COMPUTING THE DISTRIBUTIONS OF ECONOMIC MODELS VIA SIMULATION

JOHN STACHURSKI

ABSTRACT. This paper studies a Monte Carlo algorithm for computing distributions of state variables when the underlying model is a Markov process. It is shown that the L_1 error of the estimator always converges to zero with probability one, and often at a parametric rate. A related technique for computing stationary distributions is also investigated.

Journal of Economic Literature Classifications: C15, C22, C63

1. INTRODUCTION

Many models of economic and financial processes are both stochastic and dynamic. The system for the state variables often has a Markov structure, and when shocks are nondegenerate, or when the set of agents has positive measure, the distribution of the state is nondegenerate over some subset of $\mathbb{R}^n$. This distribution may indicate the dispersion of asset holdings, wealth, capital, wages or other such attributes across agents; or the probabilities of future outcomes for the state.

Date: December 15, 2005.

Key words and phrases. Distributions, Markov processes, simulation.

This project has benefitted from helpful conversations with Costas Azariadis, Vance Martin, Roberto Raimondo and Rabee Tourky, as well as financial support from Australian Research Council Grant DP0557625.

In recent years, computing these distributions themselves—rather than just moments and other summary statistics—has become an increasingly important computational task. In terms of computer time, perhaps the most important source of demand is simulation-based econometric and statistical techniques such as maximum likelihood, where distributions are evaluated numerically and then compiled into likelihood functions for optimization. Typical examples are Elerain, Chib and Shephard (2001) and Hurn, Lindsay and Martin (1999), who investigate simulation-based techniques for estimating the parameters of stochastic differential equations.

Another source of interest in distributions stems from the need to inspect the output of artificial economies. Distributions provide a complete description of event probabilities at a given point in time, or of cross-sectional outcomes in heterogeneous agent models. One example is the study of firm size dynamics, such as found in Rossi-Hansberg and Wright (2005). Much of that paper considers questions specific to distributions, including relating the weight in the tails of size distributions to human capital shares and other features.

Another example of the increased interest in computing distributions is the rapidly growing field of density forecasting. Many central banks now produce inflation density forecasts rather than point estimates. With these densities one can assess the implied likelihood of different inflation outcomes, or integrate loss functions. Predictive densities therefore permit more satisfactory evaluation of policy decisions than do moments alone.

In this paper we explore the so-called "look-ahead" estimator, a Monte Carlo method due to Glynn and Henderson (2001) for computing numerically the distributions of state variables from a given model. As with other simulation-based techniques, the method can be used to examine the predictive aspects of models too complex to admit analytical

solution. It can also be viewed as a complement to discretization techniques for solving nonlinear models, although the domain of application is not identical.

Our focus is on the global convergence properties of the look-ahead estimator, a proper understanding of which is essential for assessing numerical error. Arguably the most important global measure of error for this estimator is the L_1 distance between the estimator and target distribution. By applying a famous concentration of measure inequality due to McDiarmid, we are able to show that the L_1 error always converges to zero with probability one.

Second, we establish rates of convergence for expected L_1 and integral mean squared error for a large class of models. These rates are strictly faster than those obtained for nonparametric kernel density estimators when the latter are used to compute distributions of Markov models. As such, the estimator should prove extremely useful for simulated maximum likelihood and other computer intensive statistical techniques.

Several applications are used to illustrate the main theorems. These include a discretized diffusion processes studied by Elerain, Chib and Shephard (2001), a threshold autoregression, a model of commodity price dynamics under a rational expectations due to Samuelson (1971) and Deaton and Laroque (1992), and a simple (but nonstationary) version of Brock and Mirman's (1972) stochastic optimal growth model.

The paper is structured as follows. Section 2 gives an overview of the problem and construction of the look-ahead estimator, as well as a review of known properties. Section 3 formulates the general model and introduces the key assumptions. Section 4 considers probability one convergence for global error measures. Section 5 gives rates of convergence for global error measures. Section 6 provides applications. Proofs are given in Section 7.

2. Outline of the Problem

Let's take for now our primitive as a model which, after the relevant decision problems have been solved, can be expressed as

$$X_t = H_t(X_{t-1}, W_t), \quad X_0 = x_0 \text{ given }, \quad W_t \sim \varphi. \tag{1}$$

Here X_t takes values in $S \subset \mathbb{R}^k$ and W_t takes values in $Z \subset \mathbb{R}^j$, while H_t maps $S \times Z \to S$. We assume that the shocks $(W_t)_{t \geq 1}$ are independent over time and identically distributed (IID) with common distribution φ; and x_0 is a fixed point in S.[1]

Although (1) is a discrete time model, other models of interest include continuous time diffussions of the form

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t$$

with $X_0 = x_0$ given and $t \mapsto W_t$ a Weiner process. When solving for distributions of these models numerically, a standard technique is to discretize the time parameter along a suitably fine grid. The discretized model is in the form of (1) when H_t is appropriately defined.

When analytical results are unavailable, one can still explore the implications of (1) by computing distributions of the state variables $(X_t)_{t \geq 0}$. The distribution ψ_T of X_T provides a complete description of the probabilities implied by the model for time T events; or of the dispersion of features across the population in a heterogeneous agent model. If (1) is stationary and ergodic, another common exercise is computation of the stationary (invariant) distribution for the state, which we denote ψ_∞. The issues here are mathematically more subtle but conceptually very similar, and we discuss them in detail below.

[1]The IID restriction on the shocks and the fact that the state variable only enters with one lag may seem restrictive. In fact any vector-valued discrete time Markov process can be expressed in the form of (1) by suitably adjusting the definition of the state. In either case, the main theory in Section 3 considers a general discrete time Markov process.

2.1. **Marginal Distributions.** Let $T \in \mathbb{N}$ and let ψ_T denote the distribution of the S-valued random variable X_T defined inductively by (1). A common procedure for computing ψ_T is to first discretize the state space onto a grid of size n. One can then either derive for each t a Markov matrix on the grid which approximately represents the probabilistic dynamics in (1) and solve out for the implied distribution ψ_T^n by matrix multiplication; or apply quadrature-type techniques to approximate the relevent integral operators.

Discretization has both advantages and disadvantages. Discrete computations are usually fast, and at times globally convergent. On the other hand, bounds on the deviation of ψ_T^n from ψ_T are almost always difficult to obtain. In numerical analysis, quantitative error bounds which can be inferred from the model primitives are often as important as asymptotic convergence results.

An alternative approach is Monte Carlo simulation, which usually begins by drawing n independent observations $(X_T^1, \ldots, X_T^n)$ of the time T state by the following straightforward procedure:

for m in 1 to n **do**
    set $X = x_0$
    for t in 1 to T **do**
        draw $W \sim \varphi$ and set $X = H_t(X, W)$
    endfor
    set $X_T^m = X$
endfor

By definition each X_T^m is a draw from the target distribution ψ_T. With the sample, one can construct a histogram, an empirical distribution function, or a nonparametric kernel density estimate such as

$$f_T^n(y) := \frac{1}{n \cdot \delta_n} \sum_{m=1}^{n} K\left(\frac{y - X_T^m}{\delta_n}\right), \tag{2}$$

where K is a probability density, and the "bandwidth" parameter δ_n is chosen so that $n\delta_n \to 0$ as $n \to \infty$.

Regarding (2), it is well-known that—at least when ψ_T is a density—we always have $|f_T^n(y) - \psi_T(y)| \to 0$ as $n \to \infty$ with probability one for all $y \in S$. Further, probability one (almost sure) convergence to zero also holds for the L_1 error $\int |f_T^n - \psi_T|$, independent of the choice of kernel K (cf., e.g., Devroye and Lugosi, 2001).

On the other hand, the *finite* sample properties of f_T^n are not always good. For example, the error $\mathbf{E}|f_T^n(y) - \psi_T(y)|$ is known to be proportional asymptotically to $(n\delta_n)^{-1/2}$, and since $\delta_n \to 0$ with n at a rate that is sensitive to dimension of the state space S, the convergence rate is strictly slower that $O(n^{-1/2})$, and possibly much slower (Yakowitz, 1985). Slow convergence is common to many forms of Monte Carlo simulation.

Fast convergence of a proposed estimator such as f_T^n to ψ_T is particularly important in applications such as simulated maximum likelihood, where densities need to be computed for a large collection of parameters. Even when computing only a small number of distributions, however, convergence rates can be slow when the state space is high-dimensional, or when drawing variates from the state distributions is computationally expensive. In addition, low probability regions of the state space are rarely sampled, making it difficult to uncover features of the distribution on these sets via simulation.[2]

Speed of convergence is also an issue when one wishes to compute the expectation of loss (or utility) functions over the state space. For example, if ℓ is a loss function on S, then one often evaluates $\mathbf{E}\ell(X_T) := \int \ell(y)\psi_T(y)dy$ using the statistic $n^{-1}\sum_{m=1}^n \ell(X_T^m)$, where (X_T^m) is an IID sample as before. The worst case performance of this estimator is

[2]Another issue for f_T^n is that poor choice of bandwidth or kernels can have significant impact on rates of convergence and finite sample properties. Making good choices depends on sufficient knowledge of the target density ψ_T. Such knowledge is not always easy to acquire for marginal distributions of state variables when the information at hand consists only of the laws of motion given in (1).

in fact poor. If we restrict attention to loss functions bounded by some constant M, then, for all $n \in \mathbb{N}$,

$$\sup_{|\ell| \leq M} \left| \frac{1}{n} \sum_{m=1}^{n} \ell(X_T^m) - \int \ell(y) \psi_T(y) dy \right| = 2M \quad \text{with prob. 1.} \tag{3}$$

Here the supremum is over all Borel measurable $\ell \colon S \to \mathbb{R}$ with $|\ell| \leq M$. In other words, the worst-case error fails to decrease, let alone converge to zero, as $n \to \infty$.

