New Methods in Theory & Applications of Nonlinear Filtering

by

Andrew Papanicolaou

B. S., University of California at Santa Barbara 2003

M. S., University of Southern California 2007

A Dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Division of Applied Mathematics at Brown University

Providence, Rhode Island

May 2010

This dissertation by Andrew Papanicolaou is accepted in its present form by the Division of Applied Mathematics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Boris Rozovsky, Director

Recommended to the Graduate Council

Date _____ _____

Constantine Dafermos, Reader

Date _____ _____

George Karniadakis, Reader

Approved by the Graduate Council

Date _____ _____

Sheila Bonde, Dean of the Graduate School

Vitæ

Andrew Papanicolaou was born on 1st of January of 1980. As a youth he enjoyed jazz music and the study of its complex rhythms and harmonies. In the spring of 2003, he received a bachelors degree in mathematical sciences from the University of California at Santa Barbara, and matriculated at the University of Southern California shortly thereafter to work on a masters degree in financial mathematics. In the fall of 2007, with his masters degree completed, he arrived at Brown University to write his PhD thesis on nonlinear filtering with Boris Rozovsky.

"With Probability One"

Dedicated to My Parents

Acknowledgements

A very special thanks to my advisor for his mentorship over the past six years both at USC and here at Brown. For all my achievements thus far, he is the person to whom I give all the credit. Without his support I would never have made it this far. Special thanks to Michael Black at Brown University for introducing me to his work on human tracking with Bayesian methods. A very special thanks to the late Professor David Gottlieb whose profound wisdom in his last year encouraged me to work hard.

Preface

I certainly did not plan to earn a doctorate of applied mathematics when I started graduate school in the Fall of 2004. I enrolled in the financial mathematics program at USC with intentions of starting a career in banking as soon as I completed a Master's degree. However, things never quite work out the way you plan, and now after six years of studies I am at Brown finishing my PhD thesis on nonlinear filtering. I am happy to be finishing and will look forward to my future work as a post-doc.

Andrew Papanicolaou

Providence, Rhode Island

Contents

List of Tables

List of Figures

Abstract of "New Methods in Theory & Applications of Nonlinear Filtering"

by Andrew Papanicolaou, Ph.D., Brown University, May 2010

Hidden Markov models are used in countless signal processing problems, and the associated nonlinear filtering algorithms are used to obtain posterior distributions for the hidden states. One reason why posterior distributions are so important is because they are used to compute estimates which are optimal given a history of observed data. However, it is often the case that implementation of these algorithms is near impossible because of the curse-of-dimensionality which results from testing every possible hypothesis. This thesis explores new applications of nonlinear filtering and addresses several issues in algorithm implementation.

CHAPTER 1

Introduction

The concentration of this thesis is the various applications of nonlinear filtering and the associated challenges in algorithm implementation. There are four main chapters in this dissertation, each of which presents a self-contained analysis of a problem that is either a new and novel applications of nonlinear filtering, or a new method for deriving fast-filtering algorithms for hidden Markov models with multiple time-scales.

A nonlinear filtering algorithm is a Bayesian method for estimating the state of a hidden Markov model (HMM). For instance, let k denote a time-step and consider the Markov chain $(\theta_k)_{k=1,2,3,...}$ that takes value in the finite state-space $\mathcal{S} = \{s_1, s_2, ..., s_m\}$. There is an $m \times m$ matrix Λ of transition probabilities such that from time k to $k+1$ there is the following probability of the event $\{\theta_{k+1} = s_i\}$ given $\{\theta_k = s_j\}$,

$$\mathbb{P}(\theta_{k+1} = s_i | \theta_k = s_j) = \Lambda_{ji}$$

Now suppose that this Markov chain is unobserved or hidden, but there is an observed process Y_k that is a nonlinear function of θ_k,

$$Y_k = h(\theta_k) + \text{“iid noise”}$$

A nonlinear filtering algorithm uses Bayes rule and the Markov property to formulate the posterior distribution of θ_k given $Y_{1:k} = (Y_1, Y_2, ..., Y_k)$ in a recursive manner:

$$\pi_k^i := \mathbb{P}(\theta_k = s_i | Y_{1:k})$$

$$= \frac{\mathbb{P}(\theta_k = s_i, Y_k | Y_{1:k-1})}{\mathbb{P}(Y_k | Y_{1:k-1})} = \frac{\mathbb{P}(Y_k | \theta_k = s_i, Y_{1:k-1}) \mathbb{P}(\theta_k = s_i | Y_{1:k-1})}{\mathbb{P}(Y_k | Y_{1:k-1})}$$

$$= \frac{\mathbb{P}(Y_k | \theta_k = s_i) \mathbb{P}(\theta_k = s_i | Y_{1:k-1})}{\mathbb{P}(Y_k | Y_{1:k-1})} = \frac{\mathbb{P}(Y_k | \theta_k = s_i) \sum_j \mathbb{P}(\theta_k = s_i, \theta_{k-1} = s_j | Y_{1:k-1})}{\mathbb{P}(Y_k | Y_{1:k-1})}$$

$$= \frac{\mathbb{P}(Y_k|\theta_k = s_i)\sum_j \mathbb{P}(\theta_k = s_i|\theta_{k-1} = s_j, Y_{1:k-1})\mathbb{P}(\theta_{k-1} = s_j|Y_{1:k-1})}{\mathbb{P}(Y_k|Y_{1:k-1})}$$

$$= \frac{\mathbb{P}(Y_k|\theta_k = s_i)\sum_j \mathbb{P}(\theta_k = s_i|\theta_{k-1} = s_j)\mathbb{P}(\theta_{k-1} = s_j|Y_{1:k-1})}{\mathbb{P}(Y_k|Y_{1:k-1})}$$

$$= \frac{1}{c_k}\mathbb{P}(Y_k|\theta_k = s_i)\sum_j \Lambda_{ji}\pi^j_{k-1}$$

where c_k is a normalizing constant. This simple probabilistic recursion is the main idea behind all filtering algorithms.

Hidden Markov modeling and filtering algorithms have proven to be extremely useful in a number of applied areas including speech recognition, automatic target recognition, computer vision, and molecular biology. This thesis does considerable work with applications to human-tracking problems in computer vision, and also considers applications to finance which is an area where filtering is relatively undeveloped.

In applications, the methods used to implement the associated algorithms are often non-trivial. In fact, in many cases it is nearly impossible to implement an algorithm whose output is at all close to the posterior distribution. In the past, approximations such as the extended Kalman Filter were used, and more recently as computer technology has improved it has become more practical to use Monte Carlo methods such as the Particle filter. One of the main topics of this thesis is fast-filtering algorithms for HMMs with multiple time-scales, where it is shown by example how to construct a filter of reduced-dimension when it is not necessary to compute the distribution for part of the state associated with a fast time-scale. Such filters are considerably faster because there are fewer dimensions that need to be integrated over in updating the posterior distributions.

The most famous results in filtering theory are the Kalman Filter for linear Gaussian systems, and its counter-part from from SDE theory called the Kalman-Bucy filter. For general nonlinear systems, there is the forward Baum-Welch equation which serves as the work-horse for all nonlinear filtering and Bayesian algorithms. Other

well-known nonlinear filtering results include the Kushner equation, the Zakai equation, and the Wonham filter, as well as estimation methods such as the particle filter and Monte Carlo Markov Chain (MCMC).

The work done in this thesis examines four separate nonlinear filtering problems in close detail. For each problem, a model is derived, a filtering algorithm is formulated, and numerical experiments are carried out. Chapters 2, 3, 4, and 5 of this dissertation are self-contained works each of which focusses on a specific problem in the field of nonlinear filtering with its own unique contribution.

Chapter 2 is a novel application of filtering theory to finance and stochastic volatility, but also serves as primer for the basic tools of nonlinear filtering such as the forward Baum-Welch equation, the backward Baum-Welch equation, and the parameter re-estimation procedure given by the expectation-maximization (EM) algorithm. Historically, volatility is negatively correlated with the markets, and therefore is useful for hedging in times of uncertainty or possible downturn. One type of such hedging instruments are volatility swaps, whose returns rely on a rather crude time-series estimate of the volatility process. As an alternative instrument, swap contracts can be written for the times series of filtering estimates, and the variance of returns on these contracts will be lower than the those of a traditional swap. The main results is that there is parity as swaps that use filtering volatility exhibit a mean/variance trade-off resulting in lower expected returns.

Chapter 3 discusses a problem related to the identification/prediction of behavioral characteristics of an intelligent agent based on noisy video footage. The problem is motivated by the growing needs in surveillance for software that is capable of tracking human beings. The software is developed with the goal of tracking an array of the agent's attributes including patterns of motion (e.g. pose and position), intentions, mood swings etc. Some of the attributes, (e.g. intensions) are not directly observable, and need to be inferred from the other attributes. Others, such as pose and location, are partially observable but the observations are corrupted by noise.

The approach to identification is based on nonlinear filtering-type algorithms and optimal change-point detection for partially observable double Markov chains (DMCs). Generally speaking, nonlinear filtering for partially observable DMCs is a particular case of the HMM filtering algorithm for vector-valued Markov chains. However, one of the main obstacles to efficient performance of this filtering algorithm is the curse of dimensionality. To some degree, this problem could be addressed by the introduction of particle filters, Rao-Blackwellization, and other methods, but the high dimensionality still remains a serious problem. Introduction of DMCs is just another step in the quest for reduction of computational complexity of HMMs. It compliments other methods (e.g. particle filters and Rao-Blackwellization) when possible. This application of DMC-based nonlinear filtering to analysis of video footage demonstrates the practical potential of the approach.

Chapter 4 considers nonlinear filtering applications to target tracking based on a vector of multi-scaled models where some of the processes are rapidly mean reverting to their local equilibria. The algorithms are applicable to target tracking problems because multi-scaled models with fast mean-reversion (FMR) are a simple way to model latency in the response of tracking systems. A Kramers-Smoluchowski approximation is applied to the system of hidden states to obtain the main result, which is a simplified nonlinear filtering algorithms that is derived by exploiting the FMR structures within the HMM. Specifically, a Baum-Welch recursion of reduced dimension is derived. The simplified algorithm is implemented in numerical simulations, and its efficiency and robustness is discussed.

Chapter 5 is the most theoretical chapter in the thesis. The main part is the derivation of a fast nonlinear filtering algorithm of reduced dimension for a hidden Markov model with multiple time-scales. It is assumed that the motivation of the model and the nonlinear filtering algorithm are known, and then the bulk of the chapter focuses on the analytical/probabilistic details necessary to prove that the fast filter is optimal as the time-scale parameter increases to infinity. The main result

is a filter of reduced dimension for a particular class of multi-dimensional filtering problems where the hidden processes have multiple time scales.

As possible future topics of research, each of the four problems in this thesis can be further analyzed to obtain new and improved results. It is without a doubt that filtering applications to finance will generate interest and that further data analysis will be necessary to strengthen the claims made thus far. The application to video tracking is a good start but to date there has been no serious attempt to construct filtering algorithms with parallel implementation in mind. It will be a considerable step forward for HMMs in computer vision if the rate at which posteriors are computed is significantly increased by implementing parallel versions of the nonlinear filtering algorithms. With regard to HMMs with multiple time-scales, there is a lot of analysis to be done for models with more parameters, as many of the proofs in this thesis will require extensions by the introduction of concepts such as moving-averages and heteroskedasticity.

CHAPTER 2

Filtering for Stochastic Volatility Models & Volatility Swap Contracts

The Black-Scholes Theory says that volatility is positively correlated with the options market and is the key factor in pricing. Furthermore, since an option is an instrument whose values increases in times of uncertainty or possible downturn, volatility should be negatively correlated with the market. Indeed this trend is evident in the data of volatility resources such as the VIX. However, volatility is not a process that is directly available for quotes, and so there is demand for things such as implied volatility, volatility surface estimates, and filtering volatility estimates. In addition, after an estimate of volatility has been computed there needs to be a correct interpretation of its meaning in the marketplace.

With an understanding of volatility, an increased understanding of the brand of financial instruments known as **volatility swaps** will follow. Traditionally, a party enters into a volatility swap contract with the understanding that at specified terminal time they will receive the net difference between historical volatility and a specified strike price. For the S&P 500, such contracts are 'fairly' priced initially by the VIX which manipulates a portfolio of European call and put options to find the risk-neutral valuation of the contract's expected payoff. Thus, the VIX is a form a implied volatility because it is computed from options data. Generally speaking, the returns on volatility swap contracts give the edge to the issuer of the contract, which can be empirically verified by the historically negative returns on volatility swaps.

As an alternative to implied volatility, stochastic volatility and filtering algorithms deal directly with model parameters and observable data, and therefore return a form of historical volatility. Therefore, a time-series of filtering estimates for the volatility of the S&P 500 has fundamental differences from the well-regarded VIX,

but can provide a 'better' estimate of the average volatility over a time period than the traditional time-series estimate that is taken to be the sum of the squares on log-returns. Furthermore, upon examination of the data, one finds that the VIX and filtering estimates do have a lot in common, and in fact are a close fit if scaled properly. From this phenomenon of filtering tracking the VIX, it is clear that market forces are at work in a way that relates stochastic volatility to the VIX, and this needs to be interpreted in order to further understand the risk associated with the S&P 500 and other similar assets.

In financial math, there has been a lot of work in stochastic and implied volatility, which are described in the book by Fouque, Papanicolaou and Sircar [**19**]. Information pertaining to the VIX such as its history, the dynamics of variance swaps, and data analysis, can be found in the paper Carr and Wu [**12**], and also in the paper by Cont and Kokholm [**13**]. Basic filtering theory is covered in the book by Jazwinski [**25**] and the paper by Rabiner [**32**] presents all the algorithms for filtering and parameter re-estimation. With specific regard to content of this chapter, the previous work done on this problem is relatively sparse, but the initiating of filtering for stochastic volatility is made in the paper by Cvitanic, Rozovsky and Zaliapan [**15**] where stochastic volatility is modeled with a Cox process.

