Uniform Convergence in Extended Probability of Sub-Gradients of Convex Functions

Gordon C.R. Kemp

Department of Economics, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, United Kingdom.

Abstract

It is well known that if a sequence of stochastic convex functions on $\mathbb{R}^d$ converges in probability point-wise to some non-stochastic function then the limit function is convex and the convergence is uniform on compact sets; see Andersen and Gill (1982) and Pollard (1991). In the present paper, I establish that if the limiting function is differentiable then any sequence of measurable sub-gradients of the stochastic convex functions converges in extended probability to the gradient of the limit function uniformly on compact sets.

JEL Classification: C10.

Keywords: Convex functions, sub-gradients, convergence in probability, extended probability measure, uniform convergence

1. Introduction

Optimization based estimation and inference are very widely used in statistics and econometrics and there is a well-developed theory of their asymptotic properties; see, for example, Hayashi (2000). A key element in this theory is the uniform convergence in an appropriate sense of certain sequences of stochastic functions and consequently much attention has been paid to establishing sufficient conditions for such uniform convergence. One such set of conditions is given by Theorem II.1 of Andersen and Gill (1982) which shows that point-wise convergence in probability of stochastic convex functions implies uniform convergence on compact subsets, thus providing a stochastic version of Theorem 10.8 of Rockafellar (1970). Pollard (1991) provides a separate proof of this uniform stochastic convergence result and then uses this to establish root-n consistency and asymptotic normality of the least absolute deviations (LAD) estimator under various sets of conditions on the regressors and error terms in a linear regression model.

Here I show that if a sequence of stochastic convex functions on $\mathbb{R}^d$ converges in probability point-wise to a some non-stochastic differentiable function then any sequence of measurable sub-gradients of the stochastic convex functions converges in extended probability to the gradient of the limit function uniformly on compact sets, where the extension of the underlying probability measure is to the universal completion of the underlying σ-algebra. The proof of this result does not involve working explicitly with the functional forms of the sequence of measurable sub-gradients but instead exploits the basic characterization of sub-gradients of convex functions. I then use this result to establish the validity of a uniform local stochastic expansion of a sub-gradient given a point-wise local stochastic expansion of the objective function.

The layout of the paper is as follows. In Section 2, I present the framework and main result on uniform convergence in probability of measurable sub-gradients of stochastic convex functions. In Section 3, I show how the main result can be applied to establish the validity of an uniform local stochastic expansion of the sub-gradient of a convex objective function. Section 4 contains concluding remarks. The proof of the main result is contained in an appendix.

2. Framework and Main Result

Throughout I assume that there is an underlying probability space $(\Omega, \mathcal{F}, P)$. The completion $\mathcal{F}^{\mu}$ of $\mathcal{F}$ with respect to a probability measure μ on $(\Omega, \mathcal{F})$ is defined as the σ-algebra generated by the elements of $\mathcal{F}$ and the μ-null subsets of Ω, where A is a μ-null subset of Ω if there exists $B \in \mathcal{F}$ such that $A \subseteq B$ and $\mu(B) = 0$. The universal completion $\mathcal{F}^{u}$ of $\mathcal{F}$ is then given as $\bigcap \{\mathcal{F}^{\mu} | \mu \text{ is a probability measure on } (\Omega, \mathcal{F})\}$ and is also a σ-algebra on Ω. The extension P^{u} of P to the universal completion $\mathcal{F}^{u}$ of $\mathcal{F}$ is unique and necessarily satisfies $P^{u}(E) = P(E)$ for all $E \in \mathcal{F}$; see Stinchcombe and White (1992, pp. 498-499).

