The Discretization Filter: A Simple Way to Estimate Nonlinear State Space Models*

Leland E. Farmer†
Department of Economics
University of California, San Diego

This Version: May 12, 2017
First Version: August 2014
Link to Most Recent Version

Abstract

Existing methods for estimating nonlinear dynamic models are either too computationally complex to be of practical use, or rely on local approximations which often fail to adequately capture the nonlinear features of interest. I develop a new method, the discretization filter, for approximating the likelihood of nonlinear, non-Gaussian state space models. I apply results from the statistics literature on uniformly ergodic Markov chains to establish that the implied maximum likelihood estimator is strongly consistent, asymptotically normal, and asymptotically efficient. Through simulations I show that the discretization filter is orders of magnitude faster than alternative nonlinear techniques for the same level of approximation error and I provide practical guidelines for applied researchers. I apply my approach to estimate two models at the intersection of macroeconomics and finance: the Wu and Xia (2016) shadow rate term structure model, and the Gabaix (2012) asset pricing model of variable rare disasters. I provide the first estimates of the Gabaix model and show that the estimated model fails to identify the Great Recession as a disaster episode, suggesting the need to consider heterogeneity in the nature of disasters. My estimates of the Wu and Xia shadow rate indicate that unconventional monetary policy was more effective than previously thought.

*I am especially indebted to my thesis advisors, James D. Hamilton and Allan Timmermann, for their invaluable input, support, and meaningful discussions. I am also grateful for helpful feedback and comments from Roy Allen, Brendan Beare, Darrell Duffie, Roger E. A. Farmer, Jesús Fernández-Villaverde, Patrick Gagliardini, Émilien Gouin-Bonenfant, Lars Peter Hansen, Michael Levy, Nelson Lind, Hanno Lustig, Eric Renault, Juan Rubio-Ramirez, Andres Santos, Lawrence Schmidt, Frank Schorfheide, Neil Shephard, Kenneth J. Singleton, Alexis Toda, Harald Uhlig, Daniel Waggoner, Cynthia Wu, Fan Dora Xia, and seminar participants at UCSD, the 2014 SoFiE Summer School for Financial Econometrics at Harvard University, the 2015 California Econometrics Conference at USC, the 2016 MFM summer session for young scholars, and the 2016 Workshop on Methods and Applications for DSGE Models at the Federal Reserve Bank of Chicago. An earlier version of the current paper appeared with the title "Markov Chain Approximation and Estimation of Nonlinear, Non-Gaussian State Space Models" (Farmer 2014). All errors are my own.

†lefarmer@ucsd.edu

1 Introduction

Economists increasingly use nonlinear methods to confront their theories with data. The switch from linear to nonlinear methods is driven, in part, by increased computing power, but also by a desire to understand economic phenomena that cannot easily be captured by linear models. Examples include models which incorporate the zero lower bound on interest rates (ZLB), stochastic volatility, time-varying risk premia, Poisson jumps, credit constraints, borrowing constraints, non-convex adjustment costs, Markov-switching dynamics, and default.

Existing methods for estimating nonlinear dynamic models are either too computationally complex to be of practical use, or rely on local approximations which fail to adequately capture the nonlinear features of interest. In this paper, I develop a new method, the discretization filter, for approximating the likelihood of nonlinear, non-Gaussian state space models.

The major difficulty that arises when studying nonlinear state space models is that the likelihood cannot be evaluated recursively as it can in linear models with the Kalman filter. The discretization filter solves this problem by constructing a discrete-valued Markov chain that approximates the dynamics of the state variables. The dynamics of the system are summarized by a transition matrix as opposed to an infinite dimensional transition kernel.

When there are finitely many states, the likelihood can once again be evaluated recursively with an algorithm analogous to the Kalman filter. This computation involves a sequence of matrix multiplications which is fast and simple to implement. The discretization filter generates an approximation to the likelihood of any nonlinear, non-Gaussian state space model that can be used to estimate the model's parameters using classical or Bayesian methods.

I apply results from the statistics literature on uniformly ergodic Markov chains to establish that the implied maximum likelihood estimator is strongly consistent, asymptotically normal, and asymptotically efficient. I demonstrate through simulations that the discretization filter is orders of magnitude faster than alternative nonlinear techniques for the same level of approximation error and I provide practical guidelines for applied researchers. It is my hope that the method's simplicity will make the quantitative study of nonlinear models easier for and more accessible to applied researchers.

I apply my approach to estimate two models at the intersection of macroeconomics and finance. The first is the Gabaix (2012) asset pricing model of variable rare disasters. The second is the Wu and Xia (2016) shadow rate term structure model. Both models are inherently nonlinear and neither can be consistently estimated with linear methods.

Gabaix (2012) develops a model of asset pricing which posits that the time-varying

probability and severity of rare disasters explain why risk premia are large, volatile and time-varying. I provide the first quantitative estimates of the Gabaix model using data on equities and government bonds to identify the parameters and construct a measure of disaster risk for the U.S. economy. There have been several proposed explanations for phenomena such as the equity premium puzzle, the excess volatility puzzle, and the riskfree rate puzzle. Most existing research on this topic calibrates a model and evaluates its ability to match a few select moments of the data. In contrast, the discretization filter allows researchers to formally estimate a series of models and evaluate their relative abilities to explain the data using model comparison statistics, thus facilitating model selection.

By using a likelihood-based method for estimation, I am able to construct estimates of the hidden states relating to real and nominal risk, which allow me to study additional implications of the model not captured by calibration or moment-matching procedures. In particular, I use these estimates to construct time series for the probability of a disaster, the conditional volatility of inflation, and the expected jump in inflation in the event of a disaster for the U.S. economy. I show that the model fails to identify the Great Recession as a disaster episode, assigning less than a 5% probability to a disaster having occurred between December of 2007 and June of 2009. This is because the model requires a positive jump in inflation in the event of disaster to match an upward sloping nominal yield curve. The model is unable to match the fact that the U.S. experienced low inflation and even deflation during the Great Recession in conjunction with an upward sloping nominal yield curve. This suggests that it is important to consider heterogeneity in the nature of disasters to capture the patterns of the U.S. data.

Wu and Xia (2016) develop a tractable approximation to a shadow rate term structure model. Their model provides a description of yield curve dynamics when the economy is near the zero lower bound on interest rates and provides a way of summarizing the effects of unconventional monetary policy. I show that when the model is estimated using the discretization filter, the estimates of the shadow rate are substantially lower over the zero lower bound period than those provided in their paper. This has important implications for policy makers who use this series as an input to their decision making process. It implies, for example, that their estimates understate the effectiveness of unconventional monetary policy.

The paper is organized as follows. Section 2 reviews related literature. Section 3 explains the discretization filter. Section 4 establishes the strong consistency, asymptotic normality, and asymptotic efficiency of the approximate maximum likelihood estimator implied by the discretization filter. Section 5 provides practical implementation advice for applied researchers. Section 6 provides Monte Carlo comparisons with existing methods in the case of a linear measurement error model and a stochastic volatility model. In section 7,

I estimate the Gabaix (2012) model of variable rare disasters and illustrate a couple of its shortcomings in explaining U.S. asset pricing data. Section 8 re-examines the Wu and Xia (2016) shadow rate term structure model and constructs an updated version of their shadow rate series. Section 9 concludes.

2 Related Literature

This paper is related to the literatures on the discretization of stochastic processes, filtering algorithms for nonlinear state space models, and the statistical properties of maximum likelihood estimators for state space models.

Tauchen (1986) proposed the first method for discretizing stochastic processes with an application to first-order vector autoregressive (VAR) models. Tauchen and Hussey (1991) develop an extension of this method using quadrature formulas, but both of these methods fail to accurately approximate the dynamics of persistent processes (see Kopecky and Suen (2010)). Rouwenhorst (1995) develops a method which accurately approximates highly persistent processes. However, this method is limited to univariate first order Gaussian autoregressive (AR) models. Gospodinov and Lkhagvasuren (2014) develop a method that builds on the Rouwenhorst method to better approximate persistent Gaussian VARs by matching low order conditional moments. Most recently, Farmer and Toda (2016) develop a method for approximating general nonlinear, non-Gaussian first order Markov processes by matching conditional moments using maximum entropy.

A special case of the filtering algorithm proposed in this paper was first considered in Bucy (1969) and Bucy and Senne (1971), now referred to as the "point-mass filter." However, these papers and subsequent refinements only consider one specific method of discretizing the state process. Furthermore, none of these papers consider the asymptotic properties of estimators resulting from these filtering approximations. A comprehensive summary of filtering methods for state space models, including the point-mass filter, can be found in Chen (2003).

The theoretical results and proof techniques in this paper are most directly related to the work of Douc et al. (2004) and Douc et al. (2011). Douc et al. (2004) establish the consistency and asymptotic normality of the maximum likelihood estimator in autoregressive models with a hidden Markov regime that has a compact support. Douc et al. (2011) extend the consistency result to a setting with unbounded support. These papers build on previous work which establish asymptotic properties of the maximum likelihood estimator in several simpler state space models, Baum and Petrie (1966), Leroux (1992), Bickel and Ritov (1996), Bickel et al. (1998), Bakry et al. (1997), and Jensen and Petersen (1999).

3 The Discretization Filter

In this section I introduce the notation used in the remainder of the paper and provide a brief overview of nonlinear state space models. I then explain how the state dynamics of any nonlinear state space model can be approximated by a discrete-state Markov chain. I show how this new state space system can be used to construct an approximation to the maximum likelihood estimator for the parameters and filtering distributions of the original model.

3.1 The Setting

In what follows I restrict attention to the analysis of Hidden Markov Models (HMMs). A HMM is a special type of nonlinear state space model where the observables in any given time period are a function only of the state variables in that time period. However, the results can be generalized to the case when the observation equation additionally depends on some finite number of lags of the observables. Much of the exposition and notation follows Douc et al. (2004).

Let X_t denote the vector of hidden state variables of the state space system at time t. I assume that $\{X_t\}_{t=0}^{\infty}$ is a time-homogeneous, first-order[1], stationary Markov chain and lies in a separable, compact set $\mathcal{X}$,[2] equipped with a metrizable topology and associated Borel σ-field $\mathcal{B}(\mathcal{X})$. Let $P_\theta(x, A)$, where $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$, be the transition kernel of the Markov chain. I further assume that for all $\theta \in \Theta$ and $x \in \mathcal{X}$, each conditonal probability measure $P_\theta(x, \cdot)$ has a density $q_\theta(\cdot\,|x)$ with respect to a common finite dominating measure μ on $\mathcal{X}$.[3]

I assume that the observable sequence $\{Y_t\}_{t=1}^{\infty}$ takes values in a set $\mathcal{Y}$ that is separable and metrizable by a complete metric. I assume that for $t \geq 1$, Y_t is conditionally independent of $\{Y_s\}_{s=1}^{t-1}$ and $\{X_s\}_{s=1}^{t-1}$ given X_t. Note that this excludes models where the observation at time t depends on its own lagged values. This is purely for expositional simplicity and all of the results can be generalized to the case where Y_t depends on some fixed, finite number of lags of itself, $\{Y_{t-1}, \ldots, Y_{t-k}\}$, although this does complicate the construction of the transition matrices. I also assume that the observations conditional on any value of the state $X_t = x$, $x \in \mathcal{X}$, have a density $g_\theta(\cdot\,|x)$ with respect to a σ-finite measure ν on the Borel σ-field $\mathcal{B}(\mathcal{Y})$.

[1] Assuming that X_t is a first-order Markov chain is not restrictive, because the state space can always be redefined to include additional lags of X_t as new state variables. For example, if X_t follows an AR(2) process, one can redefine the state vector to be $(X_t, X_{t-1})'$ and recover the first-order Markov assumption.

[2] Compactness of $\mathcal{X}$ simplifies much of the notation and proofs, however many of the results can be generalized to the noncompact case using techniques developped in Douc et al. (2011)

[3] For two measures μ and ν, μ is said to *dominate* ν if for all A, $\mu(A) = 0$ implies $\nu(A) = 0$.

Define the joint process $\{Z_t\}_{t=0}^{\infty} \equiv \{(X_t, Y_t)\}_{t=0}^{\infty}$ on $\mathcal{Z} \equiv \mathcal{X} \times \mathcal{Y}$ which has transition kernel Π_θ given by

$$\Pi_\theta(z, A) = \int_A g_\theta(y' | x') \, q_\theta(x' | x) \, dx' dy'$$

for any $z \equiv (x, y) \in \mathcal{Z}$ and $A \in \mathcal{B}(\mathcal{Z})$.

I am interested in conducting estimation and inference on the finite dimensional parameter $\theta \in \Theta$ by maximum likelihood. Θ is assumed to be a compact subset of $\mathbb{R}^p$. Denote the true parameter as θ^*.

A HMM is characterized by the following two equations:

$$X_t | X_{t-1} \sim q_\theta(X_t | X_{t-1}) \tag{1}$$

$$Y_t | X_t \sim g_\theta(Y_t | X_t) \tag{2}$$

Equation (1) is the state equation, and it characterizes the distribution of the latent state next period conditional on the current state. Equation (2) is the observation, or measurement equation, and it characterizes the distribution of the observables conditional on the current state.

Let x_t and y_t denote particular realizations of the random variables X_t and Y_t. Given a sample $\{y_t\}_{t=1}^T$, the goal is to obtain estimates of the parameter vector θ and the unobserved states $\{x_t\}_{t=1}^T$, which I will denote by $\hat{\theta}_T$ and $\{\hat{x}_{t|t}\}_{t=1}^T$ respectively.[4] In order to do this, one must obtain an expression for the likelihood of the data:

$$L_T(\theta, x_0) \equiv p_\theta(\mathbf{Y}_1^T | X_0 = x_0) \tag{3}$$

where $\mathbf{Y}_1^T \equiv (Y_1, \ldots, Y_T)$, and X_0 refers to the initial condition of the state. For the remainder of the paper, the notation p_θ without explicit introduction will refer to a general density where the arguments and meaning will be clear from the context. Define the corresponding log-likelihood as

$$\ell_T(\theta, x_0) \equiv \log p_\theta(\mathbf{Y}_1^T | X_0 = x_0) \tag{4}$$

In the subsequent section, I show how to approximate equation (1) by a discrete-valued Markov chain.

3.2 Approximating the State Dynamics

The idea of discretization to alleviate computational problems in economics is not new. One of the first instances of this is Tauchen (1986). He proposes a simple way of approximating

[4] The notation $\hat{x}_{t|t}$ denotes the estimate of x_t conditional only on information through time t. Sometimes smoothed estimates of the unobserved state $\hat{x}_{t|T}$, incorporating all of the data, are of interest.

any Gaussian VAR(1) with a first-order, discrete-valued Markov chain. He then shows that this approximation does a good job of matching unconditional and conditional moments for relatively coarse discretizations. Tauchen's approximation, along with several more recent approximations proposed in the literature,[5] have been widely used to solve asset pricing and DSGE models where the ability to approximate the solutions to integral equations is of key importance.

In this paper I apply this idea of discretization to the estimation of nonlinear, non-Gaussian state space models. More specifically, I construct a discrete-valued, first-order Markov process $\{X_{t,M}\}_{t=1}^{\infty}$, whose dynamics mimic those of the original continuous-valued process $\{X_t\}_{t=1}^{\infty}$. This allows me to summarize the dynamics of the unobserved state by a finite-dimensional transition matrix $P_{\theta,M}$.[6] Note that this is fundamentally different from forecasting the next period's state by taking a local approximation around the current estimate as is done in the extended Kalman filter. My approximation method is global yet does not rely on simulation techniques.

Define a discrete set of M points in $\mathcal{X}$, $\mathcal{X}_M \equiv \{x_{m,M}\}_{m=1}^{M}$, associated with sets $\{A_{m,M}\}_{m=1}^{M}$ which partition $\mathcal{X}$, and define a transition matrix $P_{\theta,M}$ such that the mm'-th element:

$$P_{\theta,M}\left(m,m'\right) = \mathbb{P}_\theta\left(X_{t,M} = x_{m',M} \,|X_{t-1,M} = x_{m,M}\right) \tag{5}$$

corresponds to the probability of transitioning from point $x_{m,M}$ to point $x_{m',M}$ between time $t-1$ and t. The matrix $P_{\theta,M}$ is assumed to be the same for all t, and thus $X_{t,M}$ follows a first-order, time homogeneous, M-state Markov chain.

Note that each row of the matrix $P_{\theta,M}$ can be interpreted as a conditional probability distribution. Specifically, row m correponds to the distribution of $X_{t,M}$ conditional on being at point $x_{m,M}$ at time $t-1$. It is critical that these conditional distributions be good approximations to the true conditional distributions $X_t \,|X_{t-1} = x_{m,M}$.

Define $s_{t,M}$ to be the state of the approximate system at time t. In particular, I will say that the system is in state $s_{t,M} = m$ and let $\zeta_{t,M} = e_m$ when $X_{t,M} = x_{m,M}$, where e_m is the m-th column of the $(M \times M)$ identity matrix. The system outlined above is characterized by the equations:

$$\zeta_{t,M} = P'_{\theta,M}\zeta_{t-1,M} + \tilde{v}_{t,M} \tag{6}$$

$$Y_t \,|X_{t,M} \sim g_\theta\left(Y_t \,|X_{t,M}\right) \tag{7}$$

where $\tilde{v}_{t,M} = \zeta_{t,M} - \mathbb{E}_\theta\left[\zeta_{t,M} \,|\zeta_{t-1,M}\right]$ and $P'_{\theta,M}$ is the transpose of the matrix $P_{\theta,M}$. Equa-

[5] See e.g. Tauchen and Hussey (1991), Rouwenhorst (1995), Adda and Cooper (2003), Flodén (2008), Tanaka and Toda (2013), Gospodinov and Lkhagvasuren (2014), and Farmer and Toda (2016).

[6] This is similar to the idea proposed in Tauchen and Hussey (1991). However, there the primary focus was on computing conditional expectations: here it is approximating the dynamics of a state space model.

tions (6) and (7) are the state and observation equations of the new approximate model. The sequence $\{Y_t\}$ has the same distribution, conditional on the state $X_{t,M}$, as the sequence $\{Y_t\}$ generated by the original model. However, in the approximate model, the $X_{t,M}$ have been restricted to live on a discrete grid.

3.3 Evaluating the Likelihood

In the previous section, I showed how to approximate any HMM by replacing the state equation, equation (1), with a discrete-state Markov chain, equation (6). In this section, I apply the results of Hamilton (1989) to construct an approximation to the likelihood function of the HMM. Hamilton (1989) shows that when the state dynamics of a HMM are characterized by a discrete-state Markov chain, simple prediction and updating equations exist that are analogous to the Kalman filter in the linear case. I use the notation developed in Hamilton (1994). I review these results here and show how they can be used to develop an approximation to the maximum likelihood estimator for θ.

Let $\hat{\zeta}_{t,M|t} = \mathbb{E}_\theta \left[\zeta_{t,M} \left| \boldsymbol{Y}_1^t \right.\right]$ be the econometrician's best inference about the discretized state $\zeta_{t,M}$ conditional on time t information. Inuitively, $\hat{\zeta}_{t,M|t}$ is an $(M \times 1)$ vector of probabilities where each element represents the probability of being at a particular point in the state space at time t conditional on observations up to time t. The forecast of the approximate state today given the previous period's information is given by:

$$\hat{\zeta}_{t,M|t-1} = \mathbb{E}_\theta \left[\zeta_{t,M} \left| \boldsymbol{Y}_1^{t-1} \right.\right] = P'_{\theta,M} \hat{\zeta}_{t-1,M|t-1} \tag{8}$$

Also define

$$\eta_{t,M} = \begin{bmatrix} g_\theta \left(Y_t \left| X_t = x_{1,M} \right.\right) \\ \vdots \\ g_\theta \left(Y_t \left| X_t = x_{m,M} \right.\right) \end{bmatrix} \tag{9}$$

The m-th element of $\eta_{t,M}$ is the likelihood of having observed Y_t conditional on being in state m at time t, i.e. $s_{t,M} = m$.

Note that the marginal likelihood of Y_t given $\boldsymbol{Y}_1^{t-1}$ is then simply given by:

$$p_{\theta,M} \left(Y_t \left| \boldsymbol{Y}_1^{t-1} \right.\right) = 1' \left(\eta_{t,M} \odot \hat{\zeta}_{t,M|t-1}\right) \tag{10}$$

where $\odot$ is element by element multiplication of conformable matrices and 1 is an $(M \times 1)$ vector of ones. The updated inference about the state at time t is

$$\hat{\zeta}_{t,M|t} = \frac{\eta_{t,M} \odot \hat{\zeta}_{t,M|t-1}}{1' \left(\eta_{t,M} \odot \hat{\zeta}_{t,M|t-1}\right)} = \frac{\eta_{t,M} \odot \hat{\zeta}_{t,M|t-1}}{p_{\theta,M} \left(Y_t \left| \boldsymbol{Y}_1^{t-1} \right.\right)} \tag{11}$$

By iterating these equations from period 1 to the sample size T, one can obtain estimates of the filtering distributions $\left\{\hat{\zeta}_{t,M|t}\right\}_{t=1}^{T}$ and the parameters $\hat{\theta}_{T,M}$ by maximizing the log likelihood of the discretized system

$$\ell_{T,M}(\theta) = \sum_{t=1}^{T} \log p_{\theta,M}\left(Y_t \left| \mathbf{Y}_1^{t-1}\right.\right) \tag{12}$$

Alternatively, given a prior distribution for the paramter vector θ, Bayesian methods can be used to sample from its posterior distribution.

Algorithm 1 summarizes the procedure for constructing the discrete approximation to the likelihood and the filtering distributions. This can then be embedded in either a classical or Bayesian procedure for performing likelihood-based estimation.

Algorithm 1: Discretization Filter

1. **Approximate the State Dynamics:** Construct a discrete grid $\{x_{m,M}\}_{m=1}^{M}$ and its associated transition matrix $P_{\theta,M}$ using algorithm 2 in appendix B or any other method appropriate for the process X_t being considered.
2. **Initialization:** Set the initial distribution of the state $\hat{\zeta}_{0,m|0} = \pi_{\theta,M}^{X}$ or any arbitrary distribution. Set $t \rightsquigarrow 1$.
3. **Prediction:** Construct the forecast of the time t state $\hat{\zeta}_{t,M|t-1} = P_{\theta,M}' \hat{\zeta}_{t,M|t-1}$.
4. **Updating 1:** Evaluate the contemporaneous likelihood of having observed data y_t conditional on each possible value of the state, $\eta_{t,M}$, using equation (9). Compute and save the marginal likelihood of observation y_t given by equation (10).
5. **Updating 2:** Compute the time t filtered estimate of the state $\hat{\zeta}_{t,M|t}$ using (11). If $t < T$, set $t \rightsquigarrow t+1$ and go to step 3. Otherwise go to step 6.
6. **Likelihood:** Compute the approximate likelihood of the data, $\ell_{T,M}(\theta)$, using equation (12).

Note that the parameter estimates $\hat{\theta}_{T,M}$ and the log-likelihood function $\ell_{T,M}(\theta)$ are indexed by the number of discrete points M in addition to the sample size T to indicate that the estimates will depend on exactly how the space is discretized. I have omitted the explicit dependence of the likelihood function on the distribution of the initial state $x_{0,M}$. As part of the results in section 4, I will show why this initial condition is irrelevant for the asymptotic properties of $\hat{\theta}_{T,M}$.

Section 4 establishes the strong consistency, asymptotic normality, and asymptotic efficiency of the discretization filter approximation to the maximum likelihood estimator. Those who are interested in applications of the discretization filter may wish to skip ahead to section 5.

4 Asymptotic Properties of the Maximum Likelihood Estimator

In this section I establish strong consistency, asymptotic normality, and asymptotic efficiency of my proposed estimator. I consider joint asymptotics in both the sample size T and the number of discrete points M. I show that the accuracy of my approximation is governed to first order by the proximity of the infinite history filtering distributions of the approximate and true chains $X_{t,M} \left| \boldsymbol{Y}_{-\infty}^{t} \right.$ and $X_t \left| \boldsymbol{Y}_{-\infty}^{t} \right.$. The distance between these distributions is proportional to $h^*(M)$, where $h^*(M)$ is related to the approximation error between the approximate and true one-step-ahead conditional distributions of X_t. Strong consistency simply requires that $T \to \infty$ and $M \to \infty$. Asymptotic normality and asymptotic efficiency further require that $T \times h^*(M) \to 0$ as $M \to \infty$ and $T \to \infty$, i.e. that $M \to \infty$ "fast enough."

A key new theoretical contribution of my paper is to establish a rate of convergence of the ergodic distribution of the approximate discrete chain to the true ergodic distribution. This result represents a new contribution to the literature on discrete approximations of Markov chains with continuous valued states. All proofs can be found in Appendix A.

4.1 Preliminaries and Assumptions

Define the notations $\bar{\mathbb{P}}_\theta$, $\bar{\mathbb{E}}_\theta$, and $\bar{p}_\theta$ to denote probabilities, expectations, and densities evaluated under the assumption that the initial state X_0 is drawn from its ergodic distribution π_θ^X, or analogously $X_{0,M}$ from $\pi_{\theta,M}^X$ in the discrete case.