In this chapter, stochastic volatility modeling, implied volatility and nonlinear filtering come together in a way that is relatively new. A nonlinear filtering algorithm for a stochastic volatility model on a discrete space is formulated, along with associated algorithms for smoothing and parameter re-estimation. It is also shown that 30-day swap contracts offering a return on the average of the time-series of filtered volatility for S&P 500 volatility have returns with lower variance than the traditional swap contract, but the net returns have lower expectation.

This chapter thoroughly analyzes volatility swaps for the S&P 500, but future work would needs to be done such as applying the same experiments to other indices such as the DOW, the Nasdaq, the S&P 100, and various stocks for which volatility

is traded. Another possibility would be to find a portfolio which replicates the time-series filtering estimates, but it is uncertain that such an instrument is attainable.

1. Historical Volatility, Variance-Swaps, and the VIX

Let the financial markets be modeled in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where the price of a tradable asset is modeled as a stochastic processes

$$S_t = \text{the price of the stock or index at time } t \geq 0.$$

A widely accepted stochastic volatility model for S_t is the following SDE

$$dS_t = \mu_t S_t dt + \sigma_t S_t dW_t \tag{1}$$

where μ_t is a process which models annualized growth, σ_t is a process which models annualized volatility, and W_t is an independent Brownian motion. Quotes on S_t are available throughout the trading day, but the continuum of prices is not observable. For some time $T > 0$, there is a discrete set of times $(t_k)_{k=0}^{N+1}$ at which quotes are available,

$$S_{t_k} = \text{the observed quote on the asset at time } t_k \in [0, T]$$

The solution of (1) on $[t_k, t_{k+1}]$ is

$$S_{t_{k+1}} = S_{t_k} \exp\left\{ \int_{t_k}^{t_{k+1}} \left(\mu_\tau - \frac{1}{2}\sigma_\tau^2 \right) d\tau + \int_{t_k}^{t_{k+1}} \sigma_\tau dW_\tau \right\}$$

which leads to a the following time-series model of log-returns:

$$Y_k := \log(S_{t_{k+1}}/S_{t_k}) = \int_{t_k}^{t_{k+1}} \left(\mu_\tau - \frac{1}{2}\sigma_\tau^2 \right) d\tau + \int_{t_k}^{t_{k+1}} \sigma_\tau dW_\tau \tag{2}$$

For simplicity, assume that there is an increment Δt such that the quotes are given at equidistant times,

$$0 = t_0 < t_1 < ..... < t_{N+1} = T, \qquad \text{where } t_{k+1} - t_k \equiv \Delta t, \ \forall k \tag{3}$$

Define the average volatility on $[0, T]$ as

$$\sigma^2_{[0,T]} = \frac{1}{T}\int_0^T \sigma^2_\tau d\tau \tag{4}$$

and consider the time-series average of the squared of returns

$$RV_{[0,T]} := \frac{1}{T}\sum_{k=0}^{N} Y_k^2 \tag{5}$$

which converges in mean to (4) as the sample size goes to a continuum,

$$\mathbb{E}[RV_{[0,T]}] = \mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{N}\left(\int_{t_k}^{t_{k+1}} \sigma^2_\tau d\tau + o(\Delta t)\right)\right] = \mathbb{E}\left[\frac{1}{T}\int_0^T \sigma^2_\tau d\tau\right] + O(\Delta t)$$

$$= E[\sigma^2_{[0,T]}] + O(\Delta t) \to \mathbb{E}[\sigma^2_{[0,T]}] \qquad \text{as } N \nearrow \infty \tag{6}$$

The average in (5) is commonly referred to as **realized** or **historical volatility**, but in practice it is not possible to obtain a continuum of quotes in the interval $[0, T]$, and so historical volatility is only an estimate of the average volatility.

However, historical volatility is often used in a brand of derivative contracts called **variance swaps**. Variance swaps are contracts that pay the net difference of historical volatility and a pre-specified strike. For a strike-price K_{vol}, a long position in a swap on volatility pays the difference in historical volatility and the strike price,

$$\text{variance-swap payoff} = RV_{[0,T]} - K^2_{vol}$$

1.1. 'Fair' Valuation of Volatility Swaps. For a time $t > 0$, the price of the swap is dictated by the market. Initially, at time $t = 0$, by arbitrage arguments, the 'fair' price should be the discounted risk-neutral expectation of the expected payoff,

$$\text{fair swap price}|_{t=0} = e^{-rT}\left(\mathbb{E}^*_0[RV_{[0,T]}] - K^2_{vol}\right) \approx e^{-rT}\left(\mathbb{E}^*_0[\sigma^2_{[0,T]}] - K^2_{vol}\right)$$

where $r > 0$ is the risk-free rate of interest and $\mathbb{E}^*_t[\cdot]$ is the expectation under the risk-neutral measure conditioned on the market data to time t. The risk free measure is defined in terms of a Martingale,

$$\frac{d\mathbb{P}^*}{d\mathbb{P}} = M_T = \exp\left\{-\frac{1}{2}\int_0^T \left(\frac{\mu_\tau - r}{\sigma_\tau}\right)^2 d\tau - \int_0^T \frac{\mu_\tau - r}{\sigma_\tau} dW_\tau\right\}$$

under which $dS_t = rS_t dt + \sigma_t S_t dW_t^*$ where W_t^* is $\mathbb{P}^*$-Brownian motion and $dW_t^* = \frac{\mu_t - r}{\sigma_t} dt + dW_t$, and so the limit in (6) holds under the risk-neutral measure. However, it is not immediately clear how to obtain an explicit expression for this fair price. Fortunately, a clever portfolio can be constructed and arbitrate theory can be applied to obtain this price.

Suppose that at time $t = 0$ there is the following forward spot rate for S_T,

$$F = \text{spot rate for a futures contract in } S_T \text{ at time } t = 0$$

and there is a risk free rate of return r such that $F = \mathbb{E}_0^* S_T = S_0 e^{rT}$. For any strike price K, let $C(T, t; K)$ and $P(T, t; K)$ denote the risk-neutral arbitrage-free prices of a European call and put, respectively. Then, for any time $t \in [0, T]$ consider the following portfolios of out-of-the money calls and puts:

$$\text{portfolio } 1 = \{1 \text{ of each out-of-the-money European call option}\}$$

$$\text{portfolio } 2 = \{1 \text{ of each out-of-the-money European put option}\}$$

Letting the prices of these two portfolios be given by Φ_t^1 and Φ_t^2, respectively,

$$\Phi_t^1 = e^{r(T-t)} \int_F^\infty \frac{C(T, t; K)}{K^2} dK, \qquad \Phi_t^2 = e^{r(T-t)} \int_0^F \frac{P(T, t; K)}{K^2} dK$$

by arbitrage-free pricing theory, Φ_t^1 can be rewritten as

$$\Phi_t^1 = e^{r(T-t)} \int_F^\infty e^{-r(T-t)} \int_K^\infty \frac{s - K}{K^2} \rho_{T-t}(s) ds dK = \int_F^\infty \rho_{T-t}(s) \int_F^s \frac{s - K}{K^2} dK ds$$

where $\rho_{T-t}(s) = \frac{d}{ds} \mathbb{P}^*(S_T \leq s | S_t)$. Similarly for Φ_t^2, we have

$$\Phi_t^2 = e^{r(T-t)} \int_0^F e^{-r(T-t)} \int_0^K \frac{K - s}{K^2} \rho_{T-t}(s) ds dK = \int_0^F \rho_{T-t}(s) \int_s^F \frac{K - s}{K^2} dK ds$$

Summing them together and integrating over K, there is the following expression for the joint portfolio for any time,

$$\Phi_t^1 + \Phi_t^2 = \int_0^\infty \rho_{T-t}(s) \int_s^F \frac{K - s}{K^2} dK ds = \int_0^\infty \rho_{T-t}(s) \left(\log(F/s) + \frac{s - F}{F} \right) ds$$

$$= \mathbb{E}_t^*[\log(F/S_T)] + \frac{\mathbb{E}_t^* S_T - F}{F}$$

Therefore, a short position in the log-returns on a futures contract can be hedged with a short position in a futures contract and a long position in Φ_T^1 and Φ_T^2,

$$-\log(S_T/F) = -\frac{S_T - F}{F} + \Phi_T^1 + \Phi_T^2$$

In particular, at time $t = 0$, the model for the index price says that $S_T = S_0 e^{\int_0^T (r - .5\sigma_t^2)dt + \int_0^T \sigma_t dW_t^*}$ where W_t^* is Brownian motion under the risk-neutral measure, as so the expected logarithm of returns can be written as the integral of the volatility,

$$\Phi_0^1 + \Phi_0^2 = -\mathbb{E}_0^* \log(S_T/F) = -\mathbb{E}_0^* \left[\log\left(e^{-\int_0^T .5\sigma_t^2 dt + \int_0^T \sigma_t dW_t^*}\right)\right] = \frac{1}{2}\int_0^T \mathbb{E}_0^*[\sigma_t^2]dt$$

where $\mathbb{E}^*[\cdot]$ denotes the risk neutral expectation. Thus there is the following fair valuation of the average volatility on $[0, T]$,

$$\mathbb{E}_0^* \sigma_{[0,T]}^2 = \frac{2}{T}(\Phi_0^1 + \Phi_0^2)$$

and the fair price of the volatility swap with strike K_{vol} is given in terms of these portfolios:

$$\text{fair price of swap}|_{t=0} = \frac{2}{T}(\Phi_0^1 + \Phi_0^2) - K_{vol}^2$$

1.2. The VIX. The price $\frac{2}{T}(\Phi_0^1 + \Phi_0^2)$ with $T = 30$ days is what the VIX approximates by using all options data available on the S&P 500 with T-many days to expiry. In computing the VIX, the continuum of strike prices is approximated with the set of strike prices $(K_i)_{i=1,2,3,....}$ provided by the market,

$$(7) \qquad VIX_{t_k}^2 = \frac{2e^{rT}}{T}\left(\sum_{K_i > F} \frac{C(t_k + T, t_k; K_i)}{K_i^2}\Delta K_i + \sum_{K_i < F} \frac{P(t_k + T, t_k; K_i)}{K_i^2}\Delta K_i\right) - \frac{1}{T}\left(\frac{F}{K_0} - 1\right)^2$$

where $K_0 = \max\{K_i : F - K_i \geq 0\}$. The VIX is a form of **implied volatility**. Implied volatility is a value of σ_t that is obtained from options data. For instance, the VIX, or the volatility smile/skew that is obtained by inverting the Black-Scholes model on European call and put options. Implied volatility should not be confused with 'volatility surface estimates' such as the those obtained through Dupire's equation.

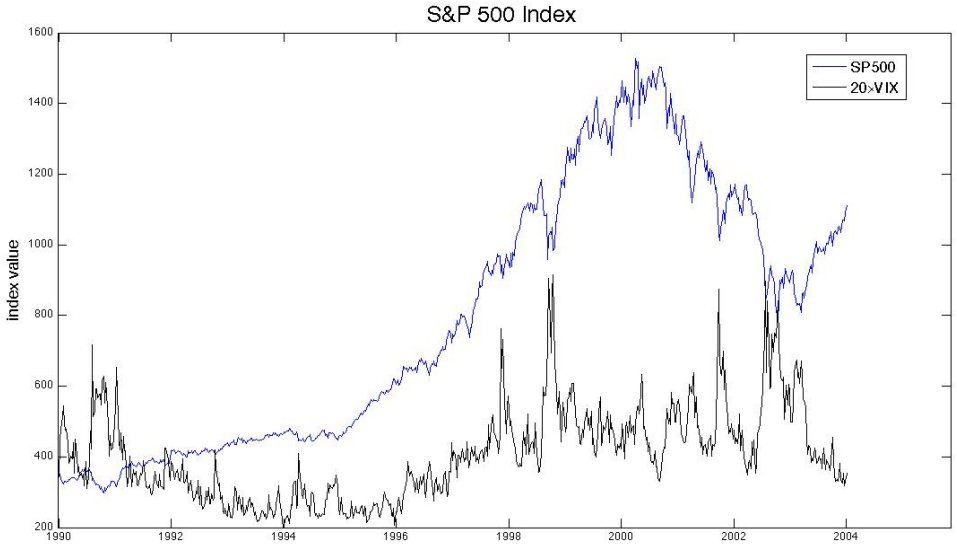

FIGURE 1. The S&P 500 from 1990 to 2004 and the VIX during the same time.

A **spot-estimate** of volatility at time t_k is an $\mathcal{F}_k$-measurable estimate of $\sigma^2_{t_k}$. As a spot -estimate, the VIX almost validates the stochastic properties of equation (2). Plugging the daily S&P 500 data and VIX data from 1990 to 2004 into equation (2), with $\mu_t \equiv \frac{1}{N+1}\sum_{k=0}^{N} Y_k$ and $\Delta t = 1/252$, there is a sequence of residuals

$$Z_k = \frac{Y_k - (\mu_{t_k} - .5VIX^2_{t_k})\Delta t}{VIX_{t_k}\sqrt{\Delta t}}$$

from which the empirical moments for Z_k can be used to validate the model:

$$
\begin{aligned}
\bar{Z} = &\quad \tfrac{1}{N+1}\textstyle\sum_{k=0}^{N} Z_k && = .0224 \\
\eta = &\quad \sqrt{\tfrac{1}{N}\textstyle\sum_{k=0}^{N}(Z_k-\bar{Z})^2} && = .7391 \\
\text{skewness} = &\quad \tfrac{1}{N}\textstyle\sum_{k=0}^{N}(Z_k-\bar{Z})^3/\eta^3 && = .0445 \\
\text{kurtosis} = &\quad \tfrac{1}{N}\textstyle\sum_{k=0}^{N}(Z_k-\bar{Z})^4/\eta^4 - 3 && = .8219
\end{aligned}
\tag{8}
$$

A spot-estimate is considered to have accounted for the risk associated with a Gaussian model if the kurtosis of Z_k is zero. If the kurtosis is greater than zero there is risk that is unaccounted for, and if the kurtosis is less than zero then there is less risk than the model suggests. In the VIX case, the standard deviation η is significantly less than one, but the kurtosis is not extremely high which means that the VIX data accounts for much of the risk associated with the S&P 500.