The term $n^{-1} \sum_{m=1}^{n} \ell(X_T^m)$ in (3) corresponds to integrating the function ℓ with respect to the empirical distribution function (EDF) associated with the sample (X_T^m). This EDF is an estimate of ψ_T. With alternative estimators of ψ_T the worst-case bound (3) converges to zero relatively quickly. We now turn to such an estimator.

2.2. **The Look-Ahead Estimator.** Improved performance of distribution estimators requires additional structure. In this paper we obtain that structure by assuming that the *conditional* distribution $P(X_{t-1}, dy)$ of X_t given X_{t-1} can be represented by density $p(X_{t-1}, y)dy$. An elementary example of when this assumption holds is provided by the Solow model

$$k_t = sAk_{t-1}^{\alpha} W_t, \quad \ln W_t \sim N(0, \sigma^2), \tag{4}$$

where k is capital, and s, α and A are positive parameters. It is clear that when k_{t-1} is taken as given, $k_t | k_{t-1}$ is lognormally distributed: $\ln k_t | k_{t-1} \sim N(\ln(sA) + \alpha \ln k_{t-1}, \sigma^2)$. Thus, $P(k_{t-1}, dy) = p(k_{t-1}, y)dy$, where

$$p(k_{t-1}, y) = (2\pi\sigma^2)^{-1/2} \frac{1}{y} \exp\left\{ \frac{-(\ln y - \ln(sA) - \alpha \ln k_{t-1}}{2\sigma^2} \right\}. \tag{5}$$

Returning to the general model, fix $T \in \mathbb{N}$ and suppose that the conditional distribution of X_T given X_{T-1} can be represented by density $p_T(X_{T-1}, y)dy$. Using p_T, Glynn and Henderson (2001) proposed the following "look-ahead" estimation scheme for ψ_T. First, generate n independent draws of the state variable as above, but this time generate

draws of X_{T-1} rather than of X_T.[3] Now calculate

$$\psi_T^n(y) := \frac{1}{n} \sum_{m=1}^{n} p_T(X_{T-1}^m, y). \tag{6}$$

That ψ_T^n is a natural estimator of ψ_T follows from the well-known Markov identity

$$\mathbf{E}\, p_T(X_{T-1}, y) = \psi_T(y), \quad \forall y \in S. \tag{7}$$

A short proof of (7) is given below, but the intuition is relatively simple: If $\psi_T(y)$ is thought of as the probability of observing y at T, then this should be equal to the probability $p_T(x, y)$ of going from x at $T-1$ to y at T, summed over x and weighted by the probability that $X_{T-1} = x$; and this is precisely the left hand side of (7).

From (6) and (7) we have $\mathbf{E}\psi_T^n(y) = \frac{1}{n} n \psi_T(y) = \psi_T(y)$ at each point y, so that ψ_T^n is pointwise unbiased. Moreover, the law of large numbers implies that, with probability one,

$$\psi_T^n(y) = \frac{1}{n} \sum_{m=1}^{n} p_T(X_{T-1}^m, y) \to \mathbf{E} p_T(X_{T-1}, y) = \psi_T(y) \tag{8}$$

as $n \to \infty$. In other words, $\psi_T^n(y)$ is a consistent estimator of $\psi_T(y)$ at each point $y \in S$.

Notice that ψ_T^n makes use of the structure of the model as embodied in p_T—a key aspect of efficient computation. In contrast to f_T^n there is no bandwidth parameter, nor any need to choose a kernel K. These two features suggest that ψ_T^n will have good finite sample properties to match the asymptotic result (8). Indeed, the Central Limit Theorem implies that when suitable second moment restrictions are satisfied, the error $\mathbf{E}|\psi_T^n(y) - \psi_T(y)|$ is asymptotically $O(n^{-1/2})$, independent of the dimension of the state space S.

[3]Hence the name "look-ahead."

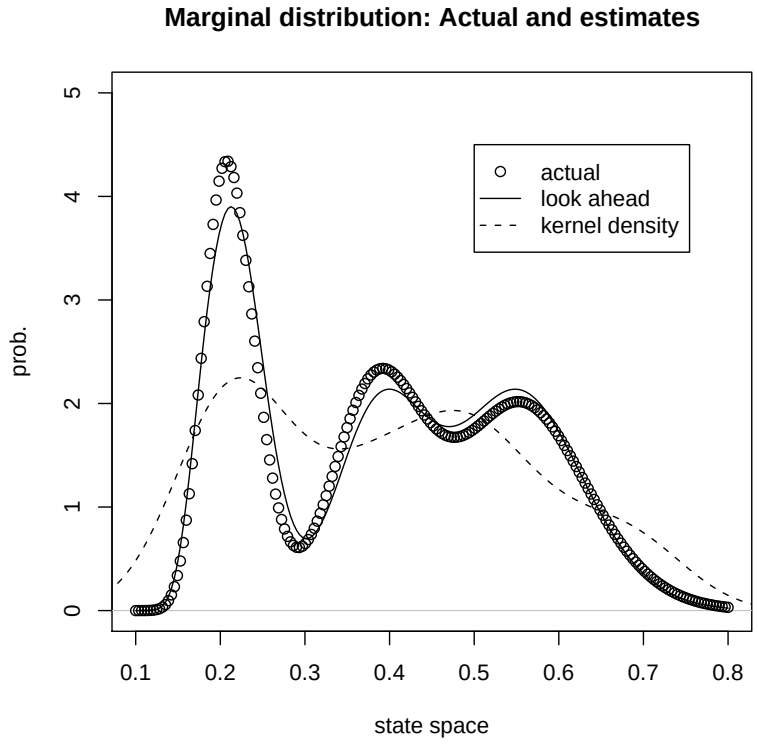

FIGURE 1. The Look-Ahead Estimator.

Figure 1 compares realizations of ψ_T^n and f_T^n with the actual time T density ψ_T for the Solow model (4).[4] We argued that the distribution for the current state k_t given k_{t-1} is the lognormal density $p(k_{t-1}, y)dy$ in (5). Given this function p, and using `samples` as the vector which contains the draws of the time $T-1$ state, the look-ahead estimate $\psi_T^n(y)$ is evaluated for Figure 1 (using the language R) by

```
look_ahead = function(y) {
  q = numeric(n)  # vector of length n
  for (i in 1:n) q[i] = p(samples[i],y)
  return( mean(q) )
}
```

[4]In the figure, the parameters are $\alpha = 0.3$, $A = 2$, $\sigma = 0.11$, and $s = 0.2$.

In Figure 1 the estimates of ψ_T are for $T = 2$. The initial condition ψ_0 has been deliberately chosen as multi-modal, making ψ_2 multi-modal and increasing the complexity of the approximation problem.[5] Despite this complexity, the combination of log-linearity and log-normality means that an analytical solution for ψ_T is also available for comparison, and this is plotted using the $\circ$ symbol. The look-ahead estimate ψ_T^n is the unbroken line. Although the sample size is tiny by Monte Carlo standards ($n = 100$), the estimator closely follows the actual density.

The broken line in Figure 1 is a kernel density estimate f_T^n of the form given in (2). In this case we are using the default algorithm in R.[6] The kernel density estimate uses the same draw of shocks as the look-ahead estimate, and the same sample size ($n = 100$). At least for this default algorithm, convergence is much slower.[7]

This paper analyzes extensively the convergence properties of the look-ahead estimator. We concentrate on global error, that is, on convergence of the *function* ψ_T^n to ψ_T. Of primary interest is the L_1 error, which is given by

$$\|\psi_T^n - \psi_T\| := \int_S |\psi_T^n(y) - \psi_T(y)| dy.$$

In contrast to the integral mean squared error, this measure is always well-defined. Further, Scheffés identity provides a natural quantitative interpretation. That is, $\|\psi_T^n - \psi_T\| = 2 \times \sup_B |\int_B \psi_T^n - \int_B \psi_T|$, where

[5]We are using $\psi_0 = (1/3)(f_1 + f_2 + f_3)$, where f_i is lognormal with parameters μ_i and σ_i; $\mu_1 = -4$, $\sigma_1 = 1$, $\mu_2 = 3$, $\sigma_2 = 1$, $\mu_3 = 7$, $\sigma_3 = 0.5$.

