A Nonparametric Simulated Maximum Likelihood Estimation Method*

Jean-David Fermanian† Bernard Salanié‡

November 4, 2003

Abstract

Existing simulation-based estimation methods are either general-purpose but asymptotically inefficient or asymptotically efficient but only suitable for restricted classes of models. This paper studies a simulated maximum-likelihood method that rests on estimating the likelihood nonparametrically on a simulated sample. We prove that this method, which can be used on very general models, is consistent and asymptotically efficient for static models. We then propose an extension to dynamic models and give some Monte-Carlo simulation results on a dynamic Tobit model.

Introduction

Many parametric estimation procedures in econometrics are based on the maximization of a criterion function. This may be the mean square error as for the least squares method, the likelihood function for maximum likelihood estimation, or the likelihood of a well-chosen pseudo-model for pseudo-maximum likelihood methods. Unfortunately, the criterion function sometimes does not have a closed-form expression. This is true, for instance, of limited-dependent variable models with lagged dependent variables, where the likelihod function and other competing criterion functions can only be written as integrals of large dimension (equal to the number of observations). Simulation-based estimation methods were devised precisely to circumvent this problem[1]. By replacing untractable expectations with their Monte-Carlo counterparts, they allow the relevant criterion functions to be computed, which has made it possible for econometricians to estimate new classes of models.

*We thank Jean-Pierre Florens, Arnoldo Frigessi, Christian Gouriéroux, Jim Heckman, Guy Laroque, Oliver Linton, Nour Meddahi, Alain Monfort, Eric Renault, Christian Robert, Neil Shephard and two referees for their comments. Remaining errors and imperfections are ours. Parts of this paper were written while Bernard Salanié was visiting the University of Chicago, which he thanks for its hospitality.

†CDC Ixis Capital Markets and CREST.

‡CREST, GRECSTA and CEPR.

[1]Early references are Lerman-Manski (1981), Pakes (1986), Laroque-Salanié (1989), McFadden (1989) and Pakes-Pollard (1989).

Simulation-based estimation methods belong to two general classes[2]. The first one consists of methods that are reasonably general-purpose but are not efficient asymptotically, even when the number of simulation draws is allowed to increase fast enough. The method of simulated moments (McFadden (1989), Pakes-Pollard (1989)) and the simulated pseudo-maximum likelihood methods (Laroque-Salanié (1989, 1993, 1994)) both belong to this class. As they rely on simulating the obvious mathematical expectation with its Monte-Carlo counterpart, they can be applied to a large class of models. However, they simulate criterion functions that (even with an infinite number of simulations) do not lead to efficient estimators. The indirect inference methods (Gouriéroux-Monfort-Renault (1993) and Smith (1993)) also belong to this first category. The second class of simulation-based estimation methods relies on simulating the likelihood function, so that the resulting estimators are asymptotically efficient (again, with an infinite number of simulations). The simulated likelihood methods (see e.g. Lee (1995)) and the method of simulated scores (Hajivassiliou-McFadden (1998)) are examples of such estimation methods. The difficulty with these methods is that as the likelihood function usually cannot be written as a function of mathematical expectations, they can only be applied to restrictive classes of models. Thus there has been a lot of emphasis on the literature on dynamic LDV models, but the methods that have been proposed only apply to models defined by linear constraints, for which several classes of efficient simulators have been devised (see, e.g., Börsch-Supan and Hajivassiliou (1993)). To the best of our knowledge, there exist few currently available methods that are both asymptotically efficient and applicable to a very wide class of econometric models. The Efficient Method of Moments (Gallant and Tauchen (1996)) is a competitor: by using the score of a well-chosen auxiliary model, an EMM estimator can become as efficient as the maximum likelihood estimator in a variety of situations. Nonetheless, this property requires that the auxiliary model encompasses the true model. Most of the time, this can be done only by considering rather intricate auxiliary models, whose number of parameters is increasing with the sample size, in the spirit of Gallant and Nychka (1987). Alternatively, GMM estimators can be asymptotically efficient using a continuum of moments based on the empirical characteristic function (Feuerverger and McDunnough (1981a,1981b), Carrasco and Florens (2002a)). Despite its generality, the latter method requires the inversion of a covariance operator in an infinite dimensional Hilbert space. This leads to some ill-posed problems whose solution involves the delicate choice of regularization parameters. See a discussion about the efficiency of these methods in Carrasco and Florens (2002b).

The purpose of this paper is to study a simulation-based estimation method, which we call the NonParametric Simulated Maximum Likelihood method, or NPSML for short. Start from a fully parametric model whose reduced form can be simulated (which is a very mild requirement). Then NPSML consists in approximating the unknown likelihood function with a kernel-based nonparametric estimator based on simulations of the endogenous variables of the model[3]. Since this strategy is applicable

[2]Gouriéroux-Monfort (1996) survey the available methods. Hajivassiliou-Ruud (1994) focusses on limited-dependent variable models, while Stern (1997) concentrates on empirical applications.

[3]We recently found out from Arnoldo Frigessi that Diggle and Gratton (1984) already proposed the

to a very wide class of models, it provides a quasi-universal simulator. Moreover, we prove in this paper that in static models, it provides consistent, asymptotically normal and asymptotically efficient estimators when the number of simulations goes to infinity and the bandwidth of the kernel estimator goes to zero. We then argue that the method can be extended to dynamic models and explain how to do so.

Section 1 presents the basic idea of the NPSML estimation method, using a static (but very general) continuous model as an application. It states our consistency and asymptotic efficiency theorems, which are proved in the appendix. Such results could be applied to estimate nonlinear simultaneous equations models. In sections 2 and 3, we show how the NPSML method can be extended to limited dependent variable models and to fully dynamic models. We then give some Monte-Carlo simulation evidence on a dynamic Tobit model in section 4.

1 NPSML for Static Models

Simulation-based methods are clearly most useful in dynamic settings. Nevertheless, it seems simpler to introduce the NPSML method and its asymptotic properties within a static model. Therefore, consider a model with reduced form

$$y = g(x, \theta, \varepsilon), \tag{1-1}$$

where

- θ, the parameter of interest, belongs to a compact set $\Theta \subset \mathbb{R}^q$,
- the observed endogenous variable y is a vector of $\mathbb{R}^m$,
- the exogenous variable x belongs to $\mathbb{R}^d$,
- $\varepsilon \in \mathbb{R}^e$ represents the disturbances.

We assume that both the function g and the distribution of the disturbances ε are fully known[4]. Thus this is a fully parametric model—only the estimation technique has a nonparametric element. To simplify exposition, we assume in this section that the distribution of y has no atoms; limited dependent variable models are discussed in section 3.

Let $(x_t, y_t)_{t=1,\dots,T}$ be an i.i.d. sample. The associated loglikelihood then is

$$L_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \ln l_t(\theta),$$

denoting $l_t(\theta)$ the density of y_t knowing (x_t, θ). We assume

NPSML estimator. However, they did not study its asymptotic properties or extend it to dynamic models.

[4]As usual, unknown parameters of the distribution of ε are integrated to θ. Moreover, lagged values of the observed endogenous variable can be subsumed within x in the usual manner. We will introduce lagged latent variables later in this paper.

Assumption L1 : the maximum likelihood estimator $\tilde{\theta}_T$ is consistent, asymptotically normal and asymptotically efficient. The true parameter θ_0 belongs to the interior of Θ. More precisely, we assume that

$$-\frac{\partial^2 L_T}{\partial\theta\partial\theta'}(\theta^*) \underset{T\to\infty}{\xrightarrow{P}} \Omega, \tag{1-2}$$

uniformly with respect to θ^* in a neighborhood of θ_0, and that

$$T^{1/2}\frac{\partial L_T}{\partial\theta}(\theta_0) \underset{T\to\infty}{\xrightarrow{D}} \mathcal{N}(0,\Omega). \tag{1-3}$$

for some positive matrix Ω.

For the class of models we are interested in, the likelihood function $l_t(\theta)$ cannot be computed in a closed form, so that it is impossible to compute the maximum likelihood estimator $\tilde{\theta}_T$. We propose instead to approximate each term $l_t(\theta)$ by a kernel estimator based on some i.i.d. simulated sample $(\varepsilon_t^s)_{s=1,\dots,S}$ drawn from the distribution[5] of ε.

Thus, denoting $y_t^s(\theta) = g(x_t,\theta,\varepsilon_t^s)$, the likelihood $l_t(\theta)$ is estimated by

$$l^S(y_t|x_t,\theta) \equiv l_t^S(\theta) \equiv \frac{1}{Sh^m}\sum_{s=1}^{S} K\left(\frac{y_t - y_t^s(\theta)}{h}\right) \tag{1-4}$$

Here, h is a bandwidth such that $h \longrightarrow 0$ when $S \longrightarrow \infty$, and K is a kernel. To simplify the presentation, assume that h is a power of S and that S is bounded above by a power of T. Under technical conditions that are stated below, $l_t^S(\theta)$ converges to $l_t(\theta)$ when the number of simulations S goes to infinity. Thus a natural idea consists in defining the NPSML estimator as the global maximizer of

$$\tilde{L}_T^S(\theta) = \frac{1}{T}\sum_{t=1}^{T} \ln l_t^S(\theta) \tag{1-5}$$

on Θ. It can be obtained using standard maximization algorithms; note that since l_t^S is a combination of kernels and the latter usually have a simple form, it is easy to feed analytical derivatives into the maximization routine to reduce the computation time if need be.

Since the logarithm has infinite derivative in zero, small simulation errors on small values of the likelihood will be amplified. Thus it is necessary to trim the smallest values of l_t^S [6]. This can be done by considering the nonparametric simulated loglikelihood

$$\tilde{L}_T^S(\theta) = \frac{1}{T}\sum_{t=1}^{T} \tau_S(l_t^S(\theta)) \ln l_t^S(\theta), \tag{1-6}$$

[5] The (ε_t^s) can also be the same for each t; the proofs of asymptotic results go through in that case.

[6] As is usual when introducing nonparametric estimators within parametric models: see Stone (1975), Bickel (1982) or more recently Ai (1997), among others.

where τ_S is a sufficiently regular function[7] such that $\tau_S(x) = 0$ if $|x| < h^\delta$ and $\tau_S(x) = 1$ if $|x| > 2h^\delta$, with $\delta > 0$.

Thus we define the NPSML estimator by

$$\hat{\theta}_T^S = \arg\max_{\theta\in\Theta} \tilde{L}_T^S(\theta)$$

We now state a set of assumptions under which it is strongly consistent when T and S go to infinity and the bandwidth h goes to zero.

The first assumption concerns the kernel. We assume

Assumption K0: the kernel K is twice continuously differentiable and has compact support.

Let ρ be the order of the kernel, i.e. $\int x_1^{\alpha_1} \dots x_m^{\alpha_m} K(x)\, dx$ is zero if $0 < \sum_{j=1}^m \alpha_j < \rho$ and nonzero if $\sum_j \alpha_j \in \{0, \rho\}$. (Classically, $\rho = 2$ for positive symmetrical kernels).

We need more assumptions on the exact likelihood function. In addition to Assumption L1, we assume

Assumption L2: $l(y|x,\theta)$ and $\partial^\rho l(y|x,\theta)/\partial y^\rho$ are bounded above on $\mathbb{R}^d \times \mathbb{R}^m \times \Theta$. There exists some $\beta > 1$ such that a.e.

$$\frac{1}{T}\sum_{t=1}^{T} |\ln l_t(\theta)|^\beta \text{ is convergent uniformly with respect to } \theta. \tag{1-7}$$

Moreover

$$E\left[\sup_\theta \left\| \frac{\partial l(Y|X,\theta)}{\partial \theta} \right\|\right] < \infty. \tag{1-8}$$

Assumption T1: there exists ν such that

$$P(\|X, Y\| > S^\nu) \ln h \underset{S\to\infty}{\longrightarrow} 0.$$

Assumption M1: there exist a function ϕ and some $s_0 \geq 0$ such that

$$h^{s_0} \sup_{\theta, \|x\| \leq S^\nu} \left\{ \left\| \frac{\partial g(x,\theta,\varepsilon)}{\partial \theta} \right\| + \left\| \frac{\partial g(x,\theta,\varepsilon)}{\partial x} \right\| \right\} \leq \phi(\varepsilon),$$

[7] One example is the continuously differentiable function defined by

$$\tau_S(x) = 4(x - h^\delta)^3/h^{3\delta} - 3(x - h^\delta)^4/h^{4\delta}$$

when $x \in [h^\delta, 2h^\delta]$. Therefore, this function τ_S is piecewise polynomial and $\|\tau_S'\|_\infty = O(h^{-\delta})$.

with $E[\phi(\varepsilon)] < \infty$, and where ν was introduced in Assumption T1.