In general, the VIX is popular because it moves upward as the S&P 500 goes down. The VIX moved opposite the S&P 500 on approximately 59% of the days from 1990 to 2004, and even more between 2003 and 2009.

2. Filtering and Smoothing Estimates of Stochastic Volatility

Stochastic volatility models introduce more precision in the understanding of the volatility process. Several models assume that σ_t is a Markov process, such as an Ornstein-Uhlenbeck or Cox-Ingersol-Ross model [**19**]. In this chapter, σ_t is taken to be a Markov chain on a finite state-space $\mathcal{S} = \{s_1, ..., s_m\}$ with a matrix $Q \in \mathbb{R}^{m\times m}$ of transition rates such that

$$
\mathbb{P}(\sigma_{t_{k+1}} = s_i | \sigma_{t_k} = s_j) = \left(e^{Q^*\Delta t}\right)_{ji} \tag{9}
$$

where Q^* denotes matrix transpose. In general, stochastic volatility modeling introduces an interesting aspect to historical volatility estimation because the process σ_t

can be treated as a Hidden Markov Model (HMM), from which filtering and smoothing estimates can be computed.

There are properties of a discrete-time/discrete-space Markov chain that are particularly nice for modeling a process that is not fully understood.

- The discrete model can be used to approximate a continuous process (such as a diffusion), but can also incorporate jumps. Both of these phenomenon are incorporated into the transition kernel $e^{Q^*\Delta t}$.
- Data is not available in continuum, so it is more natural to model discretely.
- Numerics need to be performed on a discrete model regardless of the model.

In general, a stream of volatility validates the stochastic properties of the model if the noised implied by emprical data is consistent with certain sample statistics such as mean, variance, skew and excess kurtosis. As will be seen in Section 2.4, a discrete Hidden Markov chain model for the stochastic volatility model of equation (2) is validated by the S&P 500 data.

2.1. Filtering. Suppose that σ_t is an unobserved Markov Chain in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For $t \in [0, T]$ let σ_t take value in the state-space $\mathcal{S} = \{s_1, ..., s_m\}$ and make transitions according the rate-matrix $Q \in \mathbb{R}^{m\times m}$. For the discrete data points $(t_k)_{k=0}^{N+1}$ given in (3), the observable data is the sequence $(Y_k)_{k=0}^{N}$ given by (2). For each k let $\mathcal{F}_k$ denote the filtration on $Y_{1:k}$. Given the data $Y_{1:k}$ for any $k \leq N$, the filtering distribution of σ_{t_k} is denoted with $\pi_{k|k}$,

$$\pi^i_{k|k} = \mathbb{P}(\sigma_{t_k} = s_i | Y_{1:k})$$

The filtering estimate of $\sigma^2_{t_k}$ is defined as follows:

$$\widehat{\sigma^2}_{k|k} \quad = \quad \mathbb{E}[\sigma^2_{t_k}|\mathcal{F}_k] = \sum_i s_i^2 \pi^i_{k|k}, \qquad \text{filtering estimate} \tag{10}$$

Through simple manipulation of the expectation operator, it can be shown that (10) is the optimal estimate of $\sigma^2_{t_k}$ among all $\mathcal{F}_k$-measurable functions in terms of mean-squared error (MSE),

$$\mathbb{E}\left(\widehat{\sigma^2}_{k:k} - \sigma^2_{t_k}\right)^2 \leq \mathbb{E}\left(\theta - \sigma^2_{t_k}\right)^2 \qquad \text{for all } \mathcal{F}_k\text{-measurable } \theta.$$

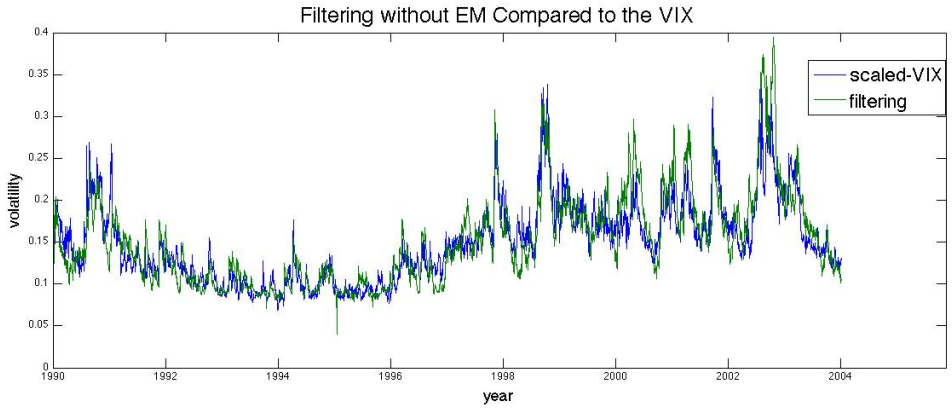

FIGURE 2. Given the S&P 500 data from 1990 to 2004, this plot show the filtering estimate $\sqrt{\mathbb{E}[\sigma^2_{t_k}|\mathcal{F}_k]}$ compared with $\eta \cdot VIX_{t_k}$ where η is the standard deviation from table (8). Although the VIX is not a true spot-estimate of volatility, it is still worthwhile to observe how it relates to filtering.

Filtering is popular because it provides online estimates and does not require any backtracking or recalibration to obtain an updated distribution as new data arrives. Instead, simple linear algebra operations are used to update the filtering distribution at each time t_k. Given the matrix Q of transition rates for σ_t, the posterior distribution is obtained recursively though the **forward Baum-Welch equation:**

$$\pi^i_{k+1|k+1} = \frac{1}{c_{k+1}} \mathbb{P}(Y_{k+1}|\sigma_{t_{k+1}} = s_i) \sum_j \left(e^{Q^* \Delta t}\right)_{ji} \pi^j_{k|k} \tag{11}$$

where c_{k+1} is a normalizing constant, and

$$\mathbb{P}(Y_k|\sigma_{t_k} = s_i) \approx \frac{1}{\sqrt{2\pi s_i^2 \Delta t}} \exp\left\{-\frac{\left(Y_k - \left(\mu_{t_k} - \frac{1}{2}s_i^2\right)\Delta t\right)^2}{2s_i^2 \Delta t}\right\}$$

For the daily S&P 500, a filtering algorithm can be written to track volatility. Assume that there are 252 trading-days per year, and let $\Lambda \in \mathbb{R}^{m \times m}$ be the exponential of Q for $\Delta t = 1/252$,

$$\Lambda = e^{Q^* \Delta t}$$

The filtering algorithm is

$$\pi^i_{k+1|k+1} = \frac{1}{c_{k+1}} \mathbb{P}(Y_{k+1}|\sigma_{t_{k+1}} = s_i) \sum_j \Lambda_{ji} \pi^j_{k|k}$$

In order to implement the algorithm there needs to some some estimates of the parameters μ, Λ and the state-space $\mathcal{S}$. Some possible estimates are $\mu = .08$, $\mathcal{S} = \{.01, ..., .5\}$, and Λ_{ji} taken to be the frequency of jumps from s_j to s_i in the VIX data:

$$\Lambda_{ji} = \frac{\sum_{k=0}^{N-1} \mathbf{1}_{\{\eta \cdot VIX_{t_{k+1}} = s_j,\ \eta \cdot VIX_{t_k} = s_j\}}}{\sum_i \text{numerator}}$$

where η is the sample standard deviation from table (8). As is seen in Figure 2, the filtering estimates with parameters estimated by frequencies of VIX jumps and the VIX itself are quite similar for the sequence of S&P 500 quotes from 1990 to 2004.

2.2. Smoothing. Given all the data $Y_{1:N}$ on $[0, T]$, estimates of σ_t for $t < T$ are referred to as smoothing estimates. If the modeling of the processes is correct, the smoothing estimate will be better than filtering because it is made with knowledge of how future events will unfold.

The smoothing distribution for σ_t given $Y_{1:N}$ is

$$\pi^i_{t|N} = \mathbb{P}(\sigma_t = s_i|Y_{1:N}) \qquad \text{for any } t \in [0, T]$$

and the smoothing estimate of σ^2_t is

$$\widehat{\sigma^2}_{t|N} = \mathbb{E}[\sigma^2_t|\mathcal{F}_N] = \sum_i s^2_i \pi^i_{t|N}, \qquad \text{smoothing estimate} \tag{12}$$

Through simple manipulation of the expectation operator (in a similar fashion to filtering), the estimate in (12) can be used to obtain a better estimate of the average volatility in (4) than the estimate provided by the historical volatility in (5),

$$\mathbb{E}\left(\frac{1}{T}\int_0^T \mathbb{E}[\sigma^2_\tau|\mathcal{F}_N]d\tau - \sigma^2_{[0,T]}\right)^2 \leq \mathbb{E}\left(RV_{[0,T]} - \sigma^2_{[0,T]}\right)^2$$

For any $t \in [t_k, t_{k+1})$, the smoothing distribution for σ_t is obtained through the duality of $\pi_{k|k}$ with a function α_t which weights the future events given the state of σ_t. This is written as

$$\pi^i_{t|N} = \alpha^i_t \sum_j \pi^j_{k|k} \left(e^{Q^*(t-t_k)}\right)_{ji}$$

with α_t given by the **backward Baum-Welch equation**

$$\alpha^i_t = \frac{1}{c_{k+1}} \sum_j \mathbb{P}(Y_{k+1}|\sigma_{t_{k+1}} = s_j) \left(e^{Q^*(t_{k+1}-t)}\right)_{ij} \alpha^j_{t_{k+1}} \qquad \forall t \in [t_k, t_{k+1}) \tag{13}$$

where c_{k+1} is the normalizing constant from the forward Baum-Welch equation (11), and the terminal condition is $\alpha_T \equiv 1$. Intuitively, α_t can be thought of as the probability of all future data conditioned on the current state of σ_t:

$$\alpha^i_t \propto \mathbb{P}(Y_{k+1:N}|\sigma_t = s_i) \qquad \text{for } t \in [t_k, t_{k+1})$$

but equation (13) scales the sequence appropriately so that $\pi_{t|N}$ sums to one.

2.3. Parameter Re-Estimation. From Sections 2.1 and 2.2 it is clear that the HMM's transition rates need to be known or well-estimated. In general, parameter estimation for HMMs needs to be something that is researched and applied with a certain degree of confidence based on a training set. However, if there is an estimate and a fairly large amount of data from the process Y_k, a parameter re-estimation method given by the EM algorithm can improve the initial parameter estimates to fit the data.

For the data points given by (3) with Δt constant, let

$$\Lambda_{ji} = \left(e^{Q^*\Delta t}\right)_{ji} \qquad , j \leq m$$

Given the collected data $Y_{1:N}$ and an initial parameter estimate Λ^0, the goal is to construct a parameter re-estimation algorithm that refines the likelihood ratio of the data with each iteration:

$$\text{for } \ell = 0, 1, 2, 3, ..., \text{ choose } \Lambda^{(\ell+1)} \text{ such that } \frac{\mathbb{P}(Y_{0:N}|\Lambda)}{\mathbb{P}(Y_{0:N}|\Lambda^{(\ell)})} \geq 1$$

Let $\vec{\sigma}$ denote the discrete path of the volatility process,

$$\vec{\sigma} = (\sigma_{t_0}, \sigma_{t_1}, ..., \sigma_{t_N}) \in \mathcal{S}^{N+1}$$

In terms of logarithms, a Jensen inequality ($\log \mathbb{E}X \geq \mathbb{E}\log(X)$) leads to a lower bound written in terms of an expectation,

$$\log \mathbb{P}(Y_{0:N}|\Lambda) - \log \mathbb{P}(Y_{0:N}|\Lambda^{(\ell)})$$

$$= \log \left(\sum_{\vec{v} \in \mathcal{S}^{N+1}} \mathbb{P}(Y_{0:N}|\vec{\sigma} = \vec{v}, \Lambda) \mathbb{P}(\vec{\sigma} = \vec{v}|\Lambda) \right) - \log \mathbb{P}(Y_{0:N}|\Lambda^{(\ell)})$$

$$= \log \left(\sum_{\vec{v} \in \mathcal{S}^{N+1}} \mathbb{P}(Y_{0:N}|\vec{\sigma} = \vec{v}, \Lambda) \mathbb{P}(\vec{\sigma} = \vec{v}|\Lambda) \frac{\mathbb{P}(\vec{\sigma} = \vec{v}|Y_{0:N}, \Lambda^{(\ell)})}{\mathbb{P}(\vec{\sigma} = \vec{v}|Y_{0:N}, \Lambda^{(\ell)})} \right) - \log \mathbb{P}(Y_{0:N}|\Lambda^{(\ell)})$$

$$= \log \left(\mathbb{E} \left[\frac{\mathbb{P}(Y_{0:N}|\vec{\sigma}, \Lambda) \mathbb{P}(\vec{\sigma}|\Lambda)}{\mathbb{P}(\vec{\sigma}|Y_{0:N}, \Lambda^{(\ell)})} \middle| \mathcal{F}_N, \Lambda^{(\ell)} \right] \right) - \log \mathbb{P}(Y_{0:N}|\Lambda^{(\ell)})$$

$$\geq \mathbb{E} \left[\log \left(\frac{\mathbb{P}(Y_{0:N}|\vec{\sigma}, \Lambda) \mathbb{P}(\vec{\sigma}|\Lambda)}{\mathbb{P}(\vec{\sigma}|Y_{0:N}, \Lambda^{(\ell)})} \right) \middle| \mathcal{F}_N, \Lambda^{(\ell)} \right] - \log \mathbb{P}(Y_{0:N}|\Lambda^{(\ell)})$$

$$= \mathbb{E} \left[\log \left(\frac{\mathbb{P}(Y_{0:N}, \vec{\sigma}|\Lambda)}{\mathbb{P}(Y_{0:N}, \vec{\sigma}|\Lambda^{(\ell)})} \right) \middle| \mathcal{F}_N, \Lambda^{(\ell)} \right] = \mathcal{M}(\Lambda, \Lambda^{(\ell)})$$

The EM algorithm chooses $\Lambda^{(\ell+1)}$ which maximizes $\mathcal{M}$,

$$\Lambda^{(\ell+1)} = \arg\max_{\Lambda} \mathcal{M}(\Lambda, \Lambda^{(\ell)}) \qquad \text{for } \ell = 0, 1, 2, 3..... \tag{14}$$

and since $\mathcal{M}(\Lambda^{(\ell)}, \Lambda^{(\ell)}) = 0$ it follows that

$$\mathbb{P}(Y_{0:N}|\Lambda^{(\ell)}) \leq \mathbb{P}(Y_{0:N}|\Lambda^{(\ell+1)}) \qquad \forall \ell \geq 0$$

which implies that $\mathcal{M}(\Lambda, \Lambda^{(\ell)}) \leq -\log \mathbb{P}(Y_{0:N}|\Lambda^{(0)}) < \infty, \ \forall \ell > 0$. Therefore, the EM algorithm produces a monotone and bounded sequence, and so it will converge. However, the limit may only be a *local* maximum.