Before continuing, it is useful to define the extension of the transition kernel $P_{\theta,M}$ to $\mathcal{X}$. For $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$, let

$$P_{\theta,M}(x, A) \equiv \sum_{m=1}^{M} \sum_{m'=1}^{M} P_{\theta,M}(m, m') \mathbb{1}\{x \in A_{m,M}\} \mathbb{1}\{x_{m',M} \in A\}$$

Similarly, define the extension of the ergodic measure $\pi_{\theta,M}^X$ to $\mathcal{X}$. For $A \in \mathcal{B}(\mathcal{X})$, let

$$\pi_{\theta,M}^X(A) \equiv \sum_{m=1}^{M} \pi_{\theta,M}^X(m) \mathbb{1}\{x_{m,M} \in A\}$$

Lastly, I define the limit as $M \to \infty$ of these objects in the natural way:

$$P_{\theta,\infty}(x, A) \equiv \lim_{M \to \infty} \sum_{m=1}^{M} \sum_{m'=1}^{M} P_{\theta,M}(m, m') \mathbb{1}\{x \in A_{m,M}\} \mathbb{1}\{x_{m',M} \in A\}$$

and

$$\pi^X_{\theta,\infty}(A) \equiv \lim_{M\to\infty} \sum_{m=1}^{M} \pi^X_{\theta,M}(m)\, \mathbb{1}\{x_{m,M} \in A\}$$

I will impose assumptions such that these limiting objects are well defined. For the remainder of the section, I will use both the versions of $P_{\theta,M}$ and $\pi^X_{\theta,M}$, defined over $\mathcal{X}$ and $\mathcal{X}_M$, interchangeably and the meaning will be clear from the context.

I now list and discuss my basic assumptions. Assumptions that overlap with Douc et al. (2004) are labeled with an A, and assumptions that are new to this paper are labeled with a B. Assumptions labeled A and B are paired by number, e.g. (A1) and (B1). Each B assumption can be thought of as an analog to the A assumption for the sequence of discrete approximations $X_{t,M}$.

(A1) (a) $0 < \sigma_- \equiv \inf_{\theta\in\Theta} \inf_{x,x'\in\mathcal{X}} q_\theta(x'|x)$ and $\sigma_+ \equiv \sup_{\theta\in\Theta} \sup_{x,x'\in\mathcal{X}} q_\theta(x'|x) < \infty$.

(b) For all $y' \in \mathcal{Y}$, $0 < \inf_{\theta\in\Theta} \int_{\mathcal{X}} g_\theta(y'|x)\,dx$ and $\sup_{\theta\in\Theta} \int_{\mathcal{X}} g_\theta(y'|x)\,dx < \infty$.

(B1) $Q^-_+ \equiv \inf_{\theta\in\Theta} \inf_{M\in\mathbb{Z}^+} \inf_{m,m',m'',m'''} \frac{P_{\theta,M}(m,m')}{P_{\theta,M}(m'',m''')} > 0$

Assumption (A1)(a) implies that there is a positive probability that the state variable can move from any part of the state space to any other part of the state space. This means that the state space $\mathcal{X}$ of the Markov chain $\{X_t\}$ is what's known as 1-small, or petite. This further implies that for all $\theta \in \Theta$, $\{X_t\}$ has a unique invariant measure π^X_θ and is uniformly ergodic (see Meyn and Tweedie (1993) for a proof).

Assumption (B1) guarantees that the discrete process $\{X_{t,M}\}$ has a unique invariant distribution $\pi^X_{\theta,M}$ and is uniformly ergodic for every value $M < \infty$. Additionally it is needed so that the bound on the mixing rate of $X_{t,M}$ is independent of M and θ. This will be satisfied for any stochastic process satisfying (A1)(a) that is approximated using the methods reviewed in section 3.2. Note that while all elements of the transition matrix $P_{\theta,M}$ converge to 0 individually as $M \to \infty$, the limits of the ratios of these elements are still well defined.

(A2) For all $\theta \in \Theta$, the transition kernel Π_θ is positive Harris recurrent and aperiodic with invariant distribution π_θ.

(B2) For all $\theta \in \Theta$, the transition kernel $\Pi_{\theta,\infty}$ is positive Harris recurrent and aperiodic with invariant distribution $\pi_{\theta,\infty}$.

These assumptions guarantee that the original joint Markov process $\{Z_t\}$ and the limiting approximating Markov chain $\{Z_{t,\infty}\}$ are themselves uniformly ergodic. Note that assumption (B2) is needed in addition to assumption (B1) to account for the limiting case of the chain.

Assumption (A2) implies that for any initial measure λ,

$$\lim_{t\to\infty} \left\| \lambda \Pi_\theta^{(t)} - \pi_\theta \right\|_{TV} = 0 \tag{13}$$

where $\|\cdot\|_{TV}$ is the total variation norm, defined for any two probability measures μ_1 and μ_2 as

$$\|\mu_1 - \mu_2\|_{TV} = \sup_A |\mu_1(A) - \mu_2(A)|$$

and $\Pi_\theta^{(t)}$ is the t-th iterate of the transition kernel Π_θ. In words, for any initial measure of the joint process $\{Z_t\}$, the probability of being in any measurable set $A \in \mathcal{B}(\mathcal{Z})$ approaches the ergodic probability of being in that set uniformly over all measurable sets A as $t \to \infty$. This convergence is also independent of the initial measure λ. An analogous property holds for the process $\{Z_{t,\infty}\}$ by assumption (B2). Developing a bound on this rate of convergence will be critical for the coming developments.

Lastly, assume that

(A3) $b_+ \equiv \sup_{\theta\in\Theta} \sup_{y_1,x} g_\theta(y_1|x) < \infty$ and $\overline{\mathbb{E}}_{\theta^*}(|\log b_-(y_1)|) < \infty$, where $b_-(y_1) \equiv \inf_{\theta\in\Theta} \int_{\mathcal{X}} g_\theta(y_1|x)\,\mu(dx)$.

(B3) $\overline{\mathbb{E}}_{\theta^*}(|\log c_-(y_1)|) < \infty$, where $c_-(y_1) \equiv \inf_{\theta\in\Theta} \inf_{M\in\mathbb{Z}^+} \inf_{1\le m\le M} \sum_{m'=1}^{M} P_{\theta,M}(m,m')\, g_\theta\left(y_1 \middle| x_{m',M}\right)$

Assumptions (A3) and (B3) are additional boundedness conditions involving the observation density g_θ which will be necessary to establish the existence of certain limits. Additional assumptions will be introduced and explained as needed.

4.2 Consistency

The proof of consistency can be broken down into two main parts. The first is to show that the approximation to the likelihood function implied by the discretization filter, properly normalized, converges to a well defined asymptotic criterion function $\ell_M(\theta)$, for fixed M, as the sample size $T \to \infty$. It is important that this convergence be uniform with respect to the parameter $\theta \in \Theta$, the initial condition $x_0 \in \mathcal{X}_M$, and the number of discrete points $M \in \mathbb{Z}^+$. This step relies largely on the analysis in Douc et al. (2004), with the additional requirement that the conditions be strengthened so that the convergence is uniform with respect to the number of discrete points M used to construct the approximation. This will be a consequence of the uniform ergodicity of the filtering distributions $\left\{X_{t,M} \middle| \mathbf{Y}_1^t\right\}_{M=1}^{\infty}$, which follows from the uniform ergodicity of the discrete Markov chains $\{X_{t,M}\}_{M=1}^{\infty}$.

The second part, which is new to this paper, is to show that this approximate limiting criterion function $\ell_M(\theta)$, which is defined for any M, converges to the true limiting criterion

function $\ell(\theta)$ as the number of points used in the approximation $M \to \infty$. I will show that this holds for any discretization method whose one-step-ahead conditional distributions $X_{t,M} | X_{t-1,M} = x$ converge in distribution to the one-step-ahead conditional distributions of the original continuous process $X_t | X_{t-1} = x$ as $M \to \infty$.

Together, these two pieces will imply that $T^{-1}\ell_{T,M}(\theta)$ converges uniformly to $\ell(\theta)$ as $T, M \to \infty$. Under some additional regularity conditions, this will imply that the estimator $\hat{\theta}_{T,M}$ converges to the true parameter θ^* almost surely as $T, M \to \infty$.

Following Douc et al. (2004), I first establish that the distribution of $X_{t,M}$ given a history of observations $\mathbf{Y}_r^s$ is itself a uniformly ergodic (inhomogeneous) Markov chain with minorizing constant independent of the parameter $\theta \in \Theta$ and the number of discrete points $M \in \mathbb{Z}^+$. This is the analogous result to Lemma 1 in their paper. Note that a Markov chain with transition kernel P_θ is said to satisfy a uniform minorization condition if there exist a probability measure μ_Q, a positive integer n, and $\epsilon > 0$ such that

$$P_\theta^{(n)}(x, A) \geq \epsilon \mu_Q(A)$$

for all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$, where $P_\theta^{(n)}$ is the n-step ahead transition kernel of the Markov chain.

Define $Q_M^- \equiv \inf_{m,m'} P_{\theta,M}(m, m')$, $Q_M^+ \equiv \sup_{m,m'} P_{\theta,M}(m, m')$, and $Q_+^- \equiv \frac{Q_M^-}{Q_M^+}$ for $M \in \mathbb{Z}^+$. I now state the first lemma

Lemma 1. *Assume (A1) and (B1). Let $s, r \in \mathbb{Z}$, with $r \leq s$, $\theta \in \Theta$, and $M \in \mathbb{Z}^+$. Under $\bar{\mathbb{P}}_\theta$, conditionally on $\mathbf{Y}_r^s$, $\{X_{t,M}\}_{t \geq r}$ is an inhomogeneous Markov chain, and for all $t > r$ there exists a function $\mu_{t,M}(\boldsymbol{y}_t^s, A)$ such that:*

(i) for any $A \in \mathcal{B}(\mathcal{X}_M)$, $\boldsymbol{y}_t^s \mapsto \mu_{t,M}(\boldsymbol{y}_t^s, A)$ is a Borel function;

(ii) for any $\boldsymbol{y}_t^s$, $\mu_{t,M}(\boldsymbol{y}_t^s, \cdot)$ is a probability measure on $\mathcal{B}(\mathcal{X}_M)$. In addition, for all $\boldsymbol{y}_t^s$ it holds that $\mu_{t,M}(\boldsymbol{y}_t^s, \cdot) \ll \mu_{c,M}$ (where $\mu_{c,M}$ is counting measure on $\mathcal{X}_M$) and for all $\mathbf{Y}_r^s$,

$$\inf_{x \in \mathcal{X}_M} \bar{\mathbb{P}}_\theta(X_{t,M} \in A | X_{t-1,M} = x, \mathbf{Y}_r^s) \geq Q_+^- \mu_{t,M}(\mathbf{Y}_t^s, A)$$

The major difference between this Lemma and the one established in Douc et al. (2004) is that for the following results, it will be crucial that the minorizing constant be the same for all M, in order to establish uniform convergence over $M \in \mathbb{Z}^+$ of the approximate likelihood function. Note that although the minorizing measure, $\mu_{t,M}(\mathbf{Y}_t^s, \cdot)$, *does* depend on both the number of points, M, and the observations the chain is conditioned on, $\mathbf{Y}_t^s$, it doesn't affect the mixing rate. The previous lemma leads to the following corollary, using standard results for uniformly minorized Markov chains (see e.g. Lindvall (1992) Sections III.9-11).

Corollary 1. *Assume (A1) and (B1). Let $r, s \in \mathbb{Z}$ with $r \leq s$, $\theta \in \Theta$, and $M \in \mathbb{Z}^+$. Then for all $t \geq r$, all probability measures μ_1 and μ_2 on $\mathcal{B}(\mathcal{X}_M)$, and all $\mathbf{Y}_r^s$,*

$$\left\| \int_{\mathcal{X}_M} \bar{\mathbb{P}}_\theta \left(X_{t,M} \in \cdot \,|X_{r,M} = x, \mathbf{Y}_r^s \right) \mu_1 (dx) - \int_{\mathcal{X}_M} \bar{\mathbb{P}}_\theta \left(X_{t,M} \in \cdot \,|X_{r,M} = x, \mathbf{Y}_r^s \right) \mu_2 (dx) \right\|_{TV} \leq \rho^{t-r}$$

where $\rho \equiv 1 - Q_+^-$.

This corollary establishes that the Markov chain "uniformly forgets" its history at an exponential rate. That is, no matter where the chain is started, it converges to its ergodic distribution exponentially fast. The fact that the bound is deterministic will be important for establishing strong consistency.

The next step consists of showing that the approximate likelihood function $\ell_{T,M}(\theta, x_{0,M})$ with an arbitrary initial condition $x_{0,M}$ stays within a deterministic bound of $\ell_{T,M}(\theta)$ where $x_{0,M}$ is drawn from its ergodic distribution.

Lemma 2. *Assume (A1)-(A2) and (B1)-(B2). Then, for all $x_{0,M} \in \mathcal{X}_M$ and $M \in \mathbb{Z}^+$,*

$$\sup_{\theta \in \Theta} |\ell_{T,M}(\theta, x_{0,M}) - \ell_{T,M}(\theta)| \leq 1/(1-\rho)^2, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

Next I show that $T^{-1}\ell_{T,M}(\theta)$ can be approximated by the sample mean of a $\bar{\mathbb{P}}_{\theta^*}$-stationary ergodic sequence of bounded random variables which has a well defined limit. To this end I first define the quantities:

$$\Delta_{t,r,M,x}(\theta) \equiv \log \bar{p}_{\theta,M}\left(Y_t \left| \mathbf{Y}_{-r}^{t-1}, X_{-r,M} = x \right. \right)$$
$$\Delta_{t,r,M}(\theta) \equiv \log \bar{p}_{\theta,M}\left(Y_t \left| \mathbf{Y}_{-r}^{t-1} \right. \right) = \int \log \bar{p}_{\theta,M}\left(Y_t \left| \mathbf{Y}_{-r}^{t-1}, X_{-r,M} = x \right. \right) \bar{\mathbb{P}}_\theta \left(dx_{-r,M} \left| \mathbf{Y}_{-r}^{t-1} \right. \right)$$

Consider the thought experiment of fixing the number of points M, but letting $T \to \infty$. Define the limiting object as

$$\ell_M(\theta) \equiv \bar{\mathbb{E}}_{\theta^*}\left[\Delta_{0,\infty,M}(\theta) \right]$$

I will show that such a limiting object is well-defined and that the sample analogue converges to this limit almost-surely. In particular, I will show that $\{\Delta_{t,r,M}\}_{r \geq 0}$ and $\{\Delta_{t,r,M,x}\}_{r \geq 0}$ converge uniformly w.r.t. $\theta \in \Theta$ $\bar{\mathbb{P}}_{\theta^*}$-a.s. by showing they are uniform Cauchy sequences.

Lemma 3. *Assume (A1)-(A3) and (B1)-(B3). Then for all $t \geq 1$, $r, r' \geq 0$, and $M \in \mathbb{Z}^+$,*

$\bar{\mathbb{P}}_{\theta^*}$-*a.s.*,

$$\sup_{\theta\in\Theta}\sup_{x,x'\in\mathcal{X}_M}\left|\Delta_{t,r,M,x}(\theta)-\Delta_{t,r',M,x'}(\theta)\right|\leq\rho^{t+\min(r,r')-1}/(1-\rho), \tag{14}$$

$$\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}_M}\left|\Delta_{t,r,M,x}(\theta)-\Delta_{t,r,M}(\theta)\right|\leq\rho^{t+r-1}/(1-\rho), \tag{15}$$

$$\sup_{\theta\in\Theta}\sup_{r\geq 0}\sup_{x\in\mathcal{X}_M}\left|\Delta_{t,r,M,x}(\theta)\right|\leq\max\left(\left|\log b_+\right|,\left|\log c_-(Y_t)\right|\right) \tag{16}$$

Equation (14) of Lemma 3 shows that $\{\Delta_{t,r,M,x}\}_{r\geq 0}$ is a uniform Cauchy sequence w.r.t. $\theta\in\Theta$ and thus converges $\bar{\mathbb{P}}_{\theta^*}$-a.s. to a limit which does not depend on the initial value x. I label this limit $\Delta_{t,\infty,M}$ and intuitively this can be thought of as $\log\bar{p}_{\theta,M}\left(Y_t\left|\boldsymbol{Y}_{-\infty}^{t-1}\right.\right)$, the marginal likelihood of an observation Y_t given an infinite history of data.

Equation (16) of Lemma 3 shows that $\{\Delta_{t,r,M,x}(\theta)\}_{r\geq 0}$ is uniformly bounded in $L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$ and thus its limit $\Delta_{t,\infty,M}(\theta)$ is also in $L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$. Furthermore, note that $\{\Delta_{t,\infty,M}(\theta)\}$ is a $\bar{\mathbb{P}}_{\theta^*}$-stationary ergodic process.

By setting $r=0$ and letting $r'\to\infty$ in equation (14), it follows that

$$\sup_{\theta\in\Theta}\left|\Delta_{t,0,M,x}(\theta)-\Delta_{t,\infty,M}(\theta)\right|\leq\rho^{t-1}/(1-\rho)$$

Furthermore, setting $r=0$ in equation (15) implies that

$$\sup_{\theta\in\Theta}\left|\Delta_{t,0,M,x}(\theta)-\Delta_{t,0,M}(\theta)\right|\leq\rho^{t-1}/(1-\rho)$$

By combining these two inequalities, applying the triangle inequality, and summing from 1 to T, I obtain Corollary 2.

Corollary 2. *Assume (A1)-(A2) and (B1)-(B2). Then*

$$\sum_{t=1}^{T}\sup_{M\in\mathbb{Z}^+}\sup_{\theta\in\Theta}\left|\Delta_{t,0,M}(\theta)-\Delta_{t,\infty,M}(\theta)\right|\leq 2/(1-\rho)^2,\quad\bar{\mathbb{P}}_{\theta^*}\text{-}a.s.$$

Corollary 2 shows that $T^{-1}\ell_{T,M}(\theta)$ can be approximated by the sample mean of a stationary ergodic sequence, uniformly w.r.t. θ. Since $\Delta_{0,\infty,M}\in L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$, the ergodic theorem implies that $T^{-1}\ell_{T,M}(\theta)\to\ell_M(\theta)$ $\bar{\mathbb{P}}_{\theta^*}$-a.s. and in $L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$ as $T\to\infty$. Note that this convergence is uniform over $M\in\mathbb{Z}^+$. This will be important when I start considering joint asymptotics in T and M.

Define $\ell(\theta)\equiv\bar{\mathbb{E}}_{\theta^*}\left[\log\bar{p}_\theta\left(Y_0\left|\boldsymbol{Y}_{-\infty}^{0}\right.\right)\right]$. The next step towards establishing consistency is to show that $\ell_M(\theta)\to\ell(\theta)$ as $M\to\infty$. The difference in these two quantities is related to the difference in the approximate and true filtering distributions for infinite histories of

observations, $X_{t,M} \left| \boldsymbol{Y}_{-\infty}^{t} \right.$ and $X_t \left| \boldsymbol{Y}_{-\infty}^{t} \right.$.

I first prove that the ergodic distribution of the approximate discrete Markov chain converges weakly to that of the original continuous Markov chain, i.e. that $X_{t,M} \xrightarrow{d} X_t$ as $M \to \infty$. Proposition 1 establishes this convergence and provides a bound on the difference between the two distributions as a function of the number of points M.

Define $\mathcal{A}$ as the collection of all continuity sets of X_t. I make one further assumption regarding the approximation quality of the sequence of transition kernels $\{P_{\theta,M}\}$.

(BT) For all $A \in \mathcal{A}$, the sequence of approximations $P_{\theta,M}$ satisfy

$$\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} \left| P_{\theta,M}(x, A) - P_\theta(x, A) \right| = O\left(h(M)\right) \tag{17}$$

where $h(M)$ satisfies $\lim_{M \to \infty} h(M) = 0$.

This assumption allows the practitioner to use *all* of the discretization methods outlined in 3.2 to construct $P_{\theta,M}$. I have chosen to illustrate the case where the Farmer and Toda (2016) method with trapezoidal quadrature rule is used. In this case, assumption (BT) is satisfied with $h(M) = M^{-2/d}$, where d is the dimension of the state space $\mathcal{X}$.[7]

Proposition 1. *Assume (A1)-(A3), (B1)-(B3), and (BT). Then it follows that for any* $A \in \mathcal{A}$,

$$\sup_{\theta \in \Theta} \left| \pi_{\theta,M}^{X}(A) - \pi_\theta^X(A) \right| = o\left(h^*(M)\right)$$

where $h^*(M)$ *satisfies* $\lim_{M \to \infty} h^*(M) = 0$. *If the transition kernel is approximated as proposed in Farmer and Toda (2016) with a trapezoidal quadrature rule,*

$$h^*(M) = M^{-(2-\delta)/d}$$

for any $\delta > 0$.

Note that even faster rates can be achieved through clever choice of the quadrature formula and the assumptions one is willing to make about the smoothness of the likelihood function.[8] By combining Proposition 1 with uniform ergodicity of $X_{t,M}$ and X_t, it can be shown that this approximation error directly translates to probabilities computed under the filtering distributions $X_{t,M} \left| \boldsymbol{Y}_{r}^{t} \right.$ and $X_t \left| \boldsymbol{Y}_{r}^{t} \right.$.

[7] For a discussion of error convergence properties see Tanaka and Toda (2015).

[8] There has been substantial research in the field of Quasi Monte-Carlo integration methods, which seek deterministic sequences to approximate high dimensional integrals which break the curse of dimensionality. These are referred to as low discrepancy sequences and their accuracy for numerical integration has been shown to depend only polynomially on the dimension d rather than exponentially. The use of these sequences to approximate the dynamics of high dimensional state processes is a promising area of study which I investigate in ongoing research. Further, there are no known convergence rates for the Tauchen or point mass filter approximations to the transition kernel and I leave this for future work.

Lemma 4. *Assume (A1)-(A3), (B1)-(B3), and (BT). Then*

$$\sup_{\theta\in\Theta}\left|\ell_M(\theta)-\ell(\theta)\right| = o\left(h^*(M)\right)$$

Combining Corollary 2, Lemma 2, and Lemma 4 leads to the following pointwise convergence result

Corollary 3. *Assume (A1)-(A3), (B1)-(B3), and (BT). Then for all sequences of initial points* $\{x_{0,M}\}$ *and* $\theta\in\Theta$*,*

$$\lim_{M,T\to\infty} T^{-1}\ell_{T,M}(\theta,x_{0,M}) = \ell(\theta), \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s. and in } L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$$

The final step before I can state the strong consistency result involves showing that $\ell_M(\theta)$ is continuous w.r.t. θ for all $M\in\mathbb{Z}^+$. This will allow me to strengthen Corollary 3 from pointwise convergence to uniform convergence in θ. Note that by (16) and the dominated convergence theorem,

$$\ell_M(\theta) = \bar{\mathbb{E}}_{\theta^*}\left[\lim_{r\to\infty}\Delta_{0,r,M,x}(\theta)\right] = \lim_{r\to\infty}\bar{\mathbb{E}}_{\theta^*}\left[\Delta_{0,r,M,x}(\theta)\right]$$

It suffices to show that $\Delta_{0,r,M,x}(\theta)$ is continuous w.r.t θ, since $\{\Delta_{0,r,M,x}(\theta)\}_{r\geq 0}$ is a uniform Cauchy sequence $\bar{\mathbb{P}}_{\theta^*}$-a.s. which is uniformly bounded in $L^1\left(\bar{\mathbb{P}}_{\theta^*}\right)$.

The following additional assumptions are needed to establish continuity

(A4) For all $x,x'\in\mathcal{X}$ and all $y'\in\mathcal{Y}$, $\theta\mapsto q_\theta(x,x')$ and $\theta\mapsto g_\theta(y'|x)$ are continuous.

(B4) For all $M\in\mathbb{Z}^+$, $x\in\mathcal{X}_M$, and $A\in\mathcal{B}(\mathcal{X}_M)$, $\theta\mapsto P_{\theta,M}(x,A)$ is continuous.

Lemma 5. *Assume (A1)-(A4), (B1)-(B4), and (BT), then*

$$\lim_{\delta\to 0}\bar{\mathbb{E}}_{\theta^*}\left[\sup_{M\in\mathbb{Z}^+}\sup_{|\theta'-\theta|\leq\delta}\left|\Delta_{t,\infty,M}(\theta')-\Delta_{t,\infty,M}(\theta)\right|\right] = 0.$$

A direct consequence of Lemma 5 is that the convergence established in Corollary 3 can be strengthened to uniform convergence in $\theta\in\Theta$.