To understand the assumptions easily, we have employed a simple codification: letter L is used for regularity conditions on the likelihood, T for the tail behavior of the underlying variables, M for the conditions on the reduced form model g, R for the rates of convergence of the parameters and K for the kernel.

More specifically, assumption L2 is necessary to bound uniformly some bias terms in kernel smoothing. It could be weakened[8]. Equation (1-7) is a uniform strong law of large numbers. Such an assumption is usual to ensure the consistency of maximum likelihood estimators. Since θ belongs to a compact subset, (1-8) mainly states the integrability of the score function. Thus, it is rather a weak assumption. Obviously, since h is power of T by assumption, $\ln h$ is of order $\ln T$. Therefore, T1 is very weak too, unless the tails of x and y are very thick. Note that if $\partial g/\partial\theta$ and $\partial g/\partial x$ are bounded in norm, then it suffices to take $s_0 = 0$ in Assumption M1. Otherwise $s_0 > 0$ is necessary because the supremum is taken over a compact set that increases in size like S^ν. M1 avoids too erratic variations of the endogenous variables y_t for small variations of exogenous variables x_t or of the parameter vector.

Our technical assumptions could be lighten by assuming the exogenous variables x belong necessarily to a bounded subset. We think it is too restrictive. Thus, in this paper, the support of the joint law of (x_t, y_t) is a priori unbounded, even if the proofs are longer and more involved.

We also need to put some assumptions on the order of the trimming function and the rate of convergence of the bandwidth to zero as the number of simulation draws goes to infinity. Trimming aside, this is the usual assumption in nonparametric density estimation.

Assumption R1: $\delta < \rho$ and $Sh^{m+2\delta}/\ln S$ goes to infinity as S goes to infinity.

We can finally state our consistency theorem.

Theorem 1.1 *Under assumptions K0, L1-L2, T1, M1 and R1, $\hat{\theta}_T^S$ is strongly consistent: almost everywhere,*

$$\hat{\theta}_T^S \underset{S,T\to\infty}{\longrightarrow} \theta_0.$$

It is easy to check that $\hat{\theta}_T^S$ is weakly consistent under the same assumptions except the second part of R1 and replacing (1-7) by $E[\sup_\theta |\ln l_t(\theta)|^\beta] < \infty$.

To apply our estimation method in practice, it is necessary to fix the values of the parameters ρ, δ, K, h, ν and S. The main difficulty consists in choosing the rates

[8]The first part of assumption L2 could be removed. The price is to replace $t = m$ by $t = 0$ in the proof of Lemma A.2. Thus, the rate of convergence of $l^S(y|x,\theta)$ would be $(Sh^{2m}/\ln S)^{1/2}$. This is nonstandard and would impose some additional constraints on h. Moreover, we would need to increase the order of the kernel ρ. This is why we choose to assume the uniform boundedness of the likelihood and its derivatives, which is satisfied most of the time.

of convergence of the number of simulations S to infinity and of the bandwidth h to zero so that the assumptions of Theorem 1.1 hold. Let us therefore assume that $S = C_1 T^a$ and $h = C_2 S^{-b}$ for some positive constants C_1 and C_2.

We would like to ensure the consistency of the NPSML estimator. We neglect assumption T1, which is bound to hold if the density of (x, y) is not too thick-tailed[9]. Then the relevant assumption is R1. This translates into

$$\delta < \rho \ , \ b < \frac{1}{m + 2\delta} \ .$$

In particular, take the usual case of a second-order kernel ($\rho = 2$). Then we can choose $\delta = 1$ for instance and the asymptotically optimal bandwidth selector for kernel density estimation, for which $b = 1/(m + 4)$, fits the bill. Thus the most natural choice for the rate of convergence of h to zero yields a consistent NPSML estimator. Note that as expected, the speed of convergence of S to infinity then is irrelevant for consistency: we only require that $a > 0$.

In order to prove the asymptotic normality and efficiency of the NPSML estimator, we need the strengthen the previous assumptions. To state them, let V_0 denote a neighborhood of θ_0.

Assumption L3: there exist $\gamma > 1$ and $\gamma' > 1$ such that

$$E\left[\sup_{\theta \in V_0} \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\|^{\gamma}\right] < \infty \quad \text{(L3a), and}$$

$$E\left[\sup_{\theta \in V_0} \left\| \frac{\partial l_t(\theta)}{\partial \theta} \right\|^{\gamma'}\right] < \infty \quad \text{(L3b).}$$

Moreover, $\partial^{\rho+1} l(y|x,\theta)/\partial\theta\partial y^{\rho}$ is bounded on $\mathbb{R}^d \times \mathbb{R}^m \times V_0$.[10]

Assumption T2: For some $\nu > 0$,

$$\left[T^{\gamma/2(\gamma-1)} + \left(\frac{T^{1/2}}{h^{\delta}}\right)^{\gamma'/(\gamma'-1)} + \left(\frac{T^{1/2}|\ln h|}{h^{(m+1+\delta)}}\right)^{\zeta/(\zeta-1)}\right] P_{\theta_0}(\|x_t, y_t\| > S^{\nu})$$

tends to zero when S and T tend to infinity, where γ and γ' (resp. ζ) were introduced in assumption L3 (resp. M2).

Assumption M2: for some $r_0 \geq 0$ and some $p_0 > 4$,

$$h^{r_0} \sup_{\theta \in V_0, \|x\| \leq S^{\nu}} \left\| \frac{\partial g(x, \theta, \varepsilon)}{\partial \theta} \right\| \leq \bar{\phi}(\varepsilon), \quad E[\bar{\phi}(\varepsilon)^{p_0}] < \infty. \tag{1-9}$$

[9] It should tend to zero more quickly than $\|(x, y)\|^{-k}$, for some $k \geq 0$, when $\|(x, y)\|$ tends to infinity.

[10] As in assumption L2, the latter condition could be removed. In this case, we have to assume $h^{\rho} S^{\nu} \to 0$, and the rates on convergence in lemma A.3 are then weakened.

Moreover, there exist a function $\bar{\psi}$ and $s_1 \geq 0$ such that, for every $\varepsilon > 0$,

$$h^{s_1} \sup_{\theta \in V_0, \|x\| \leq S^\nu} \left\{ \left\| \frac{\partial^2 g(x,\theta,\varepsilon)}{\partial^2 \theta} \right\| + \left\| \frac{\partial^2 g(x,\theta,\varepsilon)}{\partial x \partial \theta} \right\| + \left\| \frac{\partial g(x,\theta,\varepsilon)}{\partial \theta} \right\|^2 + \left\| \frac{\partial g(x,\theta,\varepsilon)}{\partial x} \right\| . \left\| \frac{\partial g(x,\theta,\varepsilon)}{\partial \theta} \right\| \right\} \leq \bar{\psi}(\varepsilon), \quad E[\bar{\psi}(\varepsilon)] < \infty.$$

Finally, for some $\zeta > 1$,

$$E\left[\sup_\theta \left\| \frac{\partial g(x_t,\theta,\varepsilon_t)}{\partial \theta} \right\|^\zeta \right] < \infty.$$

The number of simulations S has to grow sufficiently quickly with T for T2 to be satisfied. Again, if the derivatives of g are bounded, then one can take $r_0 = s_1 = 0$ in assumption M2. The assumptions on the rates of convergence also have to be strengthened:

Assumption R2:

$$T^{1/2} h^{\rho - \delta} \ln h \underset{S,T \to \infty}{\longrightarrow} 0, \quad \text{and}$$

$$T h^{-2m-2\delta-2-2r_0} \ln^2 h \ln S / S \underset{S,T \to \infty}{\longrightarrow} 0$$

To control the frequency of the trimming, we need

Assumption R3:

$$(T^{1/2} |\ln h|)^{\gamma/(\gamma-1)} P_{\theta_0}(\inf_{\theta \in V_0} l_t(\theta) \leq 2h^\delta) \underset{S,T \to \infty}{\longrightarrow} 0,$$

where γ was introduced in assumption L3.

Recall that we denote Ω the asymptotic variance-covariance matrix of the exact maximum likelihood estimator. We can finally state our asymptotic efficiency theorem.

Theorem 1.2 *Under assumptions K0, M1-M2, L1-L3, R1-R3 and T1-T2, $\hat{\theta}_T^S$ is asymptotically normal and asymptotically efficient:*

$$\sqrt{T}(\hat{\theta}_T^S - \theta_0) \underset{S,T \to \infty}{\overset{D}{\longrightarrow}} \mathcal{N}(0, \Omega), \tag{1-10}$$

where Ω has been defined in assumption L1.

To simplify the proof, we have used assumptions that are more restrictive than need be. Also, the assumptions used imply that $\hat{\theta}_T^S$ is strongly consistent, whereas convergence in probability would suffice.

When this theorem applies, $\hat{\theta}_T^S$ has the same asymptotic variance as $\tilde{\theta}_T$, the exact maximum likelihood estimator of θ_0. Thus $\hat{\theta}_T^S$ is asymptotically efficient and Ω can be estimated by

$$\hat{\Omega} = \frac{1}{T}\sum_{t=1}^{T} \tau_S(l_t^S(\hat{\theta}_T^S)) \cdot \left(\frac{\partial \ln l_t^S}{\partial \theta}(\hat{\theta}_T^S)\right) \cdot \left(\frac{\partial \ln l_t^S}{\partial \theta}(\hat{\theta}_T^S)\right)'. \tag{1-11}$$

Now consider the assumptions to get asymptotic normality. In addition to R1, R2 and R3 must also hold. R2 translates into

$$ab > \frac{1}{2(\rho-\delta)} \text{ and } 2ab(m+\delta+1+r_0)+1 < a \tag{1-12}$$

Now choose $\rho > \delta$ and some $C_0 > \frac{1}{2(\rho-\delta)}$. Moving on the hyperbola $ab = C_0$ to the zone of large a and small b will satisfy all assumptions. Thus our conditions define a nondegenerate region of the (a, b) plane. One would hope that this region intersects the line $b = 1/(m + 2\rho)$ that defines the usual asymptotically optimal bandwidth. It can be checked that such is the case if $(\rho - \delta)$ is large enough, which may imply using higher-order kernels ($\rho > 2$).

Assumption R3 is more problematic, as it should be checked on a case-by-case basis. Clearly, it holds when h goes to zero fast enough.

To illustrate the technical assumptions, we verify them for the simplest model we could imagine, say

$$y_t = x_t'\beta + \sigma\varepsilon_t, \;\; \varepsilon_t \sim \mathcal{N}(0,1), \;\; \theta = (\beta, \sigma).$$

Here, y_t is univariate, normally distributed with mean $x_t'\beta$ and variance $\sigma \in [\sigma_1, \sigma_2]$, $\sigma_1 > 0$. Since $l_t(\theta)$ is Lipschitz-continuous with respects to the parameters, some uniform strong laws of large numbers are available to get L2. Assumption T1 is true for every $\nu > 0$. To get M1 when x_t is unbounded, it is sufficient that $hS^\nu \longrightarrow 0$, which is the case with a sufficiently small ν. Thus, to satisfy R1 and to obtain the consistency of $\hat{\theta}_T^S$, set $\delta = 1$, $\rho = 2$ and $b < 1/3$. To get the asymptotic normality, note that L3, T2 and M2 are true for every choice of γ, γ' and ζ. Tedious calculations show that R3 holds if

$$T^{\frac{\gamma}{2(\gamma-1)}} |\ln h|^{(\gamma+1)/(2(\gamma-1)} h^\delta \longrightarrow 0,$$

which is satisfied for $ab > \gamma/2\delta(\gamma - 1)$. Since R2 implies $ab > 1/2$, choose $a = 6$ and $b = 1/11$ for instance. To get more realistic rates for S and h, it is necessary to increase ρ, viz to consider higher order kernels.