When maximizing the function $\mathcal{M}(\Lambda, \Lambda^{(\ell)})$, it is subject to the constraints $\sum_r \Lambda_{jr} = 1$ for all $j \leq m$. The first order conditions are then,

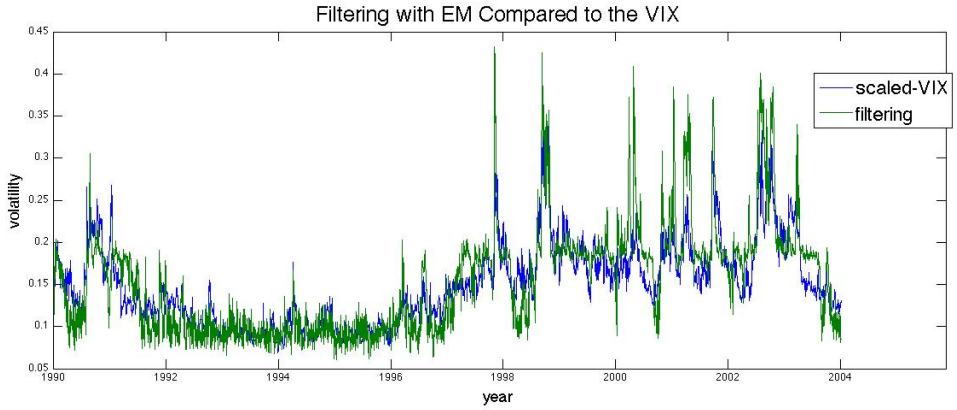

FIGURE 3. Given the S&P 500 data from 1990 to 2004, this plot shows the filtering estimate of $\sqrt{\mathbb{E}[\sigma_{t_k}^2|\mathcal{F}_k]}$ with parameters re-estimated by the EM algorithm, compared to $\eta \cdot VIX_{t_k}$ where η is the standard deviation from table (8).

$$\frac{\partial}{\partial \Lambda_{ji}}\left(\mathcal{M}(\Lambda,\Lambda^{(\ell)}) - \gamma_j \sum_r \Lambda_{jr}\right) = \frac{\partial}{\partial \Lambda_{ji}}\mathcal{M}(\Lambda,\Lambda^{(\ell)}) - \gamma_j = 0 \qquad (*)$$

Multiplying by Λ_{ji} and summing over i the expression in $(*)$ becomes

$$0 = \sum_i \Lambda_{ji}\left(\frac{\partial}{\partial \Lambda_{ji}}\mathcal{M}(\Lambda,\Lambda^{(\ell)}) - \gamma_j\right) = \sum_i \Lambda_{ji}\frac{\partial}{\partial \Lambda_{ji}}\mathcal{M}(\Lambda,\Lambda^{(\ell)}) - \gamma_j$$

which means $\gamma_j = \sum_i \Lambda_{ji}\frac{\partial}{\partial \Lambda_{ji}}M(\Lambda,\Lambda^{(\ell)})$. By multiplying $(*)$ by Λ_{ji} and then rearranging terms it is found that the optimal $\Lambda_{ji}^{(\ell+1)}$ must be chosen among the set of Λ's such that

$$\Lambda_{ji} = \frac{\Lambda_{ji}\frac{\partial}{\partial \Lambda_{ji}}\mathcal{M}(\Lambda,\Lambda^{(\ell)})}{\sum_r \Lambda_{jr}\frac{\partial}{\partial \Lambda_{jr}}\mathcal{M}(\Lambda,\Lambda^{(\ell)})} \qquad (15)$$

Now, the joint log-probability of any path $\vec{v} \in \mathcal{S}^{N+1}$ and the data $Y_{0:N}$ can be written as

$$\log \mathbb{P}(Y_{0:N}, \vec{\sigma} = \vec{v}|\Lambda) = \log \mathbb{P}(Y_0|\sigma_{t_0} = \vec{v}_0) + \log \mathbb{P}(\sigma_{t_0} = \vec{v}_0)$$

$$+ \sum_{k=0}^{N-1} \log \mathbb{P}(Y_{k+1}|\sigma_{t_{k+1}} = \vec{v}_{k+1}) + \log \mathbb{P}(\sigma_{t_{k+1}} = \vec{v}_{k+1}|\sigma_{t_k} = \vec{v}_k, \Lambda)$$

from which, the derivative of $\mathcal{M}(\Lambda, \Lambda^{(\ell)})$ with respect to Λ_{ji} can be computed as follows:

$$\frac{\partial}{\partial \Lambda_{ji}} \mathcal{M}(\Lambda, \Lambda^{(\ell)})$$

$$= \frac{\partial}{\partial \Lambda_{ji}} \mathbb{E}\left[\log \mathbb{P}(Y_{0:N}, \vec{\sigma}|\Lambda) \middle| \mathcal{F}_N, \Lambda^{(\ell)}\right] - \underbrace{\frac{\partial}{\partial \Lambda_{ji}} \mathbb{E}\left[\log \mathbb{P}(Y_{0:N}, \vec{\sigma}|\Lambda^{(\ell)}) \middle| \mathcal{F}_N, \Lambda^{(\ell)}\right]}_{=0}$$

$$= \mathbb{E}\left[\frac{\partial}{\partial \Lambda_{ji}} \log \mathbb{P}(Y_{0:N}, \vec{\sigma}|\Lambda) \middle| \mathcal{F}_N, \Lambda^{(\ell)}\right] = \mathbb{E}\left[\sum_{k=0}^{N-1} \frac{1}{\Lambda_{ji}} \mathbf{1}_{\{\sigma_{t_{k+1}}=s_i, \sigma_{t_k}=s_j\}} \middle| \mathcal{F}_N, \Lambda^{(\ell)}\right]$$

$$= \frac{1}{\Lambda_{ji}} \sum_{k=0}^{N-1} \mathbb{E}\left[\mathbf{1}_{\{\sigma_{t_{k+1}}=s_i, \sigma_{t_k}=s_j\}} \middle| \mathcal{F}_N, \Lambda^{(\ell)}\right] = \frac{1}{\Lambda_{ji}} \sum_{k=0}^{N-1} \mathbb{P}(\sigma_{t_{k+1}} = s_i, \sigma_{t_k} = s_j | Y_{0:N}, \Lambda^{(\ell)})$$

and by plugging this into equation (15), it is easily seen that the solution to (14) is

$$\Lambda_{ji}^{(\ell+1)} = \frac{\sum_{k=0}^{N-1} \mathbb{P}(\sigma_{t_{k+1}}^2 = s_i, \sigma_{t_k}^2 = s_j | Y_{0:N}, \Lambda^{(\ell)})}{\sum_i \text{numerator}} \tag{16}$$

where $\mathbb{P}(\sigma_{t_{k+1}}^2 = s_i, \sigma_{t_k}^2 = s_j | Y_{0:N}, \Lambda^{(\ell)})$ can be computed using α_t and $\pi_{k|k}$ from sections 2.1 and 2.2. It also happens that equation (16) enforces non-negativity of Λ_{ji}, which is required for well-posedness of the algorithm. Figure 3 shows the VIX overlaid by the filtering estimate of $\sigma_{t_k}^2$ with parameters re-estimated by the EM algorithm.

2.4. Comparison of Spot-Estimates. Figures 2 and 3 show plots of various spot-estimates of volatility. For any spot-estimate $\widehat{\sigma^2}_{t_k}$, the residual noise of the data process (2) can be recovered in a similar fashion as was done for the VIX and S&P 500 data in Section 1.2,

$$Z_k = \frac{Y_k - (\mu_{t_k} - .5\widehat{\sigma^2}_{t_k})\Delta t}{\sqrt{\widehat{\sigma^2}_{t_k} \Delta t}}$$

The following table shows the empirical moments obtained from recovering the noise after filtering with parameters estimated by binning the jump-frequencies of theVIX

(no EM), re-estimating with EM (with EM), and those of the VIX:

(17)

moment and formula		no EM	with EM	VIX
$\bar{Z} =$	$\frac{1}{N+1}\sum_{k=0}^{N} Z_k$	.0021	.0020	.0224
$std(Z) =$	$\sqrt{\frac{1}{N}\sum_{k=0}^{N}(Z_k-\bar{Z})^2}$	.9313	0.8770	.7391$(=\eta)$
skewness $=$	$\frac{1}{N}\sum_{k=0}^{N}(Z_k-\bar{Z})^3/std(Z)^3$	$-.1172$	-0.0875	.0445
kurtosis $=$	$\frac{1}{N}\sum_{k=0}^{N}(Z_k-\bar{Z})^4/std(Z)^4-3$	$-.0946$	-0.2214	.8219

From table (17) it is clear that filtering with parameters estimated by the jump frequencies of the VIX has kurtosis closest to zero, and filtering with parameters re-estimated by the EM has kurtosis that is more negative. This is interesting because it suggests that filtering with the EM over-fits the data compared to the filter that uses parameter estimates derived from the VIX. However, it is ad-hoc to use the VIX as a spot-estimate of volatility because it is a prediction of the 30-day average to come. A more appropriate study of filtering and the VIX is the subject of the next section.

3. Comparative Analysis of Filtering to VIX

This sections uses the methods from Sections 1 and 2 to begin a careful study of the relationship between filtering, volatility swaps and the VIX. In particular, the experiments in this section use the algorithms of Sections 2.1, 2.2, and 2.3 to obtain estimates of the average volatility of (4). Then it is argued that the VIX is the fair price of swap contracts based on these estimates, and using the S&P 500 data and the VIX data from 1990 to 2004, it is shown that such contracts would have been of lower variance than two other realized volatility swaps.

3.1. Filtering and Smoothing Estimates of Averaged Volatility. Recall from Section 1 the partition $0 = t_0 < ... < t_{N+1} = T$ of the interval $[0, T]$ and

$$FV_0 := \frac{1}{T}\sum_{k=0}^{N}\mathbb{E}[\sigma^2_{t_k}|\mathcal{F}_n] \approx \frac{1}{T}\int_0^T \sigma^2_\tau d\tau \tag{18}$$

$$SV_0 := \frac{1}{T}\sum_{k=0}^{N}\mathbb{E}[\sigma^2_{t_k}|\mathcal{F}_N] \approx \frac{1}{T}\int_0^T \sigma^2_\tau d\tau \tag{19}$$

The estimate FV in (18) is referred to as the **discrete average of filtering volatility**. Similarly, the estimate SV in (19) is referred to as the **discrete average of smoothing volatility**. Using the fact that $\sigma_t^2 dt = 2\left(\frac{dS_t}{S_t} - d\log S_t\right)$, tt is easily shown that as the partition of $[0,T]$ gets infinitesimally fine, the filtration $\mathcal{F}_N$ fills in with data on S_t, and the limit in the mean of the estimate in (19) is the average volatility:

$$\mathbb{E}_0^*\left[\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}[\sigma_{t_k}^2|\mathcal{F}_N]\right] = \mathbb{E}_0^*\left[\frac{(N+1)\Delta t}{(N+1)T}\sum_{k=0}^{N}\mathbb{E}[\sigma_{t_k}^2|\mathcal{F}_N]\right]$$

$$= \mathbb{E}_0^*\left[\frac{(N+1)}{(N+1)T}\sum_{k=0}^{N}\mathbb{E}[\sigma_{t_k}^2|\mathcal{F}_N]\Delta t\right] = \mathbb{E}_0^*\left[\frac{1}{T}\sum_{k=0}^{N}\left(\int_{t_k}^{t_{k+1}}\mathbb{E}[\sigma_\tau^2|\mathcal{F}_N]d\tau + o(\Delta t)\right)\right]$$

$$= \mathbb{E}_0^*\left[\mathbb{E}\left[\frac{1}{T}\int_0^T \sigma_\tau^2 d\tau\middle|\mathcal{F}_N\right]\right] + O(\Delta t)$$

$$= \mathbb{E}_0^*\left[\mathbb{E}\left[\frac{2}{T}\left(\int_0^T \frac{dS_\tau}{S_\tau} - \log(S_T/S_0)\right)\middle|\mathcal{F}_N\right]\right] + O(\Delta t)$$

$$\xrightarrow{dct} \mathbb{E}_0^*\left[\frac{2}{T}\left(\int_0^T \frac{dS_\tau}{S_\tau} - \log(S_T/S_0)\right)\right] \qquad \text{as } \Delta t \searrow 0$$

$$= \mathbb{E}_0^*\left[\sigma_{[0,T]}^2\right]$$

where '*dct*' stands for 'dominated convergence theorem', and $\mathbb{E}_0^*[\cdot] = \mathbb{E}[M_T \cdot\ ]$ where M_T is the same Martingale change of measure that was used in Section 1.1. Similarly, the same limit holds for the estimate in (18),

$$\mathbb{E}_0^*\left[\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}[\sigma_{t_k}^2|\mathcal{F}_k]\right] \to \mathbb{E}_0^*\left[\sigma_{[0,T]}^2\right] \qquad \text{as } \Delta t \searrow 0$$

but the proof of this convergence is significantly more complicated than the former case.