[6]The kernel K is Gaussian, and the bandwidth is selected according to the rule-of-thumb $\delta_n = 1.06 \min(\hat{\sigma}_n, \hat{R}_n/1.34) n^{-1/5}$, where $\hat{\sigma}_n$ is the sample standard deviation, and $\hat{R}_n$ is the inter-quartile range.

[7]Of course the nonparametric kernel estimator is far more general, and, moreover, careful choice of bandwidth and kernel will lead to faster convergence. The point is that when the look-ahead estimator is applicable, it *automatically* incorporates model structure, while for the kernel estimator including enough structure to obtain similar rates of convergence is in general a nontrivial exercise.

the supremum is over all Borel subsets of the state space S. It follows that if $\|\psi_T^n - \psi_T\| \leq \varepsilon$, then for any event B of interest the deviation in the probability assigned to B by the approximate density ψ_T^n from that assigned by the true density ψ_T is less than $\varepsilon/2$.

We prove for the first time that ψ_T^n always converges to ψ_T in L_1 with probability one as $n \to \infty$. The proof is based on McDairmid's famous concentration of measure inequality. In addition, we provide rates of convergence for global error measures. We prove that for a wide class of models the expected L_1 error (respectively, the integral mean square error) is $O(n^{-1/2})$ (respectively, $O(n^{-1})$). For some common models we provide upper bounds on the L_1 and integral mean square error in terms of the functions H_t and the distribution φ of the shock in the benchmark (1).

2.3. **Computation of Stationary Distributions.** In some cases the model is stationary over time ($H_t = H$ for all t) and ergodic, in the sense that the distribution ψ_t of X_t converges to some limiting distribution ψ_∞ (usually called the *stationary* or *invariant* distribution) independent of initial conditions. For such models the stationary distribution has the interpretation of long-run stochastic equilibrium, and hence is of central interest to researchers.

As Glynn and Henderson (2001) point out, the look-ahead estimator can often be applied. Precisely, let $p(X_{t-1}, y)dy$ again be the conditional density of X_t given X_{t-1} as implied by $X_t = H(X_{t-1}, W_t)$, and let $(X_1, \ldots, X_n)$ be a series drawn recursively from $X_t = H(X_{t-1}, W_t)$. They propose the estimator

$$\psi_\infty^n(y) := \frac{1}{n} \sum_{t=1}^{n} p(X_t, y). \tag{9}$$

Notice that we are now summing over time, rather than across independent samples of the state at a fixed point in time.

The intuition for ψ_∞^n is as follows. As discussed above, a stationary density for the model $X_t = H(X_{t-1}, W_t)$ is defined as a density ψ_∞ satisfying

$$\int p(x,y)\psi_\infty(x)dx = \psi_\infty(y), \quad \forall\, y \in S. \tag{10}$$

When a stationary density exists, and moreover, $\psi_t \to \psi_\infty$ in L_1 as $t \to \infty$, we also have the correlated law of large numbers result

$$\frac{1}{n}\sum_{t=1}^{n} w(X_t) \to \int w(x)\psi_\infty(x)dx \ \text{ as } n \to \infty, \tag{11}$$

where w is any measurable function with $\int w(x)\psi_\infty(x)dx$ finite, and convergence is with probability one. As a result,

$$\psi_\infty^n(y) = \frac{1}{n}\sum_{t=1}^{n} p(X_t, y) \to \int p(x,y)\psi_\infty(x)dx = \psi_\infty(y) \tag{12}$$

with probability one as $n \to \infty$. Thus the look-ahead estimator ψ_∞^n is again seen to be a very natural estimator, and Glynn and Henderson establish strong finite sample and asymptotic properties under reasonable assumptions. We extend their analysis by establishing almost sure L_1 convergence to the true density under weaker conditions than previous results.

3. The General Model

The state space is any separable and completely metrizable topological space S. Let $\mathscr{B}$ denote the Borel sets of S, and let $(S, \mathscr{B})$ be endowed with a σ-finite measure μ. Typically S is a Borel subset of $\mathbb{R}^k$, in which case μ will always be the Lebesgue measure. To emphasize this, when integrating over S with respect to μ, we write dx for $\mu(dx)$, dy for $\mu(dy)$, etc.; and $\int$ in place of $\int_S$.

As usual, $L_1(S, \mathscr{B}, \mu)$ is the set of real, $\mathscr{B}$-measurable functions $f\colon S \to \mathbb{R}$ such that f is μ-integrable. The set of *densities* on S is the set of $\psi \in L_1(S, \mathscr{B}, \mu)$ with $\psi \geq 0$ and $\int \psi d\mu = 1$. In all of what follows, $\|\cdot\|$ is the standard L_1 norm, so that $\|f\| = \int |f| d\mu$.

A *distribution* on S is a probability measure on $(S, \mathscr{B})$. A *stochastic kernel* is a family of distributions $P(x, dy)$ on S, $\forall x \in S$, with the property that $x \mapsto P(x, B)$ is Borel measurable for each $B \in \mathscr{B}$. The standard interpretation is that $P(x, dy)$ is the probability distribution of tomorrow's state given that the current state is x. For example, in the case of (1) we have

$$P_t(x, B) = \varphi\{z \in Z : H_t(x, z) \in B\}. \tag{13}$$

Although (1) is the basic model we envisage in applications, for the sake of generality we take as our formal primitive a discrete time Markov chain $(X_t)_{t \geq 0}$ on S defined by initial condition $x_0 \in S$ and stochastic kernels $(P_t)_{t \geq 1}$. That is,

$$X_0 = x_0 \text{ and then, recursively, } X_t \sim P_t(X_{t-1}, dy). \tag{14}$$

When the sequence (P_t) is defined by (13), the stochastic process (X_t) generated by (1) and the sequence defined in (14) coincide.

A more precise formulation of (14) is as follows. Given initial condition x_0 and sequence of kernels $(P_t)_{t \geq 1}$, there exists a probability space $(\Omega, \mathscr{F}, \mathbf{P})$ and a sequence of S-valued random variables $(X_t)_{t \geq 0}$ on $(\Omega, \mathscr{F}, \mathbf{P})$ with the property that $X_0 = x_0$ and

$$\mathbf{P}\{X_t \in B \,|\, \mathscr{F}_{t-1}\} = \mathbf{P}\{X_t \in B \,|\, X_{t-1}\} = P_t(X_{t-1}, B), \tag{15}$$

for all $t \geq 1$ and all $B \in \mathscr{B}$. Here $(\mathscr{F}_t)_{t \geq 0}$ is the natural filtration, so that $\mathscr{F}_t := \sigma(X_0, \ldots, X_t)$. Every discrete time Markov chain can be represented in this way, and we refer the reader to texts such as Durrett (1996, Chapter 5) for further background.

A *density kernel* p on S is a measurable map $p\colon S \times S \to [0, \infty)$ such that $p(x, y)dy$ is a density on S for every $x \in S$. We now state our main assumption, which is viewed as holding throughout the rest of the paper without need for citation.

Assumption 3.1. For each stochastic kernel P_t in $(P_t)_{t\geq 1}$, there exists a density kernel p_t which represents it. Precisely,

$$P_t(x, B) = \int_B p_t(x, y)dy, \quad \forall B \in \mathscr{B}, \ \forall\, x \in S. \tag{16}$$

At this point we can verify Equation (7). To do so, take expectations of both sides of (15) to get $\mathbf{P}\{X_t \in B\} = \mathbf{E}P_t(X_{t-1}, B)$. From this expression, (16) and Fubini's Theorem we have

$$\mathbf{P}\{X_t \in B\} = \int_B \mathbf{E}\, p_t(X_{t-1}, y)dy, \quad \forall B \in \mathscr{B}. \tag{17}$$

From (17) it is clear that the distribution of X_t is represented by density $\psi_t(y)dy := \mathbf{E}\, p_t(X_{t-1}, y)dy$, for every $t \geq 1$.

Following (6), the *T-step look-ahead* (TSLA) estimator is the random density function ψ_T^n defined by $\psi_T^n(y) := \frac{1}{n}\sum_{m=1}^n p_T(X_{T-1}^m, y)$, where $X_{T-1}^1, \ldots, X_{T-1}^n$ are IID draws from ψ_{T-1}. If $p_t = p$ for all t, then the stationary distribution look-ahead (SDLA) estimator is the random density function ψ_∞^n defined by $\psi_\infty^n(y) := \frac{1}{n}\sum_{t=1}^n p(X_t, y)$, where now we are now summing over a time series draw, rather than across independent samples of the state at a fixed point in time.[8]

4. Almost Sure Global Convergence

As discussed above, the L_1 error $\|\psi_T^n - \psi_T\|$ is arguably the most important measure of error for density estimators. Glynn and Henderson (2001) establish that the L_1 error of the TSLA ψ_T^n always converges to zero in probability and in expectation. They also prove the stronger notion of almost sure convergence when p_T is uniformly continuous and bounded on $S \times S$. In fact almost sure L_1 convergence always holds:

Theorem 4.1. *The TSLA ψ_T^n converges in L_1 to ψ_T with probability one as $n \to \infty$.*

[8]In other words, $X_1, \ldots, X_n$ obeys (14).

