Estimation of Dynamic Models with Nonparametric Simulated Maximum Likelihood

By Dennis Kristensen and Yongseok Shin*

January 2006

Abstract

We propose a simulated maximum likelihood estimator (SMLE) for general stochastic dynamic models based on nonparametric kernel methods. The method requires that, while the actual likelihood function cannot be written down, we can still simulate observations from the model. From the simulated observations, we estimate the unknown density of the model nonparametrically by kernel methods, and then obtain the SMLEs of the model parameters. Our method avoids the issue of non-identification arising from poor choice of auxiliary models in simulated methods of moments (SMM) or indirect inference. More importantly, our SMLEs achieve higher efficiency under weak regularity conditions. Finally, our method allows for potentially nonstationary processes, including time-inhomogeneous dynamics.

*Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706, USA; dkristen@ssc.wisc.edu, yshin@ssc.wisc.edu. We are grateful for the useful comments from seminar participants at Berkeley, Penn, Stanford and 2005 Econometric Society World Congress in London.

1 Introduction

We propose a general method to estimate dynamic models by maximum likelihood when the likelihood functions are of unknown forms. The method requires that, while the actual likelihood function cannot be written down, we can still simulate observations from the model. For any given parameter value, we draw N i.i.d. simulated values from the model conditional on available information, and then use the simulated data points to estimate the unknown conditional density of the model nonparametrically by kernel methods. The kernel estimate of the density will converge towards the true density as $N \to \infty$, so we can approximate the density arbitrarily well by choosing a sufficiently large N. Given this simulated density, we obtain a nonparametric simulated maximum likelihood estimator (NPSMLE) of the underlying parameters.

Our method is related to simulated methods of moments (SMM) (**??**) and indirect inference (**???**), but, unlike these approaches, is not subject to the arbitrariness involved in the selection of target moments or auxiliary models. In SMM and indirect inference, if the target moments or auxiliary models are not chosen appropriately, the parameters of interest may not be identified. Our method avoids this issue, because all the information embedded in the distribution is utilized. More importantly, under weak regularity conditions, maximum likelihood estimators enjoy higher efficiency than these estimators, and the NPSMLEs inherit this property. Finally, our method allows for potentially nonstationary processes, including time-inhomogeneous dynamics.

For specific models, a number of studies have proposed methods to approximate the likelihood or score function—e.g. **??**, **?**, **?**, **?** and **?**. These are, however, model-specific and cannot easily be adapted to other types of models. Our method, in contrast, is very general and easy to implement.

Recently, methods similar to the one proposed here have been developed. The most closely related is the NPSMLE of **?**. Their analysis, unlike ours, focuses on static models, although they do suggest generalization into dynamic models. Other papers have focused on nonparametric simulated estimation of dynamic models, such as **?** and **?**. The former uses an L_2-distance and focuses on stationary models, while the latter does not offer any theoretical results concerning the properties of the simulated estimators. In two related studies, **??** use the so-called semi-nonparametric estimation method by **??** to approximate the likelihood or score of dynamic models. However, no theoretical results

regarding the approximate estimators are given. In addition, none of the above papers are applicable to nonstationary dynamics.

In this paper, we generalize the NPSMLE of **?** to dynamic models, including nonstationary and time-inhomogeneous ones. We give primitive conditions for our NPSMLE to be consistent and have the same asymptotic distribution as the infeasible MLE. In particular, we show how the number of simulations (N) and the kernel bandwidth should be determined as the sample size grows. We also demonstrate how N affects the asymptotic variance of the NPSMLE.

One disadvantage of our proposed method is, as in **?**, **?** and **?**, the bias incurred by using kernel methods to estimate the density with a finite number of simulations. This is in contrast to, say, SMM which yields unbiased approximations. However, one can simulate one's way out of this problem by drawing a sufficiently large number of data points—that is, by letting $N \to \infty$.

The strength of our method, as is the case with **?**, lies in its general applicability and ease of implementation.

The estimation method has applications in a wide range of settings. In finance, the researcher often faces the problem of estimating a continuous-time stochastic model given discretely-observed data. The conditional density for these models are well-defined, but only in a few cases is one able to derive a closed form expression for this. So one either has to rely on approximate likelihood methods, or find alternative estimators. Our estimator allows for an easy-to-implement approximate MLE for such models.

Similar problems arise in macroeconomics where the internal logic of many dynamic stochastic general equilibrium models is so complex that a closed form expression for the transition density of the system is not available. However, given the fundamentals of the model, we can generally simulate artificial data, and hence apply our NPSMLE.

Throughout this paper, theoretical results are obtained only for cases where the transition density of the model is conditional on finite-dimensional observable variables. This limitation bites when latent variables are involved in dynamics and hence the conditioning information set grows over time. Our method can still be used in such cases, if one is willing to define a quasi-likelihood conditional on finite-dimensional observables. Obviously, as discussed in Section 2.1, the asymptotic variance has to be adjusted since we are not using the true likelihood. Extensions to methods with built-in nonlinear filters that explicitly account for latent variables are worked out in a companion paper

(**?**) based on the main results given here.

This paper is organized as follows. In the next section, we set up our framework and present the approximate density and the associated NPSMLE. In Section 3, we derive the asymptotic properties of the NPSMLE under regularity conditions. Section 4 presents specific examples, and Section 5 concludes.

2 A Simulated Maximum Likelihood Estimator

2.1 Construction of NPSMLE

Suppose that we have T observations, $(y_1, \ldots, y_T)$, $y_t \in \mathbb{R}^k$, from a parametric model with associated density $p_T\left(y_T, y_{T-1}, ..., y_1 | x_0; \theta\right)$ for some initial starting value x_0 and a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^d$. A natural estimator of θ is then the maximizer of the conditional log-likelihood,

$$\tilde{\theta} = \arg\max L_T(\theta), \quad L_T(\theta) = \sum_{t=1}^{T} \log p_t(y_t | y_{t-1}, \ldots, y_1, x_0; \theta). \tag{1}$$

Often, $p_t(y_t | y_{t-1}, \ldots, y_1, x_0; \theta)$ can be written as

$$p_t(y_t | y_{t-1}, \ldots, y_1, x_0; \theta) = p_t(y_t | x_t; \theta), \tag{2}$$

for some finite-dimensional $x_t \in \mathcal{F}(y_{t-1}, \ldots, y_1, x_0)$. We here allow the density to depend on t, such that $\{y_t\}$ is potentially time-inhomogeneous and thereby nonstationary. Suppose now that $p_t(y|x;\theta)$ is not available in explicit form and thus the maximum likelihood estimation of θ is not feasible. In this case, one may want to set up an approximate likelihood function.

We propose a method to approximate a given density $p_t(y|x;\theta)$ when one can simulate values from $p_t(\cdot|x;\theta)$. In what follows, we consider a general conditional density $p_t(\cdot|x;\theta)$ of unknown form. This could be arrived at from the above setting with $x_t \in \mathcal{F}(y_{t-1}, \ldots, y_1, x_0)$. But our approximation scheme does not rely on this assumption, and x_t can be any type of random variable, including exogenous variables.

The approximation is based on simulated values from $p_t(\cdot|x;\theta)$. Consider a fixed $t \geq 1$, $x \in \mathbb{R}^l$, and $\theta \in \Theta$. Then let $\{Y_{t,i}^{x,\theta}\}_{i=1}^N$ be N simulated i.i.d. random variables such that $Y_{t,i}^{x,\theta} \sim p_t(\cdot|x;\theta)$, $i = 1, ..., N$. The random draws are used to estimate $p_t(y|x;\theta)$

using kernel methods. Define

$$\hat{p}_t(y|x;\theta) = \frac{1}{N}\sum_{i=1}^{N} K_h(Y_{t,i}^{x,\theta} - y), \tag{NPSMLE 1}$$

where $K_h(z) = K(z/h)/h^k$, $K : \mathbb{R}^k \mapsto \mathbb{R}$ is the kernel, and $h > 0$ the bandwidth. Under regularity conditions on p_t and K, we obtain that

$$\hat{p}_t(y|x;\theta) = p_t(y|x;\theta) + O_P(1/\sqrt{Nh^k}) + O_P(h^2), \quad N \to \infty, \tag{3}$$

where the remainder terms are $o_P(1)$ if $h \to 0$ and $Nh^k \to \infty$. We can use the simulated density to obtain a simulated MLE (SMLE) of θ_0 by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \hat{L}_T(\theta), \quad \hat{L}_T(\theta) = \sum_{t=1}^{T} \log \hat{p}_t(y_t|x_t;\theta).$$

Because of (3), $\hat{L}_T(\theta) \xrightarrow{P} L_T(\theta)$ as $N \to \infty$ for any given $T \geq 1$. One would then expect that for any given $T \geq 1$, $\hat{\theta} \xrightarrow{P} \tilde{\theta}$ as $N \to \infty$ under weak smoothness conditions on $p_t(y|x;\theta)$. We will verify this claim later under regularity conditions. When finding $\hat{\theta}$ (which requires numerical optimization), it will be desirable that $\hat{L}_T(\theta)$ is continuous in θ. In the case where $Y_{t,i}^{x,\theta} = g_t(\varepsilon_i|x;\theta)$, this is ensured by continuity of $\theta \mapsto g_t(\varepsilon|x;\theta)$, and by the use of the same random draws $\{\varepsilon_i : i = 1, ..., N\}$ for all values of θ.

A disadvantage of our estimator is that for finite N and fixed $h > 0$, the simulated log-likelihood function is a biased estimate of the actual one. So to obtain consistency, one will have to let $N \to \infty$. This is in contrast to SMM type estimators where unbiased estimators of moments can be constructed, and consistency therefore obtained, for a fixed N. However, if one is willing to make a stronger assumption about the identification of the model, one can partially avoid this issue here. For example, in the stationary case, the standard identification assumption is

$$\mathbb{E}\left[\log p(y_t|x_t;\theta)\right] < \mathbb{E}\left[\log p(y_t|x_t;\theta_0)\right], \quad \theta \neq \theta_0.$$

A stronger identification condition implying the former is

$$\mathbb{E}\left[\log\left(\int K(z)p(y_t + hz|x_t;\theta)dz\right)\right] < \mathbb{E}\left[\log\left(\int K(z)p(y_t + hz|x_t;\theta_0)dz\right)\right],$$

for all $\theta \neq \theta_0$ and $h \leq \bar{h}$. Under this identification condition, one can show consistency of our estimator for any fixed $0 < h \leq \bar{h}$ as $N \to \infty$. A similar identification condition

can be found in **?**. However, for fixed $h > 0$ the resulting estimator will no longer have full efficiency; to obtain this, one has to let $h \to 0$.

This leads us to the next observation. Namely, that the use of our approximation method is not limited to actual MLEs. In many situations, one is able to define a so-called quasi- or pseudo-likelihood which, even though it is not the true likelihood, identifies the parameters of the true model. One obvious example of this is the standard regression model, where the MLE based on Gaussian errors (i.e. the least-squares estimator) proves to be robust to deviations from the normality assumption. Another example is estimation of (G)ARCH models using quasi-maximum likelihood—e.g. **?**. These are cases where the quasi-likelihood can be written explicitly. If one cannot find explicit expressions of the quasi-likelihood, one can instead employ our estimator, simulating from the quasi-model. However, note that, in this setting, the asymptotic variance has to be adjusted to accommodate for the fact that we are no longer using the true likelihood function to estimate the parameters. This obviously extends to the case of misspecified models as in **?**.

Our approximation method can also be applied to non- and semiparametric estimation problems in cases where the nonparametric component does not involve the distributions driving the model—e.g. **?**. Our asymptotic results have to be adjusted to allow for θ to be an infinite-dimensional parameter in this setting.

Discrete random variables can also be accommodated within our framework. One then simply use $K_h(Y_{t,i}^{x,\theta} - y) = \mathbb{I}\{Y_{t,i}^{x,\theta} = y\}$ for any point y in the discrete part of the support, where $\mathbb{I}\{\cdot\}$ is the indicator function. One can estimate the density for random variables with mixed continuous and discrete components by using a product kernel, as in **?**. In this case, $\hat{L}_T(\theta)$ is not differentiable w.r.t. θ, but, as shown in the following section, one can obtain the desired asymptotic properties of the NPSMLE as long as $\hat{L}_T(\theta)$ is continuous.