As an example, the S&P 500 data can be taken with the assumption that there are 252 trading-days per year and Δt= 1 day. For each $k = 0, 1, ..., N-29$, let $\mathcal{F}_k$ denote the filtration on the S&P 500 data up to time t_k, and consider estimate (18) and (19)

with $T = 30$ days. In this case, the $VIX^2_{t_k}$ equals the risk-neutral expectation of (18) and (19) plus a term of $O(\Delta t)$. Therefore, it makes sense to compare the time-series' of FV_{t_k} and SV_{t_k} to VIX_{t_k}. Figure 4 plots these three time-series for both types of parameter estimates, (i.e. both when Λ is re-estimated with the EM algorithm and when Λ is estimated with the jump-frequencies of the VIX.)

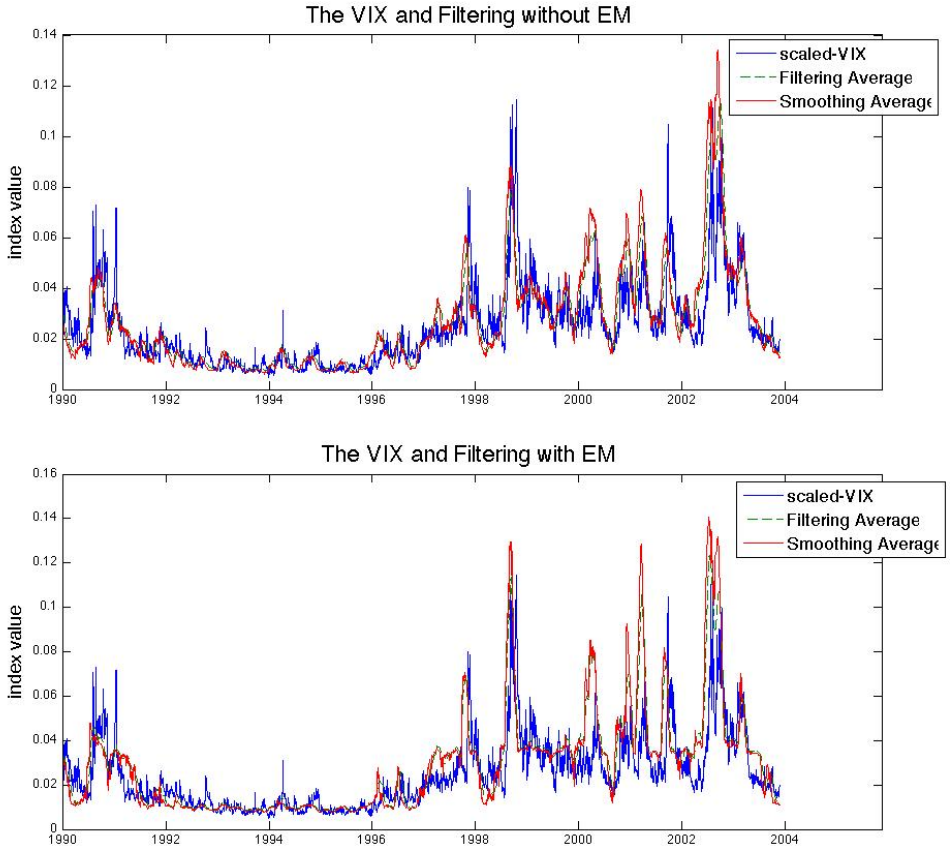

FIGURE 4. Estimates (18) and (19) compared to $\eta \cdot VIX$. The parameter $\eta \approx .73$ is the empirical standard deviation of the recovered noise under the VIX data, as was shown in tables (8) and (17).

There are noticeable differences between the plots in Figure 4 but neither estimate seems to suggest a drastically different volatility average than $\eta \cdot VIX$ had predicted. It is important to note that the VIX needed to be scaled to fit closely with FV and SV, but this is not so alarming because it is often the case that implied volatility is 1.25 to 1.35 times historical estimates. Regardless, it is clear from Figure 4 that filtering and smoothing averages of volatility are synchronized with the VIX.

3.2. Filtering and Smoothing Volatility Swaps. An appropriate way to compare the VIX and filtering is to consider swap contract based on the 30-day discrete average of filtering and smoothing volatility, called **filtering volatility swap contracts** and **smoothing volatility swap contract**, respectively. For a strike price K_{vol}, these swap contracts are defined as the net difference of FV and SV, respectively:

$$\frac{1}{30}\sum_{n=k}^{k+29}\mathbb{E}[\sigma_{t_n}^2|\mathcal{F}_n]-K_{vol}^2, \qquad \text{filtering volatility swap contract} \tag{20}$$

$$\frac{1}{30}\sum_{n=k}^{k+29}\mathbb{E}[\sigma_{t_n}^2|\mathcal{F}_{k+29}]-K_{vol}^2, \qquad \text{smoothing volatility swap contract} \tag{21}$$

Under the same risk neutral measure which swaps on 30-day historical volatility were priced, the 30-day filtering swap is priced by its risk-neutral expectation for an infinitesimally small partition, which means that the 'fair' value of the swap is initially given as follows:

$$\text{fair price of filtering vol swap}|_{t=0}=e^{-r30/252}\mathbb{E}_{t_k}^*\left[\frac{1}{30}\sum_{n=k}^{k+29}\mathbb{E}[\sigma_{t_n}^2|\mathcal{F}_n]-K_{vol}^2\right]$$

$$\approx e^{-r30/252}\mathbb{E}_{t_k}^*\left[\frac{1}{30}\int_0^{30/252}\sigma_{t_k+\tau}^2 d\tau-K_{vol}^2\right]=e^{-r30/252}\left(VIX_{t_k}^2-K_{vol}^2\right)$$

and the same idea applies to smoothing volatility swaps.

The data on the S&P 500 shows that returns on both 30-day filtering and 30-day smoothing volatility swaps have lower variance than returns on swaps based on historical volatility and priced either by the VIX or mark-to-market (a mark-to-market price for a swap is simply last month's historical volatility as the fair price for the historical volatility of the coming month.) However, filtering and smoothing swaps are sufficiently more complex because the outcome depends heavily on the parameter choice of $\Lambda=e^{Q^*\Delta t}$. In fact, if Λ is estimated on the S&P 500 data from 1990 to 2004 and then outcomes of filtering and variance swaps are computed over the same time period, it turns out that there is higher returns and less variance for

these contracts than there is for the historical volatility swap. Such an occurrence is rather alarming because it indicates a major shift in the mean-variance frontier for attainable portfolios, but since the parameters where computed in hindsight, the phenomenon can most likely by regarded as over-fitting of the data. The mean and variance of returns for the different swap contracts with time to maturity of 30-days and at-the-money strikes (i.e. return = historical vol - VIX) are shown in table (22).

(22)

Mean and Variance of Returns on 30-day at-the-money Volatility Swaps on the S&P 500

swap type:	mean:	variance:
realized vol.	-1.7092%	0.0533%
mark-to-market	$-.0050\%$	$.0626\%$
smoothing (with EM)	-1.5955%	$.0433\%$
filtering (with EM)	-1.6069%	$.0343\%$
smoothing (no EM)	-1.8193%	$.0372\%$
filtering (no EM)	-1.8565%	$.0318\%$

It is well-known that swap contracts favor the writer, and this is corroborated by Table (22) where empirical data suggests that returns on any of these four swap contracts is expected to be negative. With regard to the hypothesis of this chapter, the important thing to notice in Table 22 is parity in the form of a mean/variance trade-off (this parity does not hold for output that is corrupted by the over-fitting caused by using the EM algorithm.) For instance, it is obvious that the mark-to-market swap contract can be expected to have the least net-loss, but is clearly the most risky of all four contracts. Another example of the parity is seen by comparing filtering with EM to smoothing with EM, where it is clear that the higher expected returns of the smoothing volatility contract comes at a cost in variance.

3.3. Statistical Fitting of Swap Returns as a function of the VIX. In this last section, a statistical analysis of the VIX's ability to predict the 30-day historical

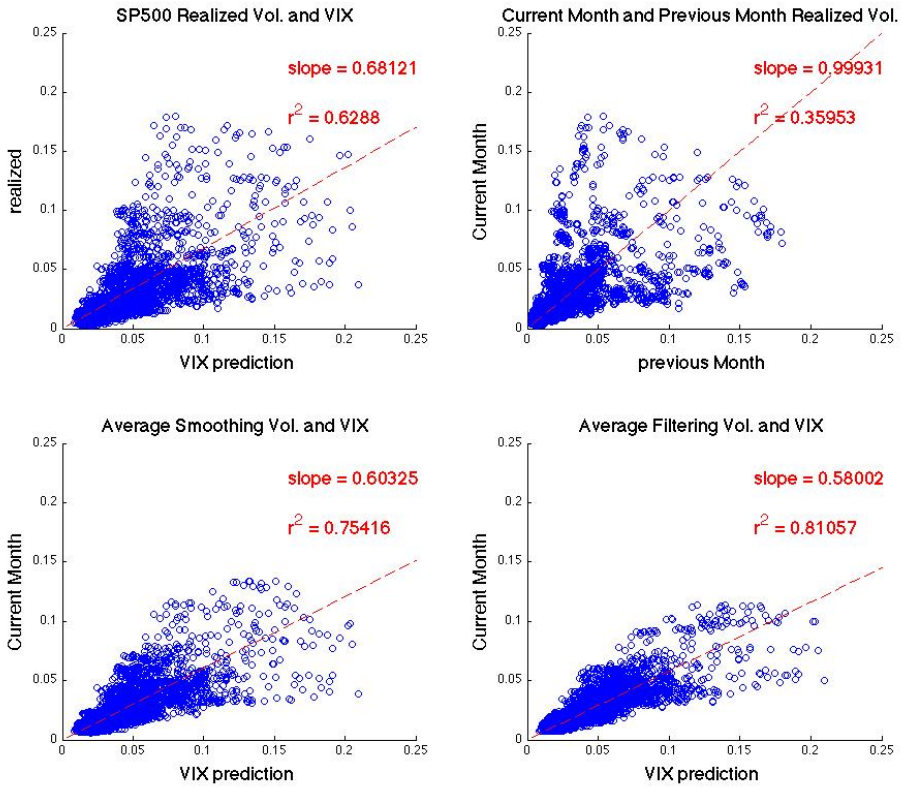

FIGURE 5. Top Left: $RV_{[0,T]}$ plotted as a function of VIX. Top Right: $RV_{[0,T]}$ for the current month plotted as a function of $RV_{[0,T]}$ for the previous month, which shows the returns on the mark-to-market swap contracts. Bottom Left: SV plotted as a function of the VIX with Λ estimated by the jump frequencies of the VIX. Bottom Right: FV plotted as a function of the VIX also with Λ estimated by the jump frequencies of the VIX.

volatility (RV), the 30-day discrete average of filtering volatility (FV), and the 30-day discrete average of smoothing volatility (SV). The analysis considers each time series as a linear function of the VIX time-series.

Consider a linear interpolation of an estimate $\widehat{\sigma^2}_{t_k}$ with its domain taken to be the VIX:

$$\hat{\ell} = \arg\min_{\ell \in \mathbb{R}} \sum_{k=0}^{N} \left| \ell \cdot VIX_{t_k}^2 - \widehat{\sigma^2}_{t_k} \right|^2 \tag{23}$$

The amount of variance in the sequence $(\widehat{\sigma^2}_{t_k})_{k=0}^N$ that is accounted for by the linear fit in (23) can be summarized by the projection error,

$$r^2 = 1 - \sqrt{\frac{\sum_{k=0}^N \left| \hat{\ell} \cdot VIX_{t_k}^2 - \widehat{\sigma^2}_{t_k} \right|^2}{\sum_{k=0}^N |\widehat{\sigma^2}_{t_k}|^2}}$$

The dotted red line in Figure 5 show this linear interpolation for estimates made using the jump frequencies of the VIX, and the dotted red line in Figure 6 shows the linear fit for estimates made with the EM re-estimation algorithm. What is important to notice in the plots is that $\hat{\ell}$ goes down as r^2 goes up, except for the cases when there is over-fitting of the EM algorithm. Loosely put, this inverse relation between $\hat{\ell}$ and r^2 suggests that the expected returns on a swap contract will go down as the amount of variance in returns that is accounted for by the VIX goes up. This trend can be related to Table (22) where a mean/variance trade-off was evident.

4. Chapter Summary

To summarize, this chapter applied the nonlinear filtering algorithms in Section 2 to the daily S&P 500 and VIX data from 1990 to 2004. It was then shown that the filtering and smoothing estimates could track the VIX to some extent, either in the sense of spot-estimates or as a 30-day average. The main result was in Section 3 where it was also shown that there could possibly be a lower variance investment if one were to enter into a swap contract that pays the net difference of the 30-day filtering volatility minus the strike price. Future work would need to explore different indices and different time periods, and if those experiments produce similar results as were found in this chapter it would on an empirical basis strengthen the claim that these are lower variance investments.