Now consider almost sure L_1 convergence for the look-ahead estimator of the stationary distribution. We require some minimal conditions on the Markov chain to ensure that its time series satisfy the strong law of large numbers. To state them, extra definitions are necessary.[9] Suppose for now that $p_t = p$ for all t. A density ψ_∞ is called *stationary* for p if (10) holds; that is, if $\int p(x, y)\psi_\infty(x)dx = \psi_\infty(y)$ holds for all $y \in S$. Let $(X_t)_{t\geq 0}$ be the Markov chain generated by p and initial condition $X_0 = x_0 \in S$. For this chain define

$$L(x_0, A) := \mathbf{P} \cup_{t\geq 1} \{X_t \in A\}.$$

The chain is called irreducible if there exists a nontrivial measure λ on $(S, \mathscr{B})$ such that $L(x_0, A) > 0$ for all $x_0 \in S$ and all $A \in \mathscr{B}$ with $\lambda(A) > 0$; and Harris recurrent if $L(x_0, A) = 1$ for all $x_0 \in A$ whenever $A \in \mathscr{B}$ and $\lambda(A) > 0$. A Harris recurrent chain with a stationary distribution is called *positive Harris*. (For Harris chains the stationary distribution is necessarily unique.)

Assumption 4.1. The model is time homogeneous: $p_t = p$ for all t. The Markov chain $(X_t)_{t\geq 0}$ generated by p is positive Harris.

This positive Harris assumption is sufficient to obtain a law of large numbers result for the series $(X_t)_{t\geq 0}$: By Meyn and Tweedie (1993, Theorem 17.1.7), if $(X_t)_{t\geq 0}$ is positive Harris with stationary distribution ψ_∞, then for every function $w\colon S \to \mathbb{R}$ with $\int |w|d\psi_\infty < \infty$ we have $\frac{1}{n}\sum_{t=1}^n w(X_t) \to \int w(x)\psi_\infty(x)dx$ almost surely as $n \to \infty$. (In fact the converse is true, in the sense that when a stationary distribution exists and the law of large numbers holds for all such w then $(X_t)_{t\geq 0}$ is positive Harris. In this sense the positive Harris assumption is minimal for our purposes.)

For positive Harris chains, Glynn and Henderson (2001) proved almost sure L_1 convergence of the SDLA ψ_∞^n to ψ_∞ when p is uniformly continuous and bounded on $S \times S$. Here we show that the same result holds

[9]See Meyn and Tweedie (1993) for further details.

under the following condition, which is weaker than uniform continuity and independent of boundedness.

Assumption 4.2. Let d metrize S. The kernel p is continuous in y uniformly in x. Precisely, for all $\varepsilon > 0$ and all $y \in S$, there is a $\delta > 0$ such that $d(y', y) < \delta$ implies $\sup_{x \in S} |p(x, y) - p(x, y')| < \varepsilon$.

Theorem 4.2. *If Assumptions 4.1 and 4.2 hold, then the SDLA ψ_∞^n converges in L_1 to ψ_∞ with probability one.*

5. Rates of Convergence

Asymptotic convergence results are reassuring, but without bounds on the rate of convergence they provide no guidance on finite sample properties, or when algorithms should be terminated. In this section we examine rates of convergence, and bounds on global error measures such as expected L_1 error or integral mean squared error.

Consider first the expected L_1 error for the TSLA ψ_T^n. In macroeconomics it is common to deal with continuous models on compact state spaces.[10] Our first result shows that for these and some related models, the expected L_1 error is $O(n^{-1/2})$.

Theorem 5.1. *If p_T is bounded by K on $S \times S$, then*

$$\mathbf{E}\|\psi_T^n - \psi_T\| \leq \sqrt{\frac{1}{n}} K \mu(S).$$

Clearly this bound is only useful when $\mu(S) < \infty$. To deal with more general state spaces, we require that the shock is additive with exponentially decreasing tails. In addition, a mild restriction is placed on the growth rate of the law of motion:

[10] See, for example, Brock and Mirman (1972), or Stokey, Lucas and Prescott (1989, Chapter 13).

Assumption 5.1. Let $S = Z = \mathbb{R}^k$, and let $X_t = g_t(X_{t-1}) + W_t$, where W_t is distributed according to some density φ on $\mathbb{R}^k$, the map $g_t \colon \mathbb{R}^k \to \mathbb{R}^k$ is measurable for all t, and, for some norm $\|\cdot\|$ on $\mathbb{R}^k$,

(i) $\exists \alpha, L > 0$ s.t. $\|g_t(x)\| \leq \alpha\|x\| + L$ for all $t \in \mathbb{N}$, all $x \in \mathbb{R}^k$; and
(ii) $\exists K, \varrho > 0$ s.t. $\varphi(z) \leq K \exp(-\varrho\|z\|^2)$ for all $z \in \mathbb{R}^k$.

Theorem 5.2. *Let (X_t) be the sequence in Assumption 5.1, where X_0 is a constant $x_0 \in S$, let ψ_T be the density of X_T, and let ψ_T^n be the TSLA of ψ_T. If Assumption 5.1 holds, then $\mathbf{E}\|\psi_T^n - \psi_T\| = O(n^{-1/2})$.*

Another common measure of global error is the integral mean square error, defined as

$$\operatorname{IMSE}(\psi_t^n) := \mathbf{E} \int [\, \psi_t^n(y) - \psi_t(y) \,]^2 \, dy, \quad t \in \mathbb{N} \cup \{\infty\}.$$

We give a condition for the integral mean square error of the TSLA to be $O(n^{-1})$. This result cannot hold in complete generality, because the IMSE is not always defined for target densities with heavy tails. We therefore impose a restriction on the tails of the family of distributions $p_t(x, y)dy$. Note that the rate $O(n^{-1})$ compares well with the optimal rate $O(n^{-4/5})$ for nonparametric kernel density estimators when the target density is twice differentiable and satisfies some tail restrictions.[11]

Theorem 5.3. *Let $(p_t)_{t \geq 1}$ be given and let $T \in \mathbb{N}$ be fixed. If ψ_T^n is the TSLA of ψ_T, then* $\operatorname{IMSE}(\psi_T^n) = O(n^{-1})$ *whenever $\int p_T(x, y)^2 \, dy$ is bounded above independent of $x \in S$. In particular,*

$$\operatorname{IMSE}(\psi_T^n) \leq \frac{1}{n} \cdot \sup_{x \in S} \int p_T(x, y)^2 \, dy. \tag{18}$$

Notice that the rate does not depend on the dimension of S, although the dimension may influence the size of the constants in the order term. We give some applications of this result in Section 6.

[11] See, for example, van der Vaart (1998, Chapter 24).

All of the preceding results pertain to the TSLA ψ_T^n. Our final result of this section shows that for the SDLA ψ_∞^n the expected L_1 error $\mathbf{E}\|\psi_\infty^n - \psi_\infty\|$ is also $O(n^{-1/2})$, at least when we restrict attention to uniformly ergodic Markov chains on finite measure spaces.

Definition 5.1. Let Assumption 4.1 hold, so that $p_t = p$ for all t. The Markov chain (X_t) generated by p and x_0 is called *uniformly ergodic* if p has a unique stationary distribution ψ_∞, and, moreover, there exist positive constants R and α, both independent of x_0, such that $\alpha < 1$ and, for all t,

$$\|\psi_t - \psi_\infty\| \leq R\alpha^t. \tag{19}$$

We note that uniform ergodicity is equivalent to aperiodicity combined with Doeblin's condition (Meyn and Tweedie, 1996, Theorem 16.0.2). A number of other useful sufficient conditions are also available, and the reader is referred to Meyn and Tweedie (1996, Chapter 16).

Theorem 5.4. *Let Assumption 4.1 hold, so that $p_t = p$ for all t, and let the Markov chain (X_t) generated by p and x_0 be uniformly ergodic. If p is bounded by K on $S \times S$, then the SDLA ψ_∞^n satisfies*

$$\mathbf{E}\|\psi_\infty^n - \psi_\infty\| \leq \sqrt{\frac{4R}{n(1-\alpha)}} K\mu(S),$$

where R and α are as in (19).

Again, this bound is only useful when $\mu(S) < \infty$. The property of $O(n^{-1/2})$ convergence for the SDLA in more general situations is left to future research.

6. Examples and Applications

6.1. **Existence of Density Kernels.** Consider Assumption 3.1, which requires that each transition probability $P_t(x, dy)$ has a density representation $p_t(x, y)dy$. When does this condition hold for the basic model

(1)? In other words, when is $P_t(x, dy)$ absolutely continuous with respect to Lebesgue measure μ for given t and x? Since

$$P_t(x, B) = \varphi\{z \in Z : H_t(x, z) \in B\} = \varphi(H_t^{-1}(x, B)), \tag{20}$$

where $H_t^{-1}(x, B) \subset Z$ is the preimage of B under $H_t(x, \cdot)$, what we require is that this inverse map pulls Lebesgue null sets back into φ null sets. If φ is itself a density, then it is sufficient that the inverse map pulls Lebesgue null sets back into Lebesgue null sets, a property known as nonsingularity.