2.2 Alternative schemes

Alternative simulation schemes can be used to approximate the log-likelihood in (1). First, in the case where (2) does not hold, one can instead simulate N i.i.d. sequences $\{Y_{t,i}^{\theta} : t = 1, ..., T\}$ from $p_T(y_T, \ldots, y_1 | x_0; \theta)$ given the initial starting value x_0, and then

calculate

$$\hat{p}_T(y_T, \ldots, y_1|x_0;\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\prod_{t=1}^{T} K_h(Y_{t,i}^{\theta} - y_t)\right]. \tag{4}$$

This will suffer from a severe curse of dimensionality with convergence rate given by $O_P(1/\sqrt{Nh^{kT}}) + O_P(h^2)$. So, a large number of simulations have to be performed to obtain a sufficient degree of accuracy. The deterioration in the convergence rate mirrors the fact that the simulations at time t are performed conditional only on x_0, not utilizing the information contained in $(y_{t-1}, ..., y_1)$. The approximation in (NPSMLE 1) on the other hand takes into account all information available at time t.

If the data-generating process is time-homogeneous such that $p_t(y|x;\theta) = p(y|x;\theta)$, and we can simulate a trajectory $\{Y_i^{\theta}, X_i^{\theta} : i = 1, ..., N\}$ from the model—as is the case with most simulation-based methods used for dynamic models, then the following alternative is available:[1]

$$\check{p}(y|x;\theta) = \frac{\sum_{i=1}^{N} K_h(Y_i^{\theta} - y)K_h(X_i^{\theta} - x)}{\sum_{i=1}^{N} K_h(X_i^{\theta} - x)}. \tag{NPSMLE 2}$$

This estimator potentially saves time on the simulation since one can use the same sample of simulated data points to approximate the density at all data points. So we only generate N values here to obtain the simulated likelihood function at a given parameter value, while in the time-inhomogeneous case we need to simulate TN values. On the other hand, the convergence of $\check{p}$ will be slowed down due to (i) the curse of dimensionality[2] and (ii) the dependence between $(Y_s^{\theta}, X_s^{\theta})$ and $(Y_t^{\theta}, X_t^{\theta})$, $s \neq t$. This is in contrast to $\hat{p}$ where the simulated values are independent. So one will have to choose N larger for the simulated density in (NPSMLE 2) relative to the one in (NPSMLE 1).

Furthermore, one will normally have to assume a stationary solution to the dynamic system under consideration for $\check{p} \xrightarrow{P} p$, and either have to start the simulation at the stationary distribution, or assume that the simulated process converges towards the stationary distribution at a suitable rate. For the latter to hold, one will need to impose some form of mixing condition on the process, as in **?**. But then a large value of N is needed to ensure that the simulated process is sufficiently close to its stationary distribution—that is, one has to allow for a "burn-in".

[1] Here and in the following we will use K to denote a generic kernel.

[2] The dimension of $(Y_t^{\theta}, X_t^{\theta})$ is greater than that of Y_t^{θ}.

The estimator in (NPSMLE 2) may work under non-stationarity. Recently, a number of papers have considered kernel estimation for nonstationary Markov processes. The kernel estimator proves to be consistent and asymptotically mixed-normally distributed when the Markov process is recurrent (**??**). However, the convergence rate will be path-dependent and relatively slow. So, for strongly dependent and nonstationary processes, it will be preferable to use the estimator in (NPSMLE 1).

The above considerations lead us to focus on (NPSMLE 1) in the remainder of this paper. The properties of (NPSMLE 2) can be obtained by following the same strategy of proof as the one we employ for (NPSMLE 1). The only difference is that, to obtain $\check{p} \xrightarrow{P} p$, one has to take into account the dependence of the simulated values. A sufficient set of conditions for $\check{p}(y|x;\theta) \xrightarrow{P} p(y|x;\theta)$ uniformly in y, x and θ when the dynamics of the parametric model is near-epoch dependent can be found in **?**, Corollary 2.

While we here focus on the kernel estimator, one can use other nonparametric density estimators as well. Examples are the semi-nonparametric estimators of **?**, **?**, **?** and **?**; the log-spline estimator of **?**; and the wavelet estimator of **?**. What is needed is that the nonparametric estimator converges towards the true density sufficiently fast.

2.3 Markov models

To illustrate the applicability of the proposed method, we set up NPSMLE for a general, fully-observed Markov model.

Markov models are widely used in economics and finance. Often, the conditional density characterizing the Markov process is not available in closed form. One then either has to use alternative estimators, which will suffer from decreased efficiency relative to the MLE, or construct an approximation of the density. For examples of the former, we refer to **?**, **?**, **?**, **?** and **?**. The latter has been pursued in the case of discretely sampled diffusion models by, amongst others, **??**, **?**, **?**, **?**, **?**, **?**, **?**, **?** and **?**.

We consider a Markov process $\{y_t\}$ with transition density $p_t(y|x)$. That is,

$$P\left(y_t \in A | y_{t-1} = x\right) = \int_A p_t(y|x)dy.$$

Let y_0 be given, and suppose we have observed T observations of the process, $y_1, \ldots, y_T$, and assume that p belongs to a parametric family, $\{p_t(y|x;\theta)|t \geq 1, \theta \in \Theta\}$, where $\Theta \subseteq$

$\mathbb{R}^d$. A natural estimator in this setting is the maximum likelihood estimator,

$$\tilde{\theta} = \arg\max_{\theta\in\Theta} L_T(\theta), \quad L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \log p_t(y_t|y_{t-1};\theta).$$

However, in many cases p is of unknown form or is difficult to evaluate directly. In this case, one may implement the NPSMLE instead based on

$$\hat{L}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \log \hat{p}_t(y_t|y_{t-1};\theta), \quad \hat{p}_t\left(y_t|y_{t-1};\theta\right) = \frac{1}{N}\sum_{i=1}^{N} K_h\left(Y_{t,i}^{y_{t-1},\theta} - y_t\right),$$

where $Y_{t,i}^{y_{t-1},\theta}$ are i.i.d. draws from $p_t(\cdot|y_{t-1};\theta)$.

3 Asymptotic Properties of the NPSMLE

Given the convergence of the simulated density towards the true density, we expect that the NPSMLE $\hat{\theta}$ will have the same asymptotic properties as the infeasible MLE $\tilde{\theta}$ for a suitably chosen sequence $N = N(T)$ and $h = h(N)$.

Throughout, we assume that the conditional distributions of the model are continuous so that standard kernel estimators can be used. One should be able to generalize the results to allow for discrete distributions by extending the lemmas in Appendix B.

3.1 Results for the general case

We give three sets of results. The first set gives very weak conditions under which there exists a sequence N such that $\hat{\theta}$ is consistent and has the same asymptotic distribution as $\tilde{\theta}$. This result is however silent about how the sequence N should be chosen as $T \to \infty$. The second set imposes further restrictions under which we are able to give more precise error bounds which are helpful when choosing N and h in practice. The third set is a further strengthening, and demonstrates how the asymptotic variance of $\hat{\theta}$ is affected by the use of simulations.

In order for $\hat{\theta}$ to be asymptotically equivalent to $\tilde{\theta}$, we obviously need to ensure that $\hat{p} \xrightarrow{P} p$ in some function norm. To establish this, we extend some standard results from the theory on kernel density estimation in Appendix B. These results are established under the following assumptions regarding the actual density and its associated model:

A.1 There exist Λ_t with $\mathbb{E}\Lambda_t^2 < \infty$ and $\beta_1 \geq 0$ such that $\|Y_t^{x,\theta} - Y_t^{x',\theta'}\| \leq \Lambda_t \left[\|x - x'\|^{\beta_1} + \|\theta - \theta'\|^{\beta_1}\right]$ for all $\theta, \theta' \in \Theta$ and $x, x' \in \mathbb{R}^l$.

A.2 The density $p_t(y|x;\theta)$ is $r \geq 2$ times continuously differentiable w.r.t. y with bounded derivatives such that

$$\max_{t\geq 1} \sup_{\theta\in\Theta} \sup_{(y,x)\in\mathbb{R}^{k+l}} \sum_{|\lambda|=r} \left| D_y^\lambda p_t(y|x;\theta) \right| < \infty.$$

A.3 The density $p_t(y|x;\theta)$ is continuous w.r.t. θ.

A.4 The density $p_t(y|x;\theta)$ is thrice continuously differentiable w.r.t. θ.

A.5 In addition to A.1, $\theta \mapsto Y^{x,\theta}$ is differentiable with its derivative, $\dot{Y}^{x,\theta}$, satisfying

$$\|\dot{Y}_t^{x,\theta} - \dot{Y}_t^{x',\theta'}\| \leq \Lambda_t \left[\|x - x'\|^{\beta_1} + \|\theta - \theta'\|^{\beta_1}\right], \quad \|\dot{Y}_t^{x,\theta}\|^2 \leq \Lambda_t \|x\|^{\beta_2},$$

for all $\theta, \theta' \in \Theta$ and $x, x' \in \mathbb{R}^l$, where $\beta_2 \geq 0$.

The first and second conditions are used to establish uniform convergence of $\hat{p}$ over x and θ. The second in conjunction with the use of higher-order kernels reduces the bias of $\hat{p}$. The third and fourth ensure that the log-likelihood is suitably bounded and sufficiently smooth. The final condition will only be needed to examine the effect that the simulations will have on the asymptotic variance of the estimator. In the case of time-homogeneous processes, the random variable $\Lambda_t = \Lambda$ will be time-invariant.

The kernel K is assumed to belong to the following class of so-called higher-order or bias-reducing kernels:

K.1 The kernel K satisfies $\int_{\mathbb{R}^k} K(z)dz = 1$; $\int_{\mathbb{R}^k} z^\lambda K(z)dz = 0$, for $1 \leq |\lambda| \leq r-1$; $\int_{\mathbb{R}^k} \|z\|^r |K(z)|dx < \infty$; $\sup_z \left[|K(z)| \max(\|z\|, 1)\right] < \infty$; K is absolutely integrable with a Fourier transform Ψ satisfying $\int_{\mathbb{R}^k} \left\{(1 + \|z\|) \sup_{b\geq 1} |\Psi(bz)|\right\} dz < \infty$.

This class of kernels were first proposed by **?**, and a discussion of the construction of specific kernels satisfying these conditions can be found in **?**. Using a kernel from this class makes it possible to reduce the bias of $\hat{p}$ and its derivatives, and thereby obtain a faster rate of convergence. The smoothness of p as measured by its number of derivatives, r, determines the degree of bias reduction.

Finally, we impose regularity conditions on the model to ensure that the actual MLE is asymptotically well-behaved. We first need to introduce the relevant terms driving the asymptotics of the MLE. Under (A.4), we can define

$$\begin{aligned} S_T(\theta) &= \frac{\partial L_T(\theta)}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \log p_t(y_t|x_t;\theta)}{\partial \theta} \in \mathbb{R}^d, \\ H_T(\theta) &= \frac{\partial^2 L_T(\theta)}{\partial \theta \partial \theta'} = \sum_{t=1}^{T} \frac{\partial^2 \log p_t(y_t|x_t;\theta)}{\partial \theta \partial \theta'} \in \mathbb{R}^{d\times d}, \\ G_{j,T}(\theta) &= \frac{\partial^3 L_T(\theta)}{\partial \theta \partial \theta' \partial \theta_j} = \sum_{t=1}^{T} \frac{\partial^3 \log p_t(y_t|x_t;\theta)}{\partial \theta \partial \theta' \partial \theta_j} \in \mathbb{R}^{d\times d}. \end{aligned}$$

The Information is then defined as

$$i_T(\theta) = \sum_{t=1}^{T} \mathbb{E}\left[\frac{\partial \log p_t(y_t|x_t;\theta)}{\partial \theta} \frac{\partial \log p_t(y_t|x_t;\theta)}{\partial \theta'}\right] = \mathbb{E}\left[H_T(\theta)\right] \in \mathbb{R}^{d\times d}.$$

We also define the diagonal matrix $\mathcal{I}_T(\theta) = \text{diag}\left\{i_T(\theta)\right\} \in \mathbb{R}^{d\times d}$, and

$$\begin{aligned} &U_T(\theta) = \mathcal{I}_T^{-1/2}(\theta) S_T(\theta), \quad V_T(\theta) = \mathcal{I}_T^{-1/2}(\theta) H_T(\theta) \mathcal{I}_T^{-1/2}(\theta), \\ &W_{j,T}(\theta) = \mathcal{I}_T^{-1/2}(\theta) G_{j,T}(\theta) \mathcal{I}_T^{-1/2}(\theta). \end{aligned}$$

We then impose the following conditions on the actual MLE:

C.1 $\Theta \subseteq \mathbb{R}^d$ is compact.

C.2 $\tilde{\theta} \xrightarrow{P} \theta_0$.

C.3 There exists a sequence $\bar{L}_T > 0$ such that:

1. $L_T(\theta)/\bar{L}_T$ is stochastically equicontinuous.
2. $\sup_{\theta\in\Theta} \sum_{t=1}^{T} \left|\log p_t(y_t|x_t;\theta)\right|^{1+\delta} / \bar{L}_T = O_P(1)$, for some $\delta > 0$.
3. $\sum_{t=1}^{T} \left\|x_t\right\|^{1+\delta} / \bar{L}_T = O_P(1)$ and $\sum_{t=1}^{T} \mathbb{E}\|\Lambda_t\| / \bar{L}_T = O_P(1)$, for some $\delta > 0$.