Proposition 2. *Assume (A1)-(A4), (B1)-(B4), and (BT). Then*

$$\lim_{M,T\to\infty}\sup_{\theta\in\Theta}\sup_{x_{0,M}\in\mathcal{X}_M}\left|T^{-1}\ell_{T,M}(\theta,x_{0,M})-\ell(\theta)\right| = 0, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

The last assumption needed to establish consistency is an identification assumption guaranteeing that θ^* is a unique maximizer of the likelihood function

(A5) $\theta = \theta^*$ if and only if

$$\bar{\mathbb{E}}_{\theta^*}\left[\log \frac{\bar{p}_{\theta^*}\left(\boldsymbol{Y}_1^t\right)}{\bar{p}_{\theta}\left(\boldsymbol{Y}_1^t\right)}\right] = 0 \qquad \text{for all } t \geq 1. \tag{18}$$

This is a high level assumption about the identification of the model. In general this is a difficult condition to verify because it relies on the ergodic distribution of the joint Markov chain $\{Z_t\}$. For a more thorough discussion on when this assumption is satisfied in the context of HMM, see Douc et al. (2011). Under the additional assumption (A5), I am ready to state my first main result, strong consistency of the maximum likelihood estimator

Theorem 1. *Assume (A1)-(A5), (B1)-(B4), and (BT). Then, for any sequence of initial points $x_{0,M} \in \mathcal{X}_M$, $\hat{\theta}_{T,M,x_{0,M}} \to \theta^*$, $\bar{\mathbb{P}}_{\theta^*}$-a.s. as $T \to \infty$ and $M \to \infty$.*

This is a powerful result. It states that the maximum likelihood estimator is not only consistent but strongly consistent. In addition, the estimator is strongly consistent *independently* of the rate at which the number of points M grows.

4.3 Asymptotic Normality

Next I turn to the asymptotic distribution of the maximum likelihood estimator. In order to establish asymptotic normality I will need additional assumptions regarding the smoothness and boundedness of first and second derivatives of the likelihood function.

Let ∇_θ and ∇_θ^2 be the gradient and the Hessian operator with respect to the parameter θ respectively. Assume there exists a positive real δ such that on $G \equiv \{\theta \in \Theta : |\theta - \theta^*| < \delta\}$, the following assumptions hold

(A6) For all $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$, the functions $\theta \mapsto q_\theta(x, x')$ and $\theta \mapsto g_\theta(y' \,|x')$ are twice continuously differentiable on G.

(A7) (a) $\sup_{\theta \in G} \sup_{x,x'} \|\nabla_\theta \log q_\theta(x, x')\| < \infty$ and $\sup_{\theta \in G} \sup_{x,x'} \left\|\nabla_\theta^2 \log q_\theta(x, x')\right\| < \infty$

(b) $\bar{\mathbb{E}}_{\theta^*}\left[\sup_{\theta \in G} \sup_x \|\nabla_\theta \log g_\theta(Y_1 \,|x)\|^2\right] < \infty$ and $\bar{\mathbb{E}}_{\theta^*}\left[\sup_{\theta \in G} \sup_x \left\|\nabla_\theta^2 \log g_\theta(Y_1 \,|x)\right\|\right] < \infty$

(A8) (a) For ν-almost all $y' \in \mathcal{Y}$ there exists a function $f_{y'} : \mathcal{X} \to \mathbb{R}^+ \in L^1(\mu)$ such that $\sup_{\theta \in G} g_\theta(y' \,|x) \leq f_{y'}(x)$.

(b) For μ-almost all $X \in \mathcal{X}$, there exist functions $f_x^1 : \mathcal{Y} \to \mathbb{R}^+$ and $f_x^2 : \mathcal{Y} \to \mathbb{R}^+$ in $L^1(\nu)$ such that $\|\nabla_\theta g_\theta(y' \,|x)\| \leq f_x^1(y')$ and $\left\|\nabla_\theta^2 g_\theta(y' \,|x)\right\| \leq f_x^2(y')$ for all $\theta \in G$.

Instead of re-establishing asymptotic normality of my proposed estimator using the techniques in Douc et al. (2004), I use Theorem 7 from their paper. I reproduce the theorem here for completeness.

Theorem 2 (Theorem 7 from Douc et al. (2004))**.** *Assume that* $\tilde{\theta}_{T,x_0}$ *is an estimator satisfying* $\ell_T\left(\tilde{\theta}_{T,x_0}, x_0\right) \geq \sup_{\theta\in\Theta} \ell_T(\theta, x_0) - R_T$ *and assumptions (A1)-(A8) hold. Then the following are true:*

(i) If $R_T = o_p(T)$ *(with* $P = \bar{\mathbb{P}}_{\theta^*}$*), then* $\tilde{\theta}_{T,x_0}$ *is consistent.*

(ii) If $R_T = O_p(1)$*, then* $T^{1/2}\left(\tilde{\theta}_{T,x_0} - \theta^*\right) = O_p(1)$*, that is the sequence* $\left\{\tilde{\theta}_{T,x_0}\right\}$ *is* $T^{1/2}-$*consistent under* $\bar{\mathbb{P}}_{\theta^*}$*.*

(iii) If $R_T = o_p(1)$*, then* $T^{1/2}\left(\tilde{\theta}_{T,x_0} - \theta^*\right) \to N\left(0, I(\theta^*)^{-1}\right)$*,* $\bar{\mathbb{P}}_{\theta^*}-$*weakly as* $T \to \infty$*.*

I derive an explicit expression for R_T as a function of M and T and provide conditions under which my proposed estimator satisfies condition (iii) of Theorem 2, which corresponds to asymptotic normality. Note that the bounds I have derived to establish consistency are not sufficient to establish asymptotic normality of my proposed estimator. I can only establish that condition (ii) of Theorem 3 is satisfied using the deterministic bounds applied thus far. To establish conditions under which (iii) is also satisfied, I use an Azuma-Hoeffding inequality derived in Douc et al. (2011). Using this new bound, I am able to state my second main result, asymptotic normality.

Theorem 3. *Assume (A1)-(A8), (B1)-(B4), (BT), and that* $I(\theta^*)$ *is positive definite. Then for any sequence of initial points* $x_{0,M} \in \mathcal{X}_M$*,*

$$\sqrt{T}\left(\hat{\theta}_{T,M,x_{0,M}} - \theta^*\right) \to N\left(0, I(\theta^*)^{-1}\right)$$

$\bar{\mathbb{P}}_{\theta^*}$*-weakly as* $T \to \infty$*,* $M \to \infty$*, and* $T \times h^*(M) \to 0$*.*

Note that this result is actually stronger than just asymptotic normality. Theorem 3 establishes that my proposed estimator and the infeasible maximum likelihood estimator are asymptotically equivalent. That is, my estimator asymptotically achieves the Cramér-Rao lower bound.

5 Recommendations for Applied Researchers

In this section I provide recommendations for how to select the grid points of the approximate finite-state Markov chain and to construct the transition matrix for the discretization filter.

5.1 Choosing the Number of Grid Points

The asymptotic theory I developed in section 4 shows that if the Farmer and Toda (2016) method with a trapezoidal quadrature rule is used to construct the transition matrix, the discretization error of the likelihood function is of the order $TM^{-2/d}$. While this is only a rate condition, I use it to recommend a rule of thumb choice for the number of points M used to construct the discretization. Setting this ratio equal to a constant and solving for M, one gets the rule of thumb

$$M = cT^{d/2} \tag{19}$$

where the constant c is a nuisance parameter. For example, if the dimension d of the state space is 1, the rule says to choose a number of points proportional to the cube root of the sample size. If $d = 2$, then the rule recommends choosing the number of points equal to the sample size. I investigate the effect of choosing different values of c on the accuracy of the approximation in section 6.

Figure 1 plots the rule-of-thumb choice for M for state spaces of dimensions 1-4, for sample sizes up to $T = 100$ and $c = 1$.

The asymptotic analysis implies that M should be chosen to be as large as possible. However, for sufficiently large computational problems, it may not be possible to choose a large number for M. An applied researcher faces a tradeoff between computation time and the accuracy of the approximation, which I will elaborate on in section 6. This rule of thumb can be thought of as a lower bound on the number of points to choose in order to retain validity of confidence intervals constructed for parameters using a normal approximation.

5.2 Selecting the Grid Points

When establishing my theoretical results, I assumed that the state space is compact. This is a convenient theoretical device that makes the proofs cleaner and more intuitive; but I conjecture that it is not necessary for my main results.[9] In general, practitioners specify state space models that take values in unbounded spaces. In this section, I address how to choose the support of the discretized probability measure when the state space is unbounded.

Consider the case where the number of discretization points, M, has been fixed and the goal is to choose the support of the discrete approximation, $\mathcal{X}_M$. In order for the discretized system to be a good approximation to the original model, the boundary points should be chosen to bracket the underlying state vector with high probability. This is analogous to picking boundary points from the tails of the ergodic distribution.

[9]The assumption of uniform ergodicity can be relaxed to geometric ergodicity, where the mixing rate of the Markov chain depends on the initial distribution. Under suitable restrictions on the initial distribution, consistency can still be established using the techniques in Douc et al. (2011). Asymptotic normality of the maximum likelihood estimator under geometric ergodicity appears to still be an open problem.

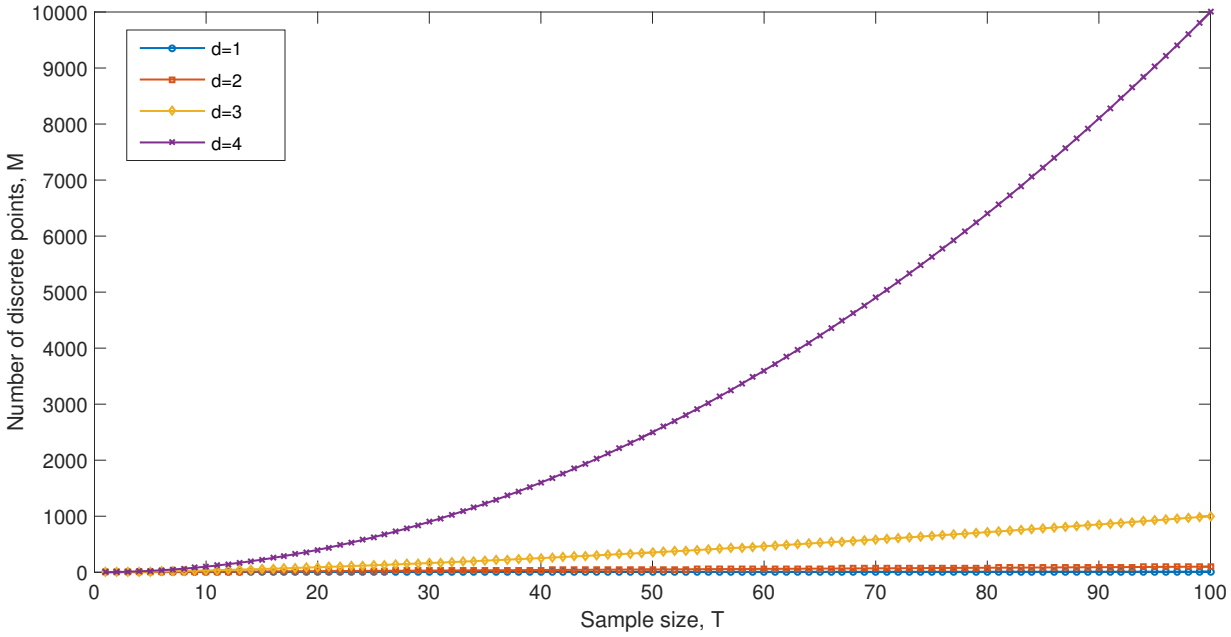

Figure 1: Rule of Thumb Choice for M

When the state follows a Gaussian VAR(1), a closed form expression for the ergodic distribution is available. Gospodinov and Lkhagvasuren (2014) provide a method to discretize Gaussian VAR(1)s that is robust to high levels of persistence. They use mixtures of Rouwenhorst (1995) approximations to match conditional moments as closely as possible. I rely on this method in section 8 for my empirical application. However, for more general time series models, no such expression exists.

Even when no expression for the unconditional distribution exists, it is often possible to compute the unconditional mean and standard deviation of the process. In this case, I recommend choosing a grid centered at the unconditional mean μ_x covering $\sqrt{M-1}$ unconditional standard deviations σ_x of the process on either side. That is, choose $\{x_{m,M}\}_{m=1}^{M}$ to be M evenly spaced points over the interval $\left[\mu_x - \sqrt{M-1}\sigma_x, \mu_x + \sqrt{M-1}\sigma_x\right]$. [10]

If the computation of unconditional moments is infeasible, I propose simulating a path of the state and discarding a fixed fraction from the beginning as burn in. If the simulated sample and burn in periods are sufficiently large, the remaining points can be treated as representative draws from the ergodic distribution. One can then estimate unconditional moments of the simulated process and use the method outlined above by replacing the population parameters μ_x and σ_x with their estimated counterparts. Alternatively, one can

[10] This is the way of constructing the grid employed in the Rouwenhorst (1995) approximation and suggested in Farmer and Toda (2016).

use empirical quantiles as the discretization points.

Consider the case when $r = 1$, that is, the state vector is one-dimensional. Suppose one simulates S points from the state equation with S_{bi} used as burn in. Denote this simulated path as $\{x_s\}_{s=1}^{S}$. Then, to construct a grid that covers the state with approximately $1-\alpha$ probability, select:

$$x_{m,M} = \hat{Q}_S\left(\tfrac{\alpha}{2} + \tfrac{m-1}{M}(1-\alpha)\right) \quad \text{for } m = 1, \ldots, M$$

where $\hat{Q}_S : (0,1) \to \mathbb{R}$ is the empirical quantile function of the sample $\{x_s\}_{s=S_{bi}}^{S}$, defined as

$$\hat{Q}_S(p) = \left\{ \inf x \in \mathbb{R} : p \leq \frac{1}{S - S_{bi}} \sum_{s=S_{bi}}^{S} \mathbb{1}\{x_s \leq x\} \right\}$$

Selecting the points in this way has the desirable property that roughly the same number of realizations of the state will fall between each pair of points.[11] By choosing α arbitrarily close to 1, it is possible to ensure that one has covered the ergodic set with any desired degree of confidence.[12] This method is also robust to skewness and fat tails in the stationary distribution.

While the simulation procedure outlined above is capable of handling very general models, it will introduce simulation error and increase the computational burden of the estimation. It is desirable to use prior knowledge of the particular model to help inform the choice of discretization whenever possible.[13]

5.3 Constructing the Transition Matrix

I recommend two ways of constructing the transition matrix for the discretization filter that are applicable to the widest range of economic models. However, there is no unique way to construct the transition matrix. [14]

First, I outline a way to extend the original method proposed by Tauchen (1986) to the nonlinear, non-Gaussian case. Create a partition of the state space $\{A_m\}_{m=1}^{M}$, where each

[11] There is no unique way to define quantile functions in the multivariate case. However, one simple way to achieve the same goal is to take the univariate empirical quantiles covering $1 - \frac{\alpha}{d}$ probability for each dimension.

[12] Of course a smaller α will require a larger number of data points for the same level of confidence in the approximation.

[13] Another possibility is to construct an ε-distinguishable set as proposed by Maliar and Maliar (2015), although this is subject to the same criticisms about introducing simulation.

[14] In addition to these two approaches, several others have been proposed in the literature: Tauchen and Hussey (1991), Rouwenhorst (1995), Adda and Cooper (2003), Flodén (2008), and Gospodinov and Lkhagvasuren (2014). However, all of these with the exception of Tauchen and Hussey (1991) only apply to linear autoregressive processes.

A_m is associated with discretization point $x_{m,M}$ for all $m = 1, \ldots, M$ (this is equivalent to intervals in the one-dimensional case). Then define:

$$P_{\theta,M}\left(m, m'\right) = \int_{A_{m'}} q_\theta\left(x \,|X_{t-1} = x_{m,M}\right) \mu\left(dx\right) \tag{20}$$

Intuitively, there are two layers of approximation in this expression. First, I am assuming that if X_{t-1} is in region A_m it is close to the point $x_{m,M}$ in the sense that the conditional distribution $q_\theta\left(X_t \,|X_{t-1}\right)$ can be well approximated by $q_\theta\left(X_t \,|X_{t-1} = x_{m,M}\right)$. Second, I am assuming that the probability of transitioning to region $A_{m'}$ from point $x_{m,M}$ is similar to the conditional density $q_\theta\left(X_t = x_{m',M} \,|X_{t-1} = x_{m,M}\right)$ over the set $A_{m'}$.

A limitation of this approach is the ability to evaluate the integrals needed to construct the transition matrix. In general, this method will only work well in practice when the A_m are hyperrectangles, and the transition density is easy to evaluate. Furthermore, there are no known results on the rate of weak convergence of the ergodic distribution of the approximate Markov chain to the that of the underlying continuous process. Since this rate is critical to obtaining asymptotic normality, researchers should be cautious about standard errors when using this approach with a small number of points.

Second, I construct the transition matrix as in Farmer and Toda (2016). They provide a general way of constructing finite-state Markov chain approximations to stochastic processes. Their method finds the discrete distribution which is "closest" to the original distribution from some prior distribution in terms of Kullback-Leibler distance, while matching a set of conditional moments of the underlying continuous distribution.

If the prior distribution is a valid quadrature formula for evaluating integrals with respsect to the original conditional density, the discrete approximation is guaranteed to converge weakly to the continuous distribution. Moreover, the rate of convergence is given by the rate of convergence of the selected quadrature formula.[15]

My Monte Carlo results in section 6 demonstrate that when the primary aim is estimation of the parameters, very coarse discretizations are adequate. This is in line with my theoretical results which show that the estimates are consistent independently of the rate at which M grows. The discretization filter has the potential to scale to higher dimensional problems by exploiting sparse grid quadrature methods (e.g. Smolyak grids), quasi-Monte Carlo methods, or the more recently proposed ε-distinguishable set method in Maliar and

[15] A special case of the discretization filter, known as the point mass filter, has been discussed at length in the computer science literature. The elements of the transition matrix are chosen to be proportional to the one-step-ahead density evaluated at the discretization points, i.e. $P_{\theta,M}\left(m, m'\right) \sim p\left(x_{m',M} \,|x_{m,M}\right)$. However, since the primary aim in the computer science literature is to filter the states, the grid is chosen to be very fine. Tensor grid product approximations quickly become intractable in higher dimensions, and for this reason the point-mass filter is infrequently used. A comprehensive survey article on the properties and applications of filtering techniques is Chen (2003).

Maliar (2015). I leave the investigation of this extension for future research.

6 Monte Carlo Evidence

In this section, I consider two simulation exercises, a linear measurement error model and a stochastic volatility model, to compare the performance of the discretization filter with exisiting alternatives.

6.1 Measuring GDP: A Linear State Space Example

I first consider a simple linear Gaussian state space model to illustrate the performance of the discretization filter in a case where the exact evaluation of the likelihood is possible using the Kalman Filter.

Aruoba et al. (2016) propose extracting a common component of the two widely available measures of GDP using a simple measurement error model in order to provide a more accurate estimate of "true" GDP. Let $\Delta\text{GDP}_{E,t}$ and $\Delta\text{GDP}_{I,t}$ denote the expenditure and income-side estimates of GDP growth respectively, and let ΔGDP_t denote true GDP growth, which is assumed to be unobserved. Consider the following state space model

$$\begin{bmatrix} \Delta\text{GDP}_{E,t} \\ \Delta\text{GDP}_{I,t} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Delta\text{GDP}_t + \begin{bmatrix} \epsilon_{E,t} \\ \epsilon_{I,t} \end{bmatrix}$$
$$\Delta\text{GDP}_t = \mu\left(1-\rho\right) + \rho\Delta\text{GDP}_{t-1} + \epsilon_{G,t}$$

where $\left(\epsilon_{G,t}, \epsilon_{E,t}, \epsilon_{I,t}\right)' \sim i.i.d.N\left(\mathbf{0}, \Sigma\right)$, with

$$\Sigma = \begin{bmatrix} \sigma_{G,G}^2 & 0 & 0 \\ 0 & \sigma_{E,E}^2 & \sigma_{E,I} \\ 0 & \sigma_{E,I} & \sigma_{I,I}^2 \end{bmatrix}$$

In their paper, Aruoba et al. (2016) also consider a more sophisticated specification of the model which allows for correlation between the measurement and state equation errors. The discretization filter can allow for this at the cost of introducing time-varying transition matrices but I omit the details for expositional simplicity. I focus on the restricted model outlined above.

I take the parameters estimated in the paper and simulate 500 samples of length $T = 204$, which is the amount of data used for estimation. For each sample, I evaluate the likelihood of the data using the Kalman filter (KF), the discretization filter (DF), and the bootstrap particle filter (PF). I examine the following two statistics for assessing the quality of the

likelihood approximation discussed in Herbst and Schorfheide (2015)

$$\hat{\Delta}_1 = \ln \hat{p}_\theta\left(\boldsymbol{Y}_1^T\right) - \ln p_\theta\left(\boldsymbol{Y}_1^T\right) \tag{21}$$

$$\hat{\Delta}_2 = \exp\left[\ln \hat{p}_\theta\left(\boldsymbol{Y}_1^T\right) - \ln p_\theta\left(\boldsymbol{Y}_1^T\right)\right] - 1 \tag{22}$$

where $\hat{p}_\theta\left(\boldsymbol{Y}_1^T\right)$ denotes the approximate likelihood computed with either the DF or the PF, and $p_\theta\left(\boldsymbol{Y}_1^T\right)$ denotes the true likelihood evaluated with the KF. Since the approximation to the likelihood provided by the PF is random, I use a 100 draws of the PF for every realization of the data. I consider several choices for the number of particles N used in the PF and for the proportionality constant used in the rule-of-thumb choice for the number of grid points M in the DF proposed in (19).

Table 1 presents the results of the simulation exercise for the accuracy of the likelihood approximations as measured by $\hat{\Delta}_1$ and $\hat{\Delta}_2$. An important distinction between the PF and the DF is that the PF approximation to the likelihood is random. It depends on the particular path that is simulated for the particles. However, the DF approximation to the likelihood is deterministic and thus has no associated sampling uncertainty for a given draw of the data.

For the PF, the bias and standard deviation of the approximations for a particular realization of the data are computed as the average value and standard deviation of the likelihood discrepancies across the 100 draws of the particles respectively. Since the DF is deterministic, there is only one value of the bias per sample realization and the standard deviation is zero. The RMSE is given by the familiar Bias^2 + Var formula. The means of these statistics are then computed as the means across randomly generated samples.

To be more precise, index a draw of the data by s and a draw of the particles by g. Define $\hat{\Delta}_{i,s,g}^{PF}$ as the value of discrepancy measure $\hat{\Delta}_i$ computed by the PF for sample s and particle draw g. Similarly, define $\hat{\Delta}_{i,s}^{DF}$ as the value of discrepancy measure $\hat{\Delta}_i$ computed by the DF for sample s. Then the PF statistics are computed as

$$\text{Mean Bias}\left(\hat{\Delta}_i^{PF}\right) = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{G}\sum_{g=1}^{G}\hat{\Delta}_{i,s,g}^{PF}\right] \tag{23}$$

$$\text{Mean Var}\left(\hat{\Delta}_i^{PF}\right) = \frac{1}{S}\sum_{s=1}^{S}\left[\hat{\Delta}_{i,s,g}^{PF} - \frac{1}{G}\sum_{g=1}^{G}\hat{\Delta}_{i,s,g}^{PF}\right]^2 \tag{24}$$

$$\text{Mean RMSE}\left(\hat{\Delta}_i^{PF}\right) = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\hat{\Delta}_{i,s,g}^{PF}\right)^2\right]^{1/2} \tag{25}$$

	Bootstrap Particle Filter					
Number of particles N	100	500	1,000	5,000	10,000	50,000
Mean Bias $\hat{\Delta}_1$	-2.484	-0.505	-0.257	-0.054	-0.029	-0.006
Mean StdD $\hat{\Delta}_1$	2.292	0.991	0.698	0.310	0.221	0.100
Mean RMSE $\hat{\Delta}_1$	3.394	1.121	0.750	0.318	0.224	0.101
Mean Bias $\hat{\Delta}_2$	-0.014	-0.003	-0.001	-0.001	-0.002	0.000
Mean StdD $\hat{\Delta}_2$	3.889	1.164	0.771	0.318	0.225	0.100
Mean RMSE $\hat{\Delta}_2$	3.933	1.171	0.775	0.320	0.226	0.101
	Discretization Filter					
Rule of thumb constant c	1/2	1	3	5	7	10
Mean Bias $\hat{\Delta}_1$	-0.405	-0.040	0.001	0.002	0.001	0.001
Mean StdD $\hat{\Delta}_1$	-	-	-	-	-	-
Mean RMSE $\hat{\Delta}_1$	1.287	0.383	0.114	0.070	0.053	0.042
Mean Bias $\hat{\Delta}_2$	0.085	0.029	0.007	0.004	0.003	0.002
Mean StdD $\hat{\Delta}_2$	-	-	-	-	-	-
Mean RMSE $\hat{\Delta}_2$	1.119	0.391	0.113	0.069	0.051	0.039

Table 1: Likelihood Discrepancies

For the DF, the mean bias and RMSE are given by

$$\text{Mean Bias}\left(\hat{\Delta}_i^{DF}\right) = \frac{1}{S}\sum_{s=1}^{S}\hat{\Delta}_{i,s}^{DF} \tag{26}$$

$$\text{Mean RMSE}\left(\hat{\Delta}_i^{DF}\right) = \frac{1}{S}\sum_{s=1}^{S}\left[\left(\hat{\Delta}_{i,s}^{DF}\right)^2\right]^{1/2} \tag{27}$$

and $\text{Mean Var}\left(\hat{\Delta}_i^{DF}\right) = 0$ for the reason explained above.