Rather than focusing on nonsingularity, we develop a sufficient condition that holds in many applications, and has the advantage of providing an explicit representation for $p_t(x, y)dy$. To start, note that $p_t(x, y)dy$ must represent the distribution of the random variable $Y := H_t(x, W)$ when $W \in Z$ is drawn according to φ. For all $y \in S$ where there is no $z \in Z$ with $H_t(x, z) = y$ we should have $p_t(x, y) = 0$. The remainder of S we denote S_x, and on this set we construct $p_t(x, y)$ by a change of variable argument. The details are in the following lemma:

Lemma 6.1. *For the model* (1)*, let Z and S be open subsets of $\mathbb{R}^k$, and let φ be a density on Z. Let $S_x := H(x, Z)$, the range of $z \mapsto H(x, z)$, and let $z \mapsto H_t(x, z)$ be one-to-one for each $x \in S_x$. Define $G_x \colon S_x \to Z$ to be the inverse mapping of this function. If G_x is a C^1 function for each $x \in S$, then Assumption 3.1 holds.*[12] *Moreover, if J_x denotes the Jacobian of G_x, then*

$$p_t(x, y) = \begin{cases} \varphi[G_x(y)] \cdot |\det J_x(y)| & \text{if } y \in S_x \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

This is an elementary change of variable result, and the proof is omitted. The following corollary helps to illustrate application of the lemma.

[12]A function f from one open subset of Euclidean space to another is called C^1 if it is continuously differentiable everywhere on its domain.

Corollary 6.1. *Assume that $Z = S = \mathbb{R}^k$, that φ is a density on Z, and that*

$$X_t = H_t(X_{t-1}, W_t) = g_t(X_{t-1}) + \Sigma_t(X_{t-1})\, W_t, \tag{22}$$

where $g_t \colon S \to S$ is any Borel measurable function, and $\Sigma_t(x)$ is an invertible $n \times n$ matrix for all t and all $x \in S$. In this case,

$$p_t(x, y) = \varphi\{\Sigma_t(x)^{-1}[y - g_t(x)]\} \cdot |\det \Sigma_t(x)^{-1}| \tag{23}$$

holds everywhere on $S \times S$.

Example 6.1. Elerain, Chib and Shephard (2001) study the continuous time diffusion processes

$$dY_t = a(t, Y_t)dt + b(t, Y_t)dW_t,$$

where Y_t is $\mathbb{R}^k$-valued, $t \mapsto W_t$ is a standard Weiner process, and b is everywhere strictly positive definite. To estimate parameters they apply the Euler–Maruyama discretization, obtaining

$$Y_t = Y_{t-1} + a(t-1, Y_{t-1}) + b(t-1, Y_{t-1})W_t,$$

where W_t is standard normal. Corollary 6.1 clearly applies, and

$$p_t(x, y) = \varphi\{b(t-1, x)^{-1}[y - x - a(t-1, x)]\} \cdot |\det b(t-1, x)^{-1}|,$$

where φ is the standard normal density.

Example 6.2. Let $Z = S = \mathbb{R}$, and consider the elementary smooth transition threshold autoregression (STAR) model

$$X_t = (\beta_0 + \beta_1 X_{t-1})(1 - G(X_{t-1})) + (\beta_0' + \beta_1' X_{t-1})G(X_{t-1}) + \sigma W_t, \tag{24}$$

where $(W_t)_{t \geq 1}$ is IID according to density φ on S, $\sigma > 0$, and $G \colon S \to [0, 1]$ is a smooth transition function, such as the logistic function, satisfying $G' > 0$, $\lim_{x \to -\infty} G(x) = 0$ and $\lim_{x \to \infty} G(x) = 1$. Evidently the conditions of Corollary 6.1 are satisfied, and from (23) we get

$$p_t(x, y) = p(x, y) = \varphi\left\{\frac{y - g(x)}{\sigma}\right\}\frac{1}{\sigma}, \tag{25}$$

where $g(x) := (\beta_0 + \beta_1 x)(1 - G(x)) + (\beta_0' + \beta_1' x)G(x)$.

Example 6.3. Next, consider the following model of a commodity market due to Samuelson (1971) and Deaton and Laroque (1992). Total supply of the commodity at time t is denoted X_t. There are two types of consumers. The first buy for consumption, and their demand is D_t. The second are speculators, who buy inventory I_t. After allowing for depreciation δ, the speculators sell their remaining stock $(1-\delta)I_t$ in the following period. The sum of this and the harvest W_{t+1} give total supply next period:

$$X_{t+1} = (1-\delta)I_t + W_{t+1} \tag{26}$$

The harvest is assumed IID with density φ on $Z := (0, \infty)$. Demand by consumers is a function $D(P)$ of the price, which in turn is solved as a rational expectations pricing functional P over the state space $S := (0, \infty)$ via abitrarge conditions. Thus, $P(X_t)$ is the price that prevails at time t, and demand by consumers is $D_t = D(P(X_t))$. Combining this with the market equilibrium condition $X_t = D_t + I_t$ and (26) we get

$$X_{t+1} = (1-\delta)[X_t - D(P(X_t))] + W_{t+1}. \tag{27}$$

From Lemma 6.1 it follows immediately that the corresponding density kernel p exists, and is given by

$$p(x, y) = \varphi\{y - (1-\delta)[x - D(P(x))]\}$$

whenever $y - (1-\delta)[x - D(P(x))] \geq 0$ and zero otherwise.

Example 6.4. Consider the optimal growth model of Brock and Mirman (1972). At t a representative household observes k_t and divides it between consumption c_t and investment x_t. Productivity A_{t+1} is then observed, and production takes place, yielding output $A_{t+1}f(x_t)$ at the start of $t+1$. Here $A_t := (1+\gamma)^t W_t$, where γ is the rate of productivity growth, and $(W_t)_{t\geq 1}$ are IID on $Z := (0, \infty)$ with density φ.

Let Π be the set of all Borel measurable $h\colon [0,\infty) \to [0,\infty)$ satisfying $0 \leq h(k) \leq k$. These are the *feasible policies*, and each defines a process

$$k_t = A_t f(h(k_{t-1})) + (1-\delta)k_{t-1}, \tag{28}$$

where $\delta \in (0,1]$ is the depreciation rate. The agent has period utility u and discount factor β; and chooses h to solve

$$\max_{h \in \Pi} \mathbf{E} \left\{ \sum_{t=0}^{\infty} \beta^t u(c_t^h) \right\}, \tag{29}$$

where $c_t^h := k_t - h(k_t)$. Let u be bounded for simplicity.[13] Let u and f both be nonnegative, differentiable, strictly increasing, with u strictly concave, $\lim_{c \to 0} u'(c) = \infty$ and $f(0) = 0$. In this case it is known that a solution h to (29) exists. Under standard conditions we also have $0 < h(k) < k$ for every $k \in S$.[14] Suppose this is the case.

Consider the optimal dynamics for k on $S := (0,\infty)$, which are given by the random sequence (28) under the optimal policy h. Since $h(k) > 0$ for all $k \in S$ and $f' > 0$ we have $f(h(k)) > 0$ for all $k \in S$. Using this fact one can verify the conditions of Lemma 6.1, and (21) gives us

$$p_t(x,y) = \varphi \left\{ \frac{y - (1-\delta)x}{(1+\gamma)^t f(h(x))} \right\} \frac{1}{(1+\gamma)^t f(h(x))} \tag{30}$$

when $y > (1-\delta)x$ and zero otherwise.

6.2. **Stationary Distributions.** Next we illustrate Assumption 4.1, which imposes Harris recurrence. In doing so, let us note that by Meyn and Tweedie (1993), Theorems 6.0.1(iii), 9.0.2 and 12.1.2(ii), if S is a subset of $\mathbb{R}^k$ which contains an open set, if p is Feller and

[13]This is assumed here only for simplicity. As is well-known, many specific models with unbounded utility can also be treated by dynamic programming on the basis of assumptions constraining maximal growth rates under the stochastic production function relative to the precise utility specification.

[14]For example, this is true when f is concave. Even when concavity fails, reasonable sufficient conditions exist. See, for example, Nishimura and Stachurski (2005).

irreducible with respect to the restriction of Lebesgue measure to S, and if the Markov chain $(X_t)_{t\geq 0}$ generated by p is tight for all the initial conditions $X_0 \equiv x_0 \in S$, then p is positive Harris.[15]

Example 6.5. Returning to the STAR model of Example 6.2, it is easy to show that if φ is standard normal, for example, then p defined in (25) is irreducible with respect to Lebesgue measure on $\mathbb{R}$. Since G is assumed continuous, p is also Feller. We now verify tightness under the hypotheses $\alpha := \max\{|\beta_1|, |\beta_1'|\} < 1$ and $\mathbf{E}|W_t| < \infty$.