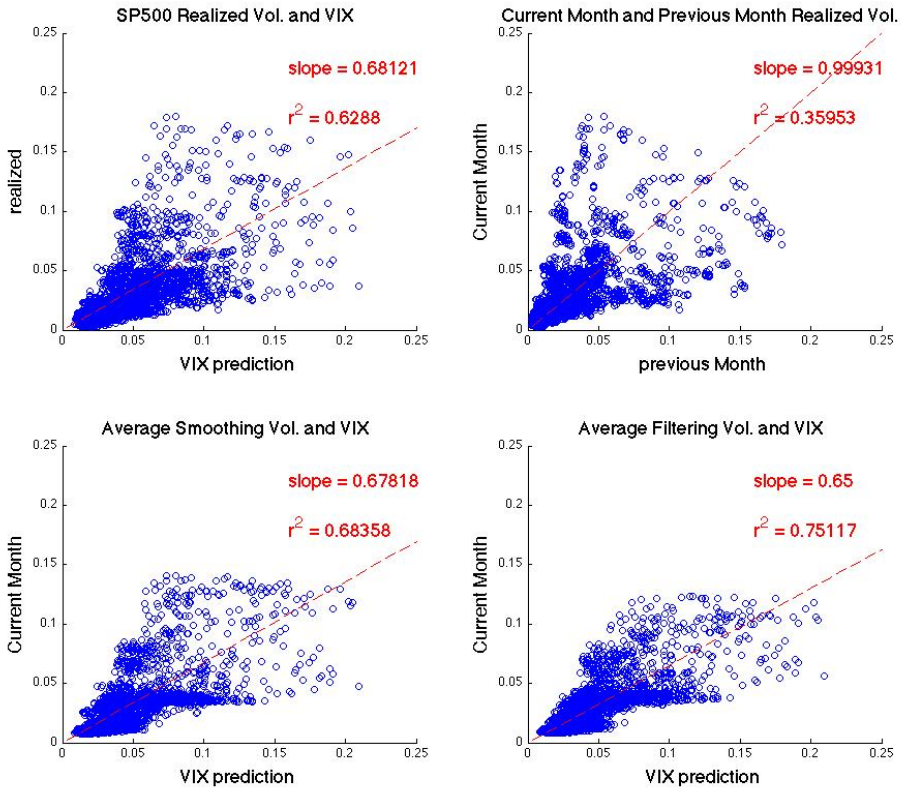

FIGURE 6. Top Left: $RV_{[0,T]}$ plotted as a function of VIX. Top Right: $RV_{[0,T]}$ for the current month plotted as a function of $RV_{[0,T]}$ for the previous month. Bottom Left: SV plotted as a function of the VIX with Λ re-estimated by EM. Bottom Right: FV plotted as a function of the VIX also with Λ re-estimated by EM.

CHAPTER 3

Tracking Intentions from Streaming Video: A Double HMM Approach

In this chapter we discuss problems related to the identification/prediction of behavioral characteristics of an intelligent agent (e.g. a human, or a robot) based on noisy video footage. The attributes of interest include: patterns of motion (e.g. pose and position), intentions, mood swings etc. Some of the attributes, (e.g. intensions) are not directly observable, and need to be inferred from the other attributes. Others, such as pose and location, are partially observable but the observations are corrupted by noise.

The problem is that the only observations available are visual, which means intentions are not directly observable and need to be inferred from the other attributes. In addition, visibility is poor making the observations very noisy which in turn makes it difficult to observe attributes such as pose and location. An agent's attributes can be modeled as a stochastic process with values in a suitable state-space, and whose temporal dynamics and observations are described by a Hidden Markov Model (HMM). In particular, a hierarchical structure for the state-space can be modeled by a Double Markov Chain (DMC) whose structure allows for an iterative conditioning of the distribution of circumstantial attributes on the values taken by more basic ones. For instance, a person's pose is a random variable selected from a set of possible poses, but the distribution with which he/she takes a particular pose depends on their intentions, which can be modeled as components of the DMC at another level in the hierarchy.

Any approach to this problem will include an algorithm which returns the "best" estimate of the attributes as information arrives. We interpret "best" to mean optimal in the sense of minimal posterior expected loss where the optimal estimate is

selected from the set of parameters that are measurable with respect to the σ-algebra generated by the observed data. In other-words, given the DMC parameters and a set of observations, the optimal estimate minimizes the posterior expectation of some loss function of the state-vector. The posterior distribution is needed to compute the posterior expectation. The computation of the posterior can be done with the Forward Baum-Welch Equations of nonlinear filtering [**25, 32**].

The use of DMCs for tracking intentions and other attributes is useful not only in modeling but also in computations. The iterative conditioning of the distribution of the state-space vector allows for samples to be drawn from the distribution of a high-dimensional state-process. For instance, if the vector of attributes is very long, the probability mass at each state-space element may be well below the threshold of machine precision, thereby causing sampling algorithms to return an error. However, the iterative structure of the DMC allows for sampling algorithms to iteratively determine the elements of a high-dimensional vector on a component-wise schedule, thereby avoiding densities or mass functions that are too small to be relevant to a machine.

Given all the necessary priors such as the parameters of the DMC, there are still several issues that make implementation of the nonlinear filter difficult. For instance it is well-known that, in general, the number of computations required to update with the forward Baum-Welch Equations is of the order of the square of the number of elements in the state-space, which is often very time-consuming to compute. This issue can be addressed with sparse vector/matrix toolboxes. In particular, by a truncation technique known as stochastic domain pursuit (SDP) [**31**], and also by particle filters [**2, 9**]. Another difficulty arises when observations are of high enough dimension (e.g. observations are 2-D images) so that the likelihood of any state-space element is below the machine precision threshold. For such cases it helps to have an observation model whose noise is of an exponential form with a statistic that is within a range that can be measured by a machine. Then, the forward Baum-Welch equation

can be rewritten in logarithmic form so that the output is significantly greater than zero.

Past work that is relevant to this paper dates back to the early 1970's when speech recognition software was developed. The algorithms present in the Hearsay II software from 1977 [**26**] are all based on hierarchical models. The 1989 paper by Rabiner [**32**] gives an overview of HMMs in speech recognition and serves as a good primer for HMMs and filtering.

Target tracking is an area closely related to the research presented in this chapter. The 1993 book by Bar-Shalom and Li [**3**] presents an introduction to interactive multiple models (IMMs), which are closely related to the DMCs. The notion of DMCs is similar to the layering of hidden parameters in the target tracking papers by Rozovsky, Tartakovsky and Yaralov [**34**] and Rozovsky and Petrov [**31, 6**]. Another important set of references from target tracking are those by Gustafsson et al [**22, 36**], in which they formulate and implement Rao-Blackwellized particle filters in tracking nautical, ground and aerial vehicles.

Computer vision is another area that is closely related to the work in this chapter, primarily through works that have sought ways to recognize pose from 2-D image data. The tracking of people in sequences of 2-D images was analyzed in the works of Black [**38, 29**] and Cremers [**14**], and in [**21**] a person's 3-D pose was shown to be recoverable from a single 2-D image. Variations in observation models such as Osher's multiplicative noise model [**37**] are also a critical part of the vision component of this work.

This chapter presents a detailed analysis of a class of nonlinear filtering problems from beginning to end. The model and filtering algorithm are presented in Section 1. DMCs for tracking intentions are introduced in Section 1.2 by hierarchically building a description of an agent's behavior, poses and actions conditioned on their intentions. In Section 1.1 a nonlinear filtering algorithm is formulated using a set of forward Baum-Welch equations that are specialized to DMCs. In Section 2 some of the numerical difficulties in computing a nonlinear filter are discussed. Section 2.1

introduces some implementation techniques that are specific to problems where the observable data is a sequence of images, and Section 2.2 discussed some techniques, such as sparse vector/matrices and stochastic domain pursuit, which save memory and processing-time when there are a lot of elements in the state-space of the DMC. Section 2.3 describes a setting in which the algorithm may be written for parallel processing, for which a particle filter is formulated. Section 2.5 explains how the iterative conditioning property of DMCs may allow for sampling from a high-dimensional state-space. Finally, Section 3 presents a detailed example with modeling, algorithm formulation, and computational results so that it can be assessed how well the algorithm works.

1. Models and Algorithms for Tracking Intentions from Image Data

As noted in the Introduction, double Markov Chains (DMCs) can be used to estimate and track the unobservable intentions of a robotic agent or person. In this setting intentions, pose, location and observations are modeled as a DMC in which the top level is observable and each lower level is an attribute that is slightly less observable. Throughout this chapter, let k be a non-negative integer that denotes time. A DMC is a stochastic process constructed in the following way:

For any integer $N > 0$ consider the set of finite spaces $\{\mathcal{S}^n\}_{n=0}^N$, and let $(\theta_k)_{k=0,1,...}$ be a stochastic process

$$\vec{\theta}_k \in \mathcal{S}^0 \times ..... \times \mathcal{S}^N$$

with k being a time index. In vector form, we write

$$\vec{\theta}_k = (\theta_k^0, ..., \theta_k^N)$$

where $\theta_k^n \in \mathcal{S}^n$ for all $n \leq N$. We say that $\vec{\theta}_k$ is a **double Markov Chain (DMC)** if for any $k > 0$ and $n \leq N$,

$$\mathbb{P}(\theta_k^n = s | \{\theta_k^\ell\}_{\ell=0}^{n-1}, \vec{\theta}_{k-1}, ..., \vec{\theta}_0) = \mathbb{P}(\theta_k^n = s | \theta_k^{n-1}, \theta_{k-1}^n)$$

Such a process can be illustrated by the following directed graph:

$$\begin{array}{ccccccccc}
\dots & \theta^0_{k-1} & \to & \theta^0_k & \to & \theta^0_{k+1} & \to & \theta^0_{k+2} & \dots \\
 & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\dots & \theta^1_{k-1} & \to & \theta^1_k & \to & \theta^1_{k+1} & \to & \theta^1_{k+2} & \dots \\
 & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\dots & \theta^2_{k-1} & \to & \theta^2_k & \to & \theta^2_{k+1} & \to & \theta^2_{k+2} & \dots \\
 & \vdots & & \vdots & & \vdots & & \vdots & \\
 & \vdots & & \vdots & & \vdots & & \vdots & \\
\dots & \theta^{N-1}_{k-1} & \to & \theta^{N-1}_k & \to & \theta^{N-1}_{k+1} & \to & \theta^{N-1}_{k+2} & \dots \\
 & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\dots & \theta^N_{k-1} & \to & \theta^N_k & \to & \theta^N_{k+1} & \to & \theta^N_{k+2} & \dots
\end{array}$$

DEFINITION 1. ***Homogeneous Double Markov Chain.*** *Let $\vec{\theta}_k = (\theta^0_k, ..., \theta^N_k)$ be a DMC. We say that $\vec{\theta}_k$ is a homogenous DMC if there is a set of transition kernels $\{\Lambda^n\}_{n=0}^N$ such that*

(1) $\mathbb{P}(\theta^0_k = s|\theta^0_{k-1} = s') = \Lambda^0(s|s'), \qquad \forall s, s' \in \mathcal{S}^0$

(2) $\mathbb{P}(\theta^n_k = s|\theta^n_{k-1} = s', \theta^{n-1}_k = r) = \Lambda^n(s|s', r)$, $\forall s, s' \in \mathcal{S}^n$, $\forall r \in \mathcal{S}^{n-1}$ *and for all* $n > 0$.

An advantage of DMCs as compared to ordinary vector-valued Markov chains are the iterative structure that the double Markov property yields in computing forward probabilities:

PROPOSITION 2. *Let $\vec{\theta}_k = (\theta^0_k, ..., \theta^N_k)$ be a homogeneous DMC with the set of transition kernels $\{\Lambda^n\}_{n=0}^N$. For all $0 \le n \le N$, define a partially-updated distribution as*

$$u^n_k(\vec{s}) = \mathbb{P}(\theta^0_k = s^0, ..., \theta^n_k = s^n, \theta^{n+1}_{k-1} = s^{n+1}, ..., \theta^N_{k-1} = s^N)$$

where $\vec{s} = (s^1, ..., s^N)$. Then there is the following iterative update scheme for each component of θ_k:

- $u^0_k(\vec{s}) = \sum_{r \in \mathcal{S}^0} \Lambda^0(s^0|r) u^N_{k-1}\left((r, s^1, ..., s^N)\right)$

- $u_k^n(\vec{s}) = \sum_{r \in \mathcal{S}^n} \Lambda^n(s^n|r, s^{n-1}) u_k^{n-1}\left((s^1, ..., s^{n-1}, r, s^{n+1}, ..., s^N)\right)$ *for* $1 \leq n \leq N$,

PROOF. (see appendix.) □

The Markov chain $\vec{\theta}_k = (\theta_k^0, ..., \theta_k^N) \in \mathcal{S}^0 \times ..... \times \mathcal{S}^N$ is defined by its initial distribution π_0 where

$$\pi_0(\vec{s}) = \mathbb{P}(\vec{\theta}_0 = \vec{s}), \qquad \forall \vec{s} \in \mathcal{S}^0 \times ..... \times \mathcal{S}^N$$

and the transition probability kernel $\mathbf{\Gamma}$ where

$$\Gamma\left(\vec{s}|\vec{r}\right) = \mathbb{P}\left(\vec{\theta}_k = \vec{s}|\vec{\theta}_{k-1} = \vec{r}\right), \qquad \text{for all } k = 1, 2, 3, 4, ...$$

Some obvious relations are

- $0 \leq \Gamma(\vec{s}|\vec{r}) \leq 1, \qquad \forall \vec{s}, \vec{r} \in \mathcal{S}^0 \times ..... \times \mathcal{S}^N$
- $\sum_{\vec{s}} \Gamma(\vec{s}|\vec{r}) = 1, \qquad \forall \vec{r}$

and any $\vec{s}, \vec{r} \in \mathcal{S}^0 \times .... \times \mathcal{S}^N$ the transition kernel of a DMC is defined as

$$\Gamma(\vec{s}|\vec{r}) = \Lambda^0(s^0|r^0)\Pi_{n=1}^N \Lambda^n(s^n|r^n, s^{n-1}) \tag{24}$$

so that

$$\mathbb{P}(\vec{\theta}_k = s_i) = \sum_j \Gamma(\vec{s}_i|\vec{s}_j)\mathbb{P}(\vec{\theta}_{k-1} = s_j) \tag{25}$$

In practice, forward computations of the distribution of a DMC will be more efficiently computed using the iterative form from proposition 2 rather than the explicit representation from equation (25). This is because the DMC structure allows for a minimization of the intra-dependence of the vector components, whereas for vector-valued Markov chains there will be an exponentially growing amount of conditioning needed with the addition of new dimensions. We use here the compact notation of Γ because it is convenient for further understanding of how these processes work. Indeed, letting $u_k(\vec{s}) = \mathbb{P}(\vec{\theta}_k = \vec{s})$ for all $\vec{s} \in \mathcal{S}^0 \times ..... \times \mathcal{S}^N$, we have the following equation matrix/vector notation the forward Kolmogorov equation:

$$\mathbf{u}_k = \mathbf{\Gamma}\mathbf{u}_{k-1}$$

where $\mathbf{u}_k = (u_k(\vec{s}_1), ..., u_k(\vec{s}_\mathbf{n}))^*$ with $\{\vec{s}_1, ..., \vec{s}_\mathbf{n}\}$ is the finite set of all vectors in the space $\mathcal{S}^0 \times ..... \times \mathcal{S}^N$.