Table 2 reports the average absolute and relative evaluation times of the likelihood function across all specifications. The absolute times are reported in seconds. For the PF, these are computed as the average across samples and particle draws. For the DF and the KF, these are simply reported as averages across the samples. The relative times are computed as the time of one evaluation of the likelihood function relative to the time it takes for the KF.

Considered together, tables 1 and 2 provide a better understanding of the tradeoff between accuracy and computational complexity that both the DF and PF exhibit. As an example, note that the evaluation of the likelihood using 100 particles for the PF and a rule of thumb constant of 7 for the DF take roughly the same amount of time, about

	Kalman Filter					
Mean Time	0.007					
	Bootstrap Particle Filter					
Number of particles N	100	500	1,000	5,000	10,000	50,000
Mean Time	0.020	0.039	0.055	0.194	0.351	1.845
Mean Relative Time	3.13	6.13	8.66	30.30	54.84	288.63
	Discretization Filter					
Rule of thumb constant c	1/2	1	3	5	7	10
Mean Time	0.009	0.009	0.011	0.014	0.020	0.032
Mean Relative Time	1.38	1.37	1.69	2.16	3.10	4.99

Table 2: Computation Time of 1 Likelihood Evaluation (in seconds)

0.02 seconds. However, the DF is 2 orders of magnitude more accurate in terms of RMSE. Similarly, consider the PF with 50,000 particles and the DF with a rule of thumb constant of 3. These are roughly the same in terms of RMSE, but the DF evaluation of the likelihood is about 170 times faster. Examining the other elements of the tables leads to a similar conclusion: the DF offers a much better tradeoff between accuracy and computation time than the PF.

6.2 Stochastic Volatility

Next, I compare the performance of different estimation procedures on a stochastic volatility model. The standard discrete time stochastic volatility model, as formulated in Taylor (1982), is given by

$$X_t = \mu(1-\rho) + \rho X_{t-1} + v_t \qquad v_t \sim \text{i.i.d. } N\left(0, \sigma^2\right) \tag{28}$$

$$Y_t = e^{X_t/2} w_t \qquad w_t \sim \text{i.i.d. } N(0,1) \tag{29}$$

Note that the measurement equation can be equivalently rewritten as:

$$\log\left(Y_t^2\right) = X_t + \log\left(w_t^2\right) \tag{30}$$

which leads to an additively separable state equation.[16] However, this simplification only applies to the most basic versions of the stochastic volatility model. I focus on results from the parameterization $\mu = -8.940$, $\rho = 0.9890$, and $\sigma = 0.1150$, which are empirical

[16] This is the specification of the observation equation I use in the EKF estimation. This can also be thought of as a misspecified Kalman filter where the measurement error is incorrectly assumed to be Gaussian.

estimates of the parameters of the stochastic volatility model on daily returns data from the DAX in Hautsch and Ou (2008). The results are not sensitive to this parameterization.

I simulate data for T = 100, 500, and 1,000 periods, and compute the likelihood of the model eight different ways: the DF using six different choices of the rule of thumb constant c, the boostrap PF with adaptive resampling using 1,000 particles, and the extended Kalman filter (EKF). Each specification is simulated 1,000 times and estimation is performed via maximum likelihood where optimization is done using MATLAB's genetic algorithm in the global optimization toolbox. The random seed used to construct the particle filter approximation is fixed for a given sample in order to make the optimization better behaved.[17]

Figures 2, 3, and 4 display the sampling distributions of the maximum likelihood estimators. The rows of each figure correspond to a particular model parameter and the columns correspond to a particular method of approximating the likelihood. A vertical line is displayed at the point of the true parameter value. All estimation using the discretization filter uses the Rouwenhorst (1995) discretization scheme.[18]

Note that for small sample sizes, $T = 100$, there is a considerable downward bias in the estimation of ρ and σ. That is, the optimization algorithm is picking values of ρ and σ extremely close to 0. This bias is most severe in the EKF estimates, especially for σ. However, this is not particularly surprising because the EKF is estimating a misspecified model, where it is treating the residual in the observation equation as a normal random variable, even though it has a $\log\left(\chi_1^2\right)$ distribution.

This bias vanishes for both the DF and the PF in the larger sample simulations and the DF appears to produce tighter estimates of all 3 parameters, especially ρ. This is due, at least in part, to the fact that the accuracy of the Rouwenhorst approximation is independent of the persistence of the AR(1) process.

I also compute the root mean squared error (RMSE) and the bias of the parameter estimates, approximating the population expectation with an average across simulations. In particular, for the i-th component of the parameter vector, I compute:

$$\text{RMSE}\left(\hat{\theta}_i\right) = \sqrt{\mathbb{E}\left[\left(\hat{\theta}_i - \theta_i\right)^2\right]} \tag{31}$$

$$\text{Bias}\left(\hat{\theta}_i\right) = \mathbb{E}\left[\theta_i\right] - \theta_i \tag{32}$$

[17] Note that traditional gradient based optimization methods are inapplicable to the PF because the likelihood function is simulated, which makes it non-differentiable. See Flury and Shephard (2011) for a more detailed discussion.

[18] Estimation was also performed using the Farmer and Toda (2016) method, the Tauchen (1986) method, and the point-mass filter. The Rouwenhorst method performs the best although the relative gains of the discretization filter are similar across all discretization methods.

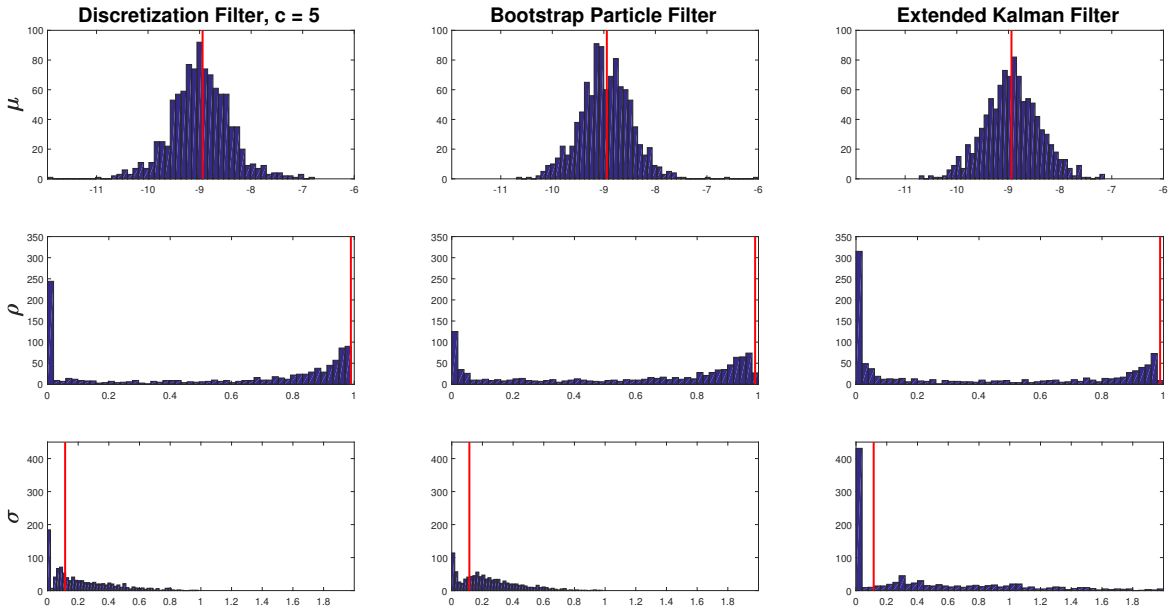

Figure 2: MLE sampling distributions for sample size $T = 100$

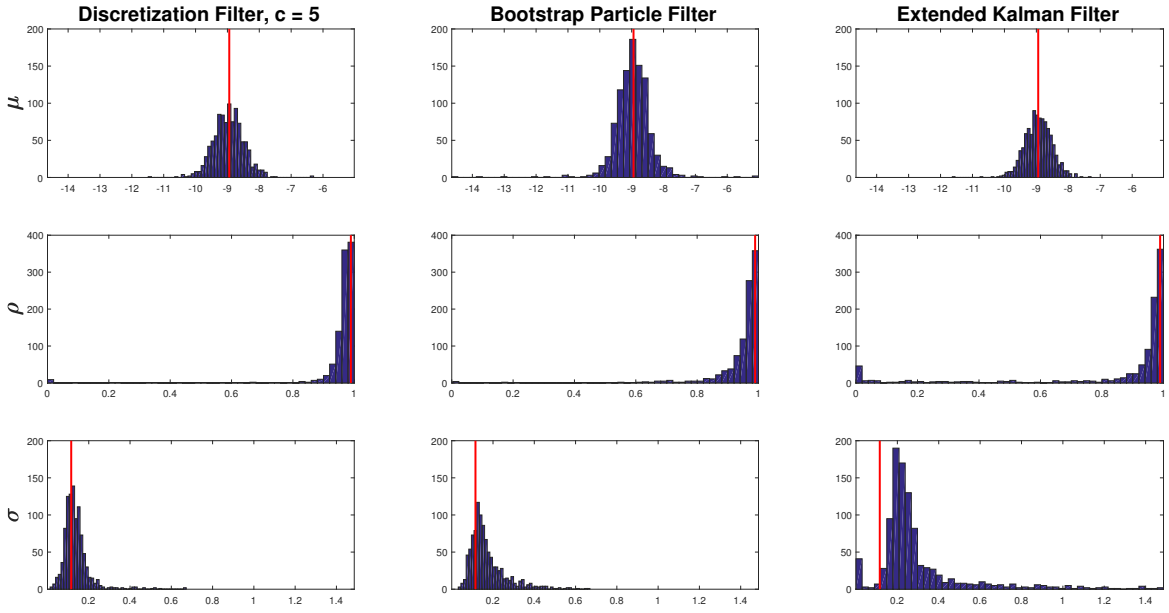

Figure 3: MLE sampling distributions for sample size $T = 500$

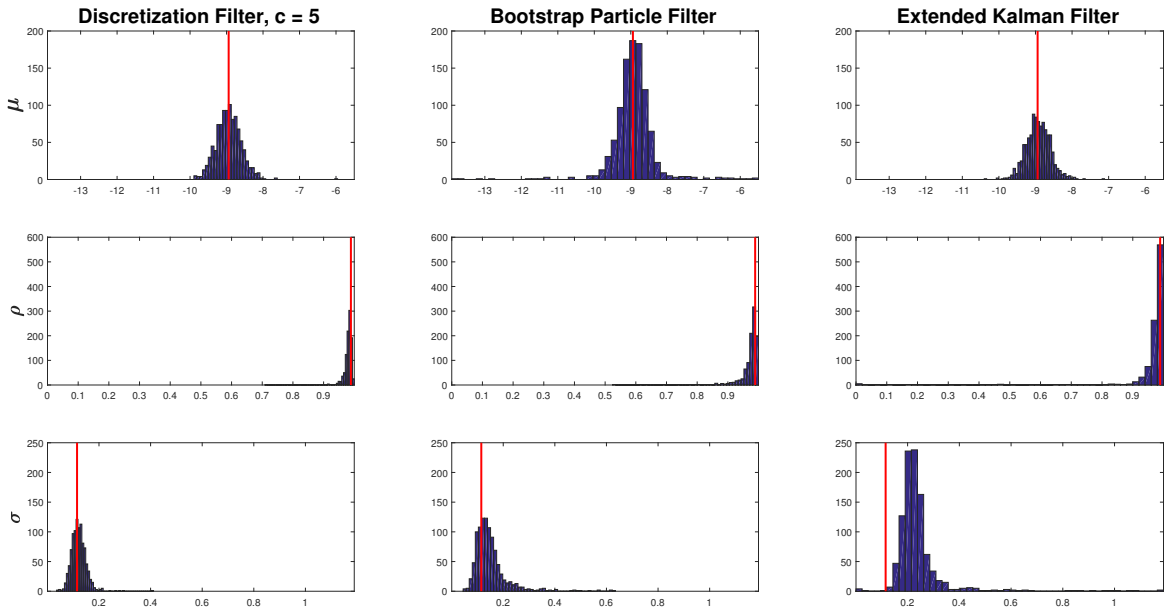

Figure 4: MLE sampling distributions for sample size $T = 1,000$

and report the results in table 3.

First consider the DF with $c = 5$ and its performance relative to the PF and the EKF. The DF and the PF are similar in terms of RMSE and bias for $T = 100$, however the DF generally outperforms the PF for the larger sample sizes. The EKF is unambiguously the worst except for estimation of the mean parameter μ. It is also interesting to note that the performance of the PF for estimating μ actually deteriorates for larger sample sizes, which seems to be evidence of sample thinning, a well known problem with importance sampling methods.

Next I examine the performance of the DF for different values of the rule of thumb constant c. For $T = 100$ and to a lesser extent for $T = 500$, the RMSE and bias actually seem to increase for larger values of c. There are a couple of possible explanations for this phenomenon. The first is that the asymptotic analysis in section 4 considers the case of a compact state space, whereas in this example as in most examples of economic interest, the state variable resides in an unbounded space. Thus, as the discretization is being constructed for larger values of c, the number of points is increasing, but so is the domain over which the approximation is constructed. This could potentially cause numerical issues for smaller sample sizes, because the discretization points cover large areas of the state space which are never visited in the sample.

A second possibility is that these larger numbers are actually more consistent with

the RMSE and bias of the infeasible maximum likelihood estimator. In other words, the misspecification caused by small values of M is actually acting as a type of regularization which is outperforming the maximum likelihood estimator for small sample sizes. Note that this phenomenon is absent for larger sample sizes, and the estimates of the RMSE and bias appear stable across all values of c.

Table 4 displays the average simulation times for all eight specifications. The differences in computational time are stark. With $c = 1$, the EKF is 32 times faster than the DF for small sample sizes and 78 times faster for large ones. However, this is at the cost of parameter estimates which are significantly less accurate for larger sample sizes. Furthermore, the EKF estimate of σ appears to be significantly biased, even asymptotically, due to the misspecification of the observation equation.

For estimates which are roughly the same accuracy for $T = 100$, the DF is an order of magnitude faster than the PF. For $T = 1,000$, the DF is between twice and three times as accurate as the particle filter while being 2 orders of magnitude faster. These results suggest that the DF is somewhere in between the EKF and the PF in terms of computational burden, while delivering accurate parameter estimates. To give the reader a rough idea, all of the simulations for the DF and the EKF ran in a matter of minutes to hours whereas the most computationally burdensome PF specification ($T = 1,000$) took almost five days to run operating in parallel on four cores. These reductions in computation time make the estimation of many dynamic macroeconomic and financial models feasible. In the case of the Gabaix (2012) rare disasters model, the estimation takes several hours running MATLAB on a standard desktop computer, whereas estimation using a PF would likely take several weeks.

Another important dimension for comparison is the accuracy of the filtered states, $\{\hat{x}_{t|t}\}_{t=1}^{T}$. I provide results on the root mean square error (RMSE) and the mean absolute error (MAE) of all the methods. For a given model specification and method, these are defined as:

$$\text{RMSE} = \left(\frac{1}{T} \sum_{t=1}^{T} \left(\hat{x}_{t|t} - x_t \right)^2 \right)^{1/2} \tag{33}$$

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} \left| \hat{x}_{t|t} - x_t \right| \tag{34}$$

I define the average RMSE (ARMSE) and average MAE (AMAE) to be the average of the RMSE and the MAE across simulations for a given method. Table 5 displays the ARMSE and AMAE of each method, where the filtering is done using the corresponding maximum likelihood estimates of the parameters for a given sample.

Root Mean Squared Error

		Discretization Filter						PF	EKF
ROT constant c		1/2	1	3	5	7	10	-	-
μ	**T = 100**	0.511	0.538	0.611	0.618	0.623	0.669	0.488	0.521
	T = 500	0.450	0.475	0.511	0.516	0.486	0.508	0.574	0.445
	T = 1,000	0.343	0.364	0.370	0.381	0.339	0.391	0.614	0.336
ρ	**T = 100**	0.598	0.584	0.603	0.617	0.630	0.637	0.572	0.727
	T = 500	0.108	0.080	0.101	0.103	0.121	0.134	0.108	0.304
	T = 1,000	0.014	0.014	0.015	0.015	0.014	0.015	0.042	0.126
σ	**T = 100**	0.228	0.225	0.236	0.238	0.246	0.251	0.244	0.621
	T = 500	0.061	0.057	0.061	0.064	0.072	0.072	0.111	0.293
	T = 1,000	0.027	0.027	0.027	0.027	0.027	0.027	0.076	0.163

Bias

		Discretization Filter						PF	EKF
ROT constant c		1/2	1	3	5	7	10	-	-
μ	**T = 100**	-0.036	-0.039	-0.035	-0.029	-0.027	-0.031	-0.028	0.005
	T = 500	-0.031	-0.027	-0.017	-0.015	-0.012	-0.018	0.003	-0.001
	T = 1,000	-0.008	0.015	0.007	0.015	0.015	0.026	0.016	0.020
ρ	**T = 100**	-0.441	-0.427	-0.455	-0.475	-0.493	-0.504	-0.442	-0.617
	T = 500	-0.030	-0.030	-0.036	-0.035	-0.038	-0.041	-0.048	-0.136
	T = 1,000	-0.008	-0.009	-0.009	-0.009	-0.009	-0.009	-0.017	-0.034
σ	**T = 100**	0.111	0.105	0.108	0.106	0.111	0.113	0.134	0.340
	T = 500	0.020	0.019	0.022	0.021	0.022	0.023	0.062	0.186
	T = 1,000	0.006	0.006	0.006	0.006	0.006	0.006	0.035	0.125

Table 3: Accuracy of Parameter Estimates

	Discretization Filter						PF	EKF
ROT constant c	1/2	1	3	5	7	10	-	-
T = 100	0.001	0.001	0.002	0.003	0.005	0.009	0.010	0.000
T = 500	0.002	0.003	0.008	0.017	0.034	0.083	0.120	0.000
T = 1,000	0.005	0.006	0.018	0.046	0.101	0.273	0.401	0.000

Table 4: Computation Time of 1 Likelihood Evaluation, in seconds

Average Root Mean Squared Error

	Discretization Filter						BPF	EKF
ROT constant c	1/2	1	3	5	7	10	-	-
T = 100	0.362	0.360	0.362	0.365	0.368	0.370	0.374	0.498
T = 500	0.378	0.379	0.385	0.388	0.390	0.391	0.383	0.465
T = 1,000	0.379	0.381	0.385	0.386	0.386	0.387	0.383	0.452

Average Absolute Mean Error

	Discretization Filter						BPF	EKF
ROT constant c	1/2	1	3	5	7	10	-	-
T = 100	0.297	0.294	0.295	0.297	0.299	0.300	0.307	0.390
T = 500	0.302	0.302	0.304	0.305	0.306	0.306	0.306	0.372
T = 1,000	0.302	0.303	0.304	0.304	0.304	0.304	0.305	0.361

Table 5: Accuracy of Filtered State Estimates

The DF and PF perform roughly the same for all sample sizes. However, keep in mind that this is for dramatically different estimation times for the parameters as discussed above. The misspecification of the measurement error distribution using the EKF translates into poor estimates of the unobserved state.

7 Variable Rare Disasters

In this section, I provide the first estimates of the Gabaix (2012) model of variable rare disasters. I show how likelihood-based estimation can be used as a model diagnosis tool. In particular, I find that (i) the estimated model fails to identify the Great Recession as a disaster episode, and (ii) the model cannot capture the change in the dynamics of the price-dividend ratio starting in the 1990s. To explain (i), the model requires a positive expected jump in inflation in the event of a disaster in order to generate an upward sloping nominal yield curve. However, during the Great Recssion, we observed a strongly upward sloping nominal yield curve in conjunction with close to zero inflation and even deflation. For (ii), the model specifies a process which is close to an AR(1) which governs the dynamics of the price-dividend ratio, while the price-dividend ratio starting the 1990s appears to exhibit a structural break both in its mean and its dynamics.

7.1 Model Setup

The model is an endowment economy where a representative agent has lifetime expected utility over consumption given by:

$$\mathbb{E}_0\left[\sum_{t=0}^{\infty} e^{-\rho t}\frac{C_t^{1-\gamma}}{1-\gamma}\right]$$

$\gamma > 0$ is the coefficient of relative risk aversion, and $\rho > 0$ is rate of time preference. Each period she receives consumption endowment C_t. For expositional purposes, I only present the version of the model with CRRA utility here, however for estimation purposes I consider the full Epstein-Zin version of the model which allows risk aversion and the IES to be independently estimated.

The endowment stream is hit by large but infrequent disasters. The dynamics of consumption are given by:

$$\Delta c_{t+1} = g_C + w_{t+1}b_{t+1} \tag{35}$$

where g_C is the normal-time growth rate of the economy, $B_{t+1}e^{g_C}$ is the growth rate if a disaster occurs ($b_{t+1} := \log B_{t+1}$), and w_{t+1} is an indicator for whether a disaster occurs at time $t+1$, which happens with probability p_t.

Consider a stock i which is a claim to a stream of dividend payments $(D_t)_{t\geq 0}$. The growth rate of its dividends is assumed to follow

$$\Delta d_{t+1} = g_D + \varepsilon_{t+1}^D + w_{t+1}f_{t+1} \tag{36}$$

where g_D is the growth rate of dividends in normal times, ε_{t+1}^D is a mean zero shock that is independent of the disaster event, and F_{t+1} ($f_{t+1} := \log F_{t+1}$) is the recovery rate of the dividend. That is, in the event of a disaster, there can be "partial default." If $F_{t+1} = 0$, the asset is completely destroyed, and if $F_{t+1} = 1$ there is no loss relative to normal times.

In contrast with some of the other more recent papers on variable rare disasters such as Wachter (2013) and Gourio (2012), the probability of a disaster is fixed in the baseline model. It is the severity of a disaster which is time-varying. The combination of variations in the disaster probability and the severity are captured by a variable called "resilience." Define resilience H_t of the asset as

$$H_t \equiv p_t\mathbb{E}_t^D\left[B_{t+1}^{-\gamma}F_{t+1} - 1\right] \tag{37}$$

Assets with high resilience are safer than assets with low resilience because they pay out more in disaster states, and thus will command lower risk premia.

As in Gabaix (2012), I split resilience into a constant part H_* and a variable part $\hat{H}_t$

with mean zero. The dynamics of $\hat{H}_t$ are assumed to follow a linearity-generating process (Gabaix 2009)

$$\hat{H}_{t+1} = \frac{1+H_*}{1+H_t} e^{-\phi_H} \hat{H}_t + \varepsilon^H_{t+1} \tag{38}$$

Linearity generating processes behave like first-order autoregressive processes close to their steady state but display nonlinear dynamics as they reach more extreme values.

Define $\delta \equiv \rho + \gamma g_C$, $h_* \equiv \log(1+H_*)$, and $\delta_i \equiv \delta - g_D - h_*$. It can be shown that the price-dividend ratio of the asset is given by

$$\frac{P_t}{D_t} = \frac{1}{1-e^{-\delta_i}} \left(1 + \frac{e^{-\delta_i - h_*} \hat{H}_t}{1 - e^{-\delta_i - \phi_H}} \right) \tag{39}$$

The unconditional equity premium for the asset is given by

$$r^e_t = \delta - H_t - p_t \mathbb{E}_t \left[1 - F_{t+1} \right] - r_f \tag{40}$$

where r_f, the risk-free rate, is given by

$$r_f = \delta - p_t \mathbb{E}_t \left[B^{-\gamma}_{t+1} - 1 \right] \tag{41}$$

Turning to the nominal side of the economy, inflation $I_t = I_* + \hat{I}_t$ is assumed to vary exogenously and its non-constant component $\hat{I}_t$ also follows a linearity-generating process. In addition, inflation jumps by an amount $J_t = J_* + \hat{J}_t$ in the event of a disaster. J_* is the baseline jump in inflation in the event of a disaster, and $\hat{J}_t$ is a mean-reverting deviation in this jump size from its baseline. Their dynamics are jointly given by

$$\hat{I}_{t+1} = \frac{1-I_*}{1-I_t} \left(e^{-\phi_I} \hat{I}_t + w_{t+1} J_t \right) + \varepsilon^I_{t+1} \tag{42}$$

$$\hat{J}_{t+1} = \frac{1-I_*}{1-I_t} e^{-\phi_J} \hat{J}_t + \varepsilon^J_{t+1} \tag{43}$$

where ε^I_{t+1} and ε^J_{t+1} are mean zero shocks which are uncorrelated with disasters, but may be correlated with each other. This allows me to define the variable π_t, the variable part of the bond premium, as

$$\pi_t \equiv \frac{p_t \mathbb{E}_t \left[B^{-\gamma}_{t+1} F_{\$,t+1} \right]}{1+H_\$} \hat{J}_t$$

π_t is what controls deviations of the slope of the nominal yield curve from its typical value, while inflation controls the level relative to the real yield.