2 Extension to Limited Dependent Variable Models

The theory can be extended to the case when the endogenous variables y_t take some discrete values with a strictly positive probability. Since limited dependent variables

models come in many guises, we illustrate the basic idea with the example of the censored regression model

$$y_t = \max(0, z_t)$$

where the univariate latent variable z_t is generated by

$$z_t = g(x_t, \theta_0, \varepsilon_t).$$

Let $z_t^s(\theta)$ be the simulated values for the latent variable. When we observe $y_t > 0$, then the likelihood can be approximated as in section 1 by

$$l_t^S(\theta) \equiv \frac{1}{Sh} \sum_{s=1}^{S} K\left(\frac{y_t - z_t^s(\theta)}{h}\right) \tag{2-13}$$

When $y_t = 0$, then we need to approximate not a density, but the probability that z_t is negative. We could just use the empirical proportion of negative $z_t^s(\theta)$, but that would not be smooth in the parameters. It seems far better to stick to our kernel-based approach and to define

$$l_t^S(\theta) \equiv \frac{1}{S} \sum_{s=1}^{S} \mathcal{K}\left(-\frac{z_t^s(\theta)}{h}\right) \tag{2-14}$$

where $\mathcal{K}$ is the integrated kernel

$$\mathcal{K}(u) = \int_{-\infty}^{u} K(v)dv.$$

Clearly, this estimator of the required probability converges faster than the estimator of the density. Therefore, the full simulated loglikelihood is

$$\begin{aligned}\tilde{L}_T^S(\theta) = \frac{1}{T} \sum_{t=1}^{T} \Bigg[& \ln\left(\frac{1}{Sh} \sum_{s=1}^{S} K\left(\frac{y_t - z_t^s(\theta)}{h}\right)\right) \mathbb{1}(y_t > 0) \\ + \quad & \ln\left(\frac{1}{S} \sum_{s=1}^{S} \mathcal{K}\left(-\frac{z_t^s(\theta)}{h}\right)\right) \mathbb{1}(y_t = 0) \Bigg] \tau_S(l_t^S(\theta)).\end{aligned}$$

The approach sketched above can obviously be generalized to other limited dependent variable models. Consider for instance the multinomial discrete choice model for which McFadden (1989) developed the simulated method of moments. For individual t, the utility of alternative $j = 1, \ldots, p$ is

$$z_{jt} = g_j(x_{jt}, \theta_0, \varepsilon_{jt})$$

and alternative j is chosen if and only if $z_{jt} \geq z_{kt}$ for all $k \neq j$. We simulate $z_{jt}^s(\theta)$ in the obvious way and we use the integrated kernel again to approximate the probability that individual t chooses alternative j by

$$\frac{1}{S} \sum_{s=1}^{S} \prod_{k \neq j} \mathcal{K}\left(\frac{z_{jt}^s(\theta) - z_{kt}^s(\theta)}{h}\right).$$

More generally, the same methodology applies when y_t is vector-valued and when its law conditionally to x_t has got some masses. By integrating the simulated likelihood for the latent model on some convenient subsets, we get the simulated likelihood for the observations (y_t, x_t). Theorems 1.1 and 1.2 continue to apply. Actually, the proofs of the consistency and the asymptotic normality of $\hat{\theta}_T^S$ are exactly the same as previously. The details are left to the reader. To deal with estimators of probability masses, we provide only the technical lemmas in the appendix (lemmas A.5 and A.6). Note that, when the endogenous variable y takes discrete values only like in the multinomial discrete choice model, m can be taken equal to zero. Thus, the conditions for consistency and asymptotic normality of the NPSML estimator are weaker in this case (assumptions R1 and R2 especially).

3 NPSML for the Dynamic Case

As mentioned before, models in which x_t contains lagged observable endogenous variables can be treated in exactly the same way (provided that ε_t is not serially correlated). However, there is a class of models for which dynamic simulations are called for. This includes

- models with both lagged observable endogenous variables and serially correlated disturbances
- models with lagged latent variables.

An example of the latter is the Tobit model with an autoregressive latent variable process (from now on the "autoregressive Tobit model"), which can be written (in its simplest form)

$$\begin{cases} y_t = \max(0, z_t) \\ z_t = a + bz_{t-1} + \sigma\varepsilon_t, \end{cases}$$

where y_t represents the observed endogenous variable, z_t is the latent variable, and ε_t is i.i.d. normal with zero mean and unit variance.

In these models, the likelihood function for observation t is a t-dimensional integral, which can very rarely be computed in closed-form. For very simple instances of these models, it is possible to apply clever tricks to use simulated maximum-likelihood or the method of simulated scores, but there exists as yet no fully general method. As we shall now see, it is possible to extend the NPSML method to these models in order to obtain a consistent and asymptotically normal estimator, with some loss in asymptotic efficiency.

We generate dynamic simulations of the autoregressive Tobit model as follows. Denote $\theta = (a, b, \sigma)$. For each t, draw ε_t^s in the assumed distribution of ε_t for $s = 1 \ldots, S$. Also draw $z_0^s(\theta)$ from the stationary distribution of the latent variable process implied by θ. Then compute recursively

$$\begin{cases} z_t^s(\theta) = a + bz_{t-1}^s(\theta) + \sigma\varepsilon_t^s \\ y_t^s(\theta) = \max(0, z_t^s(\theta)) \end{cases}$$

Given the simulated paths of the observable endogenous variables $y_t^s(\theta)$, we could proceed exactly as for the static model. This would approximate the marginals

$$l(y_t;\theta) = \int_{y_1} \dots \int_{y_{t-1}} l(y_t,\dots,y_1;\theta)$$

of the likelihood function. If the parameters are identifiable from the marginals, then given technical conditions, the maximum-likelihood estimator based on the marginals is consistent. This follows from applying Jensen's inequality to the Kullback criterion. To see this, denote E_0 the expectation with respect to the true marginal $l(y_t;\theta_0)$. Then

$$E_0 \log \frac{l(y_t;\theta)}{l(y_t;\theta_0)} \leq \log E_0 \frac{l(y_t;\theta)}{l(y_t;\theta_0)} = 0$$

so that $E_0 \log l(y_t;\theta)$ is maximal for $\theta = \theta_0$. However, this estimator uses only a small part of the information contained in the sample under study. In fact, it is easy to see that in the autoregressive Tobit model, θ is not identifiable from the marginals, as the stationary distribution of z_t is

$$N\left(\frac{a}{1-b}, \frac{\sigma^2}{1-b^2}\right)$$

which only identifies two combinations of the three parameters.

On the other hand, it is easy to see that in the autoregressive Tobit model, θ is identified from the second-order marginal $l(y_t, y_{t-1};\theta)$. This suggests an approach due to Azzalini (1983) and also applied by Laroque-Salanié (1993): we generalize our earlier procedure by choosing some integer $k > 0$ and approximating the likelihood of $(y_t,\dots,y_{t-k})$ (in a continuous model here) by

$$l_t^S(\theta) = \frac{1}{Sh^{k+1}} \sum_{s=1}^{S} K\left(\frac{y_t - y_t^s(\theta)}{h}, \dots, \frac{y_{t-k} - y_{t-k}^s(\theta)}{h}\right)$$

where K is a well-chosen $k+1$-dimensional kernel and h is a bandwidth [11]. This will allow us to approximate the marginals $l(y_t,\dots,y_{t-k};\theta)$ of the likelihood function, conditional to $(x_t,\dots,x_{t-k})$. The NPSML estimator $\hat{\theta}_T^S$ then is obtained as usual by maximizing

$$\sum_{t=1}^{T} \tau_S(l_t^S(\theta)) \ln l_t^S(\theta)$$

Note that we could also approximate the conditional form of the $(k+1)$-order likelihood

$$\log l(y_t|y_{t-1},\dots,y_{t-k}) = \log l(y_t, y_{t-1},\dots,y_{t-k}) - \log l(y_{t-1},\dots,y_{t-k}).$$

Another interesting variant which will be used in our Monte Carlo simulation circumvents the curse of dimensionality by maximizing

$$\sum_{i=1}^{k} \log l^S(y_t, y_{t-i}).$$

[11] Obviously, we could choose a different bandwidth for each marginal.

We cannot claim asymptotic efficiency with any of these dynamic methods, since we are only approximating $(k+1)$-order marginals of the likelihood function. Nevertheless, the NPSML estimator will be close to asymptotically efficient if these marginals contain about as much information as the full likelihood function. Of course, the curse of dimensionality will severely limit the possible values of k, to, say, 1 or 2. Therefore it is an empirical question whether the efficiency loss is large or not in any particular model.

This procedure clearly extends to more general models whose reduced form can be written as

$$z_t = g(x_t, z_{t-1}, \theta_0, u_t),$$

where the vector of endogenous variables z_t may contain

- observable endogenous variables y_t
- latent endogenous variables y_t^*
- disturbances ε_t.

Now u_t represents the innovations in the disturbances and is still assumed to be drawn from a known distribution $\mathcal{L}$. For instance, we might have

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \sigma u_t$$

where both ρ_1 and σ are parameters to be estimated. Dynamic simulations then are obtained, given an initializing scheme for $z_0^s(\theta)$, by computing for $s = 1, \ldots, S$ and $t = 1 \ldots, T$

$$z_t^s(\theta) = g(x_t, z_{t-1}^s(\theta), \theta, u_t^s),$$

where the u_t^s are drawn from $\mathcal{L}$. This generates the $y_t^s(\theta)$ that are the basis for the dynamic NPSML method.

Theorems 1.1 and 1.2 can be adapted to the dynamic framework; one may also prove that when the number of lags k tends to infinity at the right speed when T goes to infinity, then the NPSML estimator is asymptotically efficient. Indeed, the key tool of the proofs is lemma A.1, which can be extended to the dynamic case. There are two main caveats here. First, θ_0 must be identifiable from the marginal likelihood function $l(y_t, \ldots, y_{t-k}; \theta)$. There is clearly no general argument as to what value of k will achieve identification; this must be checked for each specific model[12]. Second, the variance-covariance matrix of the estimator must be computed as

$$V = J^{-1} I J^{-1}$$

where J is the derivative of the score and I is the outer product of the scores, corrected for serial correlation. Moreover, the notation becomes messier and the assumptions must be strengthened. There are in fact so many technical differences between the static and the dynamic cases that another paper seems to be necessary to state and prove the asymptotic properties of the dynamic NPSML estimators. Therefore we do not include the proof here. It seems more interesting to examine how well these methods perform in practice. To this end, we now turn to a Monte-Carlo simulation study of the finite-sample properties of the dynamic NPSML estimator as applied to the autoregressive Tobit model.

[12] Of course, identification always obtains for large enough T if we let k go to infinity.

Table 1: **Simulation on time series**

Parameter	True value	NPSML ($S = 50$)	NPSML ($S = 500$)
a	0.000	-0.017	-0.010
		(0.113)	(0.215)
b	0.500	0.544	0.510
		(0.184)	(0.151)
σ	1.000	0.747	0.810
		(0.133)	(0.184)

4 A Monte-Carlo Simulation Experiment

Latent variable models in which the latent variable is serially correlated are a natural area for applying NPSML, since they usually give rise to analytically untractable likelihood functions. We study here the properties of the NPSML estimators of the autoregressive Tobit model discussed above, both on time series and on panel data.

There are some choices to be made when applying the NPSML method. We used the normal kernel to nonparametrically estimate the density and we did not prewhiten the data. We chose the bandwidth h by applying Silverman's rule for the bandwidth that minimizes the mean integrated square error for a normal density. We trimmed the 5% smallest values of the likelihood. These choices could presumably be improved upon in a real-life application, but they seem to be a reasonable starting-point.

The autoregressive Tobit model is defined on time series by

$$\begin{cases} y_t = \max(0, z_t) \\ z_t = a + bz_{t-1} + \sigma\varepsilon_t, \end{cases}$$

where the true values of the parameters are $a = 0$, $b = 0.5$ and $\sigma = 1$, and ε_t is i.i.d. normal with zero mean and unit variance. This model can describe for instance the evolution of prices on a market with a regulated price floor (as with some farm products in developed economies).