The parameter space of interest, Θ, is a non-empty open convex subset of $\mathbb{R}^{d}$ for some finite d. $\{\lambda_n : \Theta \times \Omega \to \mathbb{R}\}_{n=1}^{\infty}$ is a sequence of functions such that for each $n \in \mathbb{N}$ and $\omega \in \Omega$, $\lambda_n(\cdot; \omega)$ is a proper convex function on Θ with effective domain equal to Θ. Andersen and Gill (1982, Theorem II.1) establishes that if there is a non-random function $\lambda_0(\cdot) : \Theta \to \mathbb{R}$ such that $\lambda_n(\theta, \cdot)$ converges in probability to $\lambda_0(\theta)$ point-wise for each θ in Θ then $\lambda_0(\cdot)$ is a finite convex function on Θ and, furthermore, the convergence is uniform on any non-empty compact subset Γ of Θ.

If $\lambda(\cdot)$ is a finite convex function on Θ with effective domain equal to Θ then for every $\theta_0 \in \Theta$ there exists at least one affine function $y(\cdot) : \mathbb{R}^{d} \to \mathbb{R}$ such that $y(\theta) \leq \lambda(\theta)$ for all $\theta \in \Theta$ with equality if $\theta = \theta_0$. Such an affine function can be expressed in the form

$$y(\theta) = \lambda(\theta_0) + \langle s(\theta_0), (\theta - \theta_0) \rangle, \quad \forall \theta \in \mathbb{R}^{d}, \tag{1}$$

for some $s(\theta_0) \in \mathbb{R}^{d}$, where $\langle a, b \rangle = \sum_{i=1}^{d} a_i b_i$ for any $(d \times 1)$ real-valued vectors a and b. The vector $s(\theta_0)$ is then termed a sub-gradient of $\lambda(\theta)$ at $\theta = \theta_0$. If $\lambda(\theta)$ is differentiable with respect to θ at $\theta = \theta_0$ then this affine function is unique with sub-gradient equal to the gradient $\nabla \lambda(\theta_0)$ of $\lambda(\theta)$ at $\theta = \theta_0$. The set of all sub-gradients of $\lambda(\cdot)$ at θ is called the sub-differential $\lambda(\cdot)$ at θ, denoted $\partial \lambda(\theta)$, and the correspondence which maps θ to $\partial \lambda(\theta)$ is the sub-differential map. A selection $\nabla^{\dagger} \lambda(\cdot)$ of the sub-differential map is then a single valued mapping which for each element θ of Θ selects an element $\nabla^{\dagger} \lambda(\theta) \in \partial \lambda(\theta)$. In what follows, $\mathcal{B}(\Theta)$, $\mathcal{B}(\mathbb{R})$ and $\mathcal{B}(\mathbb{R}^{d})$ are the Borel σ-algebras on Θ, $\mathbb{R}$ and $\mathbb{R}^{d}$ respectively, and $\mathcal{B}(\Theta) \otimes \mathcal{F}$ is the product σ-algebra on $\Theta \times \Omega$ generated by $(\Theta, \mathcal{B}(\Theta))$ and $(\Omega, \mathcal{F})$. A measurable sub-gradient of a $(\mathcal{B}(\Theta) \otimes \mathcal{F}) / \mathcal{B}(\mathbb{R})$-measurable function $\lambda^* : \Theta \times \Omega \to \mathbb{R}$ which is convex on Θ for all $\omega \in \Omega$ is then a $(\mathcal{B}(\Theta) \otimes \mathcal{F}) / \mathcal{B}(\mathbb{R}^d)$-measurable selection of the sub-differential map of $\lambda^*(\cdot)$.

Theorem 1. *Let Θ be a non-empty open convex subset of $\mathbb{R}^d$, where d is a finite positive integer, $\{\lambda_n : \Theta \times \Omega \to \mathbb{R}\}_{n=1}^{\infty}$ be a sequence of $(\mathcal{B}(\Theta) \otimes \mathcal{F}) / \mathcal{B}(\mathbb{R}^d)$-measurable functions such that $\lambda_n(\cdot; \omega) : \Theta \to \mathbb{R}$ is a proper convex function on Θ for each $\omega \in \Omega$ with effective domain equal to Θ, and $\lambda_0 : \Theta \to \mathbb{R}$ be a finite convex function such that $\lambda_n(\theta)$ converges in probability to $\lambda_0(\theta)$ point-wise for each θ in Θ. In addition, suppose that:*