Define $\Psi \equiv e^{-\delta}(1+H_\$)(1-I_*)$, $\tilde{\rho}_I \equiv \frac{e^{-\phi_I}+\kappa}{1-\kappa}$, and $\tilde{\rho}_J \equiv \frac{e^{-\phi_J}}{1-\kappa}$. The price of a nominal

zero-coupon bond of maturity T at time t is given by

$$Z_{\$t}(T) = (\Psi(1-\kappa))^T \times \left\{ 1 - \frac{1}{1-\kappa} \frac{1-\tilde{\rho}_I^T}{1-\tilde{\rho}_I} \left(\frac{\hat{I}_t}{1-I_*} - \kappa \right) - \frac{1}{(1-\kappa)^2} \frac{\frac{1-\tilde{\rho}_I^T}{1-\tilde{\rho}_I} - \frac{1-\tilde{\rho}_J^T}{1-\tilde{\rho}_J}}{\tilde{\rho}_I - \tilde{\rho}_J} \frac{\pi_t}{1-I_*} \right\}$$

The corresponding yield is

$$y_t(T) = -\frac{\ln Z_{\$t}(T)}{T}. \tag{44}$$

I now turn to the details of the estimation.

7.2 Estimation

I fix a subset of parameters related to the cash flow dynamics, the severity of disasters, and inflation in Table 6.

Parameters	**Values**
Growth rate of consumption and dividends	$g = g_C = g_D = 2.5\%$
Volatility of dividend growth	$\sigma_D = 11\%$
Recovery rate of C after a disaster	$\overline{B} = 0.66$
Stock's recovery rate: typical value	$F_{i*} = \overline{B} = 0.66$
Inflation: typical value	$I^* = 3.8\%$

Table 6: Calibrated Parameters

The means of consumption and dividend growth, the volatility of dividend growth, the recovery rate of consumption after a disaster, and the typical value of the stock's recovery rate are fixed to the values used in Gabaix (2012). I set the typical value of inflation to be 3.8%, which is the sample average of CPI inflation in my sample.

For my baseline estimation results, I use monthly data on the price-dividend ratio of the CRSP value-weighted portfolio, nominal yields on U.S. Treasury securities, and CPI inflation from June 1961 to December 2015. This is the longest sample for which all variables are available. The data on nominal yields are constructed as in Gürkaynak et al. (2007) and I use maturities of 3 and 6 months, 1, 2, 5, 7, and 10 years. Inflation is constructed as the 12-month change in log CPI.

The model has three state variables: resilience $\hat{H}_t$, inflation $\hat{I}_t$, and jumps in inflation $\hat{J}_t$. The mapping from resilience to the price-dividend ratio is given by (39) and the mapping from inflation and jumps in inflation to nominal yields is given by (44). I assume that the price-dividend ratio, nominal yields of all maturities, and inflation itself are observed with error with measurement errors given by $\varepsilon_{PD_{obs}} \sim N(0, \sigma_{PD_{obs}})$, $\varepsilon_{y_{obs}} \sim N(0, \sigma_{y_{obs}})$, and $\varepsilon_{I_{obs}} \sim N(0, \sigma_{I_{obs}})$ respectively. The measurement errors are assumed to be independent of each other and all other quantities in the model.

I estimate the vector of 13 parameters (10 structural and 3 measurement error variances)

$$\theta \equiv (\rho, \gamma, \psi, p, \phi_H, \sigma_I, \phi_I, J_*, \sigma_J, \phi_J, \sigma_{PD_{obs}}, \sigma_{y_{obs}}, \sigma_{I_{obs}})'$$

by maximum likelihood using the Farmer and Toda (2016) method with an 11 point grid for each state variable. This results in a total of $11^3 = 1,331$ discrete points. Table 7 shows the estimated parameters, with quasi maximum likelihood robust standard errors in parentheses, along with the calibrated values used in Gabaix (2012).

First, consider the values of the preference parameters ρ, γ, ψ. The estimated value of the rate of time preference ρ, 3.07%, is significantly lower than its calibrated value of 6.57%. In terms of annual discount factors, this translates into the difference between 0.970 and 0.936. Next, the estimated coefficient of relative risk aversion γ, 2.8, is significantly lower than its calibrated value of 4. This is heartening because traditionally asset pricing models require what are often considered unreasonably high values of risk aversion in order to match financial data. This number is more in line with typical macroeconomic calibrations of DSGE models. Lastly, the IES ψ is estimated to be 0.26, and importantly, is significantly less than 1. This is consistent with the empirical micro evidence, but at odds with values that are typically chosen in asset pricing models.

Second, the probability of a disaster is estimated to be 4.81% annually, compared to the calibrated value of 3.63% which comes from Barro and Ursúa (2008). Given that there are a very few observations of consumption disasters in the data, it seems reasonable to think that this probability may be higher than existing empirical estimates that rely on macro data.

Lastly, the estimates of the parameters governing the dynamics of inflation and jumps in inflation differ between the estimated and calibrated models. The estimated model favors more peristent and less volatile processes for both of these quantities.

I next consider some additional quantities implied by the model evaluated using both the estimated parameter values and the calibrated ones. The results are presented in table 8. A key quantity of interest is the risk-adjusted probability of a disaster, given by $p\mathbb{E}\left[B_{t+1}^{-\gamma}\right]$. This is the quantity that allows the model to match high average risk premia. The estimated

Parameters	**Estimated Values**	**Gabaix Calibration**
Time preference, ρ	3.07% (1.59%)	6.57%
Risk aversion, γ	2.812 (0.487)	4
Intertemporal elasticity of substitution, ψ	0.257 (0.101)	0.25
Probability of a disaster, p	4.81% (0.72%)	3.63%
Resilience:		
volatility, σ_F	7.0%	10%
speed of mean reversion, ϕ_H	12.48% (14.41%)	13%
Inflation:		
conditional volatility, σ_I	0.61% (3.08%)	1.5%
speed of mean reversion, ϕ_I	15.21% (5.14%)	18%
Jump in inflation:		
typical value, J_*	2.56% (0.53%)	2.1%
conditional volatility, σ_J	6.83% (6.71%)	15%
speed of mean reversion, ϕ_J	82.13% (95.33%)	92%
Volatility of measurement errors:		
price-dividend ratio, $\sigma_{PD_{obs}}$	2.62 (5.76)	-
nominal yields, $\sigma_{y_{obs}}$	0.43% (0.54%)	-
inflation, $\sigma_{I_{obs}}$	2.66% (7.14%)	-

Table 7: Model Parameters

model implies a value of 15.5%, less than 4 percentage points lower than the calibrated value of 19.2%. Given that $B_{t+1} = \overline{B}$ is fixed across both specifications, the differences in this quantity are coming from differences in the probability of a disaster and the coefficient of relative risk aversion. The higher probability of a disaster and lower value of risk aversion estimated by maximum likelihood allow the model to remain broadly consistent with a wide variety of asset pricing facts.

In particular, the model still achieves an unconditional equity premium of 3.6%, roughly half of what it is in the data, while the calibration produces 5.3%. The estimated real short-term rate is a bit high at 2.1% relative to the calibration which targets 1%, although this is consistent with the historical average of data back to 1891 is considered. The estimated model matchces the average level and volatility of the 5-year slope of the nominal yield curve produced by the calibration.

Parameters	**Estimated Values**	**Gabaix Calibration**
Ramsey discount rate, δ	12.8%	16.6%
Risk-adjusted probability of disaster, $p\mathbb{E}\left[B_{t+1}^{-\gamma}\right]$	15.5%	19.2%
Stocks:		
effective discount rate, δ_i	1.7%	5%
Stock resilience:		
typical value, H_*	8.6%	9.0%
volatility, σ_H	1.1%	1.9%
Stocks, equity premium:		
conditional on no disasters	5.3%	6.5%
unconditional	3.6%	5.3%
Real short-term rate	2.1%	1.0%
Resilience of one nominal dollar, $H_\$$	10.7%	16.0%
5-year nominal slope $y_t(5) - y_t(1)$:		
mean	0.55%	0.57%
volatility	0.81%	0.92%
Long-run − short-run yield:		
typical value, κ	3.5%	2.6%
Inflation:		
I_{**}	7.3%	6.3%
ψ_I	8.2%	13%
ψ_J	78.6%	90%
Bond risk premium:		
volatility, σ_π	0.95%	2.9%

Table 8: Implied Parameters

7.3 Implications of the Filtered State Estimates

I now focus on two implications of the model which come from the ability to examine the filtered and smoothed states implied by the estimated parameters. First, from the processes for inflation and jumps in inflation, I can back out the implied probability of a disaster having occured in any given period. Note that in the event of a disaster, the

conditional mean of inflation at time $t+1$ given information up to time t is

$$\frac{1-I_*}{1-I_t}\left(e^{-\phi_I}\hat{I}_t + J_t\right)$$

whereas in the event of no disaster, this conditional mean is

$$\frac{1-I_*}{1-I_t}e^{-\phi_I}\hat{I}_t$$

By running the discretization filter at the estimated parameter vector, I can obtain filtered and smoothed estimates of the time series $\left\{\hat{I}_t\right\}_{t=1}^{T}$ and $\{J_t\}_{t=1}^{T}$. Since the innovation to inflation each period, ε_{t+1}^{I}, has a normal distribution with standard deviation σ_I, I can compute the likelihood of having observed the value of $\hat{I}_{t+1}$ implied by these estimates in the event of a disaster and in the event of no disaster. Figure 5 plots the probability of a disaster having occurred in each period of the sample by applying this procedure using both the filtered and smoothed estimates of the states. The results using the filtered states are in blue and the results using the smoothed states are in red.

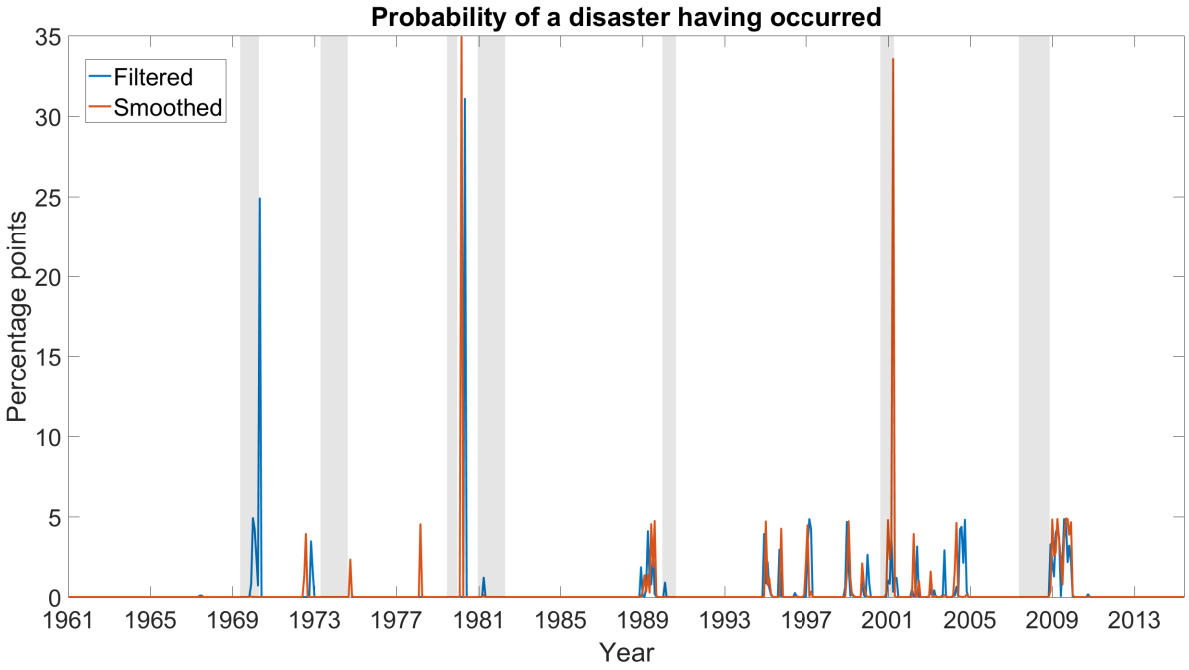

Figure 5: Disaster Probability

First, note that the filtered and smoothed estimates together identify three potential disaster episodes over the sample. I will refer to a "potential disaster episode" as a period where the estimated model assigns more than a 20% chance to a disaster having occured.

The filtered estimates identify the early 1970s and the late 1970s as potential disaster episodes. The smoothed estimates identify the same period in the late 1970s and a then a period in the early 2000s coinciding with the dot com bubble as potential disaster episodes.

While the two series do not fully agree on which periods are potential disaster episodes, they both come to the same conclusion regarding the Great Recession. At no point during the Great Recession does either series assign more than a 5% chance of a disaster having occurred. While this may seem surprising at first, upon further investigation it makes a lot of sense.

During the Great Recession, the U.S. experienced low inflation relative to the rest of the sample, and even a period of deflation. However, at the same time, the nominal yield curve was upward sloping. The Gabaix model achieves an upward sloping nominal yield curve through an expected positive jump in inflation. Since the model is being fit to both inflation data and data on nominal yields, it is trying to reconcile a period of expected low inflation / deflation with a period of upward sloping nominal yield curves but ends up splitting the difference. This suggests a shortcoming of the Gabaix framework, which is that rare disasters are typically coupled with expected increases in inflation. However, in the U.S. and many other developed countries, financial crises are typically coupled with deflation and upward sloping nominal yield curves.

Next, I examine the model's implications for the recovery rate of stocks, F_t. Recall that F_t is the fraction of its value that a stock retains in the event of a disaster. The recovery rate is an affine function of the state variable resilience $\hat{H}_t$, for which I construct filtered and smoothed estimates and plot in figure 6. As above, the results using the filtered states are in blue and the results using the smoothed states are in red. The black line is the long run average of the recovery rate, which is calibrated to be 66% as in Gabaix (2012).

What immediately jumps out is that before the late 1990s, the recovery rate is estimated to be about 20% on average, persistently low relative to its long run average of 66%. This shoots up to almost 100% during the dot com bubble and crashes back down to its long run average in the mid 2000s. It again experiences a sharp decline during the Great Recession and bounces back close to its long run average at the end of the sample. This is counterintuitive because it suggests that the model considers the Great Moderation to be particularly risky relative to the rest of the sample. On average, investors were expecting to lose about 80% of the value of their assets in the event of a disaster whereas the model implies that they should typically expect to lose 34%.

This result begins to make a lot more sense when one examines the connection between the recovery rate and the price-dividend ratio. In the Gabaix model, the price-dividend ratio of a stock is an affine function of its recovery rate. Unsurprisingly, movements in the recovery rate are closely linked to movements in the price-divdend ratio, with the only

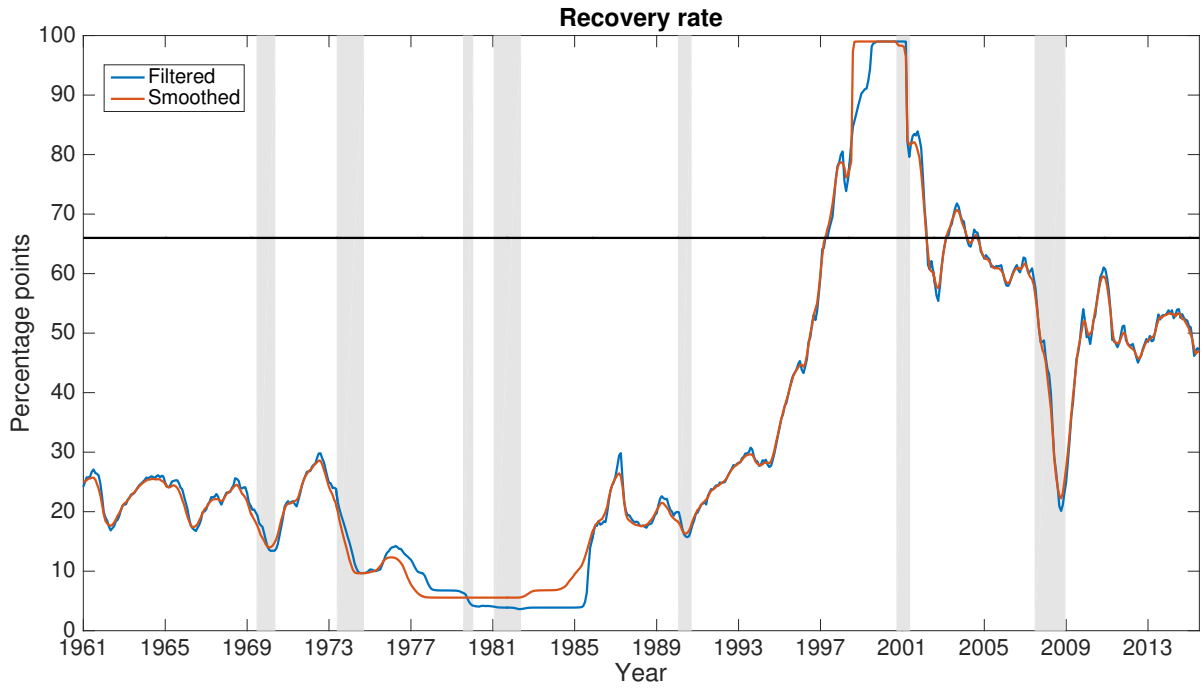

Figure 6: Recovery Rate

differences being attributed to measurement error.

The differences in the implied moments of the price-dividend ratio and stock returns are presented in table 9. The first thing that stands out from looking at this table is that the level of the price-dividend ratio implied by the estimated model parameters is about three times larger than the value implied by the calibrated model. This has a lot to do with the sample used in the estimation. The historical average of the price-dividend ratio targeted by Gabaix, 23, is computed using data that ends in 1997. However, the data I use for estimation goes all the way up to 2015, which includes the dot-com bubble and subsequent Great Recession.

	Data (Campbell Sample 1891-1997)	**Data (Estimation Sample 1961-2015)**	**Estimated Model**	**Calibrated Model**
Mean P/D	23	39.2	57.6	18.2
Std. dev. $\ln P/D$	0.33	0.40	0.21	0.30
Std. dev. of stock returns	0.18	0.15	0.11	0.15

Table 9: Stock Market Moments

The price-dividend ratio reaches a maximum value of 92 in the early 2000s and has an average value of 39 over my sample, almost twice the value targeted by Gabaix. The estimation chooses values of the structural parameters that allow the model to achieve these high values of the price-dividend ratio. The estimated model also understates the volatility of the price-dividend ratio. Unsurprisingly, given the lower volatility and higher mean of the price-dividend ratio, this results in a lower volatility of stock returns than the calibrated model, 11% vs. 15%, for the same values of the cash flow parameters.

Given the pronounced change in the both the level of the price-dividend ratio and its dynamics (sharper decreases and increases) after 1997, the fit of the Gabaix model may be greatly improved by allowing for switches in the parameters governing the long-run average, persistence, and volatility of the recovery rate. This would help produce more sensible economic estimates of the recovery rate.

The estimation of both the probability of a disaster having occurred and the recovery rate is an exercise which can only be conducted using likelihood-based estimation procedures. This highlights an advantage of likelihood-based methods over calibration and other moment-matching based methods: the ability to construct estimates of the hidden state variables. By constructing estimates of the hidden state variables, one is able to consider the model's implications for dynamics in addition to moments. The calibrated version of the Gabaix model does an excellent job of matching several moments of asset pricing data related to equities, bonds, and options. However, the estimation shows that the model also exhibits a couple of important shortcomings regarding the coupling of rare disasters with positive expected inflation and economically counterintuitive implications for the recovery rate of equity.

7.4 Model Comparison

Finally, I formally test the null hypothesis that the estimated model provides a better fit to the data than the calibrated model. To do this, I fix the parameters as calibrated in Gabaix (2012) and estimate the measurement errors using the same data as the full estimation outlined previously. Denote the resulting parameter vector as θ_0. I test the null hypothesis that $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \hat{\theta}$ using a likelihood ratio test. The likelihood ratio statistic is given by

$$LR = 2\left[\ell_{T,M}\left(\hat{\theta}\right) - \ell_{T,M}\left(\theta_0\right)\right] = 7,471$$

This is compared to a $\chi^2\left(10\right)$ because the unrestricted model has 10 extra parameters that are freely estimated. The 99% critical value for the test is 23.21 and thus the null hypothesis is overwhelmingly rejected in favor of the alternative. Again, this is unsurprising given the

calibrated model's inability to match the extreme values of the price-dividend ratio observed in the 2000s. The only way the calibrated model can rationalize these observations is by choosing unreasonably large values of the measurement error variance for the price-dividend ratio.

8 A Term Structure Model with a Zero Lower Bound

In this section, I re-estimate the term structure model proposed in Wu and Xia (2016), and provide an updated estimate of their shadow rate series with data through January 2014. Using the discretization filter I am able to replicate most of their parameter estimates. While my filtered series and the Wu and Xia estimates match closely over most of the sample, they diverge after the onset of the zero lower bound in January 2009. My estimates indicate that the shadow rate was roughly 2.2 percentage points lower in July 2012 than the Wu and Xia estimates would indicate. I conjecture that the estimates differ because the DF provides a more accurate approximation than the EKF to nonlinearities in the state space when the zero lower bound is in effect. Furthermore, the EKF estimator is in general not consistent, while the DF estimator is.

I omit details of the derivation of their shadow rate model. What is key for my purposes is that under the presence of a zero lower bound on short term interest rates, they are able to derive an approximate nonlinear state space model characterizing movements of the yield curve:

$$X_t = \mu + \rho X_{t-1} + \Sigma v_t \qquad v_t \sim \text{i.i.d. } N(0, I_3) \tag{45}$$

$$Y_{n,n+1,t} = \underline{r} + \sigma_n^{\mathbb{Q}} g\left(\frac{a_n + b_n' X_t - \underline{r}}{\sigma_n^{\mathbb{Q}}}\right) + w_t \qquad w_t \sim \text{i.i.d. } N(0, \omega) \tag{46}$$

where $Y_{n,n+1,t}$ corresponds to the one-period forward rate at time t for a loan starting at $t+n$ and maturing at $t+n+1$, and X_t is a (3×1) vector of latent factors which explain movements in the yield curve. For a derivation of the expressions for a_n, b_n, and $\sigma_n^{\mathbb{Q}}$, I refer the reader to Wu and Xia (2016).

Using their data on the 3 and 6 month, 1, 2, 5, 7, and 10 year forward rates, one has 7 observation equations, one for each observed yield maturity. They further assume that each forward rate is observed with normally distributed measurement error with the same variance ω. θ is a (22×1) vector of structural parameters.[19]

Table 10 reports maximum likelihood estimates of the parameters from the model with

[19] It has been pointed out that the results may be sensitive to the arbitrary choice of $\underline{r} = 0.25$, see Bauer and Rudebusch (2016). As a robustness check I also estimate $\underline{r}$ as a free parameter and find that it has little effect on subsequent analysis.

QMLE robust standard errors[20], using the DF with 9 discretization points along each dimension (i.e. $9^3 = 729$ total discretization points). I use the Gospodinov and Lkhagvasuren (2014) method to discretize the VAR state dynamics, which generalizes the method of Rouwenhorst (1995) to VAR(1) systems. I also include the estimates from Wu and Xia (2016) for comparison.

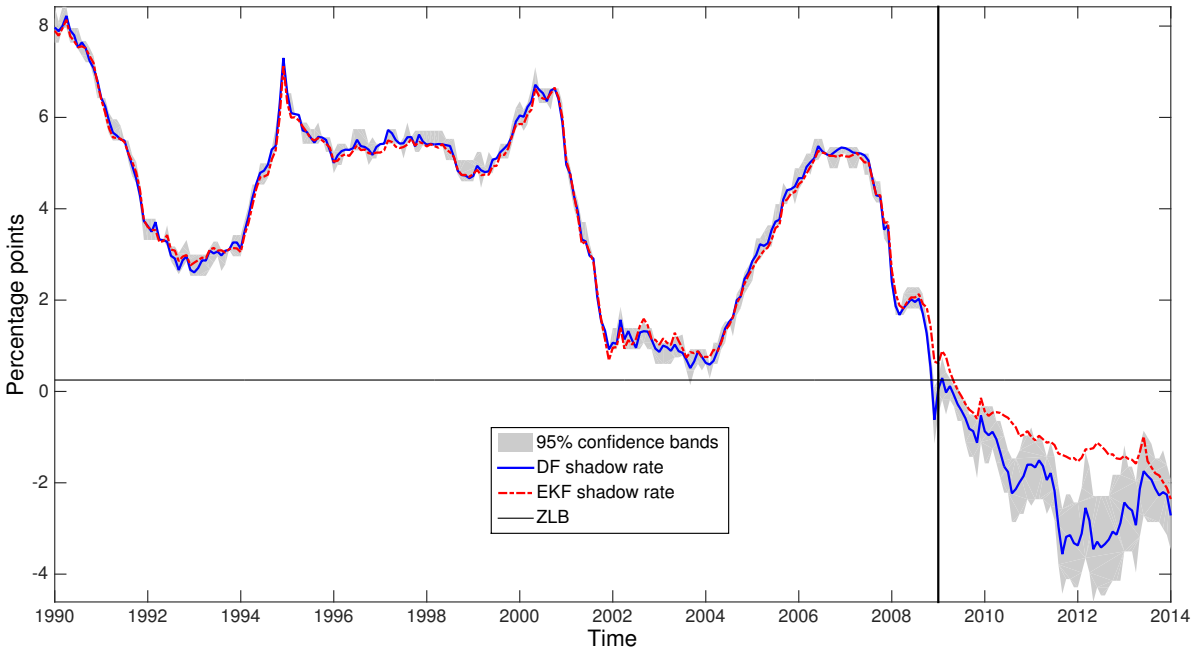

Figure 7: Shadow Rates

Though the parameter estimates obtained using the DF are similar, they produce a drastically different shadow rate series in the zero lower bound period (from about January 2009 onward).[21] These differences are illustrated in figure 7. I include 95% standard error bands for the shadow rate series estimated with the DF (where the randomness is coming from uncertainty about the state, not the parameters). This emphasizes the fact that the method used to estimate a nonlinear dynamic model can have important economic implications.