N.1 $\theta_0 \in \text{int}\Theta$.

N.2 $\mathcal{I}_T^{-1} = \mathcal{I}_T^{-1}(\theta_0) \to 0$.

N.3 $W_{j,T}(\theta) = O_P(1)$ uniformly in a neighborhood of θ_0 for $j = 1, ..., d$.

N.4 $(U_T(\theta_0), V_T(\theta_0)) \xrightarrow{d} (U_0, V_0)$ for some random variables (U_0, V_0) with V_0 being non-singular almost surely.

Assumption (C.2) gives us consistency of the actual MLE, and (C.3) is used in the proof of $\hat{\theta} \xrightarrow{P} \tilde{\theta}$, while (N.1)–(N.4) imply that the asymptotic distribution of the MLE is given by:[3]

$$\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) \xrightarrow{d} V_0^{-1} U_0.$$

Note that in the stationary case, $\bar{L}_T$ can be chosen as $\bar{L}_T = T$ given that $\mathbb{E}\,\|x_t\|^{1+\delta} < \infty$ and $\mathbb{E}\left[\sup_{\theta\in\Theta} |\log p(y_t|x_t;\theta)|^{1+\delta}\right] < \infty$. In the general case, if one can show that $\mathcal{I}_T^{-1} S_T(\theta) = O_P(1)$ uniformly in θ, then $L_T(\theta)\|\mathcal{I}_T^{-1}\|$ will be stochastically equicontinuous.

We are now ready to state our first set of results. This relies on a general result from **?**, which can be found in Proposition 8 in Appendix C. We employ this to show that $\hat{\theta}$ inherits the asymptotic properties of the actual MLE for some unspecified sequence $N \to \infty$ (Proposition 9). Under the following condition on the bandwidth, we can apply this result to our NPSMLE.

B.1 The bandwidth $h = h(N)$ satisfies: $h \to 0$ and $N^{1/2} h^{k+1} \to \infty$.

Theorem 1 *Assume that (A.1)–(A.3), (K.1), (C.1)–(C.2) and (B.1) hold. Then there exists some sequence $N(T) \to \infty$ as $T \to \infty$ such that $\hat{\theta} \xrightarrow{P} \theta_0$.*

If furthermore (A.4) and (N.1)–(N.4) hold, then there exists some (possibly different) sequence $N(T) \to \infty$ such that $\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} V_0^{-1} U_0$.

The above result only states that a sequence $N \to \infty$ exists for which the NPSMLE is consistent and has an asymptotic distribution. It is silent regarding how the econometrician should choose this sequence as $T \to \infty$.

We now derive precise conditions on N under which the NPSMLE converges towards the true MLE at a given rate. In order to obtain these, we need to strengthen the conditions used above substantially.

[3] **?** and **?** show what U_0 and V_0 look like in various cases.

First, we need to trim the approximate log-likelihood. We redefine our criterion function as

$$\hat{L}_T(\theta) = \sum_{t=1}^{T} \tau_a(\hat{p}_t(y_t|x_t;\theta)) \log \hat{p}_t(y_t|x_t;\theta),$$

where $\tau_a(\cdot)$ is continuously differentiable trimming function satisfying $\tau_a(x) = 1$ if $|x| > a$, and 0 if $|x| < a/2$, and $a = a(N) \to 0$ is a trimming sequence. One could here simply use the indicator function for the trimming, but then $\hat{L}_T(\theta)$ would no longer be differentiable. This property will be useful when using numerical optimization algorithms to solve for $\hat{\theta}$.

Furthermore, under (A.5), we can define

$$\frac{\partial \hat{p}_t(y|x;\theta)}{\partial \theta} = \frac{1}{Nh} \sum_{i=1}^{N} K_h' \left(Y_{i,t}^{x,\theta} - y \right) \dot{Y}_{i,t}^{x,\theta}, \tag{5}$$

and thereby construct an approximation of the score,

$$\begin{aligned} \hat{S}_T(\theta) \quad = \quad & \sum_{t=1}^{T} \frac{\partial \hat{p}_t(y_t|x_t;\theta)}{\partial \theta} \frac{\tau_a(\hat{p}_t(y_t|x_t;\theta))}{\hat{p}_t(y_t|x_t;\theta)} \\ & + \sum_{t=1}^{T} \frac{\partial \hat{p}_t(y_t|x_t;\theta)}{\partial \theta} \tau_a'(\hat{p}_t(y_t|x_t;\theta)) \log \hat{p}_t(y_t|x_t;\theta). \end{aligned} \tag{6}$$

We then give two results—one for the case where (A.5) does not hold and the other for the case where it does, with some additional assumptions on the bandwidth.

B.2 $\sqrt{\log(N)/N} h^{-k-1} a^{-1} \to 0$, $h^r a^{-1} \to 0$, and $N^{-2\gamma} h^{-k} \log a \to 0$ for some $\gamma > 0$.

B.3.1 $\bar{L}_T \sqrt{\log(N)/N} h^{-k-1} a^{-1} \to 0$, $\bar{L}_T h^r a^{-1} \to 0$, $\bar{L}_T N^{-2\gamma} h^{-k} \log a \to 0$, $\bar{L}_T N^{-\gamma} \to 0$, $\bar{L}_T (\log a)^{-1} \to 0$ for some $\gamma > 0$.

B.3.2 $\|\mathcal{I}_T^{1/2}\| N^{(2\gamma\beta_2 - 1)/3} \sqrt{\log N} h^{-k-2} a^{-1} \to 0$, $\|\mathcal{I}_T^{1/2}\| h^{r-1} a^{-1} \to 0$, $\|\mathcal{I}_T^{1/2}\| (\log a)^{-1} \to 0$, $\|\mathcal{I}_T^{1/2}\| N^{-2\gamma} h^{-k-1} \log a \to 0$, $\|\mathcal{I}_T^{1/2}\| N^{-\gamma} \to 0$ for some $\gamma > 0$.

Condition (B.2) is imposed when showing consistency of the NPSMLE, while either (B.3.1) or (B.3.2)—the latter in conjunction with (A.5)—will imply that the NPSMLE has the same asymptotic distribution as the MLE. Observe that in the stationary case, $\|\mathcal{I}_T^{1/2}\| = O(\sqrt{T})$ and $\bar{L}_T = O(T)$. (B.3.2) imposes weaker restrictions on N, h and a than

(B.3.1) does, but, on the other hand, we have to impose stronger smoothness conditions on the model in the form of (A.5) in order for (B.3.2) to be sufficient. Observe that the density $p_t(y|x;\theta)$ may very well be differentiable in θ while $Y_t^{\theta,x}$ is not. Our strategy of proof for the case where (A.5) is not valid is based on the following expansion:

$$L_T(\hat{\theta}) - L_T(\tilde{\theta}) = \frac{1}{2}(\hat{\theta}-\tilde{\theta})'\mathcal{I}_T^{1/2}V(\theta_0)\mathcal{I}_T^{1/2}(\hat{\theta}-\tilde{\theta}) + o_P(1).$$

One could probably weaken the restrictions on N and h in (B.3.1) by instead using arguments similar to the ones employed in **?**, Theorem 7.1 and **?**.

Theorem 2 *Assume that (A.1)–(A.3), (K.1) and (C.1)–(C.3) hold. Then* $\hat{\theta} \xrightarrow{P} \theta_0$ *for any sequences* $N \to \infty$, $h \to 0$ *satisfying (B.2).*

If furthermore (A.4) and (N.1)–(N.4) hold, then $\mathcal{I}_T^{1/2}(\hat{\theta}-\theta_0) \xrightarrow{d} V_0^{-1}U_0$ *for any sequences* $N \to \infty$, $h \to 0$ *satisfying (B.3.1). If also (A.5) holds then the result remains true for any sequences* $N \to \infty$, $h \to 0$ *satisfying (B.3.2).*

3.2 Stationary case

We give simple conditions for the NPSMLE to be consistent and asymptotically normally distributed for the case where the data generating process is stationary and ergodic.

Corollary 3 *Assume that* $\{(y_t, x_t)\}$ *is stationary and ergodic, and that (A.1)–(A.3) and (K.1) hold, and:*

(i) $|\log p(y|x;\theta)| \le b(y|x)$ *for all* $\theta \in \Theta$ *, with* $\mathbb{E}\left[b(y_t|x_t)^{1+\delta}\right] < \infty$.

(ii) $\mathbb{E}\left[\log p(y_t|x_t;\theta)\right] < \mathbb{E}\left[\log p(y_t|x_t;\theta_0)\right]$, $\forall \theta \neq \theta_0$.

(iii) $\sqrt{\log(N)/N}h^{-k-1}a^{-1} \to 0$, $h^r a^{-1} \to 0$, *and* $N^{-2\gamma}h^{-k}\log a \to 0$ *for some* $\gamma > 0$.

Then $\hat{\theta} \xrightarrow{P} \theta_0$.

If furthermore (N.1) and (A.4) hold, and:

(iv) $\mathbb{E}\left[\|\partial \log p(y|x;\theta_0)/\partial\theta\|^2\right] < \infty$.

(v) $\|\frac{\partial^2 \log p(y|x;\theta)}{\partial\theta\partial\theta'}\| \le b(y|x)$ *uniformly in a neighborhood of* θ_0 *with* $\mathbb{E}\left[b(y_t|x_t)\right] < \infty$.

(vi) $i(\theta_0) = \mathbb{E}\left[\frac{\partial^2 \log p(y_t|x_t;\theta_0)}{\partial\theta\partial\theta'}\right]$ *is nonsingular.*

(vii) $T\sqrt{\log(N)/N}h^{-k-1}a^{-1} \to 0$, $Th^r a^{-1} \to 0$, $TN^{-2\gamma}h^{-k}\log a \to 0$, $T(\log a)^{-1} \to 0$, $TN^{-\gamma} \to 0$ *for some* $\gamma > 0$.

Then $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, i(\theta_0)^{-1})$.

Finally, we demonstrate the impact of the number of simulations on the variance of $\hat{\theta}$. To this end we have to invoke the additional smoothness conditions on $Y_t^{x,\theta}$ stated in (A.5). Our result is only stated for the stationary case since it is difficult to derive the result in the general setting of nonstationary/time-inhomogeneous processes. In particular, we need to invoke U-statistics results which apparently are available only in the stationary case.

Theorem 4 *Assume that:*

(i) $\{y_t, x_t\}$ *is stationary, ergodic and* β*-mixing with exponentially decaying mixing coefficients.*

(ii) *(A.1)–(A.3) and (K.1) hold, and* $\theta \mapsto Y^{x,\theta}$ *is twice differentiable with both derivatives satisfying (A.5).*

(iii) *(i)–(vi) of Corollary 3 hold.*

(iv) $T^{1/4}\sqrt{\log(N)}N^{(2\gamma\beta_2-1)/3}h^{-k-2}a^{-1} \to 0$, $T^{1/4}h^{r-1}a^{-1} \to 0$, $\sqrt{T}N^{-2\gamma}h^{-k}\log a \to 0$, $\sqrt{T}(\log a)^{-1} \to 0$.

Then, as $T/N \to \alpha \geq 0$, $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, (1+\alpha)i(\theta_0)^{-1})$.

3.3 Estimation of variance

To do any inference in finite sample, an estimator of the asymptotic variance is needed. This can be done in several ways. A simple approach is to use numerical derivatives. Define:

$$\frac{\partial \hat{p}_t(y|x;\theta)}{\partial \theta_k} = \frac{\hat{p}_t(y|x;\theta + \varepsilon e_k) - \hat{p}_t(y|x;\theta - \varepsilon e_k)}{2\varepsilon},$$

where e_k is the kth column of the identity matrix. We have:

$$\frac{\partial \hat{p}_t(y|x;\theta)}{\partial \theta_k} - \frac{\partial p_t(y|x;\theta)}{\partial \theta_k}$$

$$= \frac{\hat{p}_t(y|x;\theta+\varepsilon e_k) - p_t(y|x;\theta+\varepsilon e_k)}{2\varepsilon} - \frac{\hat{p}_t(y|x;\theta-\varepsilon e_k) - p_t(y|x;\theta-\varepsilon e_k)}{2\varepsilon}$$
$$+ \left\{ \frac{p_t(y|x;\theta+\varepsilon e_k) - p_t(y|x;\theta-\varepsilon e_k)}{2\varepsilon} - \frac{\partial p_t(y|x;\theta)}{\partial \theta_k} \right\}.$$

Now letting $\varepsilon = \varepsilon(N) \to 0$ as $N \to \infty$ at a suitable rate, all three terms are $o_P(1)$. A consistent estimator of the second derivative can obviously be obtained in a similar fashion. These can in turn be used to construct estimators of the information and squared score.