(i) *$\nabla^\dagger \lambda_n : \Theta \times \Omega \to \mathbb{R}^d$ is a measurable sub-gradient of λ_n for each $n \in \mathbb{N}$; and*

(ii) *$\lambda_0(\theta)$ is differentiable with respect to θ for all $\theta \in \Theta$ with gradient $\nabla \lambda_0(\theta)$.*

Then:

(a) *For every non-empty compact subset Γ of Θ, $\sup_{\theta \in \Gamma} \left\| \nabla^\dagger \lambda_n(\theta, \cdot) - \nabla \lambda_0(\theta) \right\|$ is $\mathcal{F}^u / \mathcal{B}(\mathbb{R})$-measurable for each $n \in \mathbb{N}$, where $\mathcal{F}^u$ is the universal completion of $\mathcal{F}$.*

(b) $\lim_{n \to \infty} P^u \left(\sup_{\theta \in \Gamma} \left\| \nabla^\dagger \lambda_n(\theta) - \nabla \lambda_0(\theta) \right\| > \varepsilon \right) = 0$ *for all $\varepsilon > 0$, where P^u is the extension of P to $(\Omega, \mathcal{F}^u)$.*

Remark 1. Theorem 1 is a stochastic analogue of a generalization of Theorem 25.7 of Rockafellar (1970). The latter theorem establishes that if $\{\lambda_n(\cdot)\}_{n=1}^{\infty}$ is a sequence of non-random differentiable finite convex functions defined on a non-empty open convex subset Θ of $\mathbb{R}^d$ that converge point-wise to a non-random differentiable function $\lambda_0(\cdot)$ on Θ then the sequence of gradients $\{\nabla \lambda_n(\cdot)\}_{n=1}^{\infty}$ converges uniformly to $\nabla \lambda_0(\cdot)$ on any compact subset of Θ. Theorem 1 allows the functions in the sequence to be random and drops the requirement that they be differentiable, though the limit function still needs to be differentiable.

Remark 2. If the λ_n functions were differentiable on Θ for all $n \in \mathbb{N}$ and $\omega \in \Omega$ then it would only be necessary in Theorem 1 to require that $\nabla^\dagger \lambda_n(\theta)$ be $\mathcal{F} / \mathcal{B}(\mathbb{R}^d)$-measurable, i.e. a random vector, for each $\theta \in \Theta$. Corollary 25.5.1 of Rockafellar (1970) would then imply that λ_n was continuously differentiable on θ for all $\omega \in \Omega$ and hence so too was $(\lambda_n - \lambda)$. It would then be straightforward to show that $(\nabla\lambda_n - \nabla\lambda)$ converged in probability uniformly on compact subsets of Θ by using the same diagonalization method as in the proof of Theorem II.1 of Andersen and Gill (1982) but replacing the reference to Theorem 10.8 of Rockafellar (1970) by a reference to Theorem 25.7 of Rockafellar (1970). Unfortunately, in many cases of interest there is a positive probability that λ_n is not differentiable everywhere in Θ. Since Θ is an open convex set, it follows from Theorems 25.5 and 25.6 of Rockafellar (1970) that any selection of the sub-differential map of $\lambda_n(\cdot;\omega)$ will be discontinuous at those elements of Θ at which $\lambda_n(\cdot;\omega)$ is not differentiable.

Remark 3. Theorem 1 does not itself establish the existence of a measurable sub-gradient $\nabla^\dagger\lambda_n$. One case where this can be established is when λ_n depends on Ω through a $(D_n \times 1)$ vector of random variables W_n (where D_n is finite) and is jointly continuous with respect to (θ, W_n). Then $\Theta \times \mathbb{R}^{D_n}$ is σ-compact Hausdorff and the sub-differential correspondence $\partial\lambda_n : \Theta \times \mathbb{R}^{D_n} \twoheadrightarrow \mathbb{R}^d$ is non-empty compact valued so its graph is closed. Theorems 18.10 and 18.20 of Aliprantis and Border (2006) then imply that $\partial\lambda_n$ is weakly measurable. It follows by the Kuratowski–Ryll-Nardzewski theorem that $\partial\lambda_n$ has a measurable selection.