Simple algebra shows that there is a finite constant L such that

$$|g(x)| \leq \alpha|x| + L, \quad \forall x \in S. \tag{31}$$

$$\therefore \quad \mathbf{E}_{t-1}|X_t| = \mathbf{E}_{t-1}|g(X_{t-1}) + \sigma W_t| \leq \alpha|X_{t-1}| + L + \sigma \int |z|\varphi(dz).$$

$$\therefore \quad \mathbf{E}|X_t| \leq \alpha \mathbf{E}|X_{t-1}| + L', \quad L' := L + \sigma \int |z|\varphi(dz).$$

Iterating this inequality backwards in time to $t = 0$ we get

$$\mathbf{E}|X_t| \leq \alpha^t |x_0| + \frac{L'}{1-\alpha}.$$

$$\therefore \quad \sup_{t\geq 0} \mathbf{E}|X_t| \leq |x_0| + \frac{L'}{1-\alpha}.$$

Chebychev's inequality now gives

$$\mathbf{P}\{|X_t| \geq n\} \leq n^{-1} \left(|x_0| + \frac{L'}{1-\alpha} \right), \quad \forall n \in \mathbb{N}.$$

Evidently (X_t) is tight, and the STAR model is positive Harris.

Example 6.6. Consider again the stochastic growth model in Example 6.4. Let $\gamma = 0$, so that $p_t = p$ is stationary. It has been shown (Nishimura and Stachurski, 2005) that this model is positive Harris whenever the usual Inada conditions hold and both $\int x\varphi(dx)$ and $\int x^{-1}\varphi(dx)$ are finite.

[15]Recall that a collection of random variables $(X_t)_{t\geq 0}$ taking values in S is called *tight* whenever, for each $\varepsilon > 0$, there is a compact subset K of S such that $\sup_{t\geq 0} \mathbf{P}\{X_t \notin K\} \leq \varepsilon$. Recall also that p is called (weak) Feller if $x \mapsto \int h(y)p(x,y)dy$ is continuous and bounded on S whenever h is.

Now let's turn to Assumption 4.2. A special but important case is where S is an open subset of $\mathbb{R}$. For this case it is easy to see that Assumption 4.2 is satisfied whenever $p_t(x, y)$ is differentiable in y for each $(x, y) \in S \times S$, and

$$\forall y \in S, \ \exists K_y \in \mathbb{R} \text{ s.t. } \left| \frac{\partial p_t(x, y)}{\partial y} \right| \leq K_y, \ \forall x \in S. \tag{32}$$

Example 6.7. Consider the stochastic growth model of Example 6.4. Let $\ln W_t \sim N(0, 1)$, and, for simplicity, let $\delta = 1$. Notice that p_t is neither bounded nor uniformly continuous on $S \times S = (0, \infty) \times (0, \infty)$.[16] However, Assumption 4.2 holds, as can easily be verified via (32). In fact, the representation (30) and some simple calculus shows that

$$\left| \frac{\partial p_t(x, y)}{\partial y} \right| \leq K_y := \frac{1}{\sqrt{2\pi} y^2}, \quad \forall x \in S.$$

Example 6.8. In the nonlinear autoregression (24), it is clear from (25) that Assumption 4.2 holds whenever φ is differentiable on $\mathbb{R}$ and φ' is bounded.

Next we illustrate Assumption 5.1.

Example 6.9. In the STAR model $X_t = g(X_{t-1}) + W_t$, where

$$g(x) := (\beta_0 + \beta_1 x)(1 - G(x)) + (\beta_0' + \beta_1' x) G(x), \quad W_t \sim N(0, \sigma^2),$$

Assumption 5.1 is satisfied with $\alpha = \max\{|\beta_1|, |\beta_1'|\}$, $L = \max\{|\beta_0|, |\beta_0'|\}$, $K = (2\pi\sigma^2)^{-1/2}$ and $\varrho = (2\sigma^2)^{-1}$.

6.3. **Quantitative Bounds.** Finally, an application of Theorem 5.3 is given.

Proposition 6.1. *Consider the model* (22)*, where* $\Sigma_t(x)$ *is positive definite for all* t *and* x*. Let* $(p_t)_{t \geq 1}$ *be the corresponding density kernels, defined by* (23)*. Let* φ_t^x *be the density of the random term* $\Sigma_t(x) W_t$*. Let* $T \in \mathbb{N}$ *be fixed, and let* ψ_T^n *be the TSLA of* ψ_T*. If there exist constants*

[16] In fact, p_t may not be continuous when f is non-concave.

$K \geq 0$ and $\varrho > 0$ such that φ_t^x satisfies $\varphi_t^x(z) \leq K \exp(-\varrho \|z\|)$, for all $x \in S$, $t \in \mathbb{N}$ and $z \in Z$, then

$$\text{IMSE}\,(\psi_T^n) \leq \frac{K^2}{n} \frac{2\pi^{k/2}}{\Gamma(k/2)(2\varrho)^k} (k-1)! \tag{33}$$

If φ_t^x satisfies $\varphi_t^x(z) \leq K \exp(-\varrho \|z\|^2)$, for all $x \in S$, $t \in \mathbb{N}$ and $z \in Z$, then

$$\text{IMSE}\,(\psi_T^n) \leq \frac{K^2}{n} \left(\frac{\pi}{2\varrho} \right)^{k/2}. \tag{34}$$

The conditions in the proposition are just small tail assumptions for the distribution φ of W_t. The will be satisfied if, for example, $\Sigma_t(x)$ is a constant and φ is Gaussian.

7. Proofs

Proof of Theorem 4.1. The following proof draws on ideas in Devroye and Lugosi (2001, § 9.4) concerning concentration of measure inequalities. A discussion of McDairmid's inequality can be found there.

For the proof, fix $n \in \mathbb{N}$, and let S^n be the n-fold cartesian product of S with itself, a typical element of which is $x = (x_1, \ldots, x_n)$. Let $X_{T-1}^1, \ldots, X_{T-1}^n$ be IID draws from ψ_T. By McDairmid's inequality, if g is a measurable function from S^n to $\mathbb{R}$ such that

$$\sup |g(x) - g(x')| \leq c,$$

where the supremum is over all pairs x, x' in S^n which differ on at most one coordinate, then

$$\mathbf{P}\{|g(X_{T-1}) - \mathbf{E}g(X_{T-1})| \geq \varepsilon\} \leq 2 \exp\left(\frac{-2\varepsilon^2}{nc^2} \right),$$

where $g(X_{T-1}) := g(X_{T-1}^1, \ldots, X_{T-1}^n)$. Setting

$$g(x) = g(x_1, \ldots, x_n) = \int \left| \frac{1}{n} \sum_{m=1}^{n} p_T(x_m, y) - \psi_T(y) \right| dy$$

gives $g(X_{T-1}) = \|\psi_T^n - \psi_T\|$. Pick any $x, x' \in S^n$ such that x and x' differ only at the k-th coordinate. In this case $|g(x) - g(x')|$ is given by the expression

$$\left| \int \left| \frac{1}{n} \sum_{m=1}^{n} p_T(x_m, y) - \psi_T(y) \right| dy - \int \left| \frac{1}{n} \sum_{m=1}^{n} p_T(x'_m, y) - \psi_T(y) \right| dy \right|,$$

which is bounded above by

$$\int \left| \frac{1}{n} \sum_{m=1}^{n} p_T(x_m, y) - \frac{1}{n} \sum_{m=1}^{n} p_T(x'_m, y) \right| dy$$

$$= \frac{1}{n} \int |p_T(x_k, y) - p_T(x'_k, y)| dy.$$

$$\therefore \quad |g(x) - g(x')| \leq \frac{1}{n} \int |p_T(x_k, y) - p_T(x'_k, y)| dy \leq \frac{2}{n}.$$

$$\therefore \quad \mathbf{P}\{|g(X_{T-1}) - \mathbf{E}g(X_{T-1})| \geq \varepsilon\} \leq 2 \exp\left(\frac{-n\varepsilon^2}{2} \right).$$

$$\therefore \quad \mathbf{P}\{| \, \|\psi_T^n - \psi_T\| - \mathbf{E}\|\psi_T^n - \psi_T\| \, | \geq \varepsilon\} \leq 2 \exp\left(\frac{-n\varepsilon^2}{2} \right).$$

It now follows from the Borel-Cantelli Lemma that

$$\lim_{n \to \infty} | \, \|\psi_T^n - \psi_T\| - \mathbf{E}\|\psi_T^n - \psi_T\| \, | \to 0 \text{ almost surely.}$$

Thus, $\lim_{n\to\infty} \|\psi_T^n - \psi_T\| \to 0$ almost surely whenever $\mathbf{E}\|\psi_T^n - \psi_T\| \to 0$. In other words, convergence in expectation implies almost sure convergence. That convergence in expectation always holds was shown in Glynn and Henderson (2001, Theorem 4). □

Next is the proof of Theorem 4.2. By Schéffe's Lemma, $\|\psi_\infty^n - \psi_\infty\| \to 0$ whenever $\psi_\infty^n \to \psi_\infty$ pointwise. Moreover, by the LLN in Meyn and Tweedie (1993, Theorem 17.1.7), we know that at each point $y \in S$ the look-ahead estimator $\psi_\infty^n(y)$ converges to the true density $\psi_\infty(y)$ on the complement of a set E_y with $\mathbf{P}(E_y) = 0$. However, since S may be uncountable, we cannot conclude that $\psi_\infty^n \to \psi_\infty$ pointwise with probability one. Thus, to show almost sure L_1 convergence, some degree of regularity is imposed on the density kernel p to help control

the uncountable family of $\mathbf{P}$-null sets $\{E_y : y \in S\}$. This is the purpose of Assumption 4.2.