[20] See Hamilton (1994) for details.

[21] Note that once the parameter vector is estimated, I use the DF with 33 discretization points along each dimension to produce more smoothly varying filtered series. However, the qualitative difference remains even for coarser discretizations.

	Discretization Filter			Extended Kalman Filter		
1200μ	-0.2251	-0.2061	0.0256	-0.3035	-0.2381	0.0253
	(0.0767)	(1.3693)	(0.0318)	(0.1885)	(0.1815)	(0.0160)
ρ	0.9648	0.0056	0.4541	0.9638	-0.0026	0.3445
	(0.0100)	(0.0212)	(1.5863)	(0.0199)	(0.0183)	(0.4821)
	-0.0234	0.9626	0.8170	-0.0226	0.9420	1.0152
	(0.1809)	(0.0762)	(5.2167)	(0.0202)	(0.0212)	(0.5111)
	0.0046	0.0035	0.7750	0.0033	0.0028	0.8869
	(0.0023)	(0.0067)	(0.1050)	(0.0018)	(0.0019)	(0.0385)
$eig(\rho)$	0.9765+0.006i	0.9765-0.006i	0.7513	0.9832	0.9642	0.8452
$\rho^{\mathbb{Q}}$	0.9983	0	0	0.9978	0	0
	(0.0026)			(0.0003)		
	0	0.9608	1	0	0.9502	1
		(0.0121)			(0.0012)	
	0	0	0.9608	0	0	0.9502
			(0.0121)			(0.0012)
$1200\delta_0$	13.2418			13.3750		
	(2.3324)			(1.0551)		
1200Σ	0.2511			0.4160		
	(0.3467)			(0.0390)		
	-0.0535	0.2541		-0.3999	0.2445	
	(0.3483)	(0.1978)		(0.0369)	(0.0233)	
	-0.0002	0.0026	0.0338	-0.0110	0.0033	0.0390
	(0.0026)	(0.0058)	(0.0095)	(0.0069)	(0.0034)	(0.0030)
$1200\sqrt{\omega}$	0.1638			0.0893		
	(0.0403)			(0.0027)		

Table 10: Maximum likelihood parameter estimates (QMLE standard errors in parantheses)

9 Conclusion

Existing methods for estimating nonlinear dynamic models are either too computationally complex to be of practical use, or rely on local approximations which fail adequately to capture the nonlinear features of interest. In this paper, I develop a new method, the discretization filter, for approximating the likelihood of nonlinear, non-Gaussian state space models. This approximation is simple to compute and can be used to accurately estimate

a model's parameters using classical or Bayesian methods.

I apply results from the statistics literature on uniformly ergodic Markov chains to establish that the maximum likelihood estimator implied by the discretization filter is strongly consistent, asymptotically normal, and asymptotically efficient. I demonstrate through simulations that the discretization filter is orders of magnitude faster than alternative nonlinear techniques for the same level of approximation error and I provide practical guidelines for applied researchers.

I demonstrate that the filtering method used to estimate nonlinear models has sizeable effects on the accuracy of the estimated parameters and the accuracy of the filtered states. I show that these estimation differences translate into quantitatively significant economic differences using the Wu and Xia (2016) shadow rate model as an example. My findings have important implications for policy makers who use the Wu and Xia shadow rate as an input to determining the effectiveness of unconventional monetary policy. My estimation procedure leads one to conclude that the shadow rate was 2.2 percentage points lower in July 2012 than the estimates from their paper would indicate.

Additionally, I provide the first estimates of structural parameters in the Gabaix (2012) model of variable rare disasters. I show that the estimated model fails to identify the Great Recession as a disaster episode. This is due to the model's need to have a positive expected jump in inflation in the event of a disaster in order to capture an upward sloping nominal yield curve. Furthermore, I show that model fails to capture the sharp change in dynamics exhibited by the price-dividend ratio starting in the 1990s.

Going forward, I hope that economists working with nonlinear dynamic models will consider the discretization filter a valuable addition to their toolkit.

References

Adda, J. and Cooper, R. W. 2003. *Dynamic Economics: Quantitative Methods and Applications*. MIT press, Boston, MA.

Aruoba, S. B., Diebold, F. X., Nalewaik, J., Schorfheide, F., and Song, D. 2016. Improving GDP Measurement: A Measurement-Error Perspective. *Journal of Econometrics 191,* 2, 384–397.

Bakry, D., Milhaud, X., and Vandekerkhove, P. 1997. Statistique de Chaînes de Markov Cachées à Espace d'États Fini. Le Cas Non Stationnaire. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics 325,* 2, 203–206.

Barro, R. J. and Ursúa, J. F. 2008. Macroeconomic Crises Since 1870. *Brookings Papers on Economic Activity 2008,* 1, 255–350.

Bauer, M. D. and Rudebusch, G. D. 2016. Monetary Policy Expectations at the Zero Lower Bound. *Journal of Money, Credit, and Banking 48,* 7, 1439–1465.

Baum, L. E. and Petrie, T. 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Mathematical Statistics 37,* 6, 1554–1563.

Bickel, P. J. and Ritov, Y. 1996. Inference in Hidden Markov Models I: Local Asymptotic Normality in the Stationary Case. *Bernoulli 2,* 3, 199–228.

Bickel, P. J., Ritov, Y., Ryden, T., et al. 1998. Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models. *Annals of Statistics 26,* 4, 1614–1635.

Bucy, R. S. 1969. Bayes Theorem and Digital Realizations for Non-Linear Filters. *Journal of the Astronautical Sciences 17*, 80–94.

Bucy, R. S. and Senne, K. D. 1971. Digital Synthesis of Non-Linear Filters. *Automatica 7,* 3, 287–298.

Chen, Z. 2003. Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond. *Statistics 182,* 1, 1–69.

Douc, R., Moulines, E., Olsson, J., and Van Handel, R. 2011. Consistency of the Maximum Likelihood Estimator for General Hidden Markov Models. *Annals of Statistics 39,* 1, 474–513.

Douc, R., Moulines, E., and Ryden, T. 2004. Asymptotic Properties of the Maximum Likelihood Estimator in Autoregressive Models with Markov Regime. *Annals of Statistics 32,* 5, 2254–2304.

Farmer, L. E. and Toda, A. A. 2016. Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments. *Quantitative Economics, forthcoming*.

Flodén, M. 2008. A Note on the Accuracy of Markov-Chain Approximations to Highly Persistent AR(1)-Processes. *Economics Letters 99,* 3, 516–520.

Flury, T. and Shephard, N. 2011. Bayesian Inference Based Only on Simulated Likelihood: Particle Filter Analysis of Dynamic Economic Models. *Econometric Theory 27,* 05, 933–956.

Gabaix, X. 2009. Linearity-Generating Processes: A Modelling Tool Yielding Closed Forms for Asset Prices. Working Paper, New York University.

Gabaix, X. 2012. Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance. *The Quarterly Journal of Economics 127*, 645–700.

Gospodinov, N. and Lkhagvasuren, D. 2014. A Moment-Matching Method for Approximating Vector Autoregressive Processes by Finite-State Markov Chains. *Journal of Applied Econometrics 29,* 5, 843–859.

Gourio, F. 2012. Disaster Risk and Business Cycles. *American Economic Review 102,* 6, 2734–2766.

Gürkaynak, R. S., Sack, B., and Wright, J. H. 2007. The U.S. Treasury Yield Curve: 1961 to the Present. *Journal of Monetary Economics 54,* 8, 2291–2304.

Hamilton, J. D. 1989. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica 57,* 2, 357–384.

Hamilton, J. D. 1994. *Time Series Analysis*. Princeton University Press, Princeton NJ.

Hautsch, N. and Ou, Y. 2008. Stochastic Volatility Estimation Using Markov Chain Simulation. In *Applied Quantitative Finance*. Springer, Chapter 12, 249–274.

Herbst, E. P. and Schorfheide, F. 2015. *Bayesian Estimation of DSGE Models*. Princeton University Press, Princeton NJ.

Jensen, J. L. and Petersen, N. V. 1999. Asymptotic Normality of the Maximum Likelihood Estimator in State Space Models. *Annals of Statistics 27,* 2, 514–535.

Kopecky, K. A. and Suen, R. M. H. 2010. Finite State Markov-Chain Approximations to Highly Persistent Processes. *Review of Economic Dynamics 13,* 3, 701–714.

Leroux, B. G. 1992. Maximum-Likelihood Estimation for Hidden Markov Models. *Stochastic Processes and their Applications 40,* 1, 127–143.

Lindvall, T. 1992. *Lectures on the Coupling Method*. Wiley, New York, NY.

Maliar, L. and Maliar, S. 2015. Merging Simulation and Projection Approaches to Solve High-Dimensional Problems with an Application to a New Keynesian Model. *Quantitative Economics 6,* 1, 1–47.

Meyn, S. P. and Tweedie, R. L. 1993. *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

Rouwenhorst, K. G. 1995. Asset Pricing Implications of Equilibrium Business Cycle Models. In *Frontiers of Business Cycle Research*. Princeton University Press, Princeton, NJ, 294–330.

Tanaka, K. and Toda, A. A. 2013. Discrete Approximations of Continuous Distributions by Maximum Entropy. *Economics Letters 118,* 3, 445–450.

Tanaka, K. and Toda, A. A. 2015. Discretizing Distributions with Exact Moments: Error Estimate and Convergence Analysis. *SIAM Journal on Numerical Analysis 53,* 5, 2158–2177.

Tauchen, G. 1986. Finite State Markov-Chain Approximations to Univariate and Vector Autoregressions. *Economics Letters 20,* 2, 177–181.

Tauchen, G. and Hussey, R. 1991. Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models. *Econometrica 59,* 2, 371–396.

Taylor, S. 1982. Financial Returns Modelled by the Product of Two Stochastic Processes: A Study of the Daily Sugar Prices 1961-75. In *Time Series Analysis: Theory and Practice*. Vol. 1. North-Holland, Amsterdam, 203–226.

Wachter, J. A. 2013. Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility? *The Journal of Finance 68,* 3, 987–1035.

Wu, J. C. and Xia, F. D. 2016. Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound. *Journal of Money, Credit and Banking 48,* 2-3, 253–291.

A Proofs

Proof of Lemma 1. Note that by the Markov property, for $r < t \leq s$,

$$\bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \middle| \boldsymbol{X}_{r,M}^{t-1}, \boldsymbol{Y}_r^s \right) = \bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \middle| X_{t-1,M}, \boldsymbol{Y}_{t-1}^s \right)$$

Let $I_A \equiv \{ m \,|x_{m,M} \in A \}$ be the set of indices of the points in $\mathcal{X}_M$ contained in A. First consider the case where $t > s$.

$$\bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \middle| \boldsymbol{X}_{r,M}^{t-1}, \boldsymbol{Y}_r^s \right) = \bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \,| X_{t-1,M} = x \right) = \sum_{m' \in I_A} P_{\theta,M} \left(m, m' \right)$$

Next consider the case where $t \leq s$,

$$\begin{aligned}
&\bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \,| X_{t-1,M} = x, \boldsymbol{Y}_r^s \right) \\
&= \sum_{m' \in I_A} P_{\theta,M} \left(m, m' \right) \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \,| X_{t,M} = x_{m',M} \right) \times \left(\sum_{m'=1}^{M} P_{\theta,M} \left(m, m' \right) \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \,| X_{t,M} = x_{m',M} \right) \right)^{-1}
\end{aligned}$$

By assumption (B1),

$$\begin{aligned}
&\bar{\mathbb{P}}_\theta \left(X_{t,M} \in A \middle| X_{t-1,M} = x, \boldsymbol{Y}_{t-1}^s \right) \\
&\geq \sum_{m' \in I_A} Q_M^- \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \,| X_{t,M} = x_{m',M} \right) \times \left(\sum_{m'=1}^{M} Q_M^+ \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \,| X_{t,M} = x_{m',M} \right) \right)^{-1} \\
&= \frac{Q_M^-}{Q_M^+} \sum_{m' \in I_A} \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \,| X_{t,M} = x_{m',M} \right) \times \left(\sum_{m'=1}^{M} \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \,| X_{t,M} = x_{m',M} \right) \right)^{-1} \\
&\geq Q_+^- \mu_{t,M} \left(\boldsymbol{Y}_t^s, A \right)
\end{aligned}$$

where

$$\mu_{t,M} \left(\boldsymbol{Y}_t^s, A \right) \equiv \sum_{m' \in I_A} \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \,| X_{t,M} = x_{m',M} \right) \times \left(\sum_{m'=1}^{M} \bar{p}_{\theta,M} \left(\boldsymbol{Y}_t^s \,| X_{t,M} = x_{m',M} \right) \right)^{-1}$$

In the case $t > s$, it suffices to set $\mu_{t,M} \left(\boldsymbol{Y}_t^s, A \right) = \frac{\mu_{c,M}(A)}{\mu_{c,M}(\mathcal{X}_M)}$, where $\mu_{c,M}$ is counting measure on $\mathcal{X}_M$. □

Proof of Lemma 2. . Conditioning on a particular starting value $x_{0,M} \in \mathcal{X}_M$ is just a particular starting probability measure where probability 1 is assigned to that value. By

Corollary 1, it follows that

$$\left\|\mathbb{P}_\theta\left(X_{t-1,M}\in\cdot\left|\boldsymbol{Y}_0^{t-1},x_{0,M}=x_0\right.\right)-\bar{\mathbb{P}}_\theta\left(X_{t-1,M}\in\cdot\left|\boldsymbol{Y}_0^{t-1}\right.\right)\right\|_{TV}\leq\rho^{t-1}$$

Thus, for $t\geq 1$, by Corollary 1 and assumption (A3),

$$\begin{aligned}
&\left|p_{\theta,M}\left(Y_t\left|\boldsymbol{Y}_0^{t-1},X_0=x_0\right.\right)-\bar{p}_{\theta,M}\left(Y_t\left|\boldsymbol{Y}_0^{t-1}\right.\right)\right| \\
=&\left|\sum_{m=1}^{M}\sum_{m'=1}^{M}P_{\theta,M}\left(m,m'\right)g_\theta\left(Y_t\left|x_{m',M}\right.\right)\right. \\
&\left.\times\left(\mathbb{P}_\theta\left(X_{t-1,M}=x_{m,M}\left|\boldsymbol{Y}_0^{t-1},x_{0,M}=x_0\right.\right)-\bar{\mathbb{P}}_\theta\left(X_{t-1,M}=x_{m,M}\left|\boldsymbol{Y}_0^{t-1}\right.\right)\right)\right| \\
\leq&\ \rho^{t-1}\sup_{1\leq m\leq M}\sum_{m'=1}^{M}P_{\theta,M}\left(m,m'\right)g_\theta\left(Y_t\left|x_{m',M}\right.\right)
\end{aligned}$$

In addition, by assumption (B3),

$$\begin{aligned}
p_{\theta,M}\left(Y_t\left|\boldsymbol{Y}_0^{t-1},X_0=x_0\right.\right)&=\sum_{m=1}^{M}\sum_{m'=1}^{M}g_\theta\left(Y_t\left|x_{m',M}\right.\right)P_{\theta,M}\left(m,m'\right)\mathbb{P}_\theta\left(X_{t-1,M}=x_{m,M}\left|\boldsymbol{Y}_0^{t-1},x_{0,M}=x_0\right.\right) \\
&\geq\sum_{m=1}^{M}\left(\inf_{1\leq m\leq M}\sum_{m'=1}^{M}P_{\theta,M}\left(m,m'\right)g_\theta\left(Y_t\left|x_{m',M}\right.\right)\right) \\
&\quad\times\mathbb{P}_\theta\left(X_{t-1,M}=x_{m,M}\left|\boldsymbol{Y}_0^{t-1},x_{0,M}=x_0\right.\right) \\
&=\inf_{1\leq m\leq M}\sum_{m'=1}^{M}P_{\theta,M}\left(m,m'\right)g_\theta\left(Y_t\left|x_{m',M}\right.\right)
\end{aligned}$$

The same inequality also holds for $\bar{p}_{\theta,M}\left(Y_t\left|\boldsymbol{Y}_0^{t-1}\right.\right)$. It follows from the identity $|\log x-\log y|\leq|x-y|/\min(x,y)$ that

$$\begin{aligned}
&\left|\log p_{\theta,M}\left(Y_t\left|\boldsymbol{Y}_t^{s},x_{0,M}=x_0\right.\right)-\log\bar{p}_{\theta,M}\left(Y_t\left|\boldsymbol{Y}_t^{s}\right.\right)\right| \\
\leq&\ \rho^{t-1}\frac{\sup_{1\leq m\leq M}\sum_{m'=1}^{M}P_{\theta,M}\left(m,m'\right)g_\theta\left(Y_t\left|x_{m',M}\right.\right)}{\inf_{1\leq m\leq M}\sum_{m'=1}^{M}P_{\theta,M}\left(m,m'\right)g_\theta\left(Y_t\left|x_{m',M}\right.\right)} \\
\leq&\ \rho^{t-1}\frac{1}{Q_+^-} \\
\leq&\ \frac{\rho^{t-1}}{1-\rho}
\end{aligned}$$

By summing up the expression from $t = 1, \ldots, T$, we get

$$\left|\ell_{T,M}\left(\theta, x_0\right) - \ell_{T,M}\left(\theta\right)\right| \leq \sum_{t=1}^{T} \frac{\rho^{t-1}}{1-\rho} = \frac{1-\rho^{T+1}}{\left(1-\rho\right)^2} \leq \frac{1}{\left(1-\rho\right)^2}$$

Since this bound holds independently of θ and M, this concludes the proof. □

Proof of Lemma 3. Consider the first expression and let $r' \geq r$.

$$\begin{aligned}
&\bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,M} = x\right.\right) - \bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_{-r'}^{t-1}, X_{-r',M} = x'\right.\right) \\
&= \sum_{m=1}^{M}\sum_{m'=1}^{M}\sum_{m''=1}^{M} g_\theta\left(Y_t \left| x_{m'',M}\right.\right) P_{\theta,M}\left(m', m''\right) \bar{\mathbb{P}}_\theta\left(X_{t-1,M} = x_{m',M} \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,M} = x_{m,M}\right.\right) \mathbb{1}\left\{x = x_{m,M}\right\} \\
&- \sum_{m=1}^{M}\sum_{m'=1}^{M}\sum_{m''=1}^{M} g_\theta\left(Y_t \left| x_{m'',M}\right.\right) P_{\theta,M}\left(m', m''\right) \bar{\mathbb{P}}_\theta\left(X_{t-1,M} = x_{m',M} \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,M} = x_{m,M}\right.\right) \\
&\quad \bar{\mathbb{P}}_\theta\left(X_{-r,M} = x_{m,M} \left| \boldsymbol{Y}_{-r'}^{t-1}, X_{-r',M} = x'\right.\right)
\end{aligned}$$

Thus, by Corollary 1

$$\begin{aligned}
&\left|\bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,M} = x\right.\right) - \bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_{-r'}^{t-1}, X_{-r',M} = x'\right.\right)\right| \\
&= \left|\sum_{m=1}^{M}\left(\sum_{m'=1}^{M}\sum_{m''=1}^{M} g_\theta\left(Y_t \left| x_{m'',M}\right.\right) P_{\theta,M}\left(m', m''\right) \bar{\mathbb{P}}_\theta\left(X_{t-1,M} = x_{m',M} \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,M} = x_{m,M}\right.\right)\right)\right. \\
&\quad \left.\left(\mathbb{1}\left\{x = x_{m,M}\right\} - \bar{\mathbb{P}}_\theta\left(X_{-r,M} = x_{m,M} \left| \boldsymbol{Y}_{-r'}^{t-1}, X_{-r',M} = x'\right.\right)\right)\right| \\
&\leq \rho^{t+r-1} \sup_{1\leq m'\leq M} \sum_{m''=1}^{M} P_{\theta,M}\left(m', m''\right) g_\theta\left(Y_t \left| x_{m'',M}\right.\right)
\end{aligned}$$

Similary, I can obtain a lower bound on each term in the difference above as in the proof of Lemma 2,

$$\begin{aligned}
&\bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,M} = x\right.\right) \\
&= \sum_{m=1}^{M}\sum_{m'=1}^{M} g_\theta\left(Y_t \left| x_{m',M}\right.\right) P_{\theta,M}\left(m, m'\right) \bar{\mathbb{P}}_\theta\left(X_{t-1,M} = x_{m,M} \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,M} = x\right.\right) \\
&\geq \inf_{1\leq m\leq M} \sum_{m'=1}^{M} P_{\theta,M}\left(m, m'\right) g_\theta\left(Y_t \left| x_{m',M}\right.\right)
\end{aligned}$$

Using the same inequality for logs applied in the proof of Lemma 2 we obtain the desired result. An analogous expression is obtained if $r' \leq r$. The second expression of the theorem follows from setting $r = r'$ and integrating with respect to $\bar{\mathbb{P}}_\theta\left(dx_{-r,M} \left| \boldsymbol{Y}_{-r}^{t-1}\right.\right)$.

Note that by Assumption (A3),

$$c_{-}\left(Y_t\right) \leq \ \bar{p}_{\theta,M}\left(Y_t \left| \boldsymbol{Y}_{-r}^{t-1}, X_{-r,M}=x\right.\right) \leq b_{+}$$

Taking logs leads to the third inequality and concludes the proof.

□

Proof of Proposition 1. I wish to show that for any $A \in \mathcal{A}$, where $\mathcal{A}$ is the collection of continuity sets of X_t, that

$$\sup_{\theta \in \Theta}\left|\pi_{\theta,M}^{X}(A)-\pi_{\theta}^{X}(A)\right| = o\left(h^{*}(M)\right)$$

By the Portmenteau Lemma, this is equivalent to showing that $X_{t,M} \xrightarrow{d} X_t$ as $M \to \infty$. From assumption (BT), I know that for any $A \in \mathcal{A}$,

$$\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}}\left|P_{\theta,M}(x,A)-P_{\theta}(x,A)\right| = O\left(h(M)\right)$$

I will use this assumption and the fact that X_t and $X_{t,M}$ are uniformly ergodic to establish a bound on the difference in probability assigned to the set A by the approximate and true ergodic distributions.

By applying the triangle inequality twice, I can bound the expression of interest by the difference between the ergodic distribution of X_t and its T-step ahead transition kernel, the difference between $X_{t,M}$ and its T-step ahead transition kernel, and the difference between the two T-step ahead transition kernels

$$\begin{aligned}
&\sup_{\theta \in \Theta}\left|\pi_{\theta,M}^{X}(A)-\pi_{\theta}^{X}(A)\right| \\
=\ &\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}}\left|\pi_{\theta,M}^{X}(A)-\pi_{\theta}^{X}(A)\right| \\
=&\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}}\left|\pi_{\theta,M}^{X}(A)-P_{\theta,M}^{(T)}(x,A)+P_{\theta,M}^{(T)}(x,A)-P_{\theta}^{(T)}(x,A)+P_{\theta}^{(T)}(x,A)-\pi_{\theta}^{X}(A)\right| \\
\leq\ &\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}}\left|\pi_{\theta,M}^{X}(A)-P_{\theta,M}^{(T)}(x,A)\right|+\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}}\left|P_{\theta,M}^{(T)}(x,A)-P_{\theta}^{(T)}(x,A)\right| \\
&+\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}}\left|\pi_{\theta}^{X}(A)-P_{\theta}^{(T)}(x,A)\right|
\end{aligned}$$

Let ρ_1 and ρ_2 denote the uniform minorizing constants of the Markov chains X_t and $X_{t,M}$ respectively, and define $\rho_{+} \equiv \max\left(\rho_1, \rho_2\right)$. By the definition of uniform ergodicity, the first and third terms in the above expression can be bounded by their uniform minorizing

constants to the power T

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\pi^X_{\theta,M}(A)-\pi^X_{\theta}(A)\right| \\
&\leq \rho_1^T+\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P^{(T)}_{\theta,M}(x,A)-P^{(T)}_{\theta}(x,A)\right|+\rho_2^T \\
&\leq 2\rho_+^T+\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P^{(T)}_{\theta,M}(x,A)-P^{(T)}_{\theta}(x,A)\right|
\end{aligned}$$

It remains to bound the second term. Through applications of the triangle inequality and the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P^{(T)}_{\theta,M}(x,A)-P^{(T)}_{\theta}(x,A)\right| \\
=\ &\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P_{\theta,M}P^{(T-1)}_{\theta,M}(x,A)-P_{\theta,M}P^{(T-1)}_{\theta}(x,A)+P_{\theta,M}P^{(T-1)}_{\theta}(x,A)-P_{\theta}P^{(T-1)}_{\theta}(x,A)\right| \\
=\ &\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P_{\theta,M}\left(P^{(T-1)}_{\theta,M}(x,A)-P^{(T-1)}_{\theta}(x,A)\right)-\left(P_{\theta,M}-P_{\theta}\right)P^{(T-1)}_{\theta}(x,A)\right| \\
\leq\ &\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P_{\theta,M}\left(P^{(T-1)}_{\theta,M}(x,A)-P^{(T-1)}_{\theta}(x,A)\right)\right|+\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|\left(P_{\theta,M}-P_{\theta}\right)P^{(T-1)}_{\theta}(x,A)\right| \\
\leq\ &\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P_{\theta,M}(x,A)\right|\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P^{(T-1)}_{\theta,M}(x,A)-P^{(T-1)}_{\theta}(x,A)\right| \\
&+\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P_{\theta,M}(x,A)-P_{\theta}(x,A)\right|\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P^{(T-1)}_{\theta}(x,A)\right| \\
\leq\ &\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P^{(T-1)}_{\theta,M}(x,A)-P^{(T-1)}_{\theta}(x,A)\right|+\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P_{\theta,M}(x,A)-P_{\theta}(x,A)\right|
\end{aligned}$$

Applying this inequality recursively one can show that

$$\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P^{(T)}_{\theta,M}(x,A)-P^{(T)}_{\theta}(x,A)\right|\leq\ T\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|P_{\theta,M}(x,A)-P_{\theta}(x,A)\right|$$

As long as the Lebesgue measure of the discrete sets $\sup_m \lambda(A_{m,M})\to 0$ as $M\to\infty$, the set of discrete points $\{x_{m,M}\}$ will become dense in $\mathcal{X}$. That is, for any $x\in\mathcal{X}$ and $\varepsilon>0$, $\exists M>0$ and $1\leq m\leq M$ s.t. $\|x-x_{m,M}\|<\varepsilon$. Thus the error in the expression above is bounded by the quality of approximation of the marginal distributions $P_\theta(x,\cdot)$.