Remark 4. Validating the requirement that $\nabla^\dagger\lambda_n$ be $\mathcal{B}(\Theta) \otimes \mathcal{F}/\mathcal{B}(\mathbb{R}^d)$-measurable when λ_n is not differentiable everywhere in Θ needs to be done on a case-by-case basis but is often straightforward in practice. For example, in the quantile regression case when the conditional α-quantile of y given x is equal to $x'\beta_0$ for some unknown β_0 then the objective function is given by:

$$S_N(\beta;\alpha) = \sum_{i=1}^{n} \left(\alpha - \mathbb{1}_{\{y_i < x_i'\beta\}}\right)(y_i - x_i'\beta)$$

where $\mathbb{1}_{\{y_i < x_i'\beta\}}$ is the indicator function for the event $\{y_i < x_i'\beta\}$. One natural choice of sub-gradient of $S_N(\beta;\alpha)$ with respect to β is then

$$\nabla^\dagger S_N(\beta;\alpha) = -\sum_{i=1}^{n} \left[\alpha - \mathbb{1}_{\{y_i < x_i'\beta\}}\right] x_i,$$

and it is easy to verify that this choice of sub-gradient of $S_N(\beta;\alpha)$ is $\mathcal{B}(\Theta)\otimes\mathcal{F}/\mathcal{B}\left(\mathbb{R}^K\right)$-measurable, where $\Theta=\mathbb{R}^K$.

3. Relevance for Optimization Based Estimation

Suppose that Υ is a non-empty open convex subset of $\mathbb{R}^d$ and that for each $n\in\mathbb{N}$, $Q_n(\cdot)$ is a random convex function from Θ to $\mathbb{R}$ and T_n is a $(d\times 1)$ random vector. In addition, suppose that, as in Pollard (1991), there exists $\beta_0\in\Upsilon$ and a non-random symmetric positive definite $(d\times d)$ matrix A_0 such that

$$Q_n\left(\beta_0+n^{-1/2}\psi\right)=Q_n(\beta_0)+\langle T_n,\psi\rangle+\frac{1}{2}\psi' A_0\psi+o_p(1),\quad \forall\psi\in\mathbb{R}^d. \tag{2}$$

Then $\lambda_n(\psi)=Q_n\left(\beta_0+n^{-1/2}\psi\right)-Q_n(\beta_0)-\langle T_n,\psi\rangle$ is a convex function of ψ which converges in probability point-wise to the finite strictly convex function $\lambda_0(\psi)=\frac{1}{2}\psi' A_0\psi$. Theorem II.1 of Andersen and Gill (1982) implies that $\lambda_n(\cdot)$ converges uniformly in probability to $\lambda_0(\cdot)$ on any compact subset Γ of $\mathbb{R}^d$ and thus

$$\sup_{\psi\in\Gamma}\left|Q_n\left(\beta_0+n^{-1/2}\psi\right)-Q_n(\beta_0)-\langle T_n,\psi\rangle-\frac{1}{2}\psi' A_0\psi\right|=o_p(1),$$

which establishes that the local stochastic expansion given by Equation (2) is uniform on compact subsets of $\mathbb{R}^d$. The line of argument in Pollard (1991) can then be used to show that if, in addition, T_n converges in distribution to a $\mathcal{N}(0,B_0)$ random vector, where B_0 is a non-random positive definite $(d\times d)$ matrix, then

$$n^{1/2}\left(\widehat{\beta}_n-\beta_0\right)=-A_0^{-1}T_n+o_p(1), \tag{3}$$

where $\widehat{\beta}_n$ is a $\mathcal{F}/\mathcal{B}\left(\mathbb{R}^d\right)$-measurable solution to $\min_{\beta\in B} Q_n(\beta)$, and hence $n^{1/2}\left(\widehat{\beta}_n-\beta_0\right)$ converges in distribution to a $\mathcal{N}\left(0,A_0^{-1}B_0A_0^{-1}\right)$ random vector.