If the first two derivatives of $Y_t^{x,\theta}$ w.r.t. θ exist, an alternative estimator of $\partial p_t(y|x;\theta)/\partial\theta$ is given in (5), while the second derivative can be estimated by:

$$\frac{\partial^2 \hat{p}_t(y|x;\theta)}{\partial\theta\partial\theta'} = \frac{1}{Nh^2}\sum_{i=1}^{N} K_h''\left(Y_{i,t}^{x,\theta} - y\right)\left(\dot{Y}_{i,t}^{x,\theta}\right)\left(\dot{Y}_{i,t}^{x,\theta}\right)' + \frac{1}{Nh}\sum_{i=1}^{N} K_h'\left(Y_{i,t}^{x,\theta} - y\right)\ddot{Y}_{i,t}^{x,\theta}.$$

Lemma 6 shows that these are uniformly consistent estimates of the actual derivatives of the density p.

4 Applications

One of the merits of our approach is generality. In this section, however, we apply NPSMLE to simple, well-known models for expositional purposes. The first example is the short-term interest rate model of **?**. This univariate diffusion has a known transition density, and therefore has been a popular benchmark of numerous diffusion estimation strategies (**?**). We test the validity of our approach by comparing it to the true MLE and to another simulation-based method—the SMLE of **?** and **?**. In the second example, we consider a nonstationary bivariate diffusion process with Gaussian density.

4.1 Cox-Ingersoll-Ross model

? model short-term interest rates as a square-root diffusion process:

$$dy_t = \beta(\alpha - y_t)dt + \sigma\sqrt{y_t}dW_t.$$

Conveniently, the transition density for a discretely-sampled path is known. When $t > s$,

$$p(y_t, t|y_s, s) = ce^{-w-v}\left(\frac{v}{w}\right)^{q/2} I_q(2\sqrt{wv}), \tag{7}$$

where $w = cy_s e^{-\beta(t-s)}$, $v = cy_t$, $q = 2\alpha\beta/\sigma^2 - 1$, and $I_q(\cdot)$ is the modified Bessel function of the first kind of order q. We collect the unknown parameters in $\theta = (\alpha, \beta, \sigma)'$, which can be estimated with maximum likelihood.

If the proposed NPSMLE works, the distance between the true MLE and NPSMLE will be a small fraction of the distance between the true MLE and the true parameter values.

In the implementation of NPSMLE, we forgo our knowledge of (7). Given a set of parameter values and y_s, we simulate paths using the Euler scheme—c.f. **?**. We divide the interval $t - s$ into M subintervals, and recursively compute for $m = 0, \ldots, M - 1$:

$$u^i_{m+1} = u^i_m + \beta(\alpha - u^i_m)\delta + \sigma\sqrt{u^i_m}\delta^{1/2}W^i_{m+1},$$

where $u^i_0 = y_s$, $\delta = \frac{t-s}{M}$, and W^i_{m+1}'s are i.i.d. standard normal random variables. Then, we let u^i_M be $Y^{\theta,y_s}_{t,i}$—ith simulated observation of y_t. Once we have generated $Y^{\theta,y_s}_{t,i}$'s for $i = 1, \ldots, N$, then we can estimate the transition density $p(y_t, t|y_s, s; \theta)$ by:

$$\hat{p}(y_t, t|y_s, s) = \frac{1}{N}\sum_{i=1}^{N} K_h\left(Y^{\theta,y_s}_{t,i} - y_t\right).$$

It is straightforward to construct and then maximize $\hat{L}_T(\theta)$ over Θ. Here we use the Gaussian kernel.

Following the benchmark in **?**, we generate an artificial sample path of y_t according to the true transition density with $\theta = (0.06, 0.5, 0.15)$. Note that non-negativity of y_t requires $2\alpha\beta \geq \sigma^2$.

We estimate the diffusion parameters by maximum likelihood, Pedersen's simulated maximum likelihood, and our nonparametric simulated maximum likelihood. For Pedersen's method, which also involves the Euler scheme, we let $M = 8$ and $N = 256$. For NPSMLE, we pick $M = 8$, $N = 64$ and $h = 0.005$—close to what Silverman's rule of thumb suggests. This trio of estimations are performed on 512 artificial series of y_t of length $T = 1000$ each. For all three estimation schemes, we use a direct-search polytope algorithm in the optimization routine—`UMPOL` in Fortran IMSL. In the simulation stage of SMLE and NPSMLE, we adopt antithetic methods to reduce the simulation variance.

The mean and root mean squared error (RMSE) of the estimates from these 512 Monte Carlo exercises are shown in Table 1. Note that the RMSEs of the true MLEs are the distance from the true parameter values, while the RMSEs of the SMLEs measure

	True value	True MLE	Pedersen SMLE	NPSMLE
α	0.06	0.0606 (0.0063)	0.0599 (0.0026)	0.0599 (0.0024)
β	0.5	0.5420 (0.1152)	0.5337 (0.0361)	0.5351 (0.0302)
σ	0.15	0.1502 (0.0034)	0.1478 (0.0028)	0.1470 (0.0037)

Table 1: The mean estimates and the RMSEs of each scheme over 512 artificial data sets are reported. The RMSEs are in parentheses. The RMSEs of the true MLEs are the distance from the true parameter values, while the RMSEs of the SMLEs measure the distance from the true MLEs.

the distance from the true MLEs. From the table, one can see that the finite-sample performance of our NPSMLE with $N = 64$ is comparable to that of Pedersen's SMLE with $N = 256$.

4.2 Nonstationary diffusion estimation

Assume that $y_t = (r_t, s_t)$ is governed by the following nonstationary bivariate diffusion process:

$$\begin{aligned} dr_t &= \mu dt + \sigma_1 dW_{1t}, \\ ds_t &= \mu dt + \sigma_2 \rho dW_{1t} + \sigma_2 \sqrt{1-\rho^2} dW_{2t}, \end{aligned}$$

where W_{1t} and W_{2t} are standard Brownian motions and are independent of each other. These arithmetic Brownian motions have Gaussian density, and the true maximum likelihood estimation of discretely-sampled data is feasible.

In the implementation of the NPSMLE, we again pretend that we do not know the true transition density, and use the Euler scheme to simulated off paths. We let $M = 4$, $N = 128$, and $h = 0.00058$. We perform the true MLE and NPSMLE for 128 sets of artificial data of $T = 120$ each.

The mean and RMSE of the estimates from these 128 Monte Carlo exercises are shown in Table 2. Again, the RMSEs of the true MLEs are the distance from the true parameter values, while the RMSEs of the NPSMLEs measure the distance from the true MLEs. The table shows that the distance between the NPSMLE and the true MLE is small compared to the distance between the true MLE and the true parameter values with the possible exception of ρ.

	True value	True MLE	NPSMLE
μ	0.0	-0.0001 (0.0008)	-0.0001 (0.0004)
σ_1	0.005	0.0049 (0.0003)	0.0048 (0.0001)
σ_2	0.005	0.0053 (0.0004)	0.0053 (0.0002)
ρ	-0.3714	-0.3522 (0.0433)	-0.3798 (0.0570)

Table 2: The mean estimates and the RMSEs of each scheme over 128 artificial data sets are reported. The RMSEs are in parentheses. The RMSEs of the true MLEs are the distance from the true parameter values, while the RMSEs of the NPSMLEs measure the distance from the true MLEs.

5 Concluding Remarks

We have generalized the nonparametric simulated maximum likelihood estimator of **?** to deal with dynamic models, including nonstationary and time-inhomogeneous ones. Theoretical conditions in terms of the number of simulations and the bandwidth are given ensuring that the NPSMLEs inherit the asymptotic properties of the actual MLE.

This method is widely applicable, and can be implemented with ease. Two finite-sample simulation studies demonstrate that the method works well in practice.

One limitation of the paper is that we only consider cases where the transition density of the model is conditional on finite-dimensional observable variables. Extensions to methods with built-in nonlinear filters that explicitly account for latent variables are worked out in a companion paper (**?**) based on the main results given here.

A Proofs

Proof of Theorem 1 Consider any given $T \geq 1$. We have

$$|\hat{L}_T(\theta) - L_T(\theta)| \leq \sum_{t=1}^{T} |\log \hat{p}_t(y_t|x_t;\theta) - \log p_t(y_t|x_t;\theta)|.$$

The result will then follow by Proposition 9 if we can show that

$$\sup_{\theta\in\Theta} |\log \hat{p}_t(y_t|x_t;\theta) - \log p_t(y_t|x_t;\theta)| \xrightarrow{P} 0, \quad t = 1, ..., T. \tag{8}$$

For any $\gamma > 0$, $\|x_t\| \leq N^\gamma$ with probability approaching 1 (w.p.a.1), $t = 1, ..., T$, as $N \to \infty$, and we can therefore apply Lemma 5 to obtain

$$\sup_{\theta\in\Theta} |\hat{p}_t(y_t|x_t;\theta) - p_t(y_t|x_t;\theta)| \xrightarrow{P} 0, \quad t = 1, ..., T,$$

under (B.1). For a given $t \geq 1$, let $(y, x) = (y_t, x_t)$. Then, by the mean value theorem,

$$|\log \hat{p}_t(y|x;\theta) - \log p_t(y|x;\theta)| \leq \frac{1}{|(1-\lambda)\hat{p}_t(y|x;\theta) + \lambda p_t(y|x;\theta)|} |\hat{p}_t(y|x;\theta) - p_t(y|x;\theta)|,$$

for some $\lambda \in [0, 1]$, where

$$\sup_{\theta\in\Theta} \frac{1}{|(1-\lambda)\hat{p}_t(y|x;\theta) + \lambda p_t(y|x;\theta)|} \leq \sup_{\theta\in\Theta} \frac{1}{|p_t(y|x;\theta) + o_P(1)|} = \frac{1}{\inf_{\theta\in\Theta} p_t(y|x;\theta) + o_P(1)} < \infty,$$

by Lemma 5 and the fact that $\inf_{\theta\in\Theta} p_t(y|x;\theta) > 0$ since $\theta \mapsto p_t(y|x;\theta) > 0$ is continuous and Θ compact. ∎

Proof of Theorem 2 To prove consistency, we may redefine the approximate and actual likelihood functions as

$$\hat{L}_T(\theta) = \frac{1}{\bar{L}_T} \sum_{t=1}^{T} \tau_a(\hat{p}_t(y_t|x_t;\theta)) \log \hat{p}_t(y_t|x_t;\theta), \quad L_T(\theta) = \frac{1}{\bar{L}_T} \sum_{t=1}^{T} \log p_t(y_t|x_t;\theta),$$

where $\bar{L}_T$ is given in (C.3). We introduce an additional trimming function, $\tilde{\tau}_{a,t} = \tau_a(\hat{p}_t(y_t|x_t;\theta))\mathbb{I}\{\|x_t\| \leq N^\gamma\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function, and two trimming sets,

$$A_{1,t}(\varepsilon) = \{\hat{p}_t(y_t|x_t;\theta) \geq \varepsilon a, \|x_t\| \leq N^\gamma\}, \quad A_{2,t}(\varepsilon) = \{p_t(y_t|x_t;\theta) \geq \varepsilon a, \|x_t\| \leq N^\gamma\},$$

for any $\varepsilon > 0$ and some $\gamma > 0$. Defining $A_t(\varepsilon) = A_{1,t}(\varepsilon) \cap A_{2,t}(\varepsilon)$, it follows by the same arguments as in **?**, p. 588, $A_{2,t}(2\varepsilon) \subseteq A_{1,t}(\varepsilon) \subseteq A_t(\varepsilon/2)$ w.p.a.1 as $N \to \infty$ under (B.2). Thus, $\mathbb{I}_{A_{2,t}(4)} \leq \mathbb{I}_{A_{1,t}(2)} \leq \tilde{\tau}_{a,t} \leq \mathbb{I}_{A_{1,t}(1/2)} \leq \mathbb{I}_{A_t(1/4)}$.

