict convexity of  $J_n$ , this minimum is unique for  $\mathbf{t} > 0$  and we let  $\hat{\beta}_n(\omega, \mathbf{t})$  be this unique minimum for  $\mathbf{t} \in (0, L_n]$ . For  $\omega \in A_n^c$  or  $\mathbf{t} > L_n$ , we define  $\hat{\beta}_n(\omega, \mathbf{t}) = \phi_0$ , where  $\phi_0$  is any fixed point of  $\Phi$ . As for  $\mathbf{t} = 0$  and  $\omega \in A_n$ , we define

$$\hat{\beta}_n(\omega, 0) = \liminf_{\mathbf{t} \downarrow 0} \hat{\beta}_n(\omega, \mathbf{t}) \in B ,$$

where the  $\liminf$  is defined component-wise in a given coordinate system of the Euclidean space containing  $\Phi$ . Since the minimum of  $\Lambda_n(\omega, \phi, \mathbf{t})$  on  $\phi \in \Phi$  is attained within the compact set  $B$ , by continuity of  $J_n(\phi)$  and  $M_n(\omega, \phi)$  in  $\phi$ ,  $\hat{\beta}_n(\omega, 0)$  is a minimizer of  $\Lambda_n(\omega, \phi, 0)$  on  $\phi \in \Phi$ . Thus, we have defined a r.v.  $\hat{\beta}_n(\cdot, \mathbf{t})$  for any  $\mathbf{t} \geq 0$ , for which Property (b1) holds.

To conclude the proof, we show that Property (b2) holds. The continuity on  $(L_n, \infty)$  for  $\omega \in A_n$  and on  $\mathbb{R}_+$  for  $\omega \in A_n^c$  directly follows from the definition of  $\hat{\beta}_n(\omega, \mathbf{t})$ . Let us now prove that  $\hat{\beta}_n(\omega, \cdot)$  is continuous on  $(0, L_n]$  for all  $\omega \in A_n$ . Since  $J_n$  is convex, it is bounded on  $B$  and since  $\hat{\beta}_n(\omega, \mathbf{t}) \in B$ , we have  $\sup_{\mathbf{t} \in (0, L_n]} J_n(\hat{\beta}_n(\omega, \mathbf{t})) \leq \sup J_n(B) < \infty$ . Let  $\mathbf{t}$  and  $\mathbf{t}_0$  be in  $(0, L_n]$ . We have

$$\begin{aligned} \Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}), \mathbf{t}_0) &\leq \Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}), \mathbf{t}) + |\mathbf{t}_0 - \mathbf{t}| \sup J_n(B) \\ &\leq \Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}_0), \mathbf{t}) + |\mathbf{t}_0 - \mathbf{t}| \sup J_n(B) \\ &\leq \Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}_0), \mathbf{t}_0) + 2|\mathbf{t}_0 - \mathbf{t}| \sup J_n(B) . \end{aligned}$$

Since  $\Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}_0), \mathbf{t}_0) \leq \Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}), \mathbf{t}_0)$ , we get that  $\Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}), \mathbf{t}_0) \rightarrow \Lambda_n(\hat{\beta}_n(\omega, \mathbf{t}_0), \mathbf{t}_0)$  as  $\mathbf{t} \rightarrow \mathbf{t}_0$ . Since, by strict convexity of  $\Lambda_n$ ,  $\hat{\beta}_n(\omega, \mathbf{t}_0)$  is an isolated minimum of  $\Lambda_n(\cdot, \mathbf{t}_0)$ , this implies that  $\hat{\beta}_n(\omega, \mathbf{t}) \rightarrow \hat{\beta}_n(\omega, \mathbf{t}_0)$  as  $\mathbf{t} \rightarrow \mathbf{t}_0$ . The continuity of  $\hat{\beta}_n(\omega, \cdot)$  on  $(0, L_n]$  follows and the proof is achieved.  $\square$

*Proof of Proposition 1.* Let  $\epsilon > 0$  and define

$$\alpha = \inf_{\mathbf{t} \in \mathbb{T}} \left[ \inf_{d(\phi, \beta) \geq \epsilon/2} \Lambda(\phi; \mathbf{t}) - \Lambda(\beta(\mathbf{t}), \mathbf{t}) \right].$$

By (iii), we have  $\alpha > 0$ . Denote

$$A_n = \left\{ \sup_{\mathbf{t} \in \mathbb{T}} d(\widehat{\beta}_n(\mathbf{t}), \beta(\mathbf{t})) \geq \epsilon \right\} \subseteq \Omega.$$