Joint distributions of $\vec{\theta}_k$ at different times can be evaluated using only the initial distribution π_0 and the matrix of transition probabilities $\mathbf{\Gamma}$. Specifically, $\mathbf{\Gamma}$ is an $\mathbf{n} \times \mathbf{n}$ matrix whose entries are associated with pairs of state-space elements, and the probability of any path $(\vec{\theta}_0, ..., \vec{\theta}_k) = (\vec{s}_{i_0}, ..., \vec{s}_{i_k})$ can be written as a product of Γ,

$$\mathbb{P}\left(\vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right) = \Gamma(\vec{s}_{i_k}|\vec{s}_{i_{k-1}}).....\Gamma(\vec{s}_{i_1}|\vec{s}_{i_0})\pi_0\left(\vec{s}_{i_0}\right) \tag{26}$$

This formula follows from the forward Kolmogorov equation.

In addition to the *state process* given by the Markov chain $\left(\vec{\theta}_k\right)_{k=0,1,...}$ we will consider the *observation process* $(Y_k)_{k=0,1,...}$ taking values either in a finite set F or in $\mathbb{R}^d$. Often we will consider the observation process of the form

$$Y_k = h\left(\vec{\theta}_k\right) + Z_k$$

where $h(\cdot)$ is a function on $\mathcal{S}^0 \times ..... \times \mathcal{S}^N$ and $(Z_k)_{k=0,1,2,...}$ is a discrete time "white noise" (sequence of independent identically distributed random variables.)

Suppose that Z_k takes values in discrete set F, the the likelihood function of $\{\vec{\theta}_k = s_i\}$ given $\{Y_k = y_k\}$ is

$$\psi_{\vec{s}}(y_k) = \mathbb{P}\left(Y_k = y_k|\vec{\theta}_k = \vec{s}\right).$$

Assuming that $(Z_k)_{k=0,1,2,...}$ and $(\theta_k)_{k=0,1,2,...}$ are independent, for any $y_0, ..., y_k \in F$, we have

$$\mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k|\vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right) = \Pi_{\ell=0}^k \psi_{\vec{s}_\ell}\left(y_\ell\right). \tag{27}$$

Indeed, given $\left(\vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right)$, the random variables $(Y_0, ..., Y_k)$ are independent. Therefore,

$$\mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k | \vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right) =$$

$$\Pi_{\ell=0}^{k} \mathbb{P}\left(Y_\ell = y_\ell | \vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right) =$$

$$\Pi_{\ell=0}^{k} \mathbb{P}\left(Y_\ell = y_\ell | \vec{\theta}_\ell = \vec{s}_{i_\ell}\right) = \Pi_{\ell=0}^{k} \psi_{\vec{s}_\ell}\left(y_\ell\right).$$

If $(Z_k)_{k=0,1,2,...}$ is an sequence iid sequence of mean-zero Gaussian random variables in $\mathbb{R}^d$ with covariance matrix R, it is readily checked that a similar relation holds. Specifically,

$$p\left(y_0, ..., y_k | \vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right) = \Pi_{\ell=0}^{k} \psi_{\vec{s}_\ell}\left(y_\ell\right) \tag{28}$$

where $p\left(y_0, ..., y_k | \vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right)$ is a joint probability density of random variables $Y_0, ..., Y_k$ given $\vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}$. Note, that in the latter case

$$\psi_{\vec{s}}\left(y_k\right) = \frac{1}{(2\pi|R|)^{d/2}} \exp\left\{-\frac{1}{2}\left(y_k - h\left(\vec{s}\right)\right)^* R^{-1}\left(y_k - h\left(\vec{s}\right)\right)\right\}.$$

Indeed,

$$\mathbb{P}\left(Y_k \leq y | \vec{\theta}_k = \vec{s}\right) = \mathbb{P}\left(h\left(\vec{s}\right) + Z_k \leq y\right).$$

Obviously, $h\left(\vec{s}\right) + Z_k$ is Gaussian random variable with mean $h\left(\vec{s}\right)$ and covariance matrix R. Therefore, the probability density function of $\mathbb{P}\left(Y_k \leq y | \vec{\theta}_k = \vec{s}\right)$ is the Gaussian density

$$\frac{1}{(2\pi|R|)^{d/2}} \exp\left\{-\frac{1}{2}\left(y_k - h\left(\vec{s}\right)\right)^* R^{-1}\left(y_k - h\left(\vec{s}\right)\right)\right\}.$$

Relations (27), (28) hold for many types of observation processes. Moreover, as we will see below, these relations are the only properties of the observation process that matter. Therefore it is convenient to assume that (27) or (28) holds without specifying the exact structure of the observation process.

The relation (27) and (28) are often referred to as the memoryless channel property. This term is used to indicate that it is characteristic for an observation processes to depend only on the state and the noise at the current time and not on its own history. For example, (27), (28) do not hold if the observation process is of the form

$Y_k = h(\theta_k, Y_{k-1}) + Z_k$. In general, let $\mathbf{\Psi}(y_k)$ be the $\mathbf{n} \times \mathbf{n}$ diagonal matrix with the non-zero entries $\psi_i(y_k) = \psi_{\vec{s}_i}(y_k)$ where $\{\vec{s}_1, ..., \vec{s}_\mathbf{n}\}$ is the finite set of all vectors on the space $\mathcal{S}^0 \times ..... \times \mathcal{S}^N$. We will see that the triplet $\left(\pi_0, \{\Lambda^n\}_{n=0}^N, \mathbf{\Psi}\right)$ is sufficient for description of the state and observation processes $\vec{\theta}_k$ and Y_k in the same way as $\left(\pi_0, \{\Lambda^n\}_{n=0}^N\right)$ defines the double Markov chain $\vec{\theta}_k$.

1.1. Forward Baum-Welch Equation. To further reduce computational complexity as it is usually done in the theory of HMMs, we introduce a recursive linear equation known as the Forward Baum-Welch Equation. Write the joint probability of the observations and the state:

$$p\left(y_0, ..., y_k, \vec{\theta}_k = \vec{s}\right) := \partial \mathbb{P}\left(Y_0 \leq y_0, ..., Y_k \leq y_k, \vec{\theta}_k = \vec{s}\right) / \partial y_1 ... \partial y_k$$

and define

$$\phi_k(\vec{s}) = \begin{cases} \mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k, \vec{\theta}_k = \vec{s}\right) \text{ in discrete case,} \\ \\ p\left(y_0, ..., y_k, \vec{\theta}_k = \vec{s}\right) \text{ in continuous case.} \end{cases}$$

PROPOSITION 3. *Let* $\mathbf{\Phi}_k = (\phi_k(\vec{s}_1), ..., \phi_k(\vec{s}_\mathbf{n}))^*$. *The sequence* $(\phi_k)_{k=0,1,2,...}$ *verifies the following equation*

$$\phi_{k+1}(\vec{s}) = \psi_{\vec{s}}(y_{k+1}) \sum_{\vec{r}} \Gamma(\vec{s}|\vec{r}) \phi_k(\vec{r}) \tag{29}$$

or, in the matrix notation, $\mathbf{\Phi}_{k+1} = \mathbf{\Psi}(y_{k+1}) \mathbf{\Gamma} \mathbf{\Phi}_k$. *The posterior distribution of the current state given the past can be written in terms of* ϕ*:*

$$\pi_k(\vec{s}) = \mathbb{P}\left\{\vec{\theta}_k = \vec{s} | Y_0 = y_0, ..., Y_k = y_k\right\} = \phi_k(\vec{s}) / \textstyle\sum_{\vec{r}} \phi_k(\vec{r}).$$

PROOF. (see appendix.) □

Remark: The theory presented so far has been for DMCs, but could be generalized for hierarchical Markov chains of the telescoping form (i.e. intra-dependence of vector components is a function of all states that are higher in the hierarchy and not just the one state that is immediately above.) However, the telescoping structure will

exponentially increase the parameterization of the matrices $\{\Lambda^n\}_{n=0}^N$ and will put an enormous strain on the memory resources of the machine used to compute the filter.

1.2. A Tracking Intentions Example. A three-layered DMC $(\theta^0, \theta^1, \theta^2)$ can be used to model a person's intentions, actions and location. The least observable component of the double Markov model are the intentions, $\theta^0 \in \{0, ..., M\}$ with initial distribution $\rho = \{\rho_0, ..., \rho_M\}$ and θ^0 not changing for $k > 0$, that is, there are no transitions in time and so it is a static component. For example, θ^0 can be in one of two states: either benign ($\theta^0 = 0$) or hostile ($\theta^0 = 1$) with initial probabilities $\rho = (\rho_0, \rho_1)$.

The next component, θ_k^1, is the pose of the person at time k. This component takes values in a finite set of poses, $\theta_k^1 \in \{s_1, s_2, s_3, ..., s_m\}$ where $m < \infty$. Changes in pose are controlled by a Markov chain whose transition laws depend on the person's intentions:

$$\mathbb{P}(\theta_k^1 = s_i | \theta_{k-1}^1 = s_j, \theta^0 = \ell) = \Lambda(i|j, \ell), \qquad \text{for } \ell = 0, 1, 2...., M$$

The third component of the DMC, θ_k^2, is the spatial location of the person in a two dimensional set. We could also have a more elaborate model that tracks not only position, but also the dilation and rotation of the pose (this is discussed more in Section 2). To adhere to the conventions of target tracking, we denote θ_k^2 with X_k to indicate the tracking of spatial location. The evolution of X is dependent on θ_k^1 and an independent noise:

$$X_k = A(\theta_k^1, X_{k-1}) + \sigma W_k$$

where W_{k+1} are iid standard normal random variables.

A 2-D snap-shot is the observable data based on the person's visual appearance. It is denoted with Y_k and is a matrix of pixel data that is either 2-D (black and white observations) or 3-D (color observations). The observation fits a model of the form:

$$Y_k = H(\theta_k^1, X_k) + Z_k$$

TABLE 1. **Tracking Intentions Example**

intentions: θ^0 labels the intent of the person (e.g. $\{\theta^0 = 0\}$ if benign, or $\{\theta^0 = 1\}$ if hostile)

$$\mathbb{P}(\theta^0 = \ell) = \rho_\ell, \text{ for } \ell = 0, \dots M$$

↓

pose: θ^1_k labels the pose taken by the person (e.g. waiving, nodding, walking, running, jumping,.....)

$$\mathbb{P}(\theta^1_k = s_i | \theta^1_{k-1} = s_j, \theta^0 = \ell) = \Lambda(i|j, \ell)$$

for s_i, s_j in the space of poses.

↓

spatial location: X is the spatial location of the person; changes to X occur depending on θ_1 (e.g. person moves faster if in running pose, doesn't move in waiving pose)

$$X_k = A(\theta^1_k, X_{k-1}) + \sigma W_k$$

where W is an iid random variable in $\mathbb{R}^d$

↓

visual observation: Y is a snapshot of the person at time k. It depends only on the visual appearance of the person, so intentions do not have a direct effect.

$$Y_k = H(\theta^1_k, X_k) + Z_k$$

where Z is a matrix of iid standard normal noise.

where H is a known function, and Z_k is a matrix of iid Gaussian noise with covariance structure R. This example is summarized in Table 1.2.

The kernel for the forward operator semi-group for this example can be written explicitly as

$$\Gamma(x, i_1, i_0 | x', i_1', i_0') \propto \exp\left(-\frac{\|x - A(s_{i_1}, x')\|_2^2}{2\sigma^2}\right) \Lambda(i_1 | i_1', i_0) \mathbf{1}_{\{i_0 = i_0'\}}$$

where $x, x' \in \mathbb{R}^2$, i_1, i_1' are indices labeling the set of poses, and $i_0, i_0' \in \{0, \dots M\}$. The exponential arises because X is conditionally Gaussian and the indicator arises because θ_0 is constant over time. The likelihood can be written explicitly as

$$\psi_{i_1,x}(y_k) = \exp\left(-\frac{1}{2}(y_k - H(s_{i_1}, x))^* R^{-1} (y_k - H(s_{i_1}, x))\right)$$

where R is the covariance matrix of Z_k. Letting the posterior distribution of $(\theta^0, \theta_k^1, X_k)$ be written as

$$\pi_k^{i_1,i_0}(x) = \partial_x \mathbb{P}(X_k \leq x | \theta^0 = i_0, \theta_k^1 = i_1, Y_1, \dots, Y_k) \mathbb{P}(\theta^0 = i_0, \theta_k^1 = i_1 | Y_1, \dots, Y_k)$$

we can then write the Baum-Welch filtering recursion explicitly as

$$\pi_{k+1}^{i_1,i_0}(x) = \frac{1}{c_{k+1}} \psi_{i_1,x}(Y_{k+1}) (\mathbf{\Gamma} \pi_k)^{i_1}(x)$$

$$= \frac{1}{c_{k+1}} \exp\left(-\frac{1}{2}(Y_{k+1} - H(s_{i_1}, x))^* R^{-1} (Y_{k+1} - H(s_{i_1}, x))\right)$$

$$\times \sum_{i_1'} \int_{\mathbb{R}^2} \exp\left(-\frac{\|x - A(s_{i_1}, x')\|_2^2}{2\sigma^2}\right) \Lambda(i_1 | i_1', i_0) \pi_k^{i_1',i_0}(x') dx'$$

where c_{k+1} is a normalizing constant.