Now suppose that $\nabla^{\dagger} Q_n(\cdot)$ is a measurable selection of $\partial Q_n(\cdot)$. Then $\nabla^{\dagger}\lambda_n(\cdot)$, defined by

$$\nabla^{\dagger}\lambda_n(\psi) \equiv n^{-1/2}\nabla^{\dagger} Q_n\left(\beta_0 + n^{-1/2}\psi\right) - T_n, \quad \forall\psi \in \mathbb{R}^d,$$

is a measurable selection of $\partial\lambda_n(\cdot)$. Since $\lambda_0(\psi)$ is differentiable it follows from Theorem 1 above that $\sup_{\psi\in\Gamma}\left|\nabla^{\dagger}\lambda_n(\psi) - \nabla\lambda_0(\psi)\right| = o_p(1)$ and hence that

$$\sup_{\psi\in\Gamma}\left|n^{-1/2}\nabla^{\dagger} Q_n\left(\beta_0 + n^{-1/2}\psi\right) - T_n - A_0\psi\right| = o_p(1), \tag{4}$$

noting that $\nabla\lambda_0(\psi) = \frac{1}{2}\frac{\partial\left(\psi' A_0\psi\right)}{\partial\psi} = A_0\psi$.

Equations (3) and (4) then imply that $n^{-1/2}\nabla^{\dagger} Q_n\left(\widehat{\beta}_n\right) = o_p(1)$. If $\widetilde{\beta}_n$ is a root-n consistent estimator of β_0 and $\widetilde{A}_n$ is a consistent estimator of A_0 then a single Gauss-Newton step

$$\widetilde{\beta}_n^* = \widetilde{\beta}_n - \widetilde{A}_n^{-1} n^{-1}\nabla^{\dagger} Q_n\left(\widetilde{\beta}_n\right),$$

produces an estimator which is asymptotically equivalent to the full optimization estimator in that $n^{1/2}\left(\widetilde{\beta}_n^* - \widehat{\beta}_n\right) = o_p(1)$. The result in Equation (4) can also be used in the construction of score tests and generalized $C(\alpha)$ tests provided that consistent estimators of both A_0 and B_0 are available.

4. Conclusions

Following Pollard (1991) it has been well known in econometrics that point-wise convergence in probability for a sequence of stochastic convex functions to some limit function implies their uniform convergence in probability on compact sets. In the present paper, I establish that if the limiting function is differentiable then any sequence of measurable sub-gradients of the stochastic convex functions converges uniformly in extended probability on compact sets to the derivative of the limit function. I then use this result to establish the validity of a uniform local stochastic expansion of a sub-gradient of a convex objective function given a point-wise local stochastic expansion of the objective function.

Appendix A. Proof of Theorem 1

First, observe that since Γ is a compact subset of Θ then there exists $\tau > 0$ such that $N_\tau(\theta) \subset \Theta$ for all $\theta \in \Gamma$, where

$$N_\tau(\theta) = \left\{ \widetilde{\theta} \in \mathbb{R}^d : \left\| \widetilde{\theta} - \theta \right\| \leq \tau \right\}$$

is the closed ball of radius τ centered at θ. Let $A_\tau(\Gamma) = \cup_{\theta \in \Gamma} N_\tau(\theta)$; then $\Gamma \subset A_\tau(\Gamma)$ and $A_\tau(\Gamma)$ is a compact subset of Θ. Since $\lambda_n(\cdot;\omega)$ is a proper convex function on Θ for any $\omega \in \Omega$ it follows that $\lambda_n(\cdot;\omega)$ is uniformly continuous on $A_\tau(\Gamma)$ for any $\omega \in \Omega$ and hence that there exists $L_n(\Gamma;\omega) < \infty$ such that $\left|\lambda_n\left(\widetilde{\theta};\omega\right)\right| \leq L_n(\Gamma;\omega)$ for all $\widetilde{\theta} \in A_\tau(\Gamma)$. It follows by the characterization of sub-gradients given in Equation (1) that