For all  $\omega \in A_n$ , there exists  $\mathbf{t} \in \mathbb{T}$  such that  $d(\widehat{\beta}_n(\omega, \mathbf{t}), \beta(\mathbf{t})) \geq \epsilon/2$ , and thus for which  $\Lambda(\widehat{\beta}_n(\omega, \mathbf{t}), \mathbf{t}) - \Lambda(\beta(\mathbf{t}), \mathbf{t}) \geq \alpha$ . Hence, for all  $\omega \in A_n$ , we have

$$\sup_{\mathbf{t} \in \mathbb{T}} \left[ \Lambda(\widehat{\beta}_n(\omega, \mathbf{t}), \mathbf{t}) - \Lambda(\beta(\mathbf{t}), \mathbf{t}) \right] \geq \alpha.$$

Now we write, for any  $\mathbf{t}_0 \in \mathbb{T}$ ,

$$\begin{aligned} \Lambda(\widehat{\beta}_n(\mathbf{t}_0), \mathbf{t}_0) - \Lambda(\beta(\mathbf{t}_0), \mathbf{t}_0) &= \left\{ \Lambda(\widehat{\beta}_n(\mathbf{t}_0), \mathbf{t}_0) - \Lambda_n(\widehat{\beta}_n(\mathbf{t}_0), \mathbf{t}_0) \right\} \\ &\quad + \left\{ \Lambda_n(\widehat{\beta}_n(\mathbf{t}_0), \mathbf{t}_0) - \Lambda_n(\beta(\mathbf{t}_0), \mathbf{t}_0) \right\} + \left\{ \Lambda_n(\beta(\mathbf{t}_0), \mathbf{t}_0) - \Lambda(\beta(\mathbf{t}_0), \mathbf{t}_0) \right\} \\ &\leq \sup_{\phi \in \Phi} \sup_{\mathbf{t} \in \mathbb{T}} \left\{ \Lambda(\phi; \mathbf{t}) - \Lambda_n(\phi, \mathbf{t}) \right\}_+ \\ &\quad + \sup_{\mathbf{t} \in \mathbb{T}} \left\{ \Lambda_n(\widehat{\beta}_n(\mathbf{t}), \mathbf{t}) - \Lambda_n(\beta(\mathbf{t}), \mathbf{t}) \right\}_+ + \sup_{\mathbf{t} \in \mathbb{T}} \left| \Lambda_n(\beta(\mathbf{t}), \mathbf{t}) - \Lambda(\beta(\mathbf{t}), \mathbf{t}) \right|. \end{aligned}$$

Taking the sup in  $\mathbf{t}_0 \in \mathbb{T}$  we obtain that  $A_n \subseteq A_n^{(1)} \cup A_n^{(2)} \cup A_n^{(3)}$ , where  $A_n^{(1)} = \{ \sup_{\phi \in \Phi} \sup_{\mathbf{t} \in \mathbb{T}} \left\{ \Lambda(\phi; \mathbf{t}) - \Lambda_n(\phi, \mathbf{t}) \right\}_+ \geq \alpha/3 \}$ , and where  $A_n^{(2)}$  and  $A_n^{(3)}$  are defined accordingly by using the last 2 lines of the last display. Applying  $P^*(A_n) \leq P^*(A_n^{(1)}) + P^*(A_n^{(2)}) + P^*(A_n^{(3)})$ , (i), (ii) and (iv), we thus get (26), which achieves the proof.  $\square$

*Proof of Theorem 4.* We apply Proposition 1 with  $\Lambda_n$  defined by (1),  $\Lambda(\phi, \mathbf{t}) = M(\phi)$  and  $\beta(\mathbf{t}) = \beta$  for all  $\mathbf{t}$ . Let us check the conditions in Proposition 1. Since  $J_n$  is non-negative,

$$\left\{ \Lambda(\phi; \mathbf{t}) - \Lambda_n(\phi; \mathbf{t}) \right\}_+ \leq \left\{ M(\phi) - M_n(\phi) \right\}_+,$$

and Condition (i) follows from Assumption 3-(i). Condition (ii) follows from Assumption 3-(ii) and  $J_n(\beta) \rightarrow 0$ . Conditions (iii) and (iv) directly follow from Assumption 3-(iii) and Eq. (12), respectively. Hence (13) follows from (26).  $\square$