Combining this last inequality with the bounds derived above, the original expression of interest can be bounded by

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\pi^X_{\theta,M}(A)-\pi^X_{\theta}(A)\right| \\
&\leq 2\rho_+^T+T\times O\left(h(M)\right)
\end{aligned}$$

Letting T be a function of M, T_M, this means that $\exists\ 0<c<\infty$ and $\exists\ N<\infty$ such

that for all $M \geq N$,

$$\sup_{\theta \in \Theta} \left| \pi^X_{\theta,M}(A) - \pi^X_\theta(A) \right| \leq 2\rho_+^{T_M} + cT_M h(M)$$

Thus in order to control the above expression, it must be the case that T_M is chosen such that

$$2\rho_+^{T_M} + cT_M h(M) \to 0$$

as $M \to \infty$. Note that since $0 < \rho_+ < 1$, the term $2\rho_+^{T_M}$ decays exponentially fast as a function of T_M. Thus T_M can be chosen to be any function of M such that $2\rho_+^{T_M} \to 0$ and $T_M \times h(M) \to 0$ as $N \to \infty$. This will determine $h^*(M)$.

I now focus on the specific case of using the Farmer and Toda (2016) method with a trapezoidal rule quadrature rule. The trapezoidal rule has integration error which is $O\left(M^{-2/d}\right)$. Thus

$$\sup_{\theta \in \Theta} \left| \pi^X_{\theta,M}(A) - \pi^X_\theta(A) \right| \leq 2\rho_+^{T_M} + T_M \times O\left(M^{-2/d}\right)$$

This is equivalent to saying that $\exists\ 0 < c < \infty$ and $\exists\ N < \infty$ such that for all $M \geq N$,

$$\sup_{\theta \in \Theta} \left| \pi^X_{\theta,M}(A) - \pi^X_\theta(A) \right| \leq 2\rho_+^{T_M} + cT_M M^{-2/d}$$

Let $\varepsilon > 0$ and consider the sequence $T_M = M^{\varepsilon/d}$. Then

$$\begin{aligned} &2\rho_+^{T_M} + cT_M M^{-2/d} \\ =&2\rho_+^{M^{\varepsilon/d}} + cM^{\varepsilon/d} M^{-2/d} \\ =&2\rho_+^{M^{\varepsilon/d}} + cM^{(\varepsilon-2)/d} \end{aligned}$$

It is clear that the second term dominates asymptotically because it declines polynomially in M whereas the first term declines exponentially in M. This shows that

$$\sup_{\theta \in \Theta} \left| \pi^X_{\theta,M}(A) - \pi^X_\theta(A) \right| = O\left(M^{(\varepsilon-2)/d}\right)$$

This implies that for any $\delta > \varepsilon$

$$\sup_{\theta \in \Theta} \left| \pi^X_{\theta,M}(A) - \pi^X_\theta(A) \right| = o\left(M^{(\delta-2)/d}\right)$$

However since the choice of ε was arbitrary, we have that the above holds for any $\delta > 0$. This shows that for the case of the Farmer and Toda (2016) method with trapezoidal quadrature

rule, $h^*(M) = M^{(\delta-2)/d}$ for any $\delta > 0$.

□

Proof of Lemma 4. My goal is to show that the discrete approximation to the filtering distribution converges in distribution to the true filtering distribution as $M \to \infty$. Define $\boldsymbol{X}^0_{-r,M} \equiv (X_{0,M}, \ldots, X_{-r,M})$ and $\boldsymbol{X}^0_{-r} \equiv (X_0, \ldots, X_{-r})$. I will first show that $\boldsymbol{X}^0_{-r,M} \xrightarrow{d} \boldsymbol{X}^0_{-r}$ for $r \geq 0$ as $M \to \infty$. I will then show this implies that the joint distribution $\left(\boldsymbol{X}^0_{-r,M}, \boldsymbol{Y}^0_{-r}\right) \xrightarrow{d} \left(\boldsymbol{X}^0_{-r}, \boldsymbol{Y}^0_{-r}\right)$ as $M \to \infty$. This will imply my desired result, that $X_{0,M} \left| \boldsymbol{Y}^0_{-r} \right. \xrightarrow{d} X_0 \left| \boldsymbol{Y}^0_{-r} \right.$ as $M \to \infty$.

Let $f_r : \mathcal{X}^{r+1} \to \mathbb{R}$ be a bounded, continuous function. I will establish convergence in distribution by showing that the expectation of $f_r\left(\boldsymbol{X}^0_{-r,M}\right)$ converges to the expectation of $f_r\left(\boldsymbol{X}^0_{-r}\right)$ as $M \to \infty$ for any bounded, continuous f_r. Define the difference of these two expectations as

$$\Delta_E \equiv \left| \bar{\mathbb{E}}_\theta \left[f_r\left(\boldsymbol{X}^0_{-r,M}\right) \right] - \bar{\mathbb{E}}_\theta \left[f_r\left(\boldsymbol{X}^0_{-r}\right) \right] \right| \tag{47}$$

Recall the definitions of the transition kernel and ergodic distribution of the discrete approximation extended to $\mathcal{X}$

$$P_{\theta,M}(x, A) \equiv \sum_{m=1}^{M} \sum_{m'=1}^{M} P_{\theta,M}\left(m, m'\right) \mathbb{1}\left\{x \in A_{m,M}\right\} \mathbb{1}\left\{x_{m',M} \in A\right\} \tag{48}$$

$$\pi^X_{\theta,M}(A) \equiv \sum_{m=1}^{M} \pi^X_{\theta,M}(m) \mathbb{1}\left\{x_{m,M} \in A\right\} \tag{49}$$

This extended transition kernel $P_{\theta,M}$ and probability measure $\pi^X_{\theta,M}$ admit densities with respect to the measure μ on $\mathcal{X}$ which I will label as $q_{\theta,M}(\cdot\,|x) : \mathcal{X} \to \mathbb{R}$ for $x \in \mathcal{X}$, and $p_{\theta,M} : \mathcal{X} \to \mathbb{R}$. This allows me to replace summation by integration and keep the notation consistent across the discrete and continuous random variables.

I next factor the joint distribution of the sequence of $r+1$ X's into the product of the marginal distribution of the initial X and the distribution of the remaining X's conditional on the initial one. This is a straightforward application of Bayes' Rule.

$$\begin{aligned}\Delta_E = &\left| \int_{\mathcal{X}^{r+1}} f_r\left(x_0, \ldots, x_{-r}\right) p_{\theta,M}\left(x_0, \ldots, x_{-r}\right) dx_0 \cdots dx_{-r} \right. \\ &\left. - \int_{\mathcal{X}^{r+1}} f_r\left(x_0, \ldots, x_{-r}\right) p_{\theta}\left(x_0, \ldots, x_{-r}.\right) dx_0 \cdots dx_{-r} \right| \\ = &\left| \int_{\mathcal{X}^{r+1}} f_r\left(x_0, \ldots, x_{-r}\right) p_{\theta,M}\left(x_0, \ldots, x_{-r+1} \left| x_{-r} \right.\right) p_{\theta,M}\left(x_{-r}\right) dx_0 \cdots dx_{-r} \right. \\ &\left. - \int_{\mathcal{X}^{r+1}} f_r\left(x_0, \ldots, x_{-r}\right) p_{\theta}\left(x_0, \ldots, x_{-r+1} \left| x_{-r} \right.\right) p_{\theta}\left(x_{-r}\right) dx_0 \cdots dx_{-r} \right|\end{aligned}$$

Since both $X_{0,M}$ and X_0 are first order Markov processes, these distributions can be further factored into the product of the initial distribution with the sequence of r one-step-ahead conditional distributions.

$$\begin{aligned}\Delta_E = &\left| \int_{\mathcal{X}^{r+1}} f_r\left(x_0, \ldots, x_{-r}\right) q_{\theta,M}\left(x_0 \left| x_{-1} \right.\right) \cdots q_{\theta,M}\left(x_{-r+1} \left| x_{-r} \right.\right) p_{\theta,M}\left(x_{-r}\right) dx_0 \cdots dx_{-r} \right. \\ &\left. - \int_{\mathcal{X}^{r+1}} f_r\left(x_0, \ldots, x_{-r}\right) q_{\theta}\left(x_0 \left| x_{-1} \right.\right) \cdots q_{\theta}\left(x_{-r+1} \left| x_{-r} \right.\right) p_{\theta}\left(x_{-r}\right) dx_0 \cdots dx_{-r} \right|\end{aligned}$$

Before proceeding, it will be useful to define the operators associated with the transition kernels P_θ and $P_{\theta,M}$ and their r-step counterparts $P_\theta^{(r)}$ and $P_{\theta,M}^{(r)}$. For a function $f : \mathcal{X}^2 \to \mathbb{R}$, define

$$\left(P_\theta f\right)(x) \equiv \int_{\mathcal{X}} f\left(x', x\right) q_\theta\left(x' \left| x \right.\right) dx' \tag{50}$$

$$\left(P_{\theta,M} f\right)(x) \equiv \int_{\mathcal{X}} f\left(x', x\right) q_{\theta,M}\left(x' \left| x \right.\right) dx' \tag{51}$$

For $r > 1$, $0 \leq n < r$ and $f_r : \mathcal{X}^{r+1} \to \mathbb{R}$, define $f_{r-n} : \mathcal{X}^{r-n+1} \to \mathbb{R}$ as

$$f_{r-n}\left(x_0, \ldots, x_{-r+n}\right) \equiv f_r\left(x_0, \ldots, x_{-r+n}; x_{-r+n-1}, \ldots, x_{-r}\right) \tag{52}$$

where arguments after the semi-colon are held fixed. In other words, f_{r-n} can be thought of as the function f_r where the last n arguments are held fixed. This then allows me to define the i-step versions of P_θ and $P_{\theta,M}$. Define the 1-step versions as

$$\left(P_\theta^{(1)} f_1\right)(x) \equiv \left(P_\theta f_1\right)(x) = \int_{\mathcal{X}} f_r\left(x_0, x; x_{-2}, \ldots, x_{-r}\right) q_\theta\left(x_0 \left| x \right.\right) dx_0$$

$$\left(P_{\theta,M}^{(1)} f_1\right)(x) \equiv \left(P_{\theta,M} f_1\right)(x) = \int_{\mathcal{X}} f_r\left(x_0, x; x_{-2}, \ldots, x_{-r}\right) q_{\theta,M}\left(x_0 \left| x \right.\right) dx_0$$

For $i = 2, \ldots, r$, define

$$\left(P_{\theta}^{(i)} f_i\right)(x) \equiv \left(P_{\theta}\left(P_{\theta}^{(i-1)} f_{i-1}\right)\right)(x) = \int_{\mathcal{X}} \left(P_{\theta}^{(i-1)} f_{i-1}\right)(x_{-i+1}) q_{\theta}\left(x_{-i+1} \,|x\right) dx_{-i+1} \tag{53}$$

$$\left(P_{\theta,M}^{(i)} f_i\right)(x) \equiv \left(P_{\theta,M}\left(P_{\theta,M}^{(i-1)} f_{i-1}\right)\right)(x) = \int_{\mathcal{X}} \left(P_{\theta,M}^{(i-1)} f_{i-1}\right)(x_{-i+1}) q_{\theta,M}\left(x_{-i+1} \,|x\right) dx_{-i+1} \tag{54}$$

These are distinct are from what is referred to as the i-step ahead transition kernel and its associated operator. An i-step ahead transition kernel characterizes the probability of transitioning from a point in the space to any measurable set in that space i periods ahead. However, this operator implicitly characterizes the probability of moving from any point in the space to any sequence of i measurable sets. In other words, it computes probabilities over paths of the Markov chain. Note that these i-step ahead operators can be equivalently written in terms of one-step-ahead conditional densities as

$$\left(P_{\theta}^{(i)} f_i\right)(x) = \int_{\mathcal{X}^i} f_r\left(x_0, \ldots, x_{-i+1}, x; x_{-i-1}, \ldots, x_{-r}\right) q_{\theta}\left(x_0 \,|x_{-1}\right) \cdots q_{\theta}\left(x_{-i+1} \,|x\right) dx_0 \cdots dx_{-i+1}$$

$$\left(P_{\theta,M}^{(i)} f_i\right)(x) = \int_{\mathcal{X}^i} f_r\left(x_0, \ldots, x_{-i+1}, x; x_{-i-1}, \ldots, x_{-r}\right) q_{\theta,M}\left(x_0 \,|x_{-1}\right) \cdots q_{\theta,M}\left(x_{-i+1} \,|x\right) dx_0 \cdots dx_{-i+1}$$

With this new notation in hand, Δ_E can equivalenthly be rewritten in terms of the r-step operators as

$$\Delta_E = \left| \int_{\mathcal{X}} \left(P_{\theta,M}^{(r)} f_r\right)(x) p_{\theta,M}(x)\, dx - \int_{\mathcal{X}} \left(P_{\theta}^{(r)} f_r\right)(x) p_{\theta}(x)\, dx \right| \tag{55}$$

Next, I seek to establish that Δ_E can be bounded by the sum of two terms, one involving the difference in one step ahead transition kernels, the second involving the difference in $r-1$-step operators. By assumption a bound is known for the difference in integrals with respect to the one-step-ahead conditional distributions. Thus I can iteratively apply this logic to obtain a bound for Δ_E in terms of only the one-step-ahead approximation error.

Replace integration with respect to $p_{\theta,M}$ by integration with respect to p_{θ} in the first term of (55), and add and subtract the result from equation (55). Then apply the triangle

inequality to bound Δ_E by the sum of two new terms.

$$\begin{aligned}\Delta_E = &\left|\int_{\mathcal{X}}\left(P_{\theta,M}^{(r)}f_r\right)(x)p_{\theta,M}(x)\,dx - \int_{\mathcal{X}}\left(P_{\theta,M}^{(r)}f_r\right)(x)p_{\theta}(x)\,dx\right. \\ &\left. + \int_{\mathcal{X}}\left(P_{\theta,M}^{(r)}f_r\right)(x)p_{\theta}(x)\,dx - \int_{\mathcal{X}}\left(P_{\theta}^{(r)}f_r\right)(x)p_{\theta}(x)\,dx\right| \qquad (56) \\ \leq &\left|\int_{\mathcal{X}}\left(P_{\theta,M}^{(r)}f_r\right)(x)p_{\theta,M}(x)\,dx - \int_{\mathcal{X}}\left(P_{\theta,M}^{(r)}f_r\right)(x)p_{\theta}(x)\,dx\right| \\ &+ \left|\int_{\mathcal{X}}\left(P_{\theta,M}^{(r)}f_r\right)(x)p_{\theta}(x)\,dx - \int_{\mathcal{X}}\left(P_{\theta}^{(r)}f_r\right)(x)p_{\theta}(x)\,dx\right| \qquad (57)\end{aligned}$$

Consider the first term on the right hand side of inequality (57). It is simply the difference of integrals of $\left(P_{\theta,M}^{(r)}f_r\right)(x)$ with respect to $p_{\theta,M}$ and p_θ respectively. By Proposition 1, this difference is $o\left(h^*(M)\right)$.

$$\begin{aligned}&\left|\int_{\mathcal{X}}\left(P_{\theta,M}^{(r)}f_r\right)(x)p_{\theta,M}(x)\,dx - \int_{\mathcal{X}}\left(P_{\theta,M}^{(r)}f_r\right)(x)p_{\theta}(x)\,dx\right| \\ \leq &\sup_{|f|\leq 1}\left|\int_{\mathcal{X}} f(x)\,p_{\theta,M}(x)\,dx - \int_{\mathcal{X}} f(x)p_{\theta}(x)\,dx\right| \\ = &\,2\left\|\pi_{\theta,M}^X - \pi_\theta^X\right\|_{TV} \\ = &\,o\left(h^*(M)\right)\end{aligned}$$

Next consider the second term on the right hand side of inequality (57). By definition, the r-step operator can be written as the composition of the one-step-ahead operator with the $r-1$-step ahead operator.

$$\begin{aligned}&\left|\int_{\mathcal{X}}\left(P_{\theta,M}^{(r)}f_r\right)(x)p_{\theta}(x)\,dx - \int_{\mathcal{X}}\left(P_{\theta}^{(r)}f_r\right)(x)p_{\theta}(x)\,dx\right| \\ = &\left|\int_{\mathcal{X}}\left(P_{\theta,M}\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_{\theta}(x)\,dx - \int_{\mathcal{X}}\left(P_{\theta}\left(P_{\theta}^{(r-1)}f_{r-1}\right)\right)(x)p_{\theta}(x)\,dx\right| \qquad (58)\end{aligned}$$

Take the first term of equation (58), replace the first $P_{\theta,M}$ by P_θ, and add and subtract it

to equation (58). Then apply the triangle inequality again.

$$
\begin{aligned}
&\left|\int_{\mathcal{X}}\left(P_{\theta,M}\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathcal{X}}\left(P_\theta\left(P_\theta^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \\
=&\left|\int_{\mathcal{X}}\left(P_{\theta,M}\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathcal{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right. \\
&\left.+\int_{\mathcal{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathcal{X}}\left(P_\theta\left(P_\theta^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \qquad (59) \\
\leq&\left|\int_{\mathcal{X}}\left(P_{\theta,M}\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathcal{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \\
&+\left|\int_{\mathcal{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathcal{X}}\left(P_\theta\left(P_\theta^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \qquad (60)
\end{aligned}
$$

The first term of inequality (60) depends only on the approximation error of the one-step-ahead distribution, and the second term depends on the approximation error of the $r-1$-step ahead distribution. Define the function $\phi:\mathcal{X}^2\to\mathbb{R}$

$$
\phi\left(x_{-r+1},x_{-r}\right)\equiv\int_{\mathcal{X}^{r-1}}f_r\left(x_0,\ldots,x_{-r}\right)q_{\theta,M}\left(x_0\left|x_{-1}\right.\right)\cdots q_{\theta,M}\left(x_{-r+2}\left|x_{-r+1}\right.\right)dx_0\cdots dx_{-r+2} \qquad (61)
$$

Consider the first term on the right hand side of inequality (60) and substitute in the definitions of the r-step operators in terms of one-step-ahead conditional distributions. I will show that this term can be thought of as the difference in integrals of the function ϕ with respect to the one-step-ahead conditional distribution and its discrete approximation.

$$
\begin{aligned}
&\left|\int_{\mathcal{X}}\left(P_{\theta,M}\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathcal{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \\
=&\left|\int_{\mathcal{X}^{r+1}}f_r\left(x_0,\ldots,x_{-r}\right)q_{\theta,M}\left(x_0\left|x_{-1}\right.\right)\cdots q_{\theta,M}\left(x_{-r+1}\left|x_{-r}\right.\right)p_\theta\left(x_{-r}\right)dx_0\cdots dx_{-r}\right. \\
&\left.-\int_{\mathcal{X}^{r+1}}f_r\left(x_0,\ldots,x_{-r}\right)q_{\theta,M}\left(x_0\left|x_{-1}\right.\right)\cdots q_{\theta,M}\left(x_{-r+2}\left|x_{-r+1}\right.\right)q_\theta\left(x_{-r+1}\left|x_{-r}\right.\right)p_\theta\left(x_{-r}\right)dx_0\cdots dx_{-r}\right| \\
=&\left|\int_{\mathcal{X}^2}\phi\left(x_{-r+1},x_{-r}\right)q_{\theta,M}\left(x_{-r+1}\left|x_{-r}\right.\right)p_\theta\left(x_{-r}\right)dx_{-r+1}dx_{-r}\right. \\
&\left.-\int_{\mathcal{X}^2}\phi\left(x_{-r+1},x_{-r}\right)q_\theta\left(x_{-r+1}\left|x_{-r}\right.\right)p_\theta\left(x_{-r}\right)dx_{-r+1}dx_{-r}\right|
\end{aligned}
$$

The term on the right hand side of this last equality can be rewritten in terms of the

one-step-ahead operators P_θ and $P_{\theta,M}$

$$\begin{aligned}
&\left| \int_{\mathcal{X}^2} \phi\left(x_{-r+1}, x_{-r}\right) q_{\theta,M}\left(x_{-r+1} \left| x_{-r}\right.\right) p_\theta\left(x_{-r}\right) dx_{-r+1} dx_{-r} \right. \\
&\left. - \int_{\mathcal{X}^2} \phi\left(x_{-r+1}, x_{-r}\right) q_\theta\left(x_{-r+1} \left| x_{-r}\right.\right) p_\theta\left(x_{-r}\right) dx_{-r+1} dx_{-r} \right| \\
= &\left| \int_{\mathcal{X}} \left(P_{\theta,M}\phi\right)(x) p_\theta(x) dx - \int_{\mathcal{X}} \left(P_\theta \phi\right)(x) p_\theta(x) dx \right|
\end{aligned}$$

By proposition 1 the error between integrals with respect to $q_{\theta,M}$ and q_θ is $o\left(h^*\left(M\right)\right)$.

$$\begin{aligned}
&\left| \int_{\mathcal{X}} \left(P_{\theta,M}\phi\right)(x) p_\theta(x) dx - \int_{\mathcal{X}} \left(P_\theta \phi\right)(x) p_\theta(x) dx \right| \\
\leq &\sup_{|f| \leq 1} \left| \int_{\mathcal{X}} \left(P_{\theta,M} f\right)(x) p_\theta(x) dx - \int_{\mathcal{X}} \left(P_\theta f\right)(x) p_\theta(x) dx \right| \\
= &\, 2 \left\| \pi_\theta P_{\theta,M} - \pi_\theta P_\theta \right\|_{TV} \\
= &\, o\left(h^*\left(M\right)\right)
\end{aligned}$$

This leaves one term to bound to establish convergence in distribution of $\boldsymbol{X}^0_{-r,M}$ to $\boldsymbol{X}^0_{-r}$ as $M \to \infty$. Consider the second term on the right hand side of inequality (60). Similar to the above argument, it will be useful to define a new function $\varphi : \mathcal{X}^{r-1} \to \mathbb{R}$

$$\varphi\left(x_0, \ldots, x_{-r+2}\right) = \int_{\mathcal{X}^2} f_r\left(x_0, \ldots, x_{-r}\right) q_\theta\left(x_{-r+1} \left| x_{-r}\right.\right) p_\theta\left(x_{-r}\right) dx_{-r+1} dx_{-r}$$

By using Fubini's theorem, I will show that by switching the order of integration in the second term on the right hand side of inequality (60), this term can be expressed as the $(r-1)$-step operators $P_\theta^{(r-1)}$ and $P_{\theta,M}^{(r-1)}$ applied to the same function φ. I take the supremum over the conditioning value for x_{-r+2} in order to break the dependence of the terms not

captured by φ on x_{-r+1}.

$$
\begin{aligned}
&\left|\int_{\mathcal{X}}\left(P_\theta\left(P_{\theta,M}^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx-\int_{\mathcal{X}}\left(P_\theta\left(P_\theta^{(r-1)}f_{r-1}\right)\right)(x)p_\theta(x)\,dx\right| \\
=&\left|\int_{\mathcal{X}^{r+1}}f_r(x_0,\ldots,x_{-r})\,q_{\theta,M}(x_0\,|x_{-1})\cdots q_{\theta,M}(x_{-r+2}\,|x_{-r+1})\,q_\theta(x_{-r+1}\,|x_{-r})\,p_\theta(x_{-r})\,dx_0\cdots dx_{-r}\right. \\
&\left.-\int_{\mathcal{X}^{r+1}}f_r(x_0,\ldots,x_{-r})\,q_\theta(x_0\,|x_{-1})\cdots q_\theta(x_{-r+2}\,|x_{-r+1})\,q_\theta(x_{-r+1}\,|x_{-r})\,p_\theta(x_{-r})\,dx_0\cdots dx_{-r}\right| \\
\leq&\sup_{x\in\mathcal{X}}\left|\int_{\mathcal{X}^{r+1}}f_r(x_0,\ldots,x_{-r})\,q_{\theta,M}(x_0\,|x_{-1})\cdots q_{\theta,M}(x_{-r+2}\,|x)\,q_\theta(x_{-r+1}\,|x_{-r})\,p_\theta(x_{-r})\,dx_0\cdots dx_{-r}\right. \\
&\left.-\int_{\mathcal{X}^{r+1}}f_r(x_0,\ldots,x_{-r})\,q_\theta(x_0\,|x_{-1})\cdots q_\theta(x_{-r+2}\,|x)\,q_\theta(x_{-r+1}\,|x_{-r})\,p_\theta(x_{-r})\,dx_0\cdots dx_{-r}\right| \\
=&\sup_{x\in\mathcal{X}}\left|\int_{\mathcal{X}^{r-1}}\varphi(x_0,\ldots,x_{-r+2})\,q_{\theta,M}(x_0\,|x_{-1})\cdots q_{\theta,M}(x_{-r+2}\,|x)\,dx_0\cdots dx_{-r+2}\right. \\
&\left.-\int_{\mathcal{X}^{r-1}}\varphi(x_0,\ldots,x_{-r+2})\,q_{\theta,M}(x_0\,|x_{-1})\cdots q_\theta(x_{-r+2}\,|x)\,dx_0\cdots dx_{-r+2}\right|
\end{aligned}
$$

Note that the last term in the right hand side of the above equality can be thought of as the $(r-2)$-step operator applied to the function φ

$$
\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\int_{\mathcal{X}^{r-1}}\varphi(x_0,\ldots,x_{-r+2})\,q_{\theta,M}(x_0\,|x_{-1})\cdots q_{\theta,M}(x_{-r+2}\,|x)\,dx_0\cdots dx_{-r+2}\right. \\
&\left.-\int_{\mathcal{X}^{r-1}}\varphi(x_0,\ldots,x_{-r+2})\,q_{\theta,M}(x_0\,|x_{-1})\cdots q_\theta(x_{-r+2}\,|x)\,dx_0\cdots dx_{-r+2}\right| \\
=&\sup_{x\in\mathcal{X}}\left|\left(P_{\theta,M}^{(r-2)}\varphi\right)(x)-\left(P_\theta^{(r-2)}\varphi\right)(x)\right|
\end{aligned}
$$

By appyling the same logic to this component as the $(r-1)$-step ahead component, it can be shown that the second term on the right hand side of inequality (57) will be $o(r\times h^*(M))$. Combining that result with the bound on the first term on the right hand side of inequality (57) and returning to the original expression of interest, it can be seen that

$$\Delta_E\leq o(h^*(M))+o(r\times h^*(M))=o(r\times h^*(M))$$

For any fixed r, this difference converges to 0 because by assumption $h^*(M)\to 0$ as $M\to\infty$.