$$\left\langle \nabla^\dagger \lambda_n(\theta;\omega), \left(\widetilde{\theta} - \theta\right) \right\rangle \leq \lambda_n\left(\widetilde{\theta};\omega\right) - \lambda_0(\theta) \leq 2L_n(\Gamma;\omega), \quad \forall \theta \in \Gamma \,\&\, \widetilde{\theta} \in N_\tau(\theta),$$

which implies that $\sup_{\widetilde{\theta} \in N_\tau(\theta)} \left\langle \nabla^\dagger \lambda_n(\theta;\omega), \left(\widetilde{\theta} - \theta\right) \right\rangle \leq 2L_n(\Gamma;\omega)$ for all $\theta \in \Gamma$. Since $\|a\| = \sup_{\eta:\|\eta\|=1} < a, \eta >$ for any vector a then $\left\|\nabla \lambda_n^\dagger(\theta;\omega)\right\| \leq 2L_n(\Gamma;\omega)/\tau < \infty$ for all $\theta \in \Gamma$ and $\omega \in \Omega$. By a parallel argument it follows that there exists $L_0 < \infty$ such that $\|\nabla \lambda_0(\theta)\| \leq 2L_0/\tau$ for all $\theta \in \Gamma$. Together these imply that $\left\|\nabla \lambda_n^\dagger(\theta;\omega) - \nabla \lambda_0(\theta)\right\| \leq 2\left(L_0 + L_n(\Gamma;\omega)\right)/\tau$ for all $\theta \in \Gamma$ and $\omega \in \Omega$ and hence that $\sup_{\theta \in \Gamma} \left\|\nabla \lambda_n^\dagger(\theta;\omega) - \nabla \lambda_0(\theta)\right\| < \infty$ for all $\omega \in \Omega$ so $\sup_{\theta \in \Gamma} \left\|\nabla \lambda_n^\dagger(\theta,\cdot) - \nabla \lambda_0(\theta)\right\|$ is a mapping from Ω to $\mathbb{R}$. Then since $\mathbb{R}^d$ is a Polish space and Θ is an open subset of $\mathbb{R}^d$ it follows that Θ is a Lusin space and hence it is also a Souslin space. Defining the correspondence $S : \Omega \rightrightarrows \Theta$ by $S(\omega) = \Gamma$ for all $\omega \in \Omega$, it follows that the graph of S belongs to $\mathcal{B}(\Theta) \otimes \mathcal{F}$. Hence it follows from Theorem 2.17 of Stinchcombe and White (1992) that the function $h_n : \Omega \to \overline{\mathbb{R}}^d$, defined by

$$h_n(\omega) = \sup_{\theta \in S(\omega)} \left\|\nabla^\dagger \lambda_n(\theta;\omega) - \nabla \lambda_0(\theta)\right\| = \sup_{\theta \in \Gamma} \left\|\nabla^\dagger \lambda_n(\theta,\omega) - \nabla \lambda_0(\theta)\right\|,$$

is $\mathcal{F}$-analytic and hence is $\mathcal{F}^u/\mathcal{B}(\mathbb{R})$-measurable, which establishes the first result in the theorem.

Second, since $\lambda_n(\theta;\omega)$ is a convex function of $\theta \in \Theta$ for any $\omega \in \Omega$ and converges point-wise in probability to $\lambda_0(\theta)$ it follows that that $\lambda_0(\cdot)$ is convex on Θ and hence also continuous on Θ. Since $A_\tau(\Gamma)$ is a compact subset of Θ it follows that $\lambda_0(\cdot)$ is uniformly continuous on $A_\tau(\Gamma)$. Now fix $\varepsilon > 0$; then there exists $\delta > 0$ such that $\left\|\nabla\lambda_0\left(\widetilde{\theta}\right) - \nabla\lambda_0(\theta)\right\| < \varepsilon$ for all $\theta, \widetilde{\theta} \in A_\tau(\Gamma)$ satisfying $\left\|\widetilde{\theta} - \theta\right\| < \delta$. Next, fix $\theta^* \in \Gamma$, $\eta \in \mathbb{R}^d$ such that $\|\eta\| = 1$, and $\xi \in (0, \min(\tau, \delta))$. Then θ^*, $(\theta^* - \xi\eta)$ and $(\theta^* + \xi\eta)$ all belong to $A_\tau(\Gamma)$. Since $\lambda_0(\theta^* + z\eta)$ is convex in z given θ^* and η for all z such that $(\theta^* + z\eta) \in \Theta$ then