We split up into three terms,

$$\begin{aligned}\hat{L}_T(\theta) - L_T(\theta) &= \frac{1}{\bar{L}_T}\sum_{t=1}^{T}[\tau_a(\hat{p}_t(y_t|x_t;\theta)) - \tilde{\tau}_{a,t}]\log\hat{p}_t(y_t|x_t;\theta) \\ &\quad + \frac{1}{\bar{L}_T}\sum_{t=1}^{T}\tilde{\tau}_{a,t}\left[\log\hat{p}_t(y_t|x_t;\theta) - \log p_t(y_t|x_t;\theta)\right] \\ &\quad + \frac{1}{\bar{L}_T}\sum_{t=1}^{T}[\tilde{\tau}_{a,t} - 1]\log p_t(y_t|x_t;\theta) \\ &:= B_1(\theta) + B_2(\theta) + B_3(\theta),\end{aligned}$$

and then show that $\sup_{\theta\in\Theta} B_i(\theta) = o_P(1)$, $i = 1, 2, 3$. By (C.3),

$$|B_1(\theta)| \leq \frac{|\log a|}{\bar{L}_T}\sum_{t=1}^{T}\mathbb{I}\{\|x_t\| > N^\gamma\} \leq \frac{|\log a|\sum_{t=1}^{T}\|x_t\|^{1+\delta}}{\bar{L}_T N^{\gamma(1+\delta)}} \leq \frac{|\log a|}{N^{\gamma(1+\delta)}}O_P(1),$$

while by Lemma 5 and (B.2),

$$\begin{aligned}|B_2(\theta)| &\leq \frac{1}{\bar{L}_T}\sum_{t=1}^{T}\mathbb{I}_{A_t(1/4)}|\log\hat{p}_t(y_t|x_t;\theta) - \log p_t(y_t|x_t;\theta)| \\ &\leq O_P(1)\times a^{-1}\sup_{\theta\in\Theta}\sup_{y\in\mathbb{R}^k}\sup_{\|x\|\leq N^\gamma}|\hat{p}_t(y_t|x_t;\theta) - p_t(y_t|x_t;\theta)| = o_P(1).\end{aligned}$$

The final term is bounded by

$$\begin{aligned}|B_3(\theta)| &\leq \frac{1}{\bar{L}_T}\sum_{t=1}^{T}|\tilde{\tau}_{a,t} - 1||\log p_t(y_t|x_t;\theta)| \\ &\leq \frac{1}{\bar{L}_T}\sum_{t=1}^{T}\mathbb{I}\{p_t(y_t|x_t;\theta) < 4a\}|\log p_t(y_t|x_t;\theta)| \\ &\quad + \frac{1}{\bar{L}_T}\sum_{t=1}^{T}\mathbb{I}\{\|x_t\| > N^\gamma\}|\log p_t(y_t|x_t;\theta)| := B_{3,1}(\theta) + B_{3,2}(\theta).\end{aligned}$$

First, as $a \to 0$,

$$\begin{aligned}|B_{3,1}(\theta)| &\leq \frac{1}{\bar{L}_T}\sum_{t=1}^{T}\mathbb{I}\{p_t(y_t|x_t;\theta) < 4a\}\,|\log p_t(y_t|x_t;\theta)| \\ &= \frac{1}{\bar{L}_T}\sum_{t=1}^{T}\mathbb{I}\{|\log p_t(y_t|x_t;\theta)| > |\log(4a)|\}|\log p_t(y_t|x_t;\theta)| \\ &\leq \frac{|\log(4a)|^{-\delta}}{\bar{L}_T}\sum_{t=1}^{T}|\log p_t(y_t|x_t;\theta)|^{1+\delta} = |\log a|^{-\delta}O_P(1).\end{aligned}$$

Similarly,

$$\begin{aligned}
|B_{3,2}(\theta)| &\leq \frac{1}{\bar{L}_T}\sum_{t=1}^{T}\mathbb{I}\{\|x_t\|>N^{\gamma}\}|\log p_t(y_t|x_t;\theta)| \\
&\leq \left\{\frac{1}{\bar{L}_T}\sum_{t=1}^{T}\mathbb{I}\{\|x_t\|>N^{\gamma}\}\right\}^{\delta/(1+\delta)}\left\{\frac{1}{\bar{L}_T}\sum_{t=1}^{T}|\log p_t(y_t|x_t;\theta)|^{1+\delta}\right\}^{1/\delta} \\
&\leq \frac{1}{N^{\gamma(1+\delta)}}\left\{\frac{1}{\bar{L}_T}\sum_{t=1}^{T}\|x_t\|^{1+\delta}\right\}^{\delta/(1+\delta)}\left\{\frac{1}{\bar{L}_T}\sum_{t=1}^{T}|\log p_t(y_t|x_t;\theta)|^{1+\delta}\right\}^{1/\delta} \\
&= N^{-\gamma(1+\delta)}O_P(1).
\end{aligned}$$

The consistency result now follows from Theorem 10.

To show the first of the two asymptotic distribution results, we merely have to strengthen the convergence of $\hat{L}_T(\theta)$ to take place with rate $\bar{L}_T$ (c.f. Theorem 12). One can still apply the above bounds which now have to go to zero with rate $\bar{L}_T$. This is ensured by (B.3.1).

For the second asymptotic distribution result, define $\hat{S}_T(\theta)=\partial\hat{L}_T(\theta)/\partial\theta$ as:

$$\hat{S}_T(\theta) = \sum_{t=1}^{T}\frac{\tau_a(\hat{p}_t(y_t|x_t;\theta))}{\hat{p}_t(y_t|x_t;\theta)}\frac{\partial\hat{p}_t(y_t|x_t;\theta)}{\partial\theta}+\sum_{t=1}^{T}\tau_a'(\hat{p}_t(y_t|x_t;\theta))\log(\hat{p}_t(y_t|x_t;\theta))\frac{\partial\hat{p}_t(y_t|x_t;\theta)}{\partial\theta}.$$

First observe that by (N.3), $U_T(\hat{\theta})=U_T(\theta_0)+V_T(\theta_0)\mathcal{I}_T^{1/2}(\hat{\theta}-\theta_0)+o_P(1)$, where, since $\hat{S}_T(\hat{\theta})=0$,

$$\|U_T(\hat{\theta})\| = \|U_T(\hat{\theta})-\mathcal{I}_T^{-1/2}\hat{S}_T(\hat{\theta})\|\leq\|\mathcal{I}_T^{-1/2}\|\|S_T(\hat{\theta})-\hat{S}_T(\hat{\theta})\|\leq\|\mathcal{I}_T^{-1/2}\|\sup_{\theta\in\Theta}\|S_T(\theta)-\hat{S}_T(\theta)\|.$$

We obtain:

$$\begin{aligned}
&S_T(\theta)-\hat{S}_T(\theta) \\
&= \sum_{t=1}^{T}[\tau_a(\hat{p}_t(y_t|x_t;\theta))-\tilde{\tau}_{a,t}]\frac{\partial\hat{p}_t(y_t|x_t;\theta)/\partial\theta}{\hat{p}_t(y_t|x_t;\theta)}+\sum_{t=1}^{T}\tilde{\tau}_{a,t}\left\{\frac{\partial\hat{p}_t(y_t|x_t;\theta)/\partial\theta}{\hat{p}_t(y_t|x_t;\theta)}-\frac{\partial\hat{p}_t(y_t|x_t;\theta)/\partial\theta}{\hat{p}_t(y_t|x_t;\theta)}\right\} \\
&\quad+\sum_{t=1}^{T}(\tilde{\tau}_{a,t}-1)\frac{\partial\hat{p}_t(y_t|x_t;\theta)/\partial\theta}{\hat{p}_t(y_t|x_t;\theta)}+\sum_{t=1}^{T}\tau_a'(\hat{p}_t(y_t|x_t;\theta))\log(\hat{p}_t(y_t|x_t;\theta))\frac{\partial\hat{p}_t(y_t|x_t;\theta)}{\partial\theta} \\
&:= B_1(\theta)+B_2(\theta)+B_3(\theta)+B_4(\theta),
\end{aligned}$$

and then claim that $\sup_{\theta\in\Theta}\|\mathcal{I}_T^{-1/2}\|\|B_i(\theta)\|=o_P(1)$, $1\leq i\leq 4$, which will yield the desired result. On the set $\{\|x_t\|>N^{\gamma}\}$, $\partial\hat{p}_t(y_t|x_t;\theta)/\partial\theta\xrightarrow{P}\partial p_t(y_t|x_t;\theta)/\partial\theta$, where the latter is bounded, so

$$\|B_1(\theta)\|\leq\frac{C}{a}\sum_{t=1}^{T}\mathbb{I}\left\{\|x_t\|>N^{\gamma}\right\}\leq\frac{C}{aN^{\gamma(1+\delta)}}\sum_{t=1}^{T}\|x_t\|^{1+\delta}.$$

The second term is further split up:

$$\begin{aligned} B_2(\theta) &= \sum_{t=1}^{T} \frac{\tau_a(\hat{p}_t(y_t|x_t;\theta))}{\hat{p}_t(y_t|x_t;\theta)} \left\{ \frac{\partial \hat{p}_t(y_t|x_t;\theta)}{\partial \theta} - \frac{\partial p_t(y_t|x_t;\theta)}{\partial \theta} \right\} \\ &\quad + \sum_{t=1}^{T} \frac{\tau_a(\hat{p}_t(y_t|x_t;\theta))}{p_t(y_t|x_t;\theta)\hat{p}_t(y_t|x_t;\theta)} \frac{\partial p_t(y_t|x_t;\theta)}{\partial \theta} \left\{ \hat{p}_t(y_t|x_t;\theta) - p_t(y_t|x_t;\theta) \right\} \\ &:= B_{2,1}(\theta) + B_{2,2}(\theta), \end{aligned}$$

where

$$\begin{aligned} \|B_{2,1}(\theta)\| &\leq \frac{T}{a} \sup_{y\in\mathbb{R}^k} \sup_{\|x\|<N^\gamma} \sup_{\theta\in\Theta} \left\| \frac{\partial \hat{p}_t(y|x;\theta)}{\partial \theta} - \frac{\partial p_t(y|x;\theta)}{\partial \theta} \right\|, \\ \|B_{2,2}(\theta)\| &\leq \frac{CT}{a^2} \sup_{y\in\mathbb{R}^k} \sup_{\|x\|<N^\gamma} \sup_{\theta\in\Theta} \left| \hat{p}_t(y_t|x_t;\theta) - p_t(y_t|x_t;\theta) \right|. \end{aligned}$$

The third term satisfies:

$$\begin{aligned} \|B_3(\theta)\| &\leq \sum_{t=1}^{T} \mathbb{I}\left\{ p_t(y_t|x_t;\theta) < 4a \right\} \frac{\|\partial p_t(y_t|x_t;\theta)/\partial\theta\|}{p_t(y_t|x_t;\theta)} + \sum_{t=1}^{T} \mathbb{I}\left\{ \|x_t\| > N^\gamma \right\} \frac{\|\partial p_t(y_t|x_t;\theta)/\partial\theta\|}{p_t(y_t|x_t;\theta)} \\ &\leq CTa + N^{-\gamma(1+\delta)} \left\{ \sum_{t=1}^{T} \|x_t\|^{1+\delta} \right\}^{\delta/(1+\delta)} \left\{ \sum_{t=1}^{T} \frac{\|\partial p_t(y_t|x_t;\theta)/\partial\theta\|^{1+\delta}}{p_t(y_t|x_t;\theta)^{1+\delta}} \right\}^{1/\delta}, \end{aligned}$$

and the fourth term:

$$\begin{aligned} B_4(\theta) &= \sum_{t=1}^{T} \tau_a'(\hat{p}_t(y_t|x_t;\theta)) \mathbb{I}\left\{ \|x_t\| \leq N^\gamma \right\} \log(\hat{p}_t(y_t|x_t;\theta)) \left\{ \frac{\partial \hat{p}_t(y_t|x_t;\theta)}{\partial \theta} - \frac{\partial p_t(y_t|x_t;\theta)}{\partial \theta} \right\} \\ &\quad + \sum_{t=1}^{T} \tau_a'(\hat{p}_t(y_t|x_t;\theta)) \mathbb{I}\left\{ \|x_t\| \leq N^\gamma \right\} \log(\hat{p}_t(y_t|x_t;\theta)) \frac{\partial p_t(y_t|x_t;\theta)}{\partial \theta} \\ &\quad + \sum_{t=1}^{T} \tau_a'(\hat{p}_t(y_t|x_t;\theta)) \mathbb{I}\left\{ \|x_t\| > N^\gamma \right\} \log(\hat{p}_t(y_t|x_t;\theta)) \frac{\partial \hat{p}_t(y_t|x_t;\theta)}{\partial \theta} \\ &:= B_{4,1}(\theta) + B_{4,2}(\theta) + B_{4,3}(\theta). \end{aligned}$$