*Proof of Proposition 2.* Denote the left-hand side of (32) by  $U_n$  and the left-hand side of (29) by  $V_n$ . Let  $\delta > 1$  and define  $A_n = \{U_n > \delta\}$ . Then for all  $\omega \in A_n$ , we have

$$\sup_{\mathbf{t} \in \mathbb{T}} \left| G_n(\widehat{\beta}_n(\mathbf{t}), \mathbf{t}) \right| \leq 2n^{-1} \delta^{-1} U_n^2 V_n. \quad (49)$$

By (iii), using the assumed uniform  $P^*$ -consistency (26), there exist non-negative random sequences  $w_n$  and  $W_n$  such that  $w_n = o_{P^*}(1)$ ,  $W_n = O_{P^*}(1)$  and

$$\sqrt{n} \sup_{\mathbf{t} \in \mathbb{T}} \left| R_n(\widehat{\beta}_n(\mathbf{t}), \mathbf{t}) \right| \leq (U_n w_n + W_n),$$

hence, for all  $\omega \in A_n$ ,

$$n \sup_{\mathbf{t} \in \mathbb{T}} \left\{ d(\widehat{\beta}_n(\mathbf{t}), \beta(\mathbf{t})) \left| R_n(\widehat{\beta}_n(\mathbf{t}), \mathbf{t}) \right| \right\} \leq U_n (U_n w_n + W_n) \leq U_n^2 (w_n + W_n/\delta).$$

Denote the left-hand side of (27) by  $S_n$ . The last display, (49) and (28) imply that, for all  $\omega \in A_n$  and all  $\mathbf{t} \in \mathbb{T}$ ,

$$H(\widehat{\beta}_n(\mathbf{t}), \mathbf{t}) \leq S_n + U_n^2 n^{-1} \{2\delta^{-1} V_n + w_n + W_n/\delta\} .$$

Define  $B_n = \{\sup_{\mathbf{t} \in \mathbb{T}} d(\widehat{\beta}_n(\mathbf{t}), \beta(\mathbf{t})) > \epsilon\}$  where  $\epsilon$  is the positive number in Condition (ii) and denote the left-hand side of (30) by  $\alpha$ , which is positive. Then, for all  $\omega \in B_n^c$ ,  $\alpha U_n^2 \leq n \sup_{\mathbf{t} \in \mathbb{T}} H(\widehat{\beta}_n(\mathbf{t}), \mathbf{t})$ , and, using the previous display, if moreover  $\omega \in A_n$ ,

$$\alpha U_n^2 \leq n S_n + U_n^2 \{2\delta^{-1} V_n + w_n + W_n/\delta\} .$$

Using that  $P^*(B_n) \rightarrow 0$ ,  $nS_n = O_{P^*}(1)$ ,  $V_n = O_{P^*}(1)$ ,  $w_n = o_{P^*}(1)$  and  $W_n = O_{P^*}(1)$ , we easily get that  $\limsup P^*(A_n)$  can be made arbitrarily small by taking  $\delta$  large enough. Hence (32) holds.  $\square$

*Proof of Theorem 7.* Let us define  $\widehat{\mathbf{u}}_n = \sqrt{n}(\widehat{\beta}_n - \beta)$  and

$$\mathbb{L}_n(\phi, \mathbf{t}) = n \left\{ \Lambda_n(\beta(\mathbf{t}) + n^{-1/2}\phi, \mathbf{t}) - \Lambda_n(\beta(\mathbf{t}), \mathbf{t}) \right\} . \quad (50)$$

We will apply Theorem 3 with these definitions (in the case (C-2)) and thus now proceed in checking the conditions of Theorem 3 successively. Let  $K$  be a compact subset of  $\Phi$ . Using (28), (37) and (50), we get

$$\mathbb{L}_n(\phi, \mathbf{t}) = \widehat{G}_n(\phi, \mathbf{t}) + nH(\beta(\mathbf{t}) + n^{-1/2}\phi, \mathbf{t}) + \sqrt{n}\|\phi\|R_n(\beta(\mathbf{t}) + n^{-1/2}\phi, \mathbf{t}) .$$

Observe that by (34) and (35), as functions of  $(\phi, \mathbf{t})$ ,

$$nH(\beta(\mathbf{t}) + n^{-1/2}\phi, \mathbf{t}) \rightarrow \phi^T \Gamma(\mathbf{t}) \phi \quad \text{in } \ell^\infty(K \times \mathbb{T}, \mathbb{R}^p) .$$