$$\begin{aligned}\langle\nabla\lambda_0(\theta^* - \xi\eta), \eta\rangle &\leq \left[\frac{\lambda_0(\theta^*) - \lambda_0(\theta^* - \xi\eta)}{\xi}\right] \leq \langle\nabla\lambda_0(\theta^*), \eta\rangle \\ &\leq \left[\frac{\lambda_0(\theta^* + \xi\eta) - \lambda_0(\theta^*)}{\xi}\right] \leq \langle\nabla\lambda_0(\theta^* + \xi\eta), \eta\rangle.\end{aligned}$$

But since $\sup_{\widetilde{\eta}:\|\widetilde{\eta}\|=1}\left|\left\langle\nabla\lambda_0\left(\widetilde{\theta}\right), \widetilde{\eta}\right\rangle - \langle\nabla\lambda_0(\theta), \widetilde{\eta}\rangle\right| = \left\|\nabla\lambda_0\left(\widetilde{\theta}\right) - \nabla\lambda_0(\theta)\right\| < \varepsilon$ for all $\theta, \widetilde{\theta} \in A_\tau(\Gamma)$ satisfying $\left\|\widetilde{\theta} - \theta\right\| < \delta$ and since $\xi < \delta$ it follows that

$$|\langle\nabla\lambda_0(\theta^*), \eta\rangle - \langle\nabla\lambda_0(\theta^* - \xi\eta), \eta\rangle| < \varepsilon, \quad |\langle\nabla\lambda_0(\theta^*), \eta\rangle - \langle\nabla\lambda_0(\theta^* + \xi\eta), \eta\rangle| < \varepsilon. \tag{A.1}$$

In addition, since $\lambda_n(\theta^* + z\eta)$ is convex in z given θ^* and η for all z such that $(\theta^* + z\eta) \in \Theta$ then

$$\left[\frac{\lambda_n(\theta^*) - \lambda_n(\theta^* - \xi\eta)}{\xi}\right] \leq \left\langle\nabla^\dagger\lambda_n(\theta^*), \eta\right\rangle \leq \left[\frac{\lambda_n(\theta^* + \xi\eta) - \lambda_n(\theta^*)}{\xi}\right]. \tag{A.2}$$

Since $A_\tau(\Gamma)$ is a compact subset of Θ it follows from Theorem II.1 of Andersen and Gill (1982) that $\lim_{n\to\infty} P\left(\sup_{\theta\in A_\tau(\Gamma)}|\lambda_n(\theta) - \lambda_0(\theta)| \leq \varepsilon^*\right) = 1$ for any $\varepsilon^* > 0$. Since P^u is the unique extension of P to the universal completion $\mathcal{F}^u$ of $\mathcal{F}$ then $P^u(A) = P(A)$ for any $A \in \mathcal{F}$ and hence $\lim_{n\to\infty} P^u\left(\sup_{\theta\in A_\tau(\Gamma)}|\lambda_n(\theta) - \lambda_0(\theta)| \leq \varepsilon^*\right) = 1$ for any $\varepsilon^* > 0$. This in turn implies that