Using $|\tau_a'(x)\log x| \leq C|\log a| \mathbb{I}\{a/2 < x < a\}$ and the same arguments as for consistency regarding trimming sets,

$$\begin{aligned} \|B_{4,1}(\theta)\| &\leq C|\log a| \sum_{t=1}^{T} \mathbb{I}\left\{ \|x_t\| \leq N^\gamma \right\} \left\| \frac{\partial \hat{p}_t(y_t|x_t;\theta)}{\partial \theta} - \frac{\partial p_t(y_t|x_t;\theta)}{\partial \theta} \right\| \\ &\leq C|\log a| T \sup_{y\in\mathbb{R}^k} \sup_{\|x\|<N^\gamma} \sup_{\theta\in\Theta} \left\| \frac{\partial \hat{p}_t(y|x;\theta)}{\partial \theta} - \frac{\partial p_t(y|x;\theta)}{\partial \theta} \right\|, \\ \|B_{4,2}(\theta)\| &\leq \log a \sum_{t=1}^{T} \mathbb{I}\left\{ |\log(p_t(y|x;\theta))| > |\log(2a)| \right\} \frac{\partial p_t(y_t|x_t;\theta)}{\partial \theta} \leq C \sum_{t=1}^{T} \left| \log p_t(y|x;\theta) \right|, \end{aligned}$$

$$\|B_{4,3}(\theta)\| \leq C \log a \sum_{t=1}^{T} \mathbb{I}\{\|x_t\| > N^{\gamma}\} \frac{\partial \hat{p}_t(y_t|x_t;\theta)}{\partial \theta} \leq \frac{C|\log a| \sum_{t=1}^{T} \|x_t\|^{1+\delta}}{h^{k+1} N^{\gamma(1+\delta)}}.$$

One now realizes that Lemmas 5 and 6 together with the conditions in (B.3.2) ensure that all the above bounds are $o_P(\|\mathcal{I}_T^{1/2}\|)$. ∎

Proof of Theorem 4 Define $Y_i^x = Y_i^{x,\theta_0}$, and $\underline{Y}_N = \{Y_i^x : i = 1, ..., N, x \in \mathbb{R}^l\}$. Also, write $p_t = p(y_t|x_t;\theta_0)$, $\dot{p}_t = \partial p(y_t|x_t;\theta_0)/\partial\theta$, $s(y_t|x_t) = \dot{p}_t/p_t$, $\tau_{a,t} = \tau_a(\hat{p}_t(y_t|x_t;\theta_0))$, and let $p(x)$ denote the marginal density of x_t. We then have $0 = \hat{S}_T(\theta_0)/\sqrt{T} + \left\{\hat{H}_T(\bar{\theta})/T\right\}\sqrt{T}(\hat{\theta} - \theta_0)$, where $\hat{S}_T(\theta)$ is given in (6). We make a functional Taylor expansion w.r.t. $\dot{p}_t$ and p_t, yielding

$$D_{T,N} = \sum_{t=1}^{T} \tau_{a,t} \left\{ \frac{\widehat{\dot{p}}_t - \dot{p}_t}{p_t} - \frac{\dot{p}_t}{p_t^2}(\hat{p}_t - p_t) \right\},$$

where

$$\begin{aligned} &\hat{S}_T(\theta_0)/\sqrt{T} - S_T(\theta_0)/\sqrt{T} - D_{T,N}/\sqrt{T} \\ &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \tau_{a,t} \left\{ \frac{\widehat{\dot{p}}_t}{\hat{p}_t} - \frac{\dot{p}_t}{p_t} - \frac{\widehat{\dot{p}}_t - \dot{p}_t}{p_t} + \frac{\dot{p}_t}{p_t^2}(\hat{p}_t - p_t) \right\} + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \widehat{\dot{p}}_t \tau'_{a,t} \log \hat{p}_t. \end{aligned}$$

Using the same arguments as in the proof of Theorem 2, the second term can be shown to be $o_P(1)$, while the first is bounded by

$$\begin{aligned} &\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\tau_{a,t}}{|\hat{p}_t| p_t} \frac{\|\dot{p}_t\|}{p_t} \left\{ \|\widehat{\dot{p}}_t - \dot{p}_t\|^2 + \|\hat{p}_t - p_t\|^2 \right\} \\ &\leq \left\{ \frac{1}{T} \sum_{t=1}^{T} \frac{\|\dot{p}_t\|}{p_t} \right\} \sqrt{T} a^{-2} \sup_{y \in \mathbb{R}^k, \|x\| \leq N^{\gamma}} \left\{ \|\widehat{\dot{p}}(y|x) - \dot{p}(y|x)\|_{\infty}^2 + \|\hat{p}(y|x) - p(y|x)\|^2 \right\}, \end{aligned}$$

where $\frac{1}{T}\sum_{t=1}^{T} \|\dot{p}_t\|/p_t = O_P(1)$, $\sup_{\left\{y \in \mathbb{R}^k, \|x\| \leq N^{\gamma}\right\}} \|\widehat{\dot{p}}(y|x) - \dot{p}(y|x)\|_{\infty} = o_P\left(T^{-1/4}a\right)$ and $\sup_{\left\{y \in \mathbb{R}^k, \|x\| \leq N^{\gamma}\right\}} \|\hat{p}(y|x) - p(y|x)\| = o_P\left(T^{-1/4}a\right)$ under the conditions on N, h and a by Lemma 5 and 6.

Next, using standard U-statistics results for absolutely regular sequences—e.g. **?**, $D_{T,N}/\sqrt{T} = \bar{D}_N/\sqrt{T} + o_P(1)$, where,

$$\bar{D}_N = \mathbb{E}\left[\tau_{a,t} \left\{ \frac{\widehat{\dot{p}}_t - \dot{p}_t}{p_t} - \frac{\dot{p}_t}{p_t^2}(\hat{p}_t - p_t) \right\} \middle| \underline{Y}_N \right].$$

Under the conditions given in (vii), we have

$$\bar{D}_N = \mathbb{E}\left[\tau_{a,t} \left\{ \frac{\widehat{\dot{p}}_t}{p_t} - \frac{\dot{p}_t}{p_t^2}\hat{p}_t \right\} \middle| \underline{Y}_N \right]$$

$$
\begin{aligned}
&= \int\int \tau_a\left(\hat{p}(y|x)\right)\widehat{\hat{p}}(y|x)p(x)dydx - \int\int \tau_a\left(\hat{p}(y|x)\right)s(y|x)\hat{p}(y|x)p(x)dydx \\
&= \frac{1}{N}\sum_{i=1}^{N}\int \dot{Y}_i^x\left\{\frac{1}{h^{k+1}}\int K'\left(\frac{Y_i^x - y}{h}\right)dy\right\}p(x)dx \\
&\quad - \frac{1}{N}\sum_{i=1}^{N}\int\left\{\frac{1}{h^k}\int s(y|x)K\left(\frac{Y_i^x - y}{h}\right)dy\right\}p(x)dx + o_P(1) \\
&= -\frac{1}{N}\sum_{i=1}^{N}\int s(Y_i^x|x)p(x)dx + O_P(h^r).
\end{aligned}
$$

Finally,

$$
||\frac{1}{T}\hat{H}_T(\bar{\theta}) - H_0|| \leq \sup_{\theta\in\Theta}||\frac{1}{T}\hat{H}_T(\theta) - \frac{1}{T}H_T(\theta)|| + \sup_{\theta\in\Theta}||\frac{1}{T}H_T(\theta) - H\left(\theta\right)|| + \left\|H(\bar{\theta}) - H\left(\theta_0\right)\right\|.
$$

The convergence of the second and third term follows from standard ULLN results under the conditions we have imposed on $\partial^2 \log p\left(y|x;\theta\right)/\partial\theta\partial\theta'$. The first term is bounded by

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\tau_{a,t}\left\|\frac{\widehat{\ddot{p}}_t}{\hat{p}_t} - \frac{\ddot{p}_t}{p_t}\right\| + \frac{1}{T}\sum_{t=1}^{T}\tau_{a,t}\left\|\frac{\widehat{\dot{p}}_t\widehat{\dot{p}}_t'}{\hat{p}_t^2} - \frac{\dot{p}_t\dot{p}_t'}{p_t^2}\right\| \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\ddot{p}}_t\tau_{a,t}'\log\hat{p}_t + \widehat{\dot{p}}_t\widehat{\dot{p}}_t'\tau_{a,t}''\log\hat{p}_t + \frac{\widehat{\dot{p}}_t\widehat{\dot{p}}_t'}{\hat{p}_t}\tau_{a,t}'\right\| + \frac{1}{T}\sum_{t=1}^{T}|1-\tau_{a,t}|\left\|\frac{\ddot{p}_t}{p_t} - \frac{\dot{p}_t\dot{p}_t'}{p_t^2}\right\|.
\end{aligned}
$$

Following the same steps as before, each of the above terms can be shown to be $o_P(1)$ uniformly in θ. In total, with $T/N \to \alpha \geq 0$,

$$
\begin{aligned}
\sqrt{T}(\hat{\theta} - \theta_0) &= \left(H_0 + o_P(1)\right)^{-1}\left\{\frac{1}{\sqrt{T}}S_T(\theta_0) - \sqrt{\frac{T}{N}}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\int s(Y_i^x|x)p(x)dx + o_P(1)\right\} \\
&\xrightarrow{d} \mathcal{N}\left(0, (1+\alpha)i(\theta_0)^{-1}\right). \quad \blacksquare
\end{aligned}
$$

B Properties of the Simulated Density

Lemma 5 *Assume that (A.1)–(A.2) and (K.1) hold. Then for any $t \geq 1$, $\hat{p}_t$ given in (NPSMLE 1) satisfies*

$$
\begin{aligned}
&\sup_{y\in\mathbb{R}^k}\sup_{\|x\|\leq N^\gamma}\sup_{\theta\in\Theta}|\hat{p}_t(y|x;\theta) - p_t(y|x;\theta)| \\
&\quad = C_{0,1}\left(\mathbb{E}\left[\Lambda_t^2\right], K\right)O_P\left(\sqrt{\log(N)}N^{-1/2}h^{-k-1}\right) + C_{0,2}\left(K, D_y^r p_t\right)h^r
\end{aligned}
$$

for any $\gamma > 0$, and where $C_{0,i}$, $i = 1, 2$, are given in Lemma 7.

Proof This follows from Lemma 7 with $Y_i(\alpha) = Y_{t,i}^{x,\theta}$, $Z(\alpha) = 1$, $\alpha = (x,\theta)$ and $A = \mathbb{R}^l \times \Theta$. ∎

Lemma 6 *Assume that (A.1)–(A.5) and (K.1) hold. Then for any $t \geq 1$, $\partial \hat{p}_t / \partial \theta$ given in (5) satisfies*

$$\begin{aligned}
&\sup_{y \in \mathbb{R}^k} \sup_{\|x\| \leq N^\gamma} \sup_{\theta \in \Theta} \|\partial \hat{p}_t(y|x;\theta)/\partial\theta - \partial p_t(y|x;\theta)/\partial\theta\| \\
&\quad = C_{1,1}\left(\mathbb{E}\left[\Lambda_t^2\right], K\right) O_P\left(\sqrt{\log(N)} N^{(2\gamma\beta_2 - 1)/3} h^{-k-2}\right) + C_{1,2}(K, D_y^r \partial p_t/\partial\theta) h^{r-1}
\end{aligned}$$

for any $\gamma > 0$, and where $C_{1,i}$, $i = 1, 2$, are given in Lemma 7.

If furthermore, the second derivative of $Y_{t,i}^{x,\theta}$ exists, and this satisfies the same bounds as the first one, then together with $\mathbb{E}\left[\Lambda_t^4\right] < \infty$:

$$\begin{aligned}
&\sup_{y \in \mathbb{R}^k} \sup_{\|x\| \leq N^\gamma} \sup_{\theta \in \Theta} \left\|\partial^2 \hat{p}_t(y|x;\theta)/\partial\theta\partial\theta' - \partial^2 p_t(y|x;\theta)/\partial\theta\partial\theta'\right\| \\
&\quad = C_{1,1}\left(\mathbb{E}\left[\Lambda_t^4\right], K\right) O_P\left(\sqrt{\log(\nu_N)} N^{(2\gamma\beta_2 - 1)/3} h^{-k-2}\right) + C_{1,2}\left(K, D_y^r \partial p_t/\partial\theta\right) h^{r-1} \\
&\qquad + C_{2,1}\left(\mathbb{E}\left[\Lambda_t^2\right], K\right) O_P\left(\sqrt{\log(\nu_N)} N^{(2\gamma\beta_2 - 1)/3} h^{-k-3}\right) + C_{2,2}\left(K, D_y^r \partial p_t/\partial\theta\right) h^{r-2}
\end{aligned}$$

where $C_{2,i}$, $i = 1, 2$, are given in Lemma 7.