Applying (36), we obtain

$$\sup_{(\phi, \mathbf{t}) \in K \times \mathbb{T}} \sqrt{n}\|\phi\| \left| R_n(\beta(\mathbf{t}) + n^{-1/2}\phi, \mathbf{t}) \right| = o_{P^*}(1) .$$

Hence using that  $\widehat{G}_n \rightsquigarrow G$  in  $\ell^\infty(K \times \mathbb{T}, \mathbb{R}^p)$ , the three last displays yield  $\mathbb{L}_n \rightsquigarrow \mathbb{L}$  in  $\ell^\infty(K \times \mathbb{T}, \mathbb{R}^p)$ . Since  $G$  is tight in  $\ell^\infty(K \times \mathbb{T}, \mathbb{R}^p)$  by assumption,  $\mathbb{L}$  also is and thus Condition (i) holds. Conditions (ii) and (iii) hold by assumption. Applying Proposition 2, we obtain (32) and thus Condition (iv) holds. Using (33) with the above definitions, we get that Condition (v) holds.  $\square$

*Proof of Theorem 6.* We shall apply Theorem 7 for  $\Lambda_n$  given by (1) and with  $\beta(\mathbf{t}) = \beta$  for all  $\mathbf{t} \in \mathbb{T}$ . Let us check that the assumptions of this theorem hold in this context. Condition (33) and the uniform  $P^*$ -consistency (26) hold by assumption. The decomposition (28) holds with

$$G_n(\phi, \mathbf{t}) = (\phi - \beta)^T P_n \Delta + \mathbf{t} (J_n(\phi) - J_n(\beta)) \mathbb{1}(\|\phi - \beta\| \leq 1) ,$$

$$H(\phi, \mathbf{t}) = Pg(\cdot, \phi) - Pg(\cdot, \beta) - (\phi - \beta)^T P \Delta ,$$

$$R_n(\phi, \mathbf{t}) = n^{-1/2} \nu_n r(\cdot, \phi) + \mathbf{t} \|\phi - \beta\|^{-1} (J_n(\phi) - J_n(\beta)) \mathbb{1}(\|\phi - \beta\| > 1) .$$

Using (P-1) and (P-3), we have  $\sum_{k=1}^n \Delta(\xi_k) = O_P(n^{1/2})$  and, using (18), we get that Condition (i) in Theorem 7 holds. Observe that  $H(\phi, \mathbf{t})$  does not depend on  $\mathbf{t}$  and, by (P-3), we have

$$H(\phi, \mathbf{t}) = M(\phi) - M(\beta) .$$

Integrating  $x$  with respect to  $P$  in (16) and using (P-4), we get that the first derivative of  $M$  at  $\beta$  is zero and, by (P-2),

$$H(\phi, \mathbf{t}) = (\phi - \beta)^T \Gamma(\phi - \beta) + o(\|\phi - \beta\|^2) .$$

Hence Condition (ii) in Theorem 7 holds.

We have, for any sequence of positive r.v.  $(r_n)$  such that  $r_n \xrightarrow{P} 0$ ,

$$\begin{aligned} \sup_{\|\phi - \beta\| \leq r_n} \left\{ \left| n^{-1/2} \nu_n r(\cdot, \phi) \right| \right\} &\leq \frac{1 + \sqrt{n} r_n}{\sqrt{n}} \sup_{\|\phi - \beta\| \leq r_n} \left\{ \frac{|\nu_n r(\cdot, \phi)|}{1 + \sqrt{n} \|\phi - \beta\|} \right\} \\ &= o_P(n^{-1/2}) + o_P(r_n) , \end{aligned}$$

where the last equality follows from (P-4). Observing that, for  $\|\phi - \beta\| \leq r_n$  and  $r_n \leq 1$  the second term defining  $R_n$  vanishes, we obtain Condition (36) in Theorem 7.