We ran a Monte-Carlo experiment for $k = 1$. We chose a sample size of $T = 150$ and generated 500 samples of artificial data. The initial parameter values were drawn from the uniform distribution on $[-0.2, 0.2]$ for a, from the uniform distribution on $[0.25, 0.75]$ for b, and from the uniform distribution on $[0.5, 1.5]$ for σ. For each parameter vector reached during maximization, the initial value of the latent variable z_0 was drawn from the stationary distribution implied by these parameter values.

Table 1 summarizes the results for $S = 50$ and $S = 500$. The estimation takes on average two seconds for $S = 50$ and seven seconds for $S = 500$ on a Pentium 1.4 GHz PC using Gauss[13]. For each parameter, the table gives the mean estimate and

[13]All the results we give in this section neglect the few simulation runs (less than 3%) where the maximization algorithm strayed in a region of very low likelihood values; then we stopped the algorithm before taking the logarithm of the likelihood.

Table 2: **Simulation on panel data**

		$T = 2$		$T = 10$	
Parameter	True value	$S = 50$	$S = 500$	$S = 50$	$S = 500$
a	0.000	−0.013	0.031	−0.010	−0.018
		(0.014)	(0.018)	(0.008)	(0.011)
b	0.500	0.613	0.546	0.655	0.534
		(0.065)	(0.044)	(0.027)	(0.017)
σ	1.000	0.731	0.926	0.800	0.892
		(0.043)	(0.030)	(0.021)	(0.013)

Table 3: **Panel data with a random effect**

Parameter	True value	NPSML
a	0.000	−0.036
		(0.017)
b	0.500	0.585
		(0.095)
σ	1.000	0.858
		(0.020)
σ_v	0.500	0.311
		(0.121)

(between parentheses) the dispersion of the estimates. The results seem satisfactory for a and b, especially with $S = 500$. There is some downwards bias for σ, which may be due to our trimming too many observations or to too small S values that induce some not negligible biases in kernel estimates. Again, this improves with larger S.

Of course, most applications of the Tobit model use panel data. NPSML can equally well be applied to such data, although this time the asymptotic theory applies for large I (the number of individuals) and fixed T. To test the properties of NPSML on panel data, we use a sample of sizes $I = 1000$ and $T = 2$ or $T = 10$. Again, we use 500 artificial samples, except for $T = 10$ and $S = 500$, for which we only generated 250 samples[14]. The results are summarized in Table 2. We find again a (smaller) downwards bias on σ, and an upwards bias on b. These biases are a bit too large for $S = 50$, but they become acceptable for $S = 500$.

To conclude this illustrative Monte Carlo study, we add a random individual effect in the autoregressive Tobit model. Thus our model is now

$$\begin{cases} y_{it} = \max(0, z_{it}) \\ z_{it} = a + b z_{i,t-1} + v_i + \sigma \varepsilon_{it}, \end{cases}$$

where the parameters are as before except that v_i is drawn from a centered normal

[14]Each estimation takes about 30 seconds for $T = 2$ and $S = 50$, but 15 minutes for $T = 10$ and $S = 500$.

Table 4: **Simulation of the Stochastic Volatility Model**

Parameter	True value	QML	MCL	MCMC	NPSML
σ_e	0.260	0.302	0.233	0.280	0.318
		(0.17)	(0.07)	(0.07)	(0.17)
ϕ	0.950	0.906	0.930	0.920	0.913
		(0.18)	(0.10)	(0.05)	(0.10)
σ_b	0.025	–	–	–	0.022
		–	–	–	(0.003)

distribution with true standard error $\sigma_v = 0.5$. This is a much more challenging model, as we have to disentangle the serial correlation induced by $b \neq 0$ and that induced by the v_i, while only observing y_{it}. In fact, it is easy to see that the model is only identifiable for $T \geq 3$. A consequence is that we now need to choose $k \geq 2$ in our NPSML estimation. For experimental purposes, we took $T = 10$ and we maximized the simulated analog of

$$\sum_{i=1}^{I} \sum_{t=k+1}^{T} \sum_{j=1}^{k} \log l(y_{it}, y_{i,t-j})$$

using $k = 5$ and $S = 500$ simulations[15]. Each estimation now takes more than an hour, so we only used 100 artificial samples. The results are collected in Table 3. It seems to be difficult to get a good estimate of the variance of the random effect. On the other hand, the performance of the method for the estimation of a and b is similar to that exhibited in Table 2.

5 Concluding Remarks

While the NPSML method has very appealing statistical properties in static models, its real test lies in its ability to give good estimates in finite samples for dynamic models. The method works reasonably well for the autoregressive Tobit model, which is a tough test for any estimation method. The most remarkable aspect of NPSML is its generality: the method is essentially the same for all econometric models. It is very easy to implement: it takes less than 40 program lines to define the function to be maximized, and only about ten lines need to be changed when estimating a different model. Of course, such a general estimation method cannot compete with methods that exploit the particular features of a specific model, but it seems to offer a good product of performance and ease of use. Consider for instance the stochastic volatility model often used on financial data:

$$\begin{cases} y_t = \sigma_b \exp(y_t^*/2)\varepsilon_{1t} \\ y_t^* = \phi y_{t-1}^* + \sigma_e \varepsilon_{2t}, \end{cases}$$

[15] We took the initial parameter value for σ_v to be distributed uniformly from 0.2 to 0.7.

where y_t represents the observed returns and y_t^* the latent volatility, and $(\varepsilon_{1t}, \varepsilon_{2t})$ are jointly normal i.i.d. with zero mean and unit variance. Following case 5 of Jacquier-Polson-Rossi (1994), we study the model with the crucial parameter ϕ set at 0.95. Table 4 gives the results of running NPSML and three alternative estimation methods[16], in the variant in which we maximize

$$\sum_{t=11}^{T} \sum_{i=1}^{10} \log l(y_t, y_{t-i})$$

on a 500 observation sample with 500 simulations, and 500 Monte Carlo runs[17]. The estimates from NPSML appear to have roughly comparable performance to that of QML, MCL and MCMC, while these alternative estimation methods are much more difficult to implement. Presumably, the quality of NPSML estimates could be improved at little cost by adding correlations of a higher order; we haven't explored that avenue.

References

Ai, C. (1997), "A Semiparametric Maximum Likelihood Estimator", *Econometrica*, 65, 933-963.

Azzalini, A. (1983), "Maximum Likelihood Estimation of Order m for Stationary Stochastic Processes", *Biometrika*, 70, 381–387.

Bickel, P. (1982), "On Adaptive Estimation", *Annals of Statistics*, 70, 647–671.

Börsch-Supan, A. and V. Hajivassiliou (1993), "Smooth Unbiased Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models", *Journal of Econometrics*, 58, 347-368.

Bosq, D. and J.-P. Lecoutre (1987), *Théorie de l'estimation fonctionnelle*, Economica, Paris.

Carrasco, M. and J-P. Florens (2002a), "Efficient GMM Estimation Using the Empirical Characteristic Function", *mimeo*.

Carrasco, M. and J-P. Florens (2002b), "Simulation-based Method of Moments and Efficiency", *Journal of Business and Economic Statistics*, 20, 482-492.

Diggle, P. and R. Gratton (1984), "Monte Carlo Methods of Inference for Implicit Statistical Models", *Journal of the Royal Statistical Society B*, 46, 193-227.

Feuerverger, A. and P. McDunnough (1981a), "On Some Fourier methods for inference", *Journal of the Royal Statistical Society B*, 43, 20-27.

Feuerverger, A. and P. McDunnough (1981b), "On the Efficiency of Empirical Characteristic Function Procedures", *Journal of the American Statistical Association*, 78, 379-387.

[16]The simulation results for these three methods are taken from Jacquier-Polson-Rossi (1994) and Sandmann-Koopman (1998).

[17]Initial values are randomly drawn from the uniform distributions on $[0.15, 0.35]$ for σ_e, $[0.9, 0.99]$ for ϕ, and $[0.015, 0.035]$ for σ_b. Each estimation takes about 50 seconds; 16 simulation runs were discarded.

Gallant, R. and D. Nychka (1987), "Semi-nonparametric Maximum Likelihood Estimation", *Econometrica*, 55, 363-390.

Gallant, R. and G. Tauchen (1996), "Which Moments to Match?", *Econometric Theory*, 12, 657-681.

Gouriéroux, C. and A. Monfort (1996), *Simulation-based Econometric Methods*, Oxford University Press.

Gouriéroux, C., A. Monfort and E. Renault (1993), "Indirect Inference", *Journal of Applied Econometrics*, 8, S85-S118.

Hajivassiliou, V. and D. McFadden (1998), "The Method of Simulated Scores for the Estimation of Limited-Dependent Variable Models", *Econometrica*, 66, 863-896.

Hajivassiliou, V. and P. Ruud (1994), "Classical Estimation Methods for LDV Models Using Simulation", in R. Engle and D. McFadden eds, *Handbook of Econometrics, vol 4*, Elsevier.

Jacquier, E., N. Polson and P. Rossi (1994), "Bayesian Analysis of Stochastic Volatility Models", *Journal of Business and Economic Statistics*, 12, 371-417.

Laroque, G. and B. Salanié (1989), "Estimation of Multimarket Fix-price Models: An Application of Pseudo-Maximum Likelihood methods", *Econometrica*, 57, 831-860.

Laroque, G. and B. Salanié (1993), "Simulation-based Estimation of Models with Lagged Latent Variables", *Journal of Applied Econometrics*, 8, S119-S133.

Laroque, G. and B. Salanié (1994), "Estimating the Canonical Disequilibrium Model: Asymptotic Theory and Finite Sample Properties", *Journal of Econometrics*, 62, 165-210.

Lee, L.F. (1995), "Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models", *Econometric Theory*, 11, 437-483.

Lerman, S. and C. Manski (1981), "On the Use of Simulated Frequencies to Approximate Choice Probabilities", in C. Manski & D. McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press.

McFadden, D. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration", *Econometrica* , 57, 995-1026.

Pakes, A. (1986), "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks", *Econometrica*, 54, 755-84.

Pakes, A. and D. Pollard (1989), "Simulation and the Asymptotics of Optimization Estimators", *Econometrica*, 57, 1027-1057.

Sandmann, G. and S. Koopman (1998), "Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood", *Journal of Econometrics*, 87, 271-301.

Smith, A. (1993), "Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions", *Journal of Applied Econometrics*, 8, S63-S84.

Stern, S. (1997), "Simulation-based Estimation", *Journal of Economic Literature*, 35, 2006-2039.

Stone, C. (1975), "Adaptive Maximum Likelihood Estimation of a Location Parameter", *Annals of Statistics*, 3, 267–284.

Appendix: Proofs of the Asymptotic Results

Denote by Cst some "universal" positive constants (viz independent from every other quantities). By default, the expectations are taken with respects to the true law whose parameter is θ_0.

A Technical lemmas

The proofs of Theorems 1.1 and 1.2 rely on asymptotic convergence results of many kernel-based estimates. To establish them, we will repeatedly use an improved version of lemma $B.1$ of Ai (1997). The rate of convergence is a bit higher than in Ai's lemma and the result is true almost everywhere under the additional assumption (A-15) below.

Lemma A.1 *Let $(u_i)_{i\geq 1}$ be an i.i.d. sequence of realizations of a random variable u, and denote*

$$a_N(w) = N^{-1} \sum_{i=1}^{N} a_N(w, u_i)$$

a sample average of some real terms $a_N(w, u_i)$, $w \in \mathbb{R}^k$. Let h_N be a bandwidth sequence such that $h_N \to 0$ when $N \to \infty$ and such that $h_N > N^{-\pi}$ for some $\pi > 0$. Assume that for every (u, w, N, h_N),

i. $h_N^r |a_N(w, u)| < c_1(u)$ *and* $E[c_1^p(u)] < +\infty$ *for some* $r \geq 0$ *and* $p > 2$,

ii. $h_N^s \|\partial a_N(w, u)/\partial w\| < c_2(u)$ *and* $E[c_2(u)] < +\infty$ *for some* $s \geq 0$,

iii. $E[h_N^{2r} a_N^2(w, u)] \leq Cst.h_N^t$ *for some* $t \geq 0$.