Next I seek to show that $\left(\boldsymbol{X}_{-r,M}^0,\boldsymbol{Y}_{-r}^0\right)\xrightarrow{d}\left(\boldsymbol{X}_{-r}^0,\boldsymbol{Y}_{-r}^0\right)$ as $M\to\infty$. The joint density

can be written and then factored as:

$$\begin{aligned}
&p_\theta\left(X_0,\ldots,X_{-r},Y_0,\ldots,Y_{-r}\right)\\
&=p_\theta\left(Y_0,\ldots,Y_{-r}\left|X_0,\ldots,X_{-r}\right.\right)p_\theta\left(X_0,\ldots,X_{-r}\right)\\
&=g_\theta\left(Y_0\left|X_0\right.\right)\cdots g_\theta\left(Y_{-r}\left|X_{-r}\right.\right)p_\theta\left(X_0,\ldots,X_{-r}\right)
\end{aligned}$$

The same factorization can be done for the discrete approximations. Consider the expectation of an arbitrary bounded, continuous function $f:\mathcal{X}^{r+1}\times\mathcal{Y}^{r+1}\to\mathbb{R}$. In order to establish convergence in distribution it is sufficient to establish the expectation of any bounded, continuous function of the sequence of approximations converges to the expectation of the function of the limit. The difference in the expectations of the function f is given by

$$\left|\bar{\mathbb{E}}_\theta\left[f\left(\boldsymbol{X}^0_{-r,M},\boldsymbol{Y}^0_{-r}\right)\right]-\bar{\mathbb{E}}_\theta\left[f\left(\boldsymbol{X}^0_{-r},\boldsymbol{Y}^0_{-r}\right)\right]\right|$$

Define the new function $f^*:\mathcal{X}^{r+1}\to\mathbb{R}$ as:

$$f^*\left(\boldsymbol{x}^0_{-r}\right)\equiv\int_{\mathcal{Y}^{r+1}}f\left(\boldsymbol{x}^0_{-r},\boldsymbol{y}^0_{-r}\right)g_\theta\left(y_0\left|x_0\right.\right)\cdots g_\theta\left(y_{-r}\left|x_{-r}\right.\right)dy_0\cdots dy_{-r}$$

Since $g_\theta\left(\cdot\left|x\right.\right)$ is a continuous and bounded function, so is their $(r+1)$-fold product and thus their product with f. Furthermore, since integration is a continuous operator over the space $\mathcal{Y}^{r+1}$, it follows from $\boldsymbol{X}^0_{-r,M}\xrightarrow{d}\boldsymbol{X}^0_{-r}$ that $\left(\boldsymbol{X}^0_{-r,M},\boldsymbol{Y}^0_{-r}\right)\xrightarrow{d}\left(\boldsymbol{X}^0_{-r},\boldsymbol{Y}^0_{-r}\right)$ as $M\to\infty$. This implies that the filtering distribution $X_{0,M}\left|\boldsymbol{Y}^0_{-r}\right.\xrightarrow{d}X_0\left|\boldsymbol{Y}^0_{-r}\right.$ as $M\to\infty$. Making an analogous argument to that in Proposition 1, r can be chosen as a function of M, r_M, so as to maintain the convergence in distribution as both r and M go to infinity. The sufficient condition is that $r_M\times h^*\left(M\right)\to 0$ as $M\to\infty$.

Consider the initial object of interest

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\ell_M\left(\theta\right)-\ell\left(\theta\right)\right|\\
=&\sup_{\theta\in\Theta}\left|\bar{\mathbb{E}}_{\theta^*}\left[\log\bar{p}_{\theta,M}\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)-\log\bar{p}_\theta\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)\right]\right|\\
\leq&\sup_{\theta\in\Theta}\bar{\mathbb{E}}_{\theta^*}\left[\left|\log\bar{p}_{\theta,M}\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)-\log\bar{p}_\theta\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)\right|\right]\\
\leq&\sup_{\theta\in\Theta}\bar{\mathbb{E}}_{\theta^*}\left[\frac{\left|\bar{p}_{\theta,M}\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)-\bar{p}_\theta\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)\right|}{\min\left(\bar{p}_{\theta,M}\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right),\bar{p}_\theta\left(Y_0\left|\boldsymbol{Y}^0_{-\infty}\right.\right)\right)}\right]
\end{aligned}$$

This quantity converges to 0 as $M\to\infty$ due to the convergence in distribution of the filtering distributions for infinite histories. When the Farmer and Toda (2016) method with

a trapezoidal quadrature rule is used,

$$\sup_{\theta\in\Theta}\left|\ell_M(\theta)-\ell(\theta)\right| = o\left(h^*(M)\right) = o\left(M^{-(2-\delta)/d}\right)$$

for $\delta > 0$, by arguments analogous to those made in proposition 1.

□

Proof of Lemma 5. I first establish that for any fixed $x \in \mathcal{X}_M$, r, and M, $\Delta_{0,r,M,x}(\theta)$ is continuous w.r.t. θ. By definition

$$\bar{p}_{\theta,M}\left(Y_0 \left| \boldsymbol{Y}_{-r}^{-1}, X_{-r,M} = x\right.\right) = \frac{\bar{p}_{\theta,M}\left(\boldsymbol{Y}_{-r+1}^{0} \left| Y_{-r}, X_{-r,M} = x\right.\right)}{\bar{p}_{\theta,M}\left(\boldsymbol{Y}_{-r+1}^{-1} \left| Y_{-r}, X_{-r,M} = x\right.\right)}$$

Note that for $s \in \{-1, 0\}$, and assuming $x = x_{m_{-r},M}$ without loss of generality,

$$\begin{aligned}
&\bar{p}_{\theta,M}\left(\boldsymbol{Y}_{-r+1}^{s} \left| Y_{-r}, X_{-r,M} = x\right.\right) \\
&= \sum_{m_{-r},\ldots,m_s}\left[P_{\theta,M}(m_{-r}, m_{-r+1})\,\mathbb{1}\left\{x_{m_{-r}} = x\right\}\prod_{i=-r+2}^{s} P_{\theta,M}(m_{i-1}, m_i)\prod_{i=-r+1}^{s} g_\theta\left(Y_i \left| X_i = x_{m_i,M}\right.\right)\right]
\end{aligned}$$

Thus $\bar{p}_{\theta,M}\left(\boldsymbol{Y}_{-r+1}^{s} \left| Y_{-r}, X_{-r,M} = x\right.\right)$ is continuous w.r.t. θ by continuity of $P_{\theta,M}$ and g_θ. Therefore the sequence $\{\Delta_{0,r,M,x}\}$ is also continuous w.r.t. θ because it is the composition of continuous functions. Since $\{\Delta_{0,r,M,x}(\theta)\}$ converges uniformly w.r.t. $\theta \in \Theta$, $\bar{\mathbb{P}}_{\theta^*}$-a.s., $\Delta_{0,\infty,M}(\theta)$ is also continuous w.r.t. $\theta \in \Theta$, $\bar{\mathbb{P}}_{\theta^*}$-a.s. The proof follows by using Lemma 3 and the dominated convergence theorem. □

Proof of Proposition 2. Using the triangle inequality,

$$\begin{aligned}
\sup_{\theta\in\Theta}\sup_{x_0\in\mathcal{X}}\left|T^{-1}\ell_{T,M}(\theta, x_0) - \ell(\theta)\right| &= \sup_{\theta\in\Theta}\sup_{x_0\in\mathcal{X}}\left|T^{-1}\ell_{T,M}(\theta, x_0) - \ell_M(\theta) + \ell_M(\theta) - \ell(\theta)\right| \\
&\leq \sup_{\theta\in\Theta}\sup_{x_0\in\mathcal{X}}\left|T^{-1}\ell_{T,M}(\theta, x_0) - \ell_M(\theta)\right| + \sup_{\theta\in\Theta}\left|\ell_M(\theta) - \ell(\theta)\right|
\end{aligned}$$

The second term limits to 0 by Lemma 4. For the second term, note that by Lemma 2 it is sufficient to prove that

$$\limsup_{T\to\infty}\sup_{\theta\in\Theta}\sup_{M\in\mathbb{Z}^+}\left|T^{-1}\ell_{T,M}(\theta) - \ell_M(\theta)\right| = 0, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

Furthermore, since Θ is compact, this further reduces to proving that for all $\theta \in \Theta$,

$$\lim_{\delta\to 0}\limsup_{T\to\infty}\sup_{|\theta'-\theta|\leq\delta}\sup_{M\in\mathbb{Z}^+}\left|T^{-1}\ell_{T,M}(\theta') - \ell_M(\theta)\right| = 0, \quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}$$

This term can be further decomposed as

$$\begin{aligned}
&\limsup_{\delta\to 0}\limsup_{T\to\infty}\sup_{|\theta'-\theta|\leq\delta}\sup_{M\in\mathbb{Z}^+}\left|T^{-1}\ell_{T,M}\left(\theta'\right)-\ell_M\left(\theta\right)\right| \\
=&\limsup_{\delta\to 0}\limsup_{T\to\infty}\sup_{|\theta'-\theta|\leq\delta}\sup_{M\in\mathbb{Z}^+}\left|T^{-1}\ell_{T,M}\left(\theta'\right)-T^{-1}\ell_{T,M}\left(\theta\right)\right| \\
\leq& A+B+C
\end{aligned}$$

where

$$\begin{aligned}
A=&\limsup_{\delta\to 0}\limsup_{T\to\infty}\sup_{|\theta'-\theta|\leq\delta}\sup_{M\in\mathbb{Z}^+}T^{-1}\sum_{t=1}^{T}\left|\Delta_{t,0,M}\left(\theta'\right)-\Delta_{t,\infty,M}\left(\theta'\right)\right|, \\
B=&\limsup_{\delta\to 0}\limsup_{T\to\infty}\sup_{|\theta'-\theta|\leq\delta}\sup_{M\in\mathbb{Z}^+}T^{-1}\sum_{t=1}^{T}\left|\Delta_{t,\infty,M}\left(\theta'\right)-\Delta_{t,\infty,M}\left(\theta\right)\right|, \\
C=&\limsup_{T\to\infty}\sup_{M\in\mathbb{Z}^+}T^{-1}\sum_{t=1}^{T}\left|\Delta_{t,\infty,M}\left(\theta\right)-\Delta_{t,0,M}\left(\theta\right)\right|
\end{aligned}$$

Terms A and C are zero by Corollary 2, and by Lemma 5 and the ergodic theorem,

$$\begin{aligned}
B\leq&\limsup_{\delta\to 0}\limsup_{T\to\infty}T^{-1}\sum_{t=1}^{T}\sup_{|\theta'-\theta|\leq\delta}\sup_{M\in\mathbb{Z}^+}\left|\Delta_{t,\infty,M}\left(\theta'\right)-\Delta_{t,\infty,M}\left(\theta\right)\right| \\
=&\limsup_{\delta\to 0}\bar{\mathbb{E}}_{\theta^*}\left[\sup_{|\theta'-\theta|\leq\delta}\sup_{M\in\mathbb{Z}^+}\left|\Delta_{t,\infty,M}\left(\theta'\right)-\Delta_{t,\infty,M}\left(\theta\right)\right|\right] \\
=&0,\quad \bar{\mathbb{P}}_{\theta^*}\text{-a.s.}
\end{aligned}$$

□

Proof of Theorem 3. . In order to establish asymptotic normality of my proposed estimator, it is sufficient to show that $\ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_T\left(\hat{\theta}_{T,M,x_0},x_0\right)=o_P\left(1\right)$ by Theorem 7 of Douc et al. (2004). Rewriting this term

$$\begin{aligned}
&\ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_T\left(\hat{\theta}_{T,M,x_{0,M}},x_0\right) \\
=&\ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_T\left(\hat{\theta}_{T,M,x_{0,M}},x_0\right)+\ell_{T,M}\left(\hat{\theta}_{T,x_0},x_{0,M}\right)-\ell_{T,M}\left(\hat{\theta}_{T,x_0},x_{0,M}\right) \\
\leq&\ell_T\left(\hat{\theta}_{T,x_0},x_0\right)-\ell_{T,M}\left(\hat{\theta}_{T,x_0},x_{0,M}\right)+\ell_{T,M}\left(\hat{\theta}_{T,M,x_{0,M}},x_{0,M}\right)-\ell_T\left(\hat{\theta}_{T,M,x_{0,M}},x_0\right)
\end{aligned}$$

Note that it is thus sufficient to show that for any $\theta\in\Theta$, $\mathbb{P}_\theta\left(\left|\ell_{T,M}\left(\theta,x_{0,M}\right)-\ell_T\left(\theta,x_0\right)\right|\geq\varepsilon\right)\to 0$ as $T\to\infty$ and $M\to\infty$ at appropriate rates. It is possible to decompose this probability

as follows:

$$\begin{aligned}
&\mathbb{P}_\theta\left(\left|\ell_{T,M}\left(\theta, x_{0,M}\right)-\ell_T\left(\theta, x_0\right)\right| \geq \varepsilon\right) \\
=&\mathbb{P}_\theta\left(\left|\ell_{T,M}\left(\theta, x_{0,M}\right)-\ell_T\left(\theta, x_0\right)+T\ell\left(\theta\right)-T\ell\left(\theta\right)+T\ell_M\left(\theta\right)-T\ell_M\left(\theta\right)\right| \geq \varepsilon\right) \\
\leq & \mathbb{P}_\theta\left(\left|\ell_T\left(\theta, x_0\right)-T\ell\left(\theta\right)\right| \geq \frac{\varepsilon}{3}\right)+\mathbb{P}_\theta\left(\left|\ell_{T,M}\left(\theta, x_{0,M}\right)-T\ell_M\left(\theta\right)\right| \geq \frac{\varepsilon}{3}\right)+\mathbb{P}_\theta\left(T\left|\ell_M\left(\theta\right)-\ell\left(\theta\right)\right| \geq \frac{\varepsilon}{3}\right)
\end{aligned}$$

Theorem 14 from Douc et al. (2011) states that for any V_θ-uniformly ergodic state process with transition kernel P_θ, $f : \mathcal{Y}^{s+1}$ with $\|f\|_\infty < \infty$, there exists a constant $K < \infty$ such that

$$\mathbb{P}_\theta^\nu\left(\left|\sum_{t=1}^T\left\{f\left(\boldsymbol{Y}_t^{t+s}\right)-\bar{\mathbb{E}}_{\theta^*}\left[f\left(\boldsymbol{Y}_0^s\right)\right]\right\}\right| \geq \varepsilon\right) \leq K\nu\left(V\right)\exp\left[-\frac{1}{K}\left(\min\left(\frac{\varepsilon^2}{T}, \varepsilon\right)\right)\right]$$

for any initial probability measure ν and $\varepsilon > 0$. Both the original chain P_θ and each discrete chain $P_{\theta,M}$ are uniformly ergodic and thus V_θ-uniformly ergodic for $V_\theta = 1$.

Note that the first two terms are of the form considered in Theorem 14 from Douc et al. (2011). I explicity show the bound for the first term and the second term is analogous due to the uniform minorization of the sequence of discrete Markov chains for all $M \in \mathbb{Z}^+$ with the same minorizing constant

$$\begin{aligned}
&\mathbb{P}_\theta\left(\left|\ell_T\left(\theta, x_0\right)-T\ell\left(\theta\right)\right| \geq \frac{\varepsilon}{3}\right) \\
=&\mathbb{P}_\theta\left(\left|\sum_{t=1}^T\left\{\log p_\theta\left(Y_t\left|\boldsymbol{Y}_0^{t-1}, X_0=x_0\right.\right)-\ell\left(\theta\right)\right\}\right| \geq \frac{\varepsilon}{3}\right) \\
\leq & K\exp\left[-\frac{1}{K}\left(\min\left(\frac{\varepsilon^2}{9T}, \frac{\varepsilon}{3}\right)\right)\right]=o_P\left(1\right) \quad \text{with } P=\bar{\mathbb{P}}_{\theta^*}
\end{aligned}$$

For the third term, it follows from Lemma 4 that

$$\left|\ell_M\left(\theta\right)-\ell\left(\theta\right)\right|=o\left(h^*\left(M\right)\right)$$

and thus

$$T\left|\ell_M\left(\theta\right)-\ell\left(\theta\right)\right|=o\left(T \times h^*\left(M\right)\right)$$

Returning to the original expression of interest

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left|\ell_T\left(\hat{\theta}_{T,x_0}, x_0\right)-\ell_T\left(\hat{\theta}_{T,M,x_0}, x_0\right)\right| \geq \varepsilon\right) \\
\leq & \mathbb{P}_{\theta^*}\left(\left|\ell_T\left(\hat{\theta}_{T,x_0}, x_0\right)-\ell_{T,M}\left(\hat{\theta}_{T,x_0}, x_0\right)+\ell_{T,M}\left(\hat{\theta}_{T,M,x_0}, x_0\right)-\ell_T\left(\hat{\theta}_{T,M,x_0}, x_0\right)\right| \geq \varepsilon\right) \\
\leq & \mathbb{P}_{\theta^*}\left(\left|\ell_{T,M}\left(\hat{\theta}_{T,x_0}, x_0\right)-\ell_T\left(\hat{\theta}_{T,x_0}, x_0\right)\right| \geq \frac{\varepsilon}{2}\right)+\mathbb{P}_{\theta^*}\left(\left|\ell_{T,M}\left(\hat{\theta}_{T,M,x_0}, x_0\right)-\ell_T\left(\hat{\theta}_{T,M,x_0}, x_0\right)\right| \geq \frac{\varepsilon}{2}\right) \rightarrow 0
\end{aligned}$$

for $T \to \infty$, $M \to \infty$, and $T \times h^*(M) \to 0$. This ensures that my proposed estimator satisfies condition (iii) of Theorem 7 from Douc et al. (2004). □

B Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments

This appendix briefly summarizes the method for discretizting stochastic processes proposed in Farmer and Toda (2016).

Consider the time-homogeneous first-order Markov process

$$\mathbb{P}(X_t \leq x' | X_{t-1} = x) = F(x' \,|x\,),$$

where X_t is the random vector of state variables and $F(\cdot\,|x\,)$ is a cumulative distribution function (CDF) that determines the distribution of $X_t = x'$ given $X_{t-1} = x$. The dynamics of any Markov process are completely characterized by its Markov transition kernel. In the case of a discrete state space, this transition kernel is simply a matrix of transition probabilities, where each row corresponds to a conditional distribution. One can discretize the continuous process X_t by applying the Tanaka and Toda (2013) method to each conditional distribution separately.

More concretely, suppose that one has a set of grid points $D_M = \{x_m\}_{m=1}^M$ and an initial coarse approximation $Q = (q_{mm'})$, which is an $M \times M$ probability transition matrix. Additionally, suppose one wants to match some conditional moments of X_t, represented by the moment defining function $T(x)$. The exact conditional moments when the current state is $X_{t-1} = x_m$ are

$$\overline{T}_m = \mathbb{E}\left[T(X_t)\,|X_{t-1} = x_m\right] = \int T(x) dF(x\,|x_m\,),$$

where the integral is over x, fixing $X_{t-1} = x_m$. (If these moments do not have explicit expressions, highly accurate quadrature formulas can be used to compute them.) By Theorem 2.1 in Farmer and Toda (2016), these moments can be matched exactly by solving the optimization problem

$$\begin{aligned} \min_{\{p_{mm'}\}_{m'=1}^M} \quad & \sum_{m'=1}^M p_{mm'} \log \frac{p_{mm'}}{q_{mm'}} \\ \text{subject to} \quad & \sum_{m'=1}^M p_{mm'} T(x_{m'}) = \overline{T}_m, \;\; \sum_{m'=1}^M p_{mm'} = 1, \;\; p_{mm'} \geq 0 \end{aligned} \tag{62}$$

for each $m = 1, 2, \ldots, M$, or equivalently the dual problem

$$\min_{\lambda \in \mathbb{R}^L} \sum_{m'=1}^{M} q_{mm'} \mathrm{e}^{\lambda'(T(x_{m'}) - \overline{T}_m)}. \tag{63}$$

(63) has a unique solution if and only if the regularity condition

$$\overline{T}_m \in \text{int co } T(D_M) \tag{64}$$

holds. Furthermore, if the dual problem has a unique solution λ_m, then the solution to the primal problem (62) is given by

$$p_{mm'} = \frac{q_{mm'} \mathrm{e}^{\lambda_m' \left(T(x_{m'}) - \overline{T}_m\right)}}{\sum_{m'=1}^{M} q_{mm'} \mathrm{e}^{\lambda_m' \left(T(x_{m'}) - \overline{T}_m\right)}} \tag{65}$$

Lastly, define the errors associated with the moment matching as:

$$\epsilon_m \equiv \sum_{m'=1}^{M} p_{mm'} T\left(x_{m'}\right) - \overline{T}_m \tag{66}$$

The procedure for constructing the finite-state Markov chain approximation to X_t is summarized in Algorithm 2 below.

Algorithm 2: Discretization of Markov processes

1. Select a discrete set of points $D_M = \{x_m\}_{m=1}^{M}$ and an initial approximation $Q = (q_{mm'})$.
2. Select a moment defining function $T(x)$ and corresponding exact conditional moments $\left\{\overline{T}_m\right\}_{m=1}^{M}$. If necessary, approximate the exact conditional moments with highly accurate numerical integrals. Set $m \rightsquigarrow 1$ and define an error tolerance $\kappa > 0$.
3. Solve minimization problem (63) and store the resulting solution λ_m.
4. Compute ϵ_m using (66). If $\|\epsilon_m\|_\infty < \kappa$, move to step 5. If not, select a smaller set of moments to match and return to step 3.
5. Compute the conditional probabilities corresponding to row m of $P = (p_{mm'})$ using (65). Set $m \rightsquigarrow m + 1$. If $m \leq M$, move to step 3, otherwise move to step 6.
6. Collect the computed conditional probability measures in the matrix $P = (p_{mm'})$.

The resulting finite-state Markov chain approximation to X_t takes values in the set D_M and has associated transition matrix P. Since the dual problem (63) is an unconstrained convex minimization problem with a typically small number of variables, standard New-

ton type algorithms can be applied. Furthermore, since the probabilities (65) are strictly positive by construction, the transition probability matrix $P = (p_{mm'})$ is a strictly positive matrix, so the resulting Markov chain is stationary and uniformly ergodic by construction.