Lemma 7.1. *Let $B_\delta(y) := \{y' : d(y, y') < \delta\}$. If Assumption 4.2 holds then ψ_∞ is continuous on S, and ψ_∞^n is continuous on S uniformly in n, in the sense that for all $\varepsilon > 0$ and all $y \in S$ there is a $\delta > 0$ such that*

$$y' \in B_\delta(y) \quad \Longrightarrow \quad \sup_{n \in \mathbb{N}} |\psi_\infty^n(y) - \psi_\infty^n(y')| \leq \varepsilon. \tag{35}$$

Proof. Regarding the first statement, fix $\varepsilon > 0$ and $y \in S$. Choose $\delta > 0$ as in Assumption 4.2. Then for $y' \in B_\delta(y)$,

$$\begin{aligned} |\psi_\infty(y) - \psi_\infty(y')| &= \left| \int p(x, y)\psi_\infty(x)dx - \int p(x, y')\psi_\infty(x)dx \right| \\ &\leq \int |p(x, y) - p(x, y')|\psi_\infty(x)dx \leq \varepsilon. \end{aligned}$$

Regarding (35), the same argument yields a $\delta > 0$ such that for $y' \in B_\delta(y)$ we have

$$|\psi_\infty^n(y) - \psi_\infty^n(y')| \leq \frac{1}{n} \sum_{t=1}^n |p(X_t, y) - p(X_t, y')| \leq \frac{1}{n} \sum_{t=1}^n \varepsilon.$$

□

Proof of Theorem 4.2. As discussed above, it is sufficient to show that ψ_∞^n converges to ψ_∞ pointwise for all paths ω in some set $E \in \mathscr{F}$ with $\mathbf{P}(E) = 1$. So let A be a countable dense subset of S, and note by the LLN that for each $a \in A$ there is a corresponding set $E_a \subset \Omega$ with $\mathbf{P}(E_a) = 1$ and $\psi_\infty^n(a) \to \psi_\infty(a)$ on E_a. Let $E := \cap_{a \in A} E_a$. Clearly $\mathbf{P}(E) = 1$. We claim that for every path $\omega \in E$ we have $\psi_\infty^n \to \psi_\infty$ as $n \to \infty$ pointwise. To see this, fix any such path, any $y \in S$ and any $\varepsilon > 0$. By Lemma7.1 we can take a $\delta > 0$ such that $|\psi_\infty(y) - \psi_\infty(y')| < \varepsilon$ for all $y' \in B_\delta(y)$, and, in addition, (35) holds. Choose $a \in A \cap B_\delta(y)$.

By the triangle inequality, $|\psi_\infty^n(y) - \psi_\infty(y)|$ is less than

$$|\psi_\infty^n(y) - \psi_\infty^n(a)| + |\psi_\infty^n(a) - \psi_\infty(a)| + |\psi_\infty(a) - \psi_\infty(y)|$$

$$\therefore \quad |\psi_\infty^n(y) - \psi_\infty(y)| \leq 2\varepsilon + |\psi_\infty^n(a) - \psi_\infty(a)|,$$

where ε does not depend on n. Because we are considering a path in E, taking limits gives

$$\lim_{n\to\infty} |\psi_\infty^n(y) - \psi_\infty(y)| \leq 2\varepsilon.$$

Since ε is arbitrary the proof is done. □

Proof of Theorem 5.1. By Fubini's Theorem, Jensen's inequality and independence of the sequence $X_{T-1}^1, \ldots, X_{T-1}^n$ we get

$$\begin{aligned}
\mathbf{E}\|\psi_T^n - \psi_T\| &= \int \mathbf{E}|\psi_T^n(y) - \psi_T(y)|dy \\
&\leq \int \sqrt{\mathrm{Var}(\psi_T^n(y))}dy \\
&= \sqrt{\frac{1}{n}} \int \sqrt{\mathrm{Var}(p_T(X_{T-1}^m, y))}dy \\
&\leq \sqrt{\frac{1}{n}} \int \sqrt{\mathbf{E}p_T(X_{T-1}^m, y)^2}dy.
\end{aligned}$$

Since $p_T \leq K$ everywhere on $S \times S$, the bound

$$\mathbf{E}\|\psi_T^n - \psi_T\| \leq \sqrt{\frac{1}{n}} K\mu(S).$$

holds for all $n \in \mathbb{N}$. □

Next we turn to the proof of Theorem 5.2. The proof involves several lemmata.

Lemma 7.2. *If Assumption 5.1 holds, then* $\mathbf{E}\exp(r\|X_t\|) < \infty$ *for all* $r > 0$ *and all* $t \in \mathbb{N}$.

Proof. By (i) of Assumption 5.1, we have, for all $t \in \mathbb{N}$,

$$\|X_t\| \leq \alpha\|X_{t-1}\| + L + \|W_t\|.$$

$$\therefore \quad r\|X_t\| \leq r\alpha^t\|x_0\| + \sum_{i=0}^{t-1} r\alpha^i(L + \|W_{t-i}\|).$$

$$\therefore \quad \exp(r\|X_t\|) \leq \exp(r\alpha^t\|x_0\|) \prod_{i=0}^{t-1} \exp(r\alpha^i L) \prod_{i=0}^{t-1} \exp(r\alpha^i\|W_{t-i}\|)).$$

$$\therefore \quad \mathbf{E}\exp(r\|X_t\|) \leq \exp(r\alpha^t\|x_0\|) \prod_{i=0}^{t-1} \exp(r\alpha^i L) \prod_{i=0}^{t-1} \mathbf{E}\exp(r\alpha^i\|W_{t-i}\|)).$$

From (ii) of Assumption 5.1 the expectation $\mathbf{E}\exp(a\|W_t\|))$ is finite for any $a > 0$, so the right hand side of the last inequality is finite. □

Lemma 7.3. *If (ii) of Assumption 5.1 holds, then there exists a positive constant N such that $\varphi(z) \leq N\exp(-\|z\|)$ for all $z \in S$.*

Proof. Let $M := \{z : \|z\| \leq 1/\varrho\}$. For $z \notin M$ we have $\varrho\|z\| > 1$, and hence $\varrho\|z\|^2 > \|z\|$. Therefore,

$$K\exp(-\varrho\|z\|^2) \leq K\exp(-\|z\|), \quad \forall z \notin M.$$

Now set $K_0 := \sup_{z\in M} K\exp(-\varrho\|z\|^2 + \|z\|)$, so that

$$K\exp(-\varrho\|z\|^2) \leq K_0\exp(-\|z\|), \quad \forall z \in M.$$

Now setting $N := \max\{K_0, K\}$ and applying (ii) of Assumption 5.1 provides a constant with the desired property. □

Proof of Theorem 5.2. The proof of Theorem 5.1 provides the bound

$$\mathbf{E}\|\psi_T^n - \psi_T\| \leq \sqrt{\frac{1}{n}} \int \sqrt{\mathbf{E}p_T(X_{T-1}^m, y)^2}dy.$$

We must verify that the integral is finite. To this end, observe that Lemma 7.3 yields an $N < \infty$ with

$$p(X_{T-1}^m, y)^2 = \varphi(g(X_{T-1}^m) - y)^2 \leq N^2\exp(-2\|g(X_{T-1}^m) - y\|).$$

But

$$\exp(-2\|g(X_{T-1}^m) - y\|) \leq \exp(-2\|y\| + 2\|g(X_{T-1}^m)\|).$$

$$\begin{aligned}
\therefore \quad \sqrt{\mathbf{E}p(X_{T-1}^m, y)^2} &\leq N\sqrt{\mathbf{E}\exp(-2\|y\| + 2\|g(X_{T-1}^m)\|)} \\
&\leq N\exp(-\|y\|)\sqrt{\mathbf{E}\exp(2\|g(X_{T-1}^m)\|)} \\
&\leq N\exp(-\|y\|)\sqrt{\mathbf{E}\exp(2\alpha\|X_{T-1}^m\| + 2L)} \\
&\leq N\exp(-\|y\|)\exp(L)\sqrt{\mathbf{E}\exp(2\alpha\|X_{T-1}^m\|)}.
\end{aligned}$$

As a result,

$$\mathbf{E}\|\psi_T^n - \psi_T\| \leq \sqrt{\frac{1}{n}}N\int \exp(-\|y\|)dy\exp(L)\sqrt{\mathbf{E}\exp(2\alpha\|X_{T-1}^m\|)}.$$