Proof This follows from Lemma 7 with $Y_i(\alpha) = Y_{t,i}^{x,\theta}$, $Z(\alpha) = \dot{Y}_{t,i}^{x,\theta}$ and $\ddot{Y}_{t,i}^{x,\theta}$, $\alpha = (x,\theta)$ and $A = \mathbb{R}^l \times \Theta$. ∎

Lemma 7 *Assume that:*

(i) *K satisfies (K.1).*

(ii) *$\{Y_i(\alpha), Z_i(\alpha)\}$, $i = 1, ..., N$, are i.i.d. copies of $(Y(\alpha), Z(\alpha)) \sim p(y, z; \alpha)$ for any $\alpha \in A \subseteq \mathbb{R}^D$.*

(iii) *$\|Y(\alpha) - Y(\alpha')\| \leq \Lambda \|\alpha - \alpha'\|^{\beta_1}$, $\|Z(\alpha) - Z(\alpha')\| \leq \Lambda \|\alpha - \alpha'\|^{\beta_1}$ and $\|Z(\alpha)\| \leq \Lambda \|\alpha\|^{\beta_2}$ where $\beta_1 > 0$, $\beta_2 \geq 0$ and $\mathbb{E}\Lambda^2 < \infty$.*

(iv) *$G(y;\alpha) = \mathbb{E}\left[Z(\alpha)|Y(\alpha) = y\right] p(y;\alpha)$, where $p(y;\alpha) = \int p(y,z;\alpha)dz$ is r times differentiable in y, with $\sup_{\alpha \in A} \sup_{y \in \mathbb{R}^k} \sum_{|\lambda|=r} \left|D_y^\lambda G(y;\alpha)\right| < \infty$.*

(v) *$A_N = \{\alpha \in A | \|\alpha\| \leq C N^\gamma\}$ for some $C > 0$ and $\gamma \geq 0$.*

Then $D_y^\lambda \hat{G}(y;\alpha) := \frac{1}{N h^{k+|\lambda|}} \sum_{i=1}^N Z_i(\alpha) D^\lambda K\left(\frac{y - Y_i(\alpha)}{h}\right)$ satisfies:

1. $\beta_2 > 0$ *with* $\nu_N = N^{(2b+1)/(2\beta_1)+\gamma}$:

$$\begin{aligned}
&\sup_{\alpha\in A_N}\sup_{y\in\mathbb{R}^k}\left|D_y^\lambda \hat{G}(y;\alpha) - D_y^\lambda G(y;\alpha)\right| \\
&\quad = \quad C_{\lambda,1}(\mathbb{E}\Lambda^2, K) O_P\left(\sqrt{\log(\nu_N)} N^{(2\gamma\beta_2-1)/3} h^{-k-|\lambda|-1}\right) + C_{\lambda,2}(K, D_y^r G) h^{r-|\lambda|}.
\end{aligned}$$

2. $\beta_2 = 0$ *with* $\nu_N = N^{\gamma+1/(2\beta_1)}$:

$$\begin{aligned}
&\sup_{\alpha\in A_N}\sup_{y\in\mathbb{R}^k}\left|D_y^\lambda \hat{G}(y;\alpha) - D_y^\lambda G(y;\alpha)\right| \\
&\quad = \quad C_{\lambda,1}\left(\mathbb{E}\Lambda^2, K\right) O_P\left(\sqrt{\log(\nu_N)} N^{-1/2} h^{-k-1}\right) + C_{\lambda,2}(K, D_y^r G) h^r.
\end{aligned}$$

In both cases,

$$\begin{aligned}
C_{\lambda,1}\left(\mathbb{E}\Lambda^2, K\right) &= \left(\mathbb{E}\Lambda^2 + 1\right)\int (1+\|w\|)\sup_{b\geq 1}|\Psi_\lambda(bw)|\, dw, \\
C_{\lambda,2}(K, D_y^r G) &= \frac{1}{r!}\sup_{\alpha\in A}\sup_{w\in\mathbb{R}^k}\sum_{|\mu|=r}\left|D_y^\mu G(w;\alpha)\right|\int_{\mathbb{R}^k}\|z\|^r |K(z)| dz.
\end{aligned}$$

Proof The proof is a modified version of the one found in **?**, Proof of Theorem 1. By the Fourier inversion formula, we obtain

$$\begin{aligned}
&\sup_{\alpha\in A_N}\sup_{y\in\mathbb{R}^k}\left\|\hat{G}(y;\alpha) - \mathbb{E}\hat{G}(y;\alpha)\right\| \\
&\quad \leq \quad \int \sup_{\alpha\in A_N}\left\|\frac{1}{N}\sum_{j=1}^N Z_j(\alpha)\exp(iw'Y_j(\alpha)) - \mathbb{E}\left[Z(\alpha)\exp\left(iw'Y(\alpha)\right)\right]\right\|\frac{\Psi(hw)}{h^{\lambda+1}} dw. \qquad (9)
\end{aligned}$$

Define $Q_N(\alpha) = \sum_{j=1}^N Z_j(\alpha)\exp(iw'Y_j(\alpha))/N$, and $Q(\alpha) = \mathbb{E}\left[Z_j(\alpha)\exp(iw'Y_j(\alpha))\right]$. We then claim that for any $w \in \mathbb{R}^k$, in the case where $Z(\alpha)$ is unbounded,

$$\sup_{\alpha\in A_N}\|Q_N(\alpha) - Q(\alpha)\| \leq (\|w\|+1)\, C\left(\mathbb{E}\Lambda, K\right) N^{-(2\gamma\beta_2-1)/3}, \qquad (10)$$

while in the bounded case,

$$\sup_{\alpha\in A_N}\|Q_N(\alpha) - Q(\alpha)\| \leq (\|w\|+1)\, C(\mathbb{E}\Lambda, K) N^{-1/2}. \qquad (11)$$

Substituting each of these into (9) together with

$$\left\|\mathbb{E}[D_y^\lambda \hat{G}(y;\alpha)] - D_y^\lambda G(y;\alpha)\right\| \quad \leq \quad \frac{h^{r-\lambda}}{r!}\int_{\mathbb{R}k}\sum_{|\mu|=r}\left|D_y^\mu G(y-hz;\alpha)\right| z^\mu\, |K(z)|\, dz$$

$$\leq \frac{h^{r-\lambda}}{r!} \sup_{w\in\mathbb{R}^k} \sum_{|\mu|=r} \left|D_y^\mu G(w;\alpha)\right| \int_{\mathbb{R}^k} \|z\|^r |K(z)| dz$$

then yields the result.

We now prove (10). Define truncated versions $\bar{Q}_N(\alpha) = \sum_{j=1}^N \bar{Z}_j(\alpha)\exp(w'Y_j(\alpha))/N$ and $\bar{Q}(\alpha) = \mathbb{E}\left[\bar{Z}_j(\alpha)\exp(iw'Y_j(\alpha))\right]$, where $\bar{Z}_j(\alpha) = Z_j(\alpha)\mathbb{I}\left\{\|Z_j(\alpha)\| \leq N^b\right\}$ for some $b > 0$ which will be specified later. Then (10) will follow if

$$\sup_{\alpha\in A_N} \left\|Q_N(\alpha) - \bar{Q}_N(\alpha)\right\| = \mathbb{E}[\Lambda^2] O_P\left(N^{(2\gamma\beta_2-1)/3}\right), \tag{12}$$

$$\sup_{\alpha\in A_N} \left\|Q(\alpha) - \bar{Q}(\alpha)\right\| = \mathbb{E}[\Lambda^2] N^{(2\gamma\beta_2-1)/3}, \tag{13}$$

$$\sup_{\alpha\in A_N} \left\|\bar{Q}_N(\alpha) - \bar{Q}(\alpha)\right\| = C(\mathbb{E}\Lambda, K) O_P\left(N^{(2\gamma\beta_2-1)/3}\right). \tag{14}$$

For Eq. (12)–(13),

$$\begin{aligned}
\mathbb{E}\left[\sup_{\alpha\in A_N} \left\|Q_N(\alpha) - \bar{Q}_N(\alpha)\right\|\right] &\leq \mathbb{E}\left[\sup_{\alpha\in A_N} \|Z(\alpha)\|^2 \mathbb{E}\left\{\|Z(\alpha)\| > N^b\right\}\right] \\
&\leq \mathbb{E}\left[\sup_{\alpha\in A_N} \|Z(\alpha)\|^2\right]^{1/2} P\left(\sup_{\alpha\in A_N} \|Z(\alpha)\| > N^b\right)^{1/2} \\
&\leq \mathbb{E}\left[\sup_{\alpha\in A_N} \|Z(\alpha)\|^2\right] / N^b \leq \mathbb{E}\left[\Lambda^2\right] N^{2\gamma\beta_2-b},
\end{aligned}$$

and, similarly, $\sup_{\alpha\in A_N} \left\|Q(\alpha) - \bar{Q}(\alpha)\right\| \leq \mathbb{E}\left[\Lambda^2\right] N^{2\gamma\beta_2-b}$. For Eq. (14), define hypercubes

$$B_{N,k} := \left\{\alpha \in A | \, \|\alpha - \alpha_{N,k}\| \leq \frac{N^\gamma}{\nu_N}\right\}, \quad k = 1, ..., \nu_N^D,$$

where $\{\alpha_{N,k}\}$ have been chosen such that $A_N \subseteq \cup_{k=1}^{\nu_N^D} B_{N,k}$. We then have for $\alpha \in B_{N,k}$,

$$\begin{aligned}
\left\|\bar{Q}_N(\alpha) - \bar{Q}_N(\alpha_{N,k})\right\| &\leq \frac{1}{N}\sum_{j=1}^N \left\|\bar{Z}_j(\alpha)\exp(iw'Y_j(\alpha)) - \bar{Z}_j(\alpha_{N,k})\exp(iw'Y_j(\alpha_{N,k}))\right\| \\
&\leq \frac{\|w\|}{N}\sum_{j=1}^N \left\|\bar{Z}_j(\alpha)\right\| \|Y_j(\alpha) - Y_j(\alpha_{N,k})\| + \frac{1}{N}\sum_{j=1}^N \left\|\bar{Z}_j(\alpha) - \bar{Z}_j(\alpha_{N,k})\right\| \\
&\leq \frac{\|w\| \|\alpha - \alpha_{N,k}\|^{\beta_1} N^b}{N}\sum_{j=1}^N \|\Lambda_j\| + \frac{\|\alpha - \alpha_{N,k}\|^{\beta_1}}{N}\sum_{j=1}^N \|\Lambda_j\| \\
&\leq (\|w\| + 1)\left(\mathbb{E}\left[\Lambda^2\right] + o_P(1)\right)\frac{N^{b+\gamma\beta_1}}{\nu_N^{\beta_1}},
\end{aligned}$$

and similarly for $\left\|\bar{Q}(\alpha_{N,k}) - \bar{Q}(\alpha)\right\|$. Thus,

$$\left\|\bar{Q}_N(\alpha) - \bar{Q}(\alpha)\right\| \leq \left\|\bar{Q}_N(\alpha) - \bar{Q}_N(\alpha_{N,k})\right\| + \left\|\bar{Q}_N(\alpha_{N,k}) - \bar{Q}(\alpha_{N,k})\right\| + \left\|\bar{Q}(\alpha_{N,k}) - \bar{Q}(\alpha)\right\|$$

$$\leq \quad (\|w\|+1)\left(\mathbb{E}\left[\Lambda^2\right]+o_P(1)\right)\frac{N^{b+\gamma\beta_1}}{\nu_N^{\beta_1}}+\left\|\bar{Q}_N(\alpha_{N,k})-\bar{Q}(\alpha_{N,k})\right\|.$$

The last term is bounded using Bernstein's inequality for bounded zero-mean random variables,

$$\begin{aligned} P\left(\max_{1\leq k\leq \nu_N^D}\left\|\bar{Q}_N(\alpha_{N,k})-\bar{Q}(\alpha_{N,k})\right\|>M\right) &\leq \sum_{k=1}^{\nu_N^D} P\left(\left\|\bar{Q}_n(\alpha_{N,k})-\bar{Q}(\alpha_{N,k})\right\|>M\right) \\ &\leq \nu_N^D 2\exp\left[-M^2N^{1-b}\right]. \end{aligned}$$

Choosing $M=C\sqrt{\log\left(\nu_N^D\right)}N^{(1-b)/2}$, $\nu_N=\left[N^{(2b+1)/(2\beta_1)+\gamma}\right]$, and $b=2\gamma\beta_1-1/3$, we obtain the desired result.