Define $W_N = \{w \in \mathbb{R}^k; \|w\| \leq N^\nu\}$, $\nu > 0$. If

$$\sum_{N\geq 1} \left(\frac{\ln N}{N}\right)^{p/2-1} h_N^{-tp/2} < +\infty, \tag{A-15}$$

then there exists a constant C_0 such that almost everywhere, for every N,

$$\left(\frac{Nh^{2r-t}}{\ln N}\right)^{1/2} \mathbb{1}(w \in W_N) \, |a_N(w) - E[a_N(w)]| \leq C_0. \tag{A-16}$$

Moreover, replacing assumption (A-15) with

$$Nh_N^{tp/(p-2)} / \ln N \underset{N\to\infty}{\longrightarrow} +\infty, \tag{A-17}$$

a stronger result is true in probability, viz for every $\varepsilon > 0$,

$$P\left(\left(\frac{Nh^{2r-t}}{\ln N}\right)^{1/2} \mathbb{1}(w \in W_N) \, |a_N(w) - E[a_N(w)]| > \varepsilon\right) \underset{N\to\infty}{\longrightarrow} 0. \tag{A-18}$$

Proof of lemma A.1: To simplify the notation, we suppress most N subscripts from now on. The technique of proof is exactly the same as in Ai (1997) even if our result is a bit stronger and is stated a.e. Note that a.e. $\sup_N |N^{-1}\sum_{i=1}^N c_2(u_i)|$ is bounded. For some $M_N > 0$, define $d_i = \mathbb{1}(c_1(u_i) \leq M_N)$. Then $a(w) = a_1(w) + a_2(w)$ where $a_1(w)$ is a sample average of terms $(1-d_i)a(w,u_i)$ and $a_2(w)$ is a sample average of terms $d_i a(w,u_i)$, $i = 1,\ldots,N$. Then

$$P\left(\sup_{w\in W_N} |a(w) - E[a(w)]| > \varepsilon\right) \leq P\left(\sup_{w\in W_N} |a_1(w) - E[a_1(w)]| > \varepsilon/2\right)$$
$$+ \quad P\left(\sup_{w\in W_N} |a_2(w) - E[a_2(w)]| > \varepsilon/2\right) \equiv p_1 + p_2.$$

Invoking condition i, we get

$$\begin{aligned} p_1 &\leq P\left(\sup_{w\in W_N} \frac{1}{N}\sum_{i=1}^N |(1-d_i)a(w,u_i)| > \frac{\varepsilon}{4}\right) \\ &+ \quad P\left(\sup_{w\in W_N} E[(1-d_i)|a(w,u_1)|] > \frac{\varepsilon}{4}\right) \\ &\leq \quad P\left(\frac{1}{N}\sum_{i=1}^N (1-d_i)c_1(u_i) > h^r\frac{\varepsilon}{4}\right) \\ &+ \quad P\left(E\left[(1-d_i)c_1(u_i)\right] > h^r\frac{\varepsilon}{4}\right). \end{aligned} \tag{A-19}$$

By Hölder's inequality, we have

$$E[(1-d_i)c_1(u_i)] \leq P(c_1(u_i) > M_N)^{1-1/p} \cdot E[c_1^p(u_i)]^{1/p} \leq E[c_1^p(u_i)]/M_N^{p-1}.$$

Therefore, the second term of equation (A-19) is zero for N sufficiently large if $M_N^{p-1}\varepsilon h^r$ tends to the infinity when N tends to the infinity. This assumption will be satisfied with our forthcoming choices (see below). Thus we obtain for N sufficiently large

$$p_1 \leq \frac{4E[(1-d_i)c_1(u_i)]}{\varepsilon h^r} = O\left(\frac{1}{M_N^{p-1}\varepsilon h^r}\right).$$

Moreover, cover W_N classically by b_N boxes W_{jN}, $j = 1,\ldots,b_N$ of length δ_N. It is easy to choose the boxes so that $b_N \sim N^{\nu k}/\delta_N^k$. Denote w_j be the center of each box W_{jN}. If $w \in W_{jN}$, we get by assumption ii that

$$|a_2(w) - a_2(w_j)| \leq \frac{\|w - w_j\|}{Nh_N^s}\sum_{i=1}^N c_2(u_i) \leq \frac{C_1\delta_N}{h_N^s} \text{ a.e. },$$

for some constant C_1. Deduce that a.e.

$$\begin{aligned}\sup_{w\in W_N}|a_2(w)-E[a_2(w)]| &\le \max_{1\le j\le b_N}\Bigg[\sup_{w\in W_{jN}}|a_2(w)-a_2(w_j)|\\ &+ \sup_{w\in W_{jN}}|E[a_2(w)-a_2(w_j)]|+|a_2(w_j)-E[a_2(w_j)]|\Bigg]\\ &\le \frac{2C_1\delta_N}{h_N^s}+\sup_j|a_2(w_j)-E[a_2(w_j)]|.\end{aligned}$$

Applying Bernstein's inequality, we get

$$\begin{aligned}p_2 &\le P\left(\frac{2C_1\delta_N}{h_N^s}>\varepsilon/4\right)+b_N\sup_{1\le j\le b_N}P\left(|a_2(w_j)-E[a_2(w_j)]|>\varepsilon/4\right)\\ &\le P\left(C_1\frac{\delta_N}{h^s}>\varepsilon/8\right)+2b_N\exp\left(-\frac{Nh^{2r}\varepsilon^2}{C_2E[h^{2r}a_N^2(w,u)]+16M_N\varepsilon h^r}\right)\\ &\le P\left(C_1\frac{\delta_N}{h^s}>\varepsilon/8\right)+2b_N\exp\left(-\frac{Nh^{2r-t}\varepsilon^2}{C_3+16M_N\varepsilon h^{r-t}}\right)\end{aligned}$$

for some positive constants C_1, C_2 and C_3. Choosing $\varepsilon^2 = C^*h^{t-2r}\ln N/N$ and $M_N = h^{t-r}\varepsilon^{-1}$, it is easy to verify that $M_N \underset{N\to\infty}{\longrightarrow} +\infty$ under assumption (A-15). Moreover, if $\delta_N = [\varepsilon h^s/(8C_1)]\wedge 1$, then $b_N = O(N^{\bar\pi})$, $\bar\pi>0$ and

$$p_2 \le 0 + Cst.N^{\bar\pi}\exp(-Cst.C^*\ln N).$$

Thus, for C^* sufficiently large, $\sum_N p_2 < +\infty$. At last, since $M_N^{p-1}\varepsilon h^r = (\ln N/N)^{1-p/2}h^{tp/2}$, assumption (A-15) implies that $\sum_N p_1 < \infty$. Then, by Borel-Cantelli's lemma the strong uniform convergence is proved.

To state the convergence in probability (equation (A-18)), it is sufficient to prove that p_1 and p_2 tend to zero when N tends to the infinity. This is the case with our previous choices (ε, M_N) and under (A-17). □

Lemma A.1 allows us to state the strong consistency of kernel estimates uniformly with respects to the parameter θ and to some increasing compact sets of observations. It will be used repeatedly in the following three lemmas.

Lemma A.2 *Under assumptions K0, M1, L2 and R1, for every $\nu>0$, a.e.*

$$\inf\left(h^{-\rho},\left(\frac{Sh^m}{\ln S}\right)^{1/2}\right)\mathbb{1}(\|x,y\|\le S^\nu)\sup_\theta\left|l^S(y|x,\theta)-l(y|x,\theta)\right|$$

is bounded.

Proof of lemma A.2: Apply lemma A.1 with $w=(x,y,\theta)$, $u=\varepsilon$, $N=S$, $k=m+d+q$ and

$$a_N(w,u) = h^{-m}K\left(\frac{y-g(x,\theta,\varepsilon)}{h}\right).$$

Since K is bounded, we can choose $r = m$ and p arbitrarily large. Moreover

$$E[h^{2m} a_N^2(w,u)] = \int K^2\left(\frac{y - g(x,\theta,\varepsilon)}{h}\right) dP_\varepsilon = h^m \int K^2(u) l(y - hu|x,\theta)\, du.$$

By assumption L2, we can choose $t = m$. Thus, it is easy to check that assumption R1 implies A-15. Moreover

$$\begin{aligned}\left\| h^{m+1+s_0} \frac{\partial a_N(w,u)}{\partial w'} \right\| &= h^{s_0} \left\| K'\left(\frac{y - g(x,\theta,\varepsilon)}{h}\right) \cdot \left[-\frac{\partial g(x,\theta,\varepsilon)}{\partial x'}, 1, -\frac{\partial g(x,\theta,\varepsilon)}{\partial \theta'} \right] \right\| \\ &\leq \quad Cst \|K'\|_\infty (1 + \phi(\varepsilon))\end{aligned}$$

belongs to L^1. Hence, under R1,

$$\left(\frac{S h^m}{\ln S}\right)^{1/2} \mathbb{1}(\|x,y\| \leq S^\nu) \sup_\theta \left\| l^S(y|x,\theta) - E[l^S(y|x,\theta)] \right\|$$

is bounded a.e. It remains to deal with the bias term. A Taylor expansion provides

$$E[l^S(y|x,\theta)] = l(y|x,\theta) + \frac{(-h)^\rho}{\rho!} \int \frac{\partial^\rho l(y - \theta_t^* h t|x,\theta)}{\partial^\rho y} K(t) t^\rho \, dt,$$

where $\theta_t^* \in [0,1]$. Since $d^\rho l(\cdot|x,\theta)$ is uniformly bounded (assumption L2),

$$\sup_{(x,y,\theta)} h^{-\rho} \left| E[l^S(y|x,\theta)] - l(y|x,\theta) \right|$$

is bounded, proving the result. □

Lemma A.3 *Under assumptions K0, M2 and L3, for every $\nu > 0$, we have*

$$\inf\left(h^{-\rho}, \left(\frac{S h^{2m+2+2r_0}}{\ln S}\right)^{1/2} \right) \mathbb{1}(\|x,y\| \leq S^\nu) \sup_\theta \left\| \frac{\partial l^S(y|x,\theta)}{\partial \theta} - \frac{\partial l(y|x,\theta)}{\partial \theta} \right\|$$

is bounded a.e.

Proof of lemma A.3: Apply lemma A.1 with $w = (x,y,\theta)$, $u = \varepsilon$, $N = S$, $k = m + d + q$ and

$$\begin{aligned} a_N(w,u) &= h^{-m} \frac{\partial}{\partial \theta_k} K\left(\frac{y - g(x,\theta,\varepsilon)}{h})\right) \\ &= \quad -h^{-m-1} \frac{\partial g(x,\theta,\varepsilon)}{\partial \theta_k} K'\left(\frac{y - g(x,\theta,\varepsilon)}{h}\right), \end{aligned}$$

for each $k = 1, \ldots, q$. Set $r = m + 1 + r_0$ and $s = m + 2$. Since we can impose $\|x\| \leq S^\nu$,

$$h^{m+1+r_0} |a_N(w,u_i)| \leq \|K'\|_\infty \bar{\phi}(\varepsilon_i),$$

and $E[\bar{\phi}(\varepsilon_i)^{p_0}] < \infty$, $p_0 > 2$, lemma A.1 is valid with $p = p_0$. Moreover,

$$\begin{aligned} E[h^{2m+2+2r_0} a_N^2(w,u)] &= h^{2r_0} \int K'\left(\frac{y-g(x,\theta,\varepsilon)}{h}\right)^2 \left(\frac{\partial g(x,\theta,\varepsilon)}{\partial \theta_k}\right)^2 dP_\varepsilon \\ &\leq \|K'\|_\infty^2 E[\bar{\phi}(\varepsilon)^2] < \infty. \end{aligned}$$