Here the expectation on the right is finite from Lemma 7.2. □

Proof of Theorem 5.3. Since the TSLA is unbiased and $\{X_{T-1}^1, \ldots, X_{T-1}^n\}$ are independent, we have

$$\begin{aligned}
\text{IMSE}\,(\psi_T^n) &= \int \mathbf{E}[\psi_T^n(y) - \psi_T(y)]^2\, dy \\
&= \int \text{Var}(\psi_T^n(y))\, dy = \frac{1}{n}\int \text{Var}(p_T(X_{T-1}^m, y))\, dy.
\end{aligned}$$

But evidently

$$\int \text{Var}\, p_T(X_{T-1}^m, y)\, dy \leq \int \mathbf{E}p_T(X_{T-1}^m, y)^2\, dy = \mathbf{E}\int p_T(X_{T-1}^m, y)^2\, dy.$$

$$\therefore \quad \text{IMSE}\,(\psi_T^n) \leq \frac{1}{n}\mathbf{E}\int p_T(X_{T-1}^m, y)^2\, dy. \tag{36}$$

The result (18) now follows. □

Proof of Theorem 5.4. By Fubini's theorem,

$$\mathbf{E}\|\psi_\infty^n - \psi_\infty\| = \int \mathbf{E}|\psi_\infty^n(y) - \psi_\infty(y)|dy,$$

so it suffices to show that for any $y \in S$,

$$(\mathbf{E}|\psi_\infty^n(y) - \psi_\infty(y)|)^2 \leq \frac{4RK^2}{n(1-\alpha)}. \tag{37}$$

Our first observation is that, by Jensen's inequality,

$$(\mathbf{E}|\psi_\infty^n(y) - \psi_\infty(y)|)^2 \leq \mathbf{E}(\psi_\infty^n(y) - \psi_\infty(y))^2. \tag{38}$$

Letting $g(x) := p(x, y) - \psi_\infty(y)$, we can rewrite the right hand side of (38) as

$$\begin{aligned}
\mathbf{E}\left(\frac{1}{n}\sum_{t=1}^{n} g(X_t)\right)^2 &= \frac{1}{n^2}\sum_{1\le s,t\le n} \mathbf{E}\, g(X_s)g(X_t) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{E}\, g(X_t)g(X_t) + \frac{2}{n^2}\sum_{1\le i<j\le n} \mathbf{E}\, g(X_s)g(X_t) \\
&\le \frac{2}{n^2}\sum_{1\le i\le j\le n} \mathbf{E}\, g(X_s)g(X_t) \\
&= \frac{2}{n^2}\sum_{k=0}^{n-1}\sum_{i=1}^{n-k} \mathbf{E}\, g(X_t)g(X_{t+k}).
\end{aligned}$$

Our next step is to consider the terms $\mathbf{E}\, g(X_t)g(X_{t+k})$. In doing so, we use the following result, which can be established from (15), the monotone class theorem and a simple inductive argument (see Durrett, 1996, § 5.1): For any bounded Borel measurable real function h on S and any $t, k \in \mathbb{N}$ we have

$$\mathbf{E}(h(X_{t+k}) \,|\, \mathscr{F}_t) = \int h(z)p^k(X_t, z)dz, \tag{39}$$

where $p^k(x, z)dz$ is the distribution of the state k periods hence when the current state is x, defined inductively by

$$p^1 := p, \quad p^k(x, z) := \int p^{k-1}(x, z')p(z', z)dz'.$$

As a result,

$$\begin{aligned}
\mathbf{E}\, g(X_t)g(X_{t+k}) &= \mathbf{E}(\mathbf{E}(g(X_t)g(X_{t+k}) \,|\, \mathscr{F}_t)) \\
&= \mathbf{E}(g(X_t)\mathbf{E}(g(X_{t+k}) \,|\, \mathscr{F}_t)) = \mathbf{E}\left(g(X_t)\int g(z)p^k(X_t, z)dz\right).
\end{aligned}$$

$$\therefore \quad \mathbf{E}\, g(X_t)g(X_{t+k}) \le \sup_{x\in S} |g(x)| \times \left|\int g(z)p^k(x, z)dz\right|. \tag{40}$$

Pick any $x \in S$. On one hand,

$$|g(x)| = |p(x,y) - \psi_\infty(y)|$$
$$\leq K + |\psi_\infty(y)| = K + \left| \int p(x,y)\psi_\infty(x)dx \right| \leq 2K.$$

On the other hand,

$$\begin{aligned}
\left| \int g(z)p^k(x,z)dz \right| &= \left| \int p(z,y)p^k(x,z)dz - \psi_\infty(y) \right| \\
&= \left| \int p(z,y)p^k(x,z)dz - \int p(z,y)\psi_\infty(z)dz \right| \\
&\leq \int |p(z,y)| \times |p^k(x,z) - \psi_\infty(z)|dz \\
&\leq K \int |p^k(x,z) - \psi_\infty(z)|dz.
\end{aligned}$$

Moreover, by uniform ergodicity,

$$\int |p^k(x,z) - \psi_\infty(z)|dz \leq R\alpha^k.$$

Putting these bounds together with (40) and using the fact that $x \in S$ was arbitrary, we obtain

$$\mathbf{E}\, g(X_t)g(X_{t+k}) \leq 2K^2 R\alpha^k.$$

$$\therefore \quad \sum_{k=0}^{n-1}\sum_{i=1}^{n-k} \mathbf{E}\, g(X_t)g(X_{t+k}) \leq \sum_{k=0}^{n-1}\sum_{i=1}^{n-k} 2K^2 R\alpha^k \leq n\frac{2K^2 R}{1-\alpha}.$$

$$\therefore \quad \frac{2}{n^2}\sum_{k=0}^{n-1}\sum_{i=1}^{n-k} \mathbf{E}\, g(X_t)g(X_{t+k}) \leq \frac{4K^2 R}{n(1-\alpha)}.$$

The proof is now done. □

Proof of Proposition 6.1. We apply (18). Under the first condition we have

$$\int p_T(x,y)^2\, dy = \int \varphi_t^x[y - g_T(x)]^2 dy$$
$$\leq \int K^2 \exp(-2\varrho\|y - g_T(x)\|)dy = \int K^2 \exp(-2\varrho\|z\|)dz.$$

Direct integration gives

$$\int \exp(-2\varrho\|x\|)dx = \frac{2\pi^{k/2}}{\Gamma(k/2)(2\varrho)^k}(k-1)!\,,$$

from which (33) now follows. The proof for the second case is essentially identical, this time using

$$\int \exp(-2\varrho\|x\|^2)dx = \left(\frac{\pi}{2\varrho}\right)^{k/2}.$$

□

References

[1] Brock, W. A. and L. Mirman (1972): "Optimal Economic Growth and Uncertainty: The Discounted Case," *Journal of Economic Theory,* 4, 479–513.

[2] Deaton, A. and G. Laroque (1992): "On the Behavior of Commodity Prices," *Review of Economic Studies,* 59 (1), 1–23.

[3] Devroye, L. and G. Lugosi (2001): *Combinatorial Methods in Density Estimation,* Springer-Verlag, New York.

[4] Durrett, R. (1996): *Probability: Theory and Examples,* Second Edition, Duxbury Press, New York.

[5] Elerain, O., S. Chib and N. Shephard (2001): "Likelihood Inference for Discretely Observed Nonlinear Diffusions," *Econometrica,* 69 (4), 959–993.

[6] Glynn, P. W. and S. G. Henderson (2001): "Computing Densities for Markov Chains via Simulation," *Mathematics of Operations Research,* 26, 375–400.

[7] Hurn, A.S., K.A. Lindsay and V.L. Martin (2003): "On the Efficiacy of Simulated Maximum Likelihood for Estimating the Parameters of Stochastic Differential Equations," *Journal of Time Series Analysis,* 24 (1), 45–63.

[8] Meyn, S. P. and Tweedie, R. L. (1993): *Markov Chains and Stochastic Stability,* Springer-Verlag: London.

[9] Nishimura, K. and J. Stachurski (2005): "Stability of Stochastic Optimal Growth Models: A New Approach," *Journal of Economic Theory*, 122 (1), 100–118.

[10] Rossi-Hansberg, E. and M.L.J. Wright (2005): "Establishment Size Dynamics in the Aggregate Economy," manuscript.

[11] Samuelson, P.A. (1971): "Stochastic Speculative Price," *Proceedings of the National Academy of Science,* 68 (2), 335–337.

[12] Stokey, N. L., R. E. Lucas and E. C. Prescott (1989): *Recursive Methods in Economic Dynamics,* Harvard University Press, Massachusetts.

[13] van der Vaart, A.W. (1998): *Asymptotic Statistics,* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge.

[14] Yakowitz, S.J. (1985): "Nonparametric Density Estimation, Prediction and Regression for Markov Sequences," *Journal of the American Statistical Association,* Vol. 80, No. 389, 215–221.

Department of Economics, University of Melbourne, VIC 3010, Australia, *j.stachurski@econ.unimelb.edu.au*