The proof of (11) follows along the same lines but without truncation such that

$$\begin{aligned} \|Q_N(\alpha)-Q_N(\alpha_{N,k})\| &\leq \frac{(\|w\|+1)N^{\gamma\beta_1}}{\nu_N^{\beta_1}}\left(\mathbb{E}\left[\Lambda^2\right]+o_P(1)\right), \\ P\left(\max_{1\leq k\leq \nu_N^D}\|Q_N(\alpha_{N,k})-Q(\alpha_{N,k})\|>M\right) &\leq \nu_N^D 2\exp\left[-\frac{M^2N}{\bar{Z}}\right], \end{aligned}$$

where $\bar{Z}$ is the upper bound of $\sup_\alpha Z(\alpha)$. Choosing $M=C\sqrt{\log\left(\nu_N^D\right)}N^{-1/2}$ and $\nu_N=\left[N^{\gamma+1/(2\beta_1)}\right]$, we obtain the desired result. ∎

C Some General Results for Approximate Estimators

Consider $\hat{\theta}=\arg\max_{\theta\in\Theta}\hat{L}_T(\theta)$ and $\tilde{\theta}=\arg\max_{\theta\in\Theta}L_T(\theta)$, where $L_T(\theta)$ is the true but infeasible criterion function, not necessarily the log-likelihood, and $\hat{L}_T(\theta)=\hat{L}_{T,N}(\theta)$ is an approximation of $L_T(\theta)$. In the following we give sufficient conditions for the approximate estimator, $\hat{\theta}$, to be asymptotically equivalent to $\tilde{\theta}$.

C.1 Existence of N

Proposition 8 *Assume that (C.1) holds and for a given $T\geq 1$:*

(i) *$\theta\mapsto L_T(\theta)$ is continuous.*

(ii) *$\theta\mapsto\hat{L}_T(\theta)$ is continuous.*

(iii) *$\sup_{\theta\in\Theta}|\hat{L}_T(\theta)-L_T(\theta)|\xrightarrow{P}0$ as $N\to\infty$.*

Then $\hat{\theta}\xrightarrow{P}\tilde{\theta}$ as $N\to\infty$ for this $T\geq 1$.

Proof See **?**, Theorem 1.

Proposition 9 *Assume that the assumptions of Proposition 8 are satisfied together with (C.2). Then there exists a sequence $N \to \infty$ such that $\hat{\theta} \xrightarrow{P} \theta_0$.*

Assume furthermore that, for some random variable Z and some sequence of positive definite matrices $\mathcal{I}_T^{-1} \to 0$, $\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) \xrightarrow{d} Z$. Then there exists a sequence $N \to \infty$ such that $\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) \to Z$.

Proof To show consistency, take any positive sequence $\{\delta_T\}$ with $\delta_T \to 0$. For each $T \geq 1$, choose $N \geq 1$ such that $\|\hat{\theta} - \tilde{\theta}\| = O_P(\delta_T)$, which is possible due to Proposition 8. Thus, $\|\hat{\theta} - \theta_0\| \leq \|\hat{\theta} - \tilde{\theta}\| + \|\tilde{\theta} - \theta_0\| = O_P(\delta_T) + o_P(1) = o_P(1)$, by our choice of $\{\delta_T\}$ and (C.2).

The weak convergence result is shown similarly. For each $T \geq 1$, choose $N(T) \geq 1$ such that $\|\mathcal{I}_T^{1/2}(\hat{\theta} - \tilde{\theta})\| = O_P(\delta_T)$, and we proceed as before. ∎

C.2 Conditions on N for Consistency

Theorem 10 *Assume that:*

(i) *$\theta \mapsto L_T(\theta)$ is stochastically equicontinuous.*

(ii) *$\tilde{\theta} \xrightarrow{P} \theta_0$.*

(iii) *There exists a sequence $N = N(T) \to \infty$ such that $\sup_{\theta \in \Theta} |\hat{L}_T(\theta) - L_T(\theta)| \xrightarrow{P} 0$ as $T \to \infty$.*

Then $\hat{\theta} \xrightarrow{P} \theta_0$.

Proof We wish to show that for any $\varepsilon > 0$, $P(\|\hat{\theta} - \tilde{\theta}\| > \varepsilon) \to 0$ as $T \to \infty$. Let $\varepsilon > 0$ be given. Then by (i) there exists a $\delta > 0$ such that, $\|\theta - \tilde{\theta}\| > \varepsilon$ implies $|L_T(\theta) - L_T(\tilde{\theta})| \geq \delta$ with probability tending to 1. Thus, as $T \to \infty$,

$$P(\|\hat{\theta} - \tilde{\theta}\| > \varepsilon) \leq P(|L_T(\hat{\theta}) - L_T(\tilde{\theta})| \geq \delta).$$

We then have to show that the RHS converges to zero. Since $\tilde{\theta}$ is the maximizer of $L_T(\theta)$, we know that $L_T(\tilde{\theta}) \geq L_T(\hat{\theta})$. Thus,

$$|L_T(\tilde{\theta}) - L_T(\hat{\theta})| \quad = \quad L_T(\tilde{\theta}) - L_T(\hat{\theta}) = \{\hat{L}_T(\hat{\theta}) - L_T(\hat{\theta})\} + \{L_T(\tilde{\theta}) - \hat{L}_T(\hat{\theta})\},$$

where, by (iii), $\hat{L}_T(\hat{\theta}) - L_T(\hat{\theta}) \leq |L_T(\hat{\theta}) - \hat{L}_T(\hat{\theta})| \leq \sup_{\theta\in\Theta} |L_T(\theta) - \hat{L}_T(\theta)| = o_P(1)$, while, by the definition of $\hat{\theta}$ and again using (iii),

$$L_T(\tilde{\theta}) - \hat{L}_T(\hat{\theta}) \leq L_T(\tilde{\theta}) - \hat{L}_T(\tilde{\theta}) \leq \sup_{\theta\in\Theta} |L_T(\theta) - \hat{L}_T(\theta)| = o_P(1).$$

In conclusion, $L_T(\hat{\theta}) \xrightarrow{P} L_T(\tilde{\theta})$ as desired. ∎

More primitive conditions for (i) above can be arrived at by utilizing the results of **?**.

Corollary 11 *Assume that (ii) and (iii) of Theorem 10 hold. Further assume that Θ is compact and there exists a non-random, equicontinuous function $\bar{L}_T(\theta)$ such that $\sup_{\theta\in\Theta} |L_T(\theta) - \bar{L}_T(\theta)| \xrightarrow{P} 0$. Then $\hat{\theta} \xrightarrow{P} \theta_0$ for this sequence N.*

C.3 Conditions on N for Convergence Rate

Theorem 12 *Assume that (C.2) and (N.1)–(N.4) hold together with:*

(i) *$\theta \mapsto L_T(\theta)$ is thrice differentiable in a neighborhood $\mathcal{N}$ of θ_0.*

(ii) *There exists a sequence $N = N(T) \to \infty$ such that $\hat{\theta} \xrightarrow{P} \theta_0$ and $\sup_{\theta\in\mathcal{N}} |\hat{L}_T(\theta) - L_T(\theta)| = O_P(b_T)$ for some $b_T \to 0$ in $\mathcal{N}$ as $T \to \infty$.*

Then $\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) \xrightarrow{d} V^{-1}U$ and $\mathcal{I}_T^{1/2}(\hat{\theta} - \tilde{\theta}_T) = O_P(\sqrt{b_T})$ for this sequence N. In particular, $\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} V^{-1}U$.

Proof Using standard arguments, one can show that under (C.2) and (N.1)–(N.4),

$$\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) = V_T^{-1}(\theta_0)U_T(\theta_0) + o_P(1) \xrightarrow{d} V^{-1}U;$$

see e.g. **?**. Next, by a standard Taylor expansion, we have w.p.a.1,

$$L_T(\tilde{\theta}) - L_T(\hat{\theta}) \quad = \quad \frac{1}{2}(\tilde{\theta} - \hat{\theta})' H_T(\bar{\theta})(\tilde{\theta} - \hat{\theta}) = \frac{1}{2}(\tilde{\theta} - \hat{\theta})' \mathcal{I}_T^{1/2} V_T(\bar{\theta}) \mathcal{I}_T^{1/2}(\tilde{\theta} - \hat{\theta}),$$

for some $\bar{\theta} \in [\tilde{\theta}, \hat{\theta}]$, since $S_T(\tilde{\theta}) = 0$ by the definition of $\tilde{\theta}$. By another Taylor-expansion,

$$\|V_T(\bar{\theta}) - V_T(\theta_0)\| \leq W_T(\breve{\theta}) \times \|\bar{\theta} - \theta_0\|,$$

for some $\breve{\theta} \in [\theta_0, \bar{\theta}]$. The first term on the RHS is $O_P(1)$ by (N.3), while the second is $o_P(1)$ since $\tilde{\theta}$ and $\hat{\theta}$, and thereby $\bar{\theta}$, converge towards θ_0. In total,

$$L_T(\tilde{\theta}) - L_T(\hat{\theta}) = \frac{1}{2}(\tilde{\theta} - \hat{\theta})' \mathcal{I}_T^{1/2} V_T(\theta_0) \mathcal{I}_T^{1/2}(\tilde{\theta} - \hat{\theta}) + o_P(1).$$

We now use the exact same argument as before for the LHS. Since $\tilde{\theta}$ is the maximizer of $L_T(\theta)$, we know that $L_T(\tilde{\theta}) \geq L_T(\hat{\theta})$. Thus,

$$L_T(\tilde{\theta}) - L_T(\hat{\theta}) = \{L_T(\tilde{\theta}) - \hat{L}_T(\hat{\theta})\} + \{\hat{L}_T(\hat{\theta}) - L_T(\hat{\theta})\},$$

where, by (ii), $\hat{L}_T(\hat{\theta}) - L_T(\hat{\theta}) \leq \sup_{\theta \in \mathcal{N}} |\hat{L}_T(\theta) - L_T(\theta)| = O_P(b_T)$, while, by the definition of $\tilde{\theta}$ and (ii),

$$L_T(\tilde{\theta}) - \hat{L}_T(\hat{\theta}) \leq L_T(\tilde{\theta}) - \hat{L}_T(\tilde{\theta}) \leq \sup_{\theta \in \mathcal{N}} |L_T(\theta) - \hat{L}_T(\theta)| = O_P(b_T).$$

We have now shown that $\frac{1}{2}(\tilde{\theta} - \hat{\theta})' \mathcal{I}_T^{1/2} V_T(\theta_0) \mathcal{I}_T^{1/2} (\tilde{\theta} - \hat{\theta}) = O_P(b_T)$. On the other hand,

$$\frac{1}{2}(\tilde{\theta} - \hat{\theta})' \mathcal{I}_T^{1/2} V_T(\theta_0) \mathcal{I}_T^{1/2} (\tilde{\theta} - \hat{\theta}) \xrightarrow{d} \|\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0)\|_V^2,$$

where $\|\theta\|_V^2 = \theta' V \theta$ is an Euclidean norm since V is nonsingular. This is equivalent to $\|\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0)\|^2 = O_P(b_T)$. The second result now follows from Slutzky's theorem. ∎

Theorem 13 *Assume that (C.2) and (N.1)–(N.4) hold together with*

(i) $\theta \mapsto L_T(\theta)$ *is thrice differentiable in a neighbourhood* $\mathcal{N}$ *of* θ_0.

(ii) $\theta \mapsto \hat{L}_T(\theta)$ *is once differentiable in a neighborhood* $\mathcal{N}$ *of* θ_0.

(iii) *There exists a sequence* $N = N(T) \to \infty$ *such that* $\hat{\theta} \xrightarrow{P} \theta_0$ *and* $\sup_{\theta \in \mathcal{N}} \|\hat{S}_T(\theta) - S_T(\theta)\| = o_P(\|\mathcal{I}_T^{1/2}\|)$ *as* $T \to \infty$.

Then $\mathcal{I}_T^{1/2}(\tilde{\theta} - \theta_0) \xrightarrow{d} V^{-1}U$ *and* $\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} V^{-1}U$ *for this sequence* N.

Proof By a standard Taylor expansion,

$$\mathcal{I}_T^{-1/2} S_T(\hat{\theta}) = \mathcal{I}_T^{-1/2} S_T(\theta_0) + \mathcal{I}_T^{-1/2} H_T(\theta_0) \mathcal{I}_T^{-1/2} \; \mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) + o_P(1),$$

where

$$\|\mathcal{I}_T^{-1/2} S_T(\hat{\theta})\| = \|\mathcal{I}_T^{-1/2}\{S_T(\hat{\theta}) - \hat{S}_T(\hat{\theta})\}\| \leq \sup_{\theta \in \mathcal{N}} \|\mathcal{I}_T^{-1/2}\{S_T(\theta) - \hat{S}_T(\theta)\}\| = o_P(1).$$

Thus, $\mathcal{I}_T^{1/2}(\hat{\theta} - \theta_0) = V_T^{-1}(\theta_0) \{U_T(\theta_0) + o_P(1)\} \xrightarrow{d} V^{-1}U$. ∎