Defining  $\widehat{G}_n$  as in (37) gives

$$\widehat{G}_n(\phi, \mathbf{t}) = \phi^T (\sqrt{n} P_n \Delta) + \mathbf{t} \left[ n J_n(\beta + n^{-1/2} \phi) - n J_n(\beta) \right] .$$

Using (P-1) and (P-3), we have that  $\sqrt{n} P_n \Delta$  converge in distribution to  $W$  and, by (19), for any compact  $K \subset \mathbb{R}^p$   $\widehat{G}_n \rightsquigarrow G$  in  $\ell^\infty(K \times \mathbb{T}, \mathbb{R}^p)$ , where  $G(\phi, \mathbf{t}) = \phi^T W + \mathbf{t} J_\infty(\phi)$ . This definition of  $G$  and (38) gives (20). Hence Theorem 7 yields (21).  $\square$

*Proof of Lemma 1.* We have, for all  $\phi \in \mathbb{R}^p$ ,

$$\left| \sum_{k=1}^p |\phi_k|^\gamma - \sum_{k=1}^p |\beta_k|^\gamma \right| \leq C (\|\phi - \beta\|^\gamma + \|\phi - \beta\|) ,$$

where  $C$  only depends on  $\beta$  and  $\gamma > 0$ . The bound (23) follows directly for  $\gamma \geq 1$ . For  $\gamma < 1$ , one obtains

$$n \left| J_n^{(\gamma)}(\phi) - J_n^{(\gamma)}(\beta) \right| \leq C' \left( (\sqrt{n} \|\phi - \beta\|)^\gamma + n^{\gamma/2} \|\phi - \beta\| \right) ,$$

and (23) follows by observing that  $a^\gamma \leq 1 + a$  for  $a \geq 0$ , and  $n^{\gamma/2} \leq n^{1/2}$ .

Relation (24) is easily obtained by using the Taylor expansion, valid for  $x \neq 0$ ,  $|x+y|^\gamma = |x|^\gamma + \gamma|x|^{\gamma-1} \text{sgn}(x)y + O(y^2)$ , which concludes the proof.  $\square$

*Proof of Theorem 1.* As  $\phi \mapsto M_n(\phi) = \frac{1}{n} \sum_{k=1}^n (y_k - \mathbf{x}_k^T \phi)^2$  is a convex function, we apply Theorem 5. In fact, by Assumption 1-(i),  $M_n$  is strictly convex for  $n$  large enough, and hence the more precise Assertion (c) applies. We now show that Assumption 4-(i) holds.

$$M_n(\phi) - M_n(\beta) = (\phi - \beta)^T C_n(\phi - \beta) - \frac{2}{n} \varepsilon_n^T \mathbf{X}_n(\phi - \beta) \quad (51)$$

where  $\varepsilon_n = Y_n - \mathbf{X}_n \beta$ . Since

$$\mathbb{E} \|\mathbf{X}_n^T \varepsilon_n\|^2 = \mathbb{E} [\text{Tr}(\varepsilon_n^T \mathbf{X}_n \mathbf{X}_n^T \varepsilon_n)] = \text{Tr}[\mathbf{X}_n \mathbf{X}_n^T] = O(n) ,$$

by Assumption 1-(i), it comes  $-\frac{2}{n} \varepsilon_n^T \mathbf{X}_n(\phi - \beta) = O_P(n^{-1/2})$ . And furthermore, by Assumption 1-(i) :

$$M_n(\phi) - M_n(\beta) \rightarrow_P (\phi - \beta)^T C(\phi - \beta) = \Delta(\phi) .$$

Since  $C$  is positive-definite,  $\Delta$  is strictly convex and Assumption 4-(ii) holds. By definition of  $\widehat{\beta}_n(\mathbf{t})$ , (12) holds. Finally, the condition  $J_n(\beta) \rightarrow 0$  holds, as the penalty is defined

by  $J_n(\beta) = \lambda_n \|\beta\|_1$ , with  $\|\cdot\|_1$  denoting the  $\ell^1$  norm. The uniform consistency on every compact set follows as an application of Theorem 5.  $\square$

*Proof of Theorem 2.* We apply Theorem 7 with  $\mathbb{T}$  a compact subset of  $\mathbb{R}_+$ . By definition of  $\widehat{\beta}_n(\mathbf{t})$ , condition (33) holds. We just obtained uniform consistency in Theorem 1. Using (51), we have the decomposition (28) of  $\Lambda_n(\phi, \mathbf{t})$ , with

$$G_n(\phi, \mathbf{t}) = -2n^{-1/2}U_n^T(\phi - \beta) + \mathbf{t}\lambda_n(\|\phi\|_1 - \|\beta\|_1),$$

$$H(\phi, \mathbf{t}) = (\phi - \beta)^T C(\phi - \beta) \text{ and } R_n(\phi, \mathbf{t}) = \|\phi - \beta\|^{-1}(\phi - \beta)^T(C_n - C)(\phi - \beta),$$

where  $U_n = n^{-1/2}\mathbf{X}_n^T \varepsilon_n$  and  $\lambda_n = n^{-1/2}$ , by Assumption 2-(iii).