Then, set $t = 0$. Clearly, assumption ii of lemma A.1 is satisfied with $s = m + 2 + s_1$. Hence, since $p_0 > 4$, we have

$$\sum_{S \geq 1} \left(\frac{\ln S}{S}\right)^{p_0/2-1} < +\infty, \tag{A-20}$$

Then (A-15) is satisfied and

$$\left(\frac{h^{2m+2+2r_0} S}{\ln S}\right)^{1/2} \mathbf{1}(\|x,y\| \leq S^\nu) \sup_\theta \left\| \frac{\partial l^S(y|x,\theta)}{\partial \theta} - E[\frac{\partial l^S(y|x,\theta)}{\partial \theta}] \right\|$$

is bounded a.e. To deal with the bias, a Taylor expansion provides as previously

$$E[\frac{\partial l^S(y|x,\theta)}{\partial \theta}] = \frac{\partial l(y|x,\theta)}{\partial \theta} + \frac{(-h)^\rho}{\rho!} \int \frac{\partial^{\rho+1}}{\partial \theta \partial^\rho y} l(y - \theta_t^* ht|x,\theta) K(t) t^\rho \, dt,$$

where $\theta_t^* \in [0,1]$. Since $\partial^{\rho+1} l(y|x,\theta)/\partial\theta\partial^\rho y$ is uniformly bounded,

$$\sup_{(x,y,\theta)} h^{-\rho} \left\| E[\frac{\partial l^S(y|x,\theta)}{\partial \theta}] - \frac{\partial l(y|x,\theta)}{\partial \theta} \right\| \text{ is bounded,}$$

proving the result. □

Lemma A.4 *Under (1-8), we have a.e.*

$$\left(\frac{T}{\ln T}\right)^{1/2} \sup_\theta \left\{ \frac{1}{T} \sum_{t=1}^T [1 - \tau_S(l_t(\theta))] - E[1 - \tau_S(l_t(\theta))] \right\} \text{ is bounded .} \tag{A-21}$$

Proof of lemma A.4: Apply lemma A.1 with $w = \theta$, $u = (x,y)$, $N = T$, $k = q$ and

$$a_N(\theta) = T^{-1} \sum_{t=1}^T a_N(\theta, u_t), \qquad a_N(\theta, u_t) = [1 - \tau_S(l_t(\theta))] .$$

Recall that $h = h(T)$ is a sequence which tends to zero when $T \longrightarrow +\infty$. Since a_N is bounded, choose $r = 0$ and p arbitrarily large. Moreover, we can choose $s = \delta$ since

$$\left\| \frac{\partial a_N(\theta,u)}{\partial \theta} \right\| = \left\| \tau_S'(l(y|x,\theta)) \frac{\partial l(y|x,\theta)}{\partial \theta} \right\| \leq Cst.h^{-\delta} \sup_\theta \left\| \frac{\partial l(y|x,\theta)}{\partial \theta} \right\| .$$

Finally, note that $E[h^{2r} a_N^2(\theta,u)] = O(1)$ and set $t = 0$. Then (A-15) is satisfied and a.e.

$$\left(\frac{T}{\ln T}\right)^{1/2} \sup_\theta \left\{ T^{-1} \sum_{t=1}^T [1 - \tau_S(l_t(\theta))] - E[(1 - \tau_S(l_t(\theta)))] \right\} \text{ is bounded ,} \tag{A-22}$$

proving the result. □

Now, let us assume that y_t takes only some discrete values $(a_1, \ldots, a_r)$, and there exists a one-dimensional latent variable Z_t such that, for every $k = 1, \ldots, r$,

$$P(Y_t = a_k | x_t, \theta) = P(Z_t \in I_k(x_t, \theta)),$$

for some intervals $I_k(x_t, \theta) = (\alpha_k(x_t, \theta), \beta_k(x_t, \theta))$. Obviously, the latent variable Z_t has a density with respects to the Lebesgue measure. Moreover, the underlying latent model now is

$$z_t = g(x_t, \theta, \varepsilon_t). \tag{A-23}$$

This allows us to simulate some samples $(z_t^s(\theta))_{s=1,\ldots,S}$, with the previous notations. We consider the simulated likelihood

$$\begin{aligned} \bar{l}^S(a_k|x_t, \theta) = \frac{1}{S} \sum_{s=1}^{S} \left\{ \mathcal{K}\left(\frac{\beta_k(x_t, \theta) - z_t^s(\theta)}{h} \right) \right. \\ \left. - \quad \mathcal{K}\left(\frac{\alpha_k(x_t, \theta) - z_t^s(\theta)}{h} \right) \right\} \approx P(Y_t = a_k | x_t, \theta), \end{aligned} \tag{A-24}$$

where $\mathcal{K}$ is an integrated kernel. Note that equation (A-24) holds even if the bounds of $I_k(x_t, \theta)$ are infinite.

Lemma A.5 *Assuming that y_t takes a finite number of values $(a_1, \ldots, a_r)$ only, under assumptions K1, M1, L2 and R1 (with $m = 0$), for every $\nu > 0$, a.e.*

$$\inf\left(h^{-\rho}, \left(\frac{S}{\ln S} \right)^{1/2} \right) \mathbb{1}(\|x\| \leq S^{\nu}) \sup_{\theta, y} \left| \bar{l}^S(y|x, \theta) - P(Y_t = y | x, \theta) \right|$$

is bounded.

Proof of lemma A.5: Recall that $Y_t = a_k$ if and only if $Z_t \in (\alpha_k(x_t, \theta), \beta_k(x_t, \theta))$. Apply lemma A.1 with $w = (x, \alpha, \beta, \theta)$, $(\alpha, \beta) \in \{(\alpha_k, \beta_k), k = 1, \ldots, r\}$, $u = \varepsilon$, $N = S$, $k = d + 2 + q$ and

$$a_N(w, u) = \mathcal{K}\left(\frac{\beta - g(x, \theta, \varepsilon)}{h} \right) - \mathcal{K}\left(\frac{\alpha - g(x, \theta, \varepsilon)}{h} \right) \equiv a_{N,1}(w, u) - a_{N,2}(w, u).$$

Since $\mathcal{K}$ is bounded, we can choose $r = 0$ and p arbitrarily large. Moreover

$$\begin{aligned} E[a_{N,1}^2(w, u)] &= \int \mathcal{K}^2\left(\frac{\beta - z}{h} \right) dP_{Z_t|x,\theta}(z) \\ &= \frac{1}{h} \int 2K\mathcal{K}\left(\frac{\beta - z}{h} \right) P(Z_t \leq z | x, \theta)\, dz, \end{aligned}$$

by an integration by parts with respects to z. Since the cdf of Z_t is bounded by one, we get $E[a_{N,1}^2(w, u)] \leq Cst$. Obviously, the same result is true for $a_N(w, u)$. Thus, set $t = 0$. As previously, we get

$$\begin{aligned} \left\| h^{1+s_0} \frac{\partial a_{N,1}(w, u)}{\partial w'} \right\| &= h^{s_0} \left\| \mathcal{K}'\left(\frac{\beta - g(x, \theta, \varepsilon)}{h} \right) \cdot \left[-\frac{\partial g(x, \theta, \varepsilon)}{\partial x'}, 1, -\frac{\partial g(x, \theta, \varepsilon)}{\partial \theta'} \right] \right\| \\ &\leq \quad Cst \|\mathcal{K}'\|_{\infty} (1 + \phi(\varepsilon)) \end{aligned}$$

belongs to L^1. Hence, under R1,

$$\left(\frac{S}{\ln S}\right)^{1/2} \mathbb{1}(\|x\| \leq S^\nu) \sup_{\theta,y} \left\|\bar{l}^S(y|x,\theta) - E[\bar{l}^S(y|x,\theta)]\right\|$$

is bounded a.e. It remains to deal with the bias term. By an integration by parts, the first term of $E[\bar{l}^S(y|x,\theta)]$ is

$$\begin{aligned}
&\frac{1}{h}\int P(Z_t \leq z|x,\theta)\, K\left(\frac{\beta - z}{h}\right) dz \\
&\quad = \quad P(Z_t \leq \beta|x_t,\theta) + \frac{(-h)^\rho}{\rho!}\int \frac{\partial^\rho l_Z(\beta - \theta_t^* ht|x,\theta)}{\partial^\rho z} K(t) t^\rho \, dt,
\end{aligned}$$

where $\theta_t^* \in [0,1]$. By assumption L2 applied to the latent likelihood l_Z, we get

$$E[\bar{l}^S(y|x,\theta)] = P(Z_t \in [\alpha,\beta]|x_t,\theta) + o(h^\rho),$$

proving the result. □

Similarly, we can easily prove

Lemma A.6 *Under assumptions K1, M2 and L3, for every $\nu > 0$, we have a.e.*

$$\inf\left(h^{-\rho}, \left(\frac{Sh^{2+2r_0}}{\ln S}\right)^{1/2}\right) \mathbb{1}(\|x\| \leq S^\nu) \sup_{\theta,y} \left\|\frac{\partial \bar{l}^S(y|x,\theta)}{\partial\theta} - \frac{\partial P(Y_t = y|x,\theta)}{\partial\theta}\right\|$$

is bounded.

B Proof of Theorem 1.1

In this proof, $\sup_\theta$ means the supremum over $\theta \in \Theta$. A simple splitting of the simulated loglikelihood provides

$$\begin{aligned}
\tilde{L}_T^S(\theta) - L_T(\theta) &= \frac{1}{T}\sum_{t=1}^T \mathbb{1}\left(\|x_t, y_t\| > S^\nu\right) \tau_S(l_t^S(\theta)) \ln l_t^S(\theta) \\
&- \quad \frac{1}{T}\sum_{t=1}^T \mathbb{1}\left(\|x_t, y_t\| > S^\nu\right) \ln l_t(\theta) \\
&+ \quad \frac{1}{T}\sum_{t=1}^T \mathbb{1}\left(\|x_t, y_t\| \leq S^\nu\right) \tau_S(l_t^S(\theta)) \left[\ln l_t^S(\theta) - \ln l_t(\theta)\right] \\
&+ \quad \frac{1}{T}\sum_{t=1}^T \mathbb{1}\left(\|x_t, y_t\| \leq S^\nu\right) \left[\tau_S(l_t^S(\theta)) - 1\right] \ln l_t(\theta) \equiv T_1 + T_2 + T_3 + T_4.
\end{aligned}$$

The proof is completed if we show that $\sup_{\theta\in\Theta} |\tilde{L}_T^S(\theta) - L_T(\theta)|$ tends to zero a.e. when S and T tend to the infinity.

Study of $T3$: Invoking lemma A.2, we have almost surely

$$\inf\left\{h^{-\rho},\left(\frac{Sh^m}{\ln S}\right)^{1/2}\right\}\mathbb{1}(\|x_t,y_t\|\leq S^\nu)\sup_\theta\left|l_t^S(\theta)-l_t(\theta)\right|$$

is bounded. Note that

$$\tau_S(l_t^S(\theta))\left|\ln l_t^S(\theta)-\ln l_t(\theta)\right|\leq\frac{1}{l_t^*(\theta)}|l_t^S(\theta)-l_t(\theta)|,$$

where $l_t^*(\theta)$ lies between $l_t^S(\theta)$ and $l_t(\theta)$. Moreover, if $l_t^S(\theta)$ tends to $l_t(\theta)$ faster than h^δ, then $\tau_S(l_t^S(\theta))>0$ implies that $|l_t^*(\theta)|\geq C.h^\delta$ for some constant C. Hence, since $\delta<\rho$ and $Sh^{m+2\delta}/\ln S\to\infty$, we have a.e. uniformly with respect to θ,

$$|T_3|\leq Cst.h^{-\delta}\left\{h^\rho\vee\left(\frac{Sh^m}{\ln S}\right)^{-1/2}\right\}\equiv O(u_S), \tag{B-25}$$

which tends to zero when $S\to\infty$.