$$\begin{aligned}\lim_{n\to\infty} P^u\left(\sup_{\theta\in\Gamma,\widetilde{\eta}:\|\widetilde{\eta}\|=1}\left|\left[\frac{\lambda_n(\theta) - \lambda_n(\theta - \xi\widetilde{\eta})}{\xi}\right] - \left[\frac{\lambda_0(\theta) - \lambda_0(\theta - \xi\widetilde{\eta})}{\xi}\right]\right| \leq \varepsilon\right) = 1, \\ \lim_{n\to\infty} P^u\left(\sup_{\theta\in\Gamma,\widetilde{\eta}:\|\widetilde{\eta}\|=1}\left|\left[\frac{\lambda_n(\theta + \xi\widetilde{\eta}) - \lambda_n(\theta)}{\xi}\right] - \left[\frac{\lambda_0(\theta + \xi\widetilde{\eta}) - \lambda_0(\theta)}{\xi}\right]\right| \leq \varepsilon\right) = 1.\end{aligned}$$

Combined with Equation (A.1) these imply that,

$$\lim_{n\to\infty} P^u \left(\sup_{\theta\in\Gamma,\widetilde{\eta}:\|\widetilde{\eta}\|=1} \left| \left[\frac{\lambda_n(\theta) - \lambda_n(\theta - \xi\widetilde{\eta})}{\xi} \right] - \langle \nabla\lambda_0(\theta), \widetilde{\eta} \rangle \right| \le 2\varepsilon \right) = 1,$$
$$\lim_{n\to\infty} P^u \left(\sup_{\theta\in\Gamma,\widetilde{\eta}:\|\widetilde{\eta}\|=1} \left| \left[\frac{\lambda_n(\theta + \xi\widetilde{\eta}) - \lambda_n(\theta)}{\xi} \right] - \langle \nabla\lambda_0(\theta), \widetilde{\eta} \rangle \right| \le 2\varepsilon \right) = 1.$$

Now for any quadruplet (a, b, c, d) such that $a \le b \le c$, $|a - d| \le 2\varepsilon$ and $|c - d| \le 2\varepsilon$ it follows by the triangle inequality that $|b - d| \le 2\varepsilon$. Hence it follows from Equation (A.2) that

$$\lim_{n\to\infty} P^u \left(\sup_{\theta\in\Gamma,\widetilde{\eta}:\|\widetilde{\eta}\|=1} \left| \left\langle \nabla^\dagger\lambda_n(\theta), \widetilde{\eta} \right\rangle - \langle \nabla\lambda_0(\theta), \widetilde{\eta} \rangle \right| \le 2\varepsilon \right) = 1.$$

But

$$\sup_{\theta\in\Gamma,\widetilde{\eta}:\|\widetilde{\eta}\|=1} \left| \left\langle \nabla^\dagger\lambda_n(\theta), \widetilde{\eta} \right\rangle - \langle \nabla\lambda_0(\theta), \widetilde{\eta} \rangle \right| = \sup_{\theta\in\Gamma} \left\| \nabla^\dagger\lambda_n(\theta) - \nabla\lambda_0(\theta) \right\|,$$

so it follows in turn that

$$\lim_{n\to\infty} P^u \left(\sup_{\theta\in\Gamma} \left\| \nabla^\dagger\lambda_n(\theta) - \nabla\lambda_0(\theta) \right\| \le 2\varepsilon \right) = 1.$$

Since $\varepsilon > 0$ was arbitrary this establishes the third result in the theorem. ■

References

Aliprantis, C.D., Border, K.C., 2006. Infinite Dimensional Analysis: A Hitchhiker's Guide. 3rd ed., Springer, Berlin.

Andersen, P., Gill, R., 1982. Cox's regression model for counting processes: A large sample study. Annals of Statistics 10, 1100–1120. doi:10.1214/aos/1176345976.

Hayashi, F., 2000. Econometrics. Princeton University Press, Princeton.

Pollard, D., 1991. Asymptotics for least absolute deviations regression estimators. Econometric Theory 7, 186–199. doi:10.1017/s0266466600004394.

Rockafellar, R., 1970. Convex Analysis. Princeton University Press, Princeton.

Stinchcombe, M.B., White, H., 1992. Some measurability results for extrema of random functions over random sets. The Review of Economic Studies 59, 495–514. doi:10.2307/2297861.