The sequence  $\{U_n\}$  converges in distribution to  $U \sim \mathcal{N}(0, \sigma^2 C)$  by the Lindeberg-Feller theorem and Assumption 2. We have, for all  $\phi \in \mathbb{R}^p$  and  $\mathbf{t} \in \mathbb{T}$ ,  $n|G_n(\phi, \mathbf{t})| \leq \sqrt{n}U_n\|\phi - \beta\| + \mathbf{t}\sqrt{n}\|\phi\|_1 - \|\beta\|_1 \leq \|\phi - \beta\|(O_P(\sqrt{n}) + c\sqrt{n})$ , where  $c$  is a positive constant. Hence  $G_n$  satisfies (29).

Conditions (34) and (35) on  $H$  are immediately verified by taking  $\Gamma(\mathbf{t}) = C$ , for all  $\mathbf{t} \in \mathbb{T}$  and using Assumption 2-(i).

Observe that  $|R_n(\phi, \mathbf{t})| \leq \rho(C_n - C)\|\phi - \beta\|$  where  $\rho(C_n - C)$  is the spectral radius of  $(C_n - C)$ . Since  $C_n \xrightarrow{P} C$ ,  $\rho(C_n - C) = o_P(1)$  and  $\sup\{R_n(\phi, \mathbf{t}), \phi \in \Phi, \|\phi - \beta\| \leq r_n\} = o_P(r_n)$ . Condition (36) on  $R_n$  follows.

As in (37), we define

$$\widehat{G}_n(\phi, \mathbf{t}) = nG_n(\beta + n^{-1/2}\phi, \mathbf{t}) = -2U_n^T\phi + \mathbf{t}n^{1/2}\sum_{j=1}^p\left\{\left|\beta_j + n^{-1/2}\phi_j\right| - |\beta_j|\right\}.$$

For any compact  $K \subseteq \mathbb{R}^p$ , let  $f$  map  $u \in \mathbb{R}^p$  to  $f[u] \in \ell^\infty(K \times \mathbb{T})$ , defined by  $f[u](\phi, \mathbf{t}) = u^T\phi$ . The map  $f$  is continuous and by the continuous mapping theorem,  $f(U_n)$  converges to  $f(U)$  in  $\ell^\infty(K \times \mathbb{T})$ . From this and (24) with  $\gamma = 1$ , it follows that  $\widehat{G}_n$  converges to  $G$  in  $\ell^\infty(K \times \mathbb{T})$ , where

$$G(\phi, \mathbf{t}) = -2U^T\phi + \mathbf{t}\sum_{j=1}^p\left\{\phi_j \operatorname{sgn}(\beta_j)\mathbb{1}_{\{\beta_j \neq 0\}} + |\phi_j|\mathbb{1}_{\{\beta_j = 0\}}\right\}.$$

By Assumption 1-(i) one has  $\mathbb{L}(\phi, \mathbf{t}) \geq c_1\|\phi\|^2 + c_2\|\phi\|$  for all  $\phi \in \mathbb{R}^p$  and  $\mathbf{t} \in \mathbb{T}$ , with  $c_1 > 0$  and  $c_2$  a finite random variable. Since  $\mathbb{L}(0, \mathbf{t}) = 0$ , we get  $0 \geq \mathbb{L}(\widehat{u}(\mathbf{t}), \mathbf{t}) \geq c_1\|\widehat{u}(\mathbf{t})\|^2 + c_2\|\widehat{u}(\mathbf{t})\|$  thus  $\widehat{u}(\mathbf{t}) \leq -\frac{c_2}{c_1}$ . Condition (ii) of Theorem 3 follows immediately and so does Condition (iii) of Theorem 3, observing that  $\mathbb{L}(\phi, \mathbf{t})$  is continuous in  $(\phi, \mathbf{t})$  and strictly convex in  $\phi$ . The convergence (6) follows as an application of Theorem 7.  $\square$

RENAULT DREAM-DTAA, TECHNOCENTRE GUYANCOURT, 1, AVENUE DU GOLF, 78288 GUYANCOURT, FRANCE.

*E-mail address:* jean-francois.germain@renault.com

INSTITUT TELECOM, TELECOM PARISTECH, LTCI CNRS, 46, RUE BARRAULT, 75634 PARIS CEDEX 13, FRANCE

*E-mail address:* roueff@telecom-paristech.fr