Study of T_4: Obviously, we have

$$\begin{aligned}T_4=&\frac{1}{T}\sum_{t=1}^T\mathbb{1}\left(\|x_t,y_t\|\leq S^\nu\right)\left[\tau_S(l_t^S(\theta))-\tau_S(l_t(\theta))\right]\ln l_t(\theta)\\+&\quad\frac{1}{T}\sum_{t=1}^T\mathbb{1}\left(\|x_t,y_t\|\leq S^\nu\right)\left[\tau_S(l_t(\theta))-1\right]\ln l_t(\theta)\equiv T_{41}+T_{42}.\end{aligned}$$

Since $h^\delta\|\tau_S'\|_\infty$ is bounded, deduce from lemma A.2 that a.e.

$$\sup_\theta\sup_{x_t,y_t,\|x_t,y_t\|\leq S^\nu}|\ \tau_S(l_t^S(\theta))-\tau_S(l_t(\theta))\ |=O\left(h^{-\delta}\left\{h^\rho\vee\left(\frac{Sh^m}{\ln S}\right)^{-1/2}\right\}\right)=O(u_S),$$

which tends to zero. Hence, since $\sup_\theta T^{-1}\sum_{t=1}^T|\ln l_t(\theta)|$ is bounded a.e. as a consequence of Assumption (1-7), then a.e.

$$\sup_\theta|T_{41}|=O(u_S).\sup_\theta\frac{1}{T}\sum_{t=1}^T|\ln l_t(\theta)|=O(u_S)\underset{S\to\infty}{\longrightarrow}0.$$

Moreover, by Hölder's inequality, we have for each θ,

$$|T_{42}|\leq\left[\frac{1}{T}\sum_{t=1}^T[1-\tau_S(l_t(\theta))]\right]^{1/\alpha}\cdot\left[\frac{1}{T}\sum_{t=1}^T|\ln l_t(\theta)|^\beta\right]^{1/\beta}, \tag{B-26}$$

where $\alpha^{-1}+\beta^{-1}=1$, $\alpha>1$, $\beta>1$. Thus, by assumption (1-7) and lemma A.4, we have a.e.

$$\sup_\theta|T_{42}|\leq Cst.\left[\sup_\theta P(l_t(\theta)\leq 2h^\delta)+\left(\frac{\ln T}{T}\right)^{1/2}\right]^{1/\alpha}, \tag{B-27}$$

which tends to zero when $h \longrightarrow 0$ and $T \longrightarrow +\infty$.

Study of T_1: Note that $|\tau_S(x)\ln x| \leq \mathbb{1}(x > h^\delta)|\ln x|$ and that $l_t^S(\theta) \leq \|K\|/h^m$. Thus, since the logarithmic function is monotonic,

$$\sup_\theta \left|\tau_S(l_t^S(\theta))\ln l_t^S(\theta)\right| \leq \sup_{l_t^S(\theta)\in[h^\delta,\|K\|/h^m]} \left|\tau_S(l_t^S(\theta))\ln l_t^S(\theta)\right| \leq \left|\ln\left(\frac{\|K\|}{h^m}\right)\right| \vee |\ln h^\delta| = O(\ln h).$$

Thus,

$$\sup_\theta |T_1| \leq Cst.|\ln h| \cdot \frac{1}{T}\sum_{t=1}^T \mathbb{1}\left(\|x_t, y_t\| > S^\nu\right).$$

Using Hoeffding's inequality (Bosq and Lecoutre (1987)), for every $\varepsilon > 0$,

$$\begin{aligned} P\left(\sup_{S\leq T^\kappa}\left|\frac{1}{T}\sum_{t=1}^T \mathbb{1}\left(\|x_t, y_t\| \geq S^\nu\right) - E[\mathbb{1}(\|x_t, y_t\| > S^\nu)]\right| > \varepsilon\right) \\ \leq \quad 2T^\kappa \sup_{S\leq T^\kappa} \exp(-2T\varepsilon^2). \end{aligned}$$

By Borel-Cantelli's lemma, and setting $\varepsilon^2 = C^* \ln T/T$, it is easy to see that a.e.

$$\sup_{S\leq T^\kappa}\left|\frac{1}{T}\sum_{t=1}^T \mathbb{1}\left(\|x_t, y_t\| \geq S^\nu\right) - P_{\theta_0}\left(\|x_t, y_t\| > S^\nu\right)\right| = O\left(\left(\frac{\ln T}{T}\right)^{1/2}\right).$$

Because h is a power of T, $\ln h = O(\ln T)$. Then, deduce from assumption T1 that a.e.

$$\sup_\theta |T_1| \underset{S,T\to\infty}{\longrightarrow} 0.$$

Study of T_2: Note that, by Hölder's inequality, we have

$$|T_2| \leq \left[\frac{1}{T}\sum_{t=1}^T \mathbb{1}\left(\|x_t, y_t\| > S^\nu\right)\right]^{1/\alpha} \cdot \left[\frac{1}{T}\sum_{t=1}^T |\ln l_t(\theta)|^\beta\right]^{1/\beta}.$$

Then, invoking assumption (1-7), this term can be dealt like T_1, viz $\sup_\theta |T_2|$ tends to zero a.e. □

C Proof of Theorem 1.2

Now, we seek to state the asymptotic normality of $\hat\theta_T^S$. Note that

$$\frac{\partial L_T}{\partial\theta}(\hat\theta_T^S) = \frac{\partial L_T}{\partial\theta}(\theta_0) + \frac{\partial^2 L_T}{\partial\theta\partial\theta'}(\theta^*)(\hat\theta_T^S - \theta_0) \text{ and } \frac{\partial \tilde L_T^S}{\partial\theta}(\hat\theta_T^S) = 0,$$

where θ^* lies between θ_0 and $\hat\theta_T^S$. Thus,

$$T^{1/2}(\hat\theta_T^S - \theta_0) = \left(-\frac{\partial^2 L_T}{\partial\theta\partial\theta'}(\theta^*)\right)^{-1} \cdot \left\{T^{1/2}\frac{\partial L_T}{\partial\theta}(\theta_0) + T^{1/2}\frac{\partial(\tilde L_T^S - L_T)}{\partial\theta}(\hat\theta_T^S)\right\}. \quad \text{(C-28)}$$

The assumptions of Theorem 1.2 contain those of Theorem 1.1, so that $\hat{\theta}_S^T$ is strongly consistent. Given assumption L1, it is sufficient to prove that

$$T^{1/2}\frac{\partial(\tilde{L}_T^S - L_T)}{\partial\theta}(\theta) \tag{C-29}$$

tends to zero in probability uniformly with respect to θ belonging to a neighborhood V_0 of θ_0, or more precisely that, for every $\varepsilon > 0$,

$$P\left(\sup_{\theta\in V_0}\left\|T^{1/2}\frac{\partial(\tilde{L}_T^S - L_T)(\theta)}{\partial\theta}\right\| > \varepsilon\right) \underset{S,T\to\infty}{\longrightarrow} 0. \tag{C-30}$$

In this proof, θ belongs to V_0. Particularly, $\sup_\theta$ means $\sup_{\theta\in V_0}$. It is sufficient to verify (C-30). Some obvious calculations provide

$$\begin{aligned}
T^{1/2}\frac{\partial(\tilde{L}_T^S - L_T)(\theta)}{\partial\theta} &= T^{-1/2}\sum_{t=1}^{T}\tau_S(l_t^S(\theta))\left[\frac{\partial l_t^S(\theta)}{\partial\theta} - \frac{\partial l_t(\theta)}{\partial\theta}\right]\frac{1}{l_t^S(\theta)} \\
&+ T^{-1/2}\sum_{t=1}^{T}\tau_S(l_t^S(\theta))\frac{(l_t - l_t^S)(\theta)}{l_t^S(\theta)}\cdot\frac{\partial \ln l_t(\theta)}{\partial\theta} \\
&+ T^{-1/2}\sum_{t=1}^{T}[\tau_S(l_t^S(\theta)) - 1]\frac{\partial \ln l_t(\theta)}{\partial\theta} \\
&+ T^{-1/2}\sum_{t=1}^{T}\tau_S'(l_t^S(\theta))\frac{\partial l_t^S(\theta)}{\partial\theta}\ln l_t^S(\theta) \equiv A_1 + A_2 + A_3 + A_4.
\end{aligned}$$

Study of A_1: Note that, for every θ and every realization,

$$\left|\frac{\tau_S(l_t^S(\theta))}{l_t^S(\theta)}\right| \le h^{-\delta}.$$

Applying lemma A.3, we obtain for every $\varepsilon > 0$,

$$
\begin{aligned}
P(\sup_\theta \|A_1\| > \varepsilon) \leq\ & P\left(\sup_\theta \left\| \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| \leq S^\nu) \tau_S(l_t^S(\theta)) \left[\frac{\partial l_t^S(\theta)}{\partial \theta} - \frac{\partial l_t(\theta)}{\partial \theta}\right] \frac{1}{l_t^S(\theta)} \right\| > T^{1/2}\varepsilon/2\right) \\
+\ & P\left(\sup_\theta \left\| \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \tau_S(l_t^S(\theta)) \left[\frac{\partial l_t^S(\theta)}{\partial \theta} - \frac{\partial l_t(\theta)}{\partial \theta}\right] \frac{1}{l_t^S(\theta)} \right\| > T^{1/2}\varepsilon/2\right) \\
\leq\ & P\left(\sup_\theta \sup_{x_t, y_t, \|x_t, y_t\| \leq S^\nu} \left\| \frac{\partial l_t^S(\theta)}{\partial \theta} - \frac{\partial l_t(\theta)}{\partial \theta} \right\| > T^{-1/2} h^\delta \varepsilon/2\right) \\
+\ & P\left(\sup_\theta \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \left\| \frac{\partial l_t^S(\theta)}{\partial \theta} - \frac{\partial l_t(\theta)}{\partial \theta} \right\| > T^{1/2} h^\delta \varepsilon/2\right) \\
\leq\ & P\left(Cst \left\{ h^\rho \vee \left(\frac{\ln S}{S h^{2m+2+2r_0}}\right)^{1/2} \right\} > T^{-1/2} h^\delta \varepsilon/2\right) \\
+\ & P\left(\sup_\theta \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \left\| \frac{\partial l_t(\theta)}{\partial \theta} \right\| > T^{1/2} h^\delta \varepsilon/4\right) \\
+\ & P\left(\sup_\theta \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \left\| \frac{\partial l_t^S(\theta)}{\partial \theta} \right\| > T^{1/2} h^\delta \varepsilon/4\right) \equiv P_{11} + P_{12} + P_{13}.
\end{aligned}
$$

The first term P_{11} is zero for T sufficiently large under assumption R2. Moreover, for every positive constant C, we have

$$
\begin{aligned}
P_{12} \leq\ & P\left(\sup_\theta \left(\sum_{t=1}^T \left\| \frac{\partial l_t(\theta)}{\partial \theta} \right\|^{\gamma'}\right)^{1/\gamma'} \cdot \left(\sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu)\right)^{1-1/\gamma'} > T^{1/2} h^\delta \varepsilon/4\right) \\
\leq\ & P\left(\frac{1}{T}\sum_{t=1}^T \sup_\theta \left\| \frac{\partial l_t(\theta)}{\partial \theta} \right\|^{\gamma'} \geq C^{\gamma'}\right) \\
+\ & P\left(\left(\frac{1}{T}\sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu)\right)^{1-1/\gamma'} > T^{-1/2} h^\delta \varepsilon/(4C)\right) \\
\leq\ & \frac{E\left[\sup_\theta \left\| \frac{\partial l_t(\theta)}{\partial \theta} \right\|^{\gamma'}\right]}{C^{\gamma'}} + \frac{(4C)^{\gamma'/(\gamma'-1)} P(\|x_t, y_t\| > S^\nu)}{(T^{-1/2} h^\delta \varepsilon)^{\gamma'/(\gamma'-1)}},
\end{aligned}
$$

which tends to zero under assumptions L3 and T2. To deal with P_{13}, note that

$$
\left\| \frac{\partial l_t^S}{\partial \theta} \right\| \leq \|K'\|_\infty h^{-m-1} \cdot \frac{1}{S} \sum_{s=1}^S \left\| \frac{\partial g(x_t, \theta, \varepsilon_t^s)}{\partial \theta} \right\|.
$$

The application of assumption M2 provides

$$\begin{aligned}
P_{13} \leq P\Bigg(& \sup_\theta \left(\frac{1}{ST} \sum_{t=1}^{T} \sum_{s=1}^{S} \left\| \frac{\partial g(x_t, \theta, \varepsilon_t^s)}{\partial \theta} \right\|^\zeta \right)^{1/\zeta} \cdot \left(\frac{1}{ST} \sum_{t=1}^{T} \sum_{s=1}^{S} \mathbb{1}(\|x_t, y_t\| \right. \\
> & \left. S^\nu))^{1-1/\zeta} > T^{-1/2} h^{m+1+\delta} \varepsilon / 4 \right) \\
\leq & P\left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(\|x_t, y_t\| > S^\nu) > \left(T^{-1/2} h^{m+1+\delta} \varepsilon / (4C) \right)^{\zeta/(\zeta-1)} \right) \\
+ & P\left(\sup_\theta \left(\frac{1}{ST} \sum_{t=1}^{T} \sum_{s=1}^{S} \left\| \frac{\partial g(x_t, \theta, \varepsilon_t^s)}{\partial \theta} \right\|^\zeta \right)^{1/\zeta} > C \right) \\
\leq & \left(\frac{4CT^{1/2}}{\varepsilon h^{(m+1+\delta)}} \right)^{\zeta/(\zeta-1)} P(\|x_t, y_t\| > S^\nu) + \frac{1}{C^\zeta} E\left[\sup_\theta \left\| \frac{\partial g(x_t, \theta, \varepsilon_t^s)}{\partial \theta} \right\|^\zeta \right],
\end{aligned}$$

for some arbitrarily large constant C. Thus, P_{13} tends to zero under assumption T2.

Study of A_2: For each θ and each realization, we have

$$\begin{aligned}
\|A_2\| \leq & T^{-1/2} \sum_{t=1}^{T} |(l_t^S - l_t)(\theta)| \mathbb{1}(\|x_t, y_t\| \leq S^\nu) \left| \frac{\tau_S(l_t^S(\theta))}{l_t^S(\theta)} \right| \cdot \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\| \\
+ & T^{-1/2} \sum_{t=1}^{T} \left(|l_t^S(\theta)| + l_t(\theta) \right) \mathbb{1}(\|x_t, y_t\| > S^\nu) \left| \frac{\tau_S(l_t^S(\theta))}{l_t^S(\theta)} \right| \cdot \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\| \equiv A_{21} + A_{22}.
\end{aligned}$$

Applying lemma A.2, it is easy to see that, for all $\varepsilon > 0$,

$$\begin{aligned}
P(\sup_\theta \|A_{21}\| \geq \varepsilon) \leq & P\left(T^{-1/2} \sup_\theta \sum_{t=1}^{T} |(l_t^S - l_t)(\theta)| \mathbb{1}(\|x_t, y_t\| \leq S^\nu) \cdot \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\| > h^\delta \varepsilon \right) \\
\leq & P\left(T^{-1/2} \left\{ h^{\rho-\delta} \vee \left(\frac{\ln S}{S h^{m+2\delta}} \right)^{1/2} \right\} \sup_\theta \sum_{t=1}^{T} \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\| > Cst.\varepsilon \right) \\
\leq & P\left(\sup_\theta \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\|^\gamma \right)^{1/\gamma} > Cst.\varepsilon \lambda_T^S \right)
\end{aligned}$$

where $\lambda_T^S \to \infty$ when S and T tend to the infinity (by assumption R2). Hence, by assumption L3, we have

$$P(\sup_\theta \|A_{21}\| > \varepsilon) \underset{S,T \to \infty}{\longrightarrow} 0.$$

Moreover,

$$\begin{aligned}
\sup_\theta \|A_{22}\| &\leq \sup_\theta T^{-1/2} \sum_{t=1}^T \left[1 + Cst.h^{-\delta} l_t(\theta)\right] \mathbb{1}(\|x_t, y_t\| > S^\nu) \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\| \\
&\leq T^{1/2} \left(\frac{1}{T} \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \right)^{1-1/\gamma} \cdot \left(\sup_\theta \frac{1}{T} \sum_{t=1}^T \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\|^\gamma \right)^{1/\gamma} \\
&+ Cst.T^{1/2} h^{-\delta} \left(\frac{1}{T} \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \right)^{1-1/\gamma'} \cdot \sup_\theta \left(\frac{1}{T} \sum_{t=1}^T \left\| \frac{\partial l(y|x, \theta)}{\partial \theta} \right\|^{\gamma'} \right)^{1/\gamma'}.
\end{aligned}$$

Therefore, using assumption L3,

$$\begin{aligned}
P(\sup_\theta \|A_{22}\| > \varepsilon) &\leq P\left(\left(\frac{1}{T} \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \right)^{1-1/\gamma} > Cst.T^{-1/2}\varepsilon \right) \\
&+ P\left(\left(\frac{1}{T} \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \right)^{1-1/\gamma'} > Cst.T^{-1/2} h^\delta \varepsilon \right) + o_P(1) \\
&\leq Cst \frac{T^{\gamma/2(\gamma-1)}}{\varepsilon^{\gamma/(\gamma-1)}} P\left(\|x_t, y_t\| > S^\nu\right) + Cst \frac{(T^{1/2}h^{-\delta})^{\gamma'/(\gamma'-1)}}{\varepsilon^{\gamma'/(\gamma'-1)}} P\left(\|x_t, y_t\| > S^\nu\right) + o_P(1).
\end{aligned}$$

Thus, invoking assumption T2, $\sup_\theta \|A_{22}\|$ tends to zero in probability.

Study of A_3: Thanks to assumption L3 and Hölder's inequality, note that

$$\begin{aligned}
\|A_3\| &\leq T^{-1/2} \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| \leq S^\nu) \left| \tau_S(l_t^S(\theta)) - \tau_S(l_t(\theta)) \right| \cdot \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\| \\
&+ T^{-1/2} \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\| + T^{-1/2} \sum_{t=1}^T [1 - \tau_S(l_t(\theta))] \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\| \\
&\leq T^{1/2} \left\{ Cst.h^{-\delta} \left[\frac{1}{T} \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| \leq S^\nu) \left| l_t^S(\theta) - l_t(\theta) \right|^{\gamma/(\gamma-1)} \right]^{1-1/\gamma} \right. \\
&+ \left[\frac{1}{T} \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \right]^{1-1/\gamma} \\
&+ \left. \left[\frac{1}{T} \sum_{t=1}^T [1 - \tau_S(l_t(\theta))] \right]^{1-1/\gamma} \right\} \cdot \left[\frac{1}{T} \sum_{t=1}^T \left\| \frac{\partial \ln l_t(\theta)}{\partial \theta} \right\|^\gamma \right]^{1/\gamma} \qquad \text{(C-31)}
\end{aligned}$$

Applying lemma A.2, the first term is bounded in probability by

$$Cst.T^{1/2} h^{-\delta} \left[\left(\frac{\ln S}{S h^m} \right)^{1/2} + h^\rho \right],$$

which tends to zero by assumption R2. Moreover, note that

$$\sup_\theta \frac{1}{T}\sum_{t=1}^{T}[1-\tau_S(l_t(\theta))] \leq \sup_\theta \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}(l_t(\theta)\leq 2h^\delta) \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}(l_t^*(\theta_0)\leq 2h^\delta), \quad \text{(C-32)}$$

where $l_t^*(\theta_0)=\inf_{\theta\in V_0} l_t(\theta)$. Thus,

$$\begin{aligned} & P\left(\sup_\theta T^{1/2}\left[\frac{1}{T}\sum_{t=1}^{T}[1-\tau_S(l_t(\theta))]\right]^{1-1/\gamma} > \varepsilon\right) \\ \leq\ & P\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}(l_t^*(\theta_0)\leq 2h^\delta) > (\varepsilon T^{-1/2})^{\gamma/(\gamma-1)}\right) \\ \leq\ & \varepsilon^{-\gamma/(\gamma-1)}T^{\gamma/2(\gamma-1)}P(l_t^*(\theta_0)\leq 2h^\delta), \end{aligned}$$

which tends to zero by assumption R3.

It remains to deal with the second term of (C-31), which is of the same order as

$$T^{1/2}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}(\|x_t,y_t\|>S^\nu)\right]^{1-1/\gamma}.$$

But, for every $\eta>0$,

$$\begin{aligned} & P\left(T^{1/2}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}(\|x_t,y_t\|>S^\nu)\right]^{1-1/\gamma} > \eta\right) \leq \left(\eta T^{-1/2}\right)^{\gamma/(1-\gamma)} P(\|X,Y\|>S^\nu) \\ =\ & O\left(T^{\gamma/(2\gamma-2)}P(\|X,Y\|>S^\nu)\right), \end{aligned}$$

which tends to zero when $(S,T)\longrightarrow\infty$, by assumption T2.

Study of A_4: Let us split A_4 as

$$\begin{aligned} A_4 = & T^{-1/2}\sum_{t=1}^{T}\tau_S'(l_t^S(\theta))\ln l_t^S(\theta)\left(\frac{\partial l_t^S(\theta)}{\partial\theta}-\frac{\partial l_t(\theta)}{\partial\theta}\right)\mathbb{1}(\|x_t,y_t\|\leq S^\nu) \\ + & T^{-1/2}\sum_{t=1}^{T}\tau_S'(l_t^S(\theta))\ln l_t^S(\theta)\frac{\partial l_t(\theta)}{\partial\theta}\mathbb{1}(\|x_t,y_t\|\leq S^\nu) \\ + & T^{-1/2}\sum_{t=1}^{T}\tau_S'(l_t^S(\theta))\ln l_t^S(\theta)\frac{\partial l_t^S(\theta)}{\partial\theta}\mathbb{1}(\|x_t,y_t\|> S^\nu) \equiv A_{41}+A_{42}+A_{43}. \end{aligned}$$

Since τ_S' is a polynomial supported by $[h^\delta,2h^\delta]$, we have for every $x>0$,

$$0\leq \tau_S'(x)|\ln x| \leq Cst.h^{-\delta}|\ln h|\bar\tau_S(x),$$

where $(\bar{\tau}_S)_{S\geq 1}$ is a bounded sequence of polynomials supported by $[h^\delta, 2h^\delta]$. Invoking lemma A.3, we obtain that

$$P(\sup_\theta \|A_{41}\| > \varepsilon) \leq P\left(\sup_\theta \sup_{x_t,y_t,\|x_t,y_t\|\leq S^\nu} \left\| \frac{\partial l_t^S(\theta)}{\partial\theta} - \frac{\partial l_t(\theta)}{\partial\theta} \right\| > Cst.\varepsilon T^{-1/2} h^\delta / |\ln h| \right)$$
$$\leq \quad P\left(\sup_\theta T^{1/2} |\ln h| \left\{ h^{\rho-\delta} \vee \left(\frac{\ln S}{S h^{2m+2\delta+2+2r_0}} \right)^{1/2} \right\} > Cst.\varepsilon \right)$$

which is zero for S sufficiently large, thanks to assumption R2.

Since the functions $(\bar{\tau}_S)_{S\geq 1}$ can be dealt exactly like $(1-\tau_S)_{S\geq 1}$, the term A_{42} is bounded like for A_3. Therefore, for S sufficiently large

$$\|A_{42}\| \leq Cst.T^{1/2} h^{-\delta} |\ln h| \sup_\theta \frac{1}{T} \sum_{t=1}^T l_t(\theta) \bar{\tau}_S(l_t(\theta)) \left\| \frac{\partial \ln l(y_t|x_t,\theta)}{\partial\theta} \right\|$$
$$\leq \quad Cst.|\ln h| T^{1/2} \sup_\theta \left(\frac{1}{T} \sum_{t=1}^T \left\| \frac{\partial \ln l(y_t|x_t,\theta)}{\partial\theta} \right\|^\gamma \right)^{1/\gamma} \cdot \left(\frac{1}{T} \sum_{t=1}^T \mathbb{1}(l_t^*(\theta_0) \leq 2h^\delta) \right)^{1-1/\gamma},$$

where $l_t^* = \inf_{\theta\in V_0} l_t(\theta)$. Thus, for every $\varepsilon > 0$,

$$P\left(\sup_\theta \|A_{42}\| > \varepsilon\right) \leq Cst.(\varepsilon^{-1}|\ln h| T^{1/2})^{\gamma/(\gamma-1)} P_{\theta_0}(l_t^*(\theta_0) \leq 2h^\delta) + o_P(1),$$

which tends to zero under R3.

Finally, note that

$$\sup_\theta \|A_{43}\| \leq Cst \cdot \frac{|\ln h|}{h^\delta} T^{-1/2} \sum_{t=1}^T \mathbb{1}(\|x_t, y_t\| > S^\nu) \left\| \frac{\partial l_t^S(\theta)}{\partial\theta} \right\|,$$

that can be dealt exactly like P_{13}. This proves the result. □