2. Filtering Techniques for Image Data Applications

In this section we present some numerical techniques that allow an effective computation of the nonlinear filtering algorithm. It is well-known that the nonlinear filtering algorithm becomes difficult to implement as the size of the state-space grows. However, there are some good ways to implement the algorithm that are not obvious from the theoretical derivation of the filter and will reduce computational cost. In addition, we address the issues that arise when the likelihood functions ψ take values below machine precision, which indeed happens when our observation process Y_k is a 2-D image.

2.1. Scaling Procedure for Small Likelihoods. Consider the tracking intentions model that has been presented in the examples to this point, with $R = \gamma^2 I_{n\times m}$. A pressing issue when applying the filtering algorithm to image data is the exponentially small likelihood probabilities. For instance, when each observation is an $n \times m$ image, we can compute the likelihood of the event $\{(\theta_k^1, X_k) = (s_i, x)\}$ given any individual value of Y_k at a given pixel:

$$P(Y_k^{r_1 r_2} | \theta_k = s_i, X_k = x)$$

$$= \frac{1}{c} \exp\left(-\frac{1}{2} \frac{(Y_k^{r_1 r_2} - H^{r_1 r_2}(s_i, x)))^2}{\gamma^2} \right), \ \forall r_1 \leq n, r_2 \leq m$$

where $H^{r_1 r_2}(s_i, x)$ is the $(r_1, r_2)^{th}$ pixel of the image $H(s_i, x)$ and $Y_k^{r_1 r_2}$ is the $(r_1, r_2)^{th}$ pixel of the observation Y_k, and c is a normalizing constant.

The scalability issue arises when we try to compute the likelihood given the entire image. To get a sense of the size of these likelihoods, consider the log of the likelihood function for any (s_i, x):

$$\frac{1}{nm} \log \psi_{i,x}(Y_k) = \frac{1}{nm} \log P(Y_k | \theta_k^1 = s_i, X_k = x)$$

$$= \frac{1}{nm} \log \left(\prod_{r_1, r_2} \exp \left\{ -\frac{1}{2} \left(\frac{Y_k^{r_1 r_2} - H^{r_1 r_2}(s_i, x)}{\gamma} \right)^2 \right\} \right)$$

$$= -\frac{1}{2nm} \sum_{r_1, r_2} \left(\frac{Y_k^{r_1 r_2} - H^{r_1 r_2}(s_i, x)}{\gamma} \right)^2$$

$$= -\frac{1}{2nm} \sum_{r_1, r_2} \left(\frac{Y_k^{r_1 r_2} - H^{r_1 r_2}(\theta_k^1, X_k) + H^{r_1 r_2}(\theta_k^1, X_k) - H^{r_1 r_2}(s_i, x)}{\gamma} \right)^2$$

$$= -\frac{1}{2nm} \sum_{r_1, r_2} \left(Z_k^{r_1 r_2} + \frac{H^{r_1 r_2}(\theta_k^1, X_k) - H^{r_1 r_2}(s_i, x)}{\gamma} \right)^2$$

$$= -\frac{1}{2nm} \sum_{r_1, r_2} (Z_k^{r_1 r_2})^2 - \frac{1}{\gamma nm} \sum_{r_1, r_2} Z_k^{r_1 r_2} \left(H^{r_1 r_2}(\theta_k, X_k) - H^{r_1 r_2}(s_i, x) \right)$$

$$-\frac{1}{2nm} \sum_{r_1, r_2} \left(\frac{H^{r_1 r_2}(\theta_k^1, X_k) - H^{r_1 r_2}(s_i, x)}{\gamma} \right)^2$$

Now let's define the energy of any image produced by a state-space element (s_i, x) as:

$$\mathcal{E}_{nm}(s_i, x) = \frac{1}{nm} \sum_{r_1, r_2} H^{r_1, r_2}(s_i, x)^2$$

If we assume that the energy of any possible image remains bounded by a positive constant $\alpha \leq 1$ as we increase the resolution, we have $\mathcal{E}_{nm}(s_i, x) \leq \alpha$ for all (s_i, x) and $nm > 0$. Then we have the following decomposition:

$$\frac{1}{nm} \log \psi_{i,x}(Y_k) = -\frac{1}{2} - \frac{1}{\gamma} \varphi_{i,x}(k) - c_{i,x}(k)$$

(1) For any (s_i, x), $c_{i,x}(k)$ is the random variable defined as

$$c_{i,x}(k) = \frac{1}{2nm} \sum_{r_1, r_2} \left(\frac{H^{r_1 r_2}(\theta_k^1, X_k) - H^{r_1 r_2}(s_i, x)}{\gamma} \right)^2 \geq 0,$$

(2) For any (s_i, x) the middle term is the random variable defined as

$$\varphi_{i,x}(k) = \frac{1}{nm} \sum_{r_1,r_2} Z_k^{r_1 r_2} \left(H^{r_1 r_2}(\theta_k^1, X_k) - H^{r_1 r_2}(s_i, x) \right)$$

and its second moment approaches zero as follows:

$$\mathbb{E}\varphi_{i,x}^2(k) = \mathbb{E}\left(\frac{1}{nm} \sum_{r_1,r_2} Z_k^{r_1 r_2} \left(H^{r_1 r_2}(\theta_k^1, X_k) - H^{r_1 r_2}(s_i, x) \right) \right)^2$$

$$= \frac{2}{(nm)^2} \sum_{r_1,r_2} \mathbb{E}\left[\left(H^{r_1 r_2}(\theta_k^1, X_k) - H^{r_1 r_2}(s_i, x) \right)^2 \right] \underbrace{\mathbb{E}\left[(Z_k^{r_1 r_2})^2 \right]}_{=1}$$

where we have applied the independence of each $Z_k^{r_1 r_2}$ from all other random variables,

$$= \frac{2}{(nm)^2} \sum_{r_1,r_2} \mathbb{E}\left[\left(H^{r_1 r_2}(\theta_k^1, X_k) - H^{r_1 r_2}(s_i, x) \right)^2 \right]$$

$$\leq \frac{4}{nm} \mathbb{E}\left[\mathcal{E}_{nm}(\theta_k^1, X_k) \right] + \frac{4}{nm} \mathcal{E}_{nm}(s_i, x) \leq \frac{8\alpha}{nm}$$

Now applying the Chebyshev Inequality with some small $\delta > 0$, we have

$$\mathbb{P}\left(\frac{1}{nm} \log \psi_{i,x}(Y_k) + \frac{1}{2} - c_{i,x}(k) \geq \delta \right) \leq \frac{1}{\delta^2} \mathbb{E}\left(\frac{1}{nm} \log \psi_{i,x}(Y_k) + \frac{1}{2} - c_{i,x}(k) \right)^2$$

$$= \frac{1}{\delta^2} \mathbb{E}\left(-\frac{1}{2nm} \sum_{r_1,r_2} (Z_k^{r_1 r_2})^2 + \frac{1}{2} - \frac{1}{\gamma} \varphi_{i,x}(k) \right)^2$$

$$\leq \frac{2}{\delta^2} \mathbb{E}\left(-\frac{1}{2nm} \sum_{r_1,r_2} (Z_k^{r_1 r_2})^2 + \frac{1}{2} \right)^2 + \frac{2}{\delta^2} \mathbb{E}\left(\frac{1}{\gamma} \varphi_{i,x}(k) \right)^2$$

$$\leq \frac{2}{\delta^2} \frac{\kappa}{nm} + \frac{2}{\delta^2} \frac{8\alpha}{nm}$$

where κ is a constant such that

$$\mathbb{E}\left(-\frac{1}{2nm}\sum_{r_1,r_2}(Z_k^{r_1r_2})^2+\frac{1}{2}\right)^2=\frac{\kappa}{nm}$$

Since the bound on the probability is very small for large nm, we have

$$\mathbb{P}\left(\psi_{i,x}(Y_k)\geq e^{(\delta-1/2)nm}\right)\leq\mathbb{P}\left(\psi_{i,x}(Y_k)\geq e^{(\delta-1/2)nm-c_{i,x}(k)nm}\right)\leq\frac{2}{\delta^2}\frac{\kappa+8\alpha}{nm},$$

So if we choose $\delta=1/4$, we have exponentially small likelihoods:

$$\mathbb{P}(\psi_{i,x}(Y_k)\geq e^{-nm/4})\leq\frac{\mathcal{C}}{nm}$$

where $\mathcal{C}$ is some fixed constant.

Indeed, even low resolution images may have tens of thousands of pixels which will make the likelihood of any state (including the most likely state) extremely small. Therefore, a program that directly applies the recursive formula from Section 1 will return an error:

$$\pi_{k+1}^i(x)=\frac{\psi_{i,x}(Y_{k+1})(\mathbf{\Gamma}\pi_k)^i(x)}{\sum_j\int\psi_{j,x'}(Y_{k+1})(\mathbf{\Gamma}\pi_k)^j(x')dx'}\approx\frac{0}{0}$$

which is undefined. Therefore, models involving such likelihoods should use the scaled nonlinear filtering algorithm presented in the table on page (45).

2.2. Filtering With Limited Computing Resources. The nonlinear filtering algorithm can be very slow, even on the most modern processors. The fact is that the algorithm has so many operations to carry out in order to update the posterior, and even though the recursion from Section 1 is easy to comprehend, the programming involved can be quite intricate.

Considering again the tracking intentions example, let the hidden states (θ_k^1,X_k) take values in a state-space $\Omega=\mathcal{S}^1\times\mathbb{R}^d$, and let $|\Omega|$ denote the number of elements in the state-space. In general, the application of the forward operator Γ to the prior distribution of (θ_k^1,X_k) will require at least $\mathcal{O}(|\Omega|^2)$ operations. In addition, if the number of operations needed to compute the likelihood of each state-space

TABLE 2. **Scaled Nonlinear Filtering Algorithm**

step 1: input data and the old posterior, Y_{k+1} and π_k.
step 2: predict and correct the distribution of (θ, X), $$\tilde{\pi}^i_{k+1}(x) = \log\left((\mathbf{\Gamma}\pi_k)^i(x)\right) - \frac{\|Y_k - H(s_i, x)\|_2^2}{2\sigma^2}$$
step 3: normalize the distribution of (θ, X), $$\text{define: } \beta = \max_{i,x} \tilde{\pi}^i_{k+1}(x)$$ $$\text{set: } \tilde{\pi}^i_{k+1}(x) = \tilde{\pi}^i_{k+1}(x) - \beta, \ \forall i, x$$ and then take exponentials and normalize: $$\pi^i_{k+1}(x) = \frac{\exp(\tilde{\pi}^i_{k+1}(x))}{\sum_j \int \exp(\tilde{\pi}^j_{k+1}(x'))dx'}$$
step 4: return the updated posterior, π_{k+1}, and any estimates such as the MAP or posterior mean.

element is $O(|\Omega|)$ or greater, then the total machine cost for computing likelihoods for every state-space element will also be greater or equal to $\mathcal{O}(|\Omega|^2)$. Often times a computational cost of order squared can be too much for a machine to handle. Therefore, the implementation of the nonlinear filtering algorithm depends heavily on our ability to interpret Γ as a sparse (or banded) matrix and then only evaluate the likelihoods of the state-space elements which have significant probability. These ideas are what is referred to as stochastic domain pursuit (SDP) [**31**], and can be written explicitly as the following recursive updating formula:

$$\tilde{\pi}^i_{k+1}(x) = \sum_{j,x' \in \mathcal{A}} \Gamma(i, x|j, x')\pi^j_k(x'), \qquad \text{where } \mathcal{A} = \{j, x' : \Gamma(i, x|j, x') \geq \delta \text{ and } \pi^j_k(x') \geq \delta\}$$

$$\pi^i_{k+1}(x) = \begin{cases} \frac{1}{c_{k+1}}\psi_{i,x}(Y_{k+1})\tilde{\pi}^i_{k+1}(x), & \text{if } \tilde{\pi}^i_{k+1}(x) \geq \log\epsilon \\ 0, & \text{otherwise} \end{cases}$$

where $\epsilon > 0$ and $\delta > 0$ are two relatively small thresholds, and c_{k+1} is a normalizing constant. Obviously the choice of thresholds is non-trivial, but there are numerous examples of DMCs where SDP can be readily implemented. For instance, when observations are scalar and X_k is conditionally Gaussian, its conditional transition probability matrix takes a banded form and the whole recursion will require only $\mathcal{O}(|\Omega|)$ operations. The results of Petrov and Rozovsky [**31**] implemented SDP in HMM models for tracking aircraft in a noisy medium.

Aside from SDP, another fast way to implement the algorithm is to use particles. A particle filter consists of two parts: Sequential Monte Carlo (SMC) and Sampling Importance Resampling (SIR). In Section 2.4 a particle algorithm will be developed in more detail. For more on SMC and SIR see the book by Asmussen and Glynn [**2**] or the book by Cappe, Moulines and Ryden [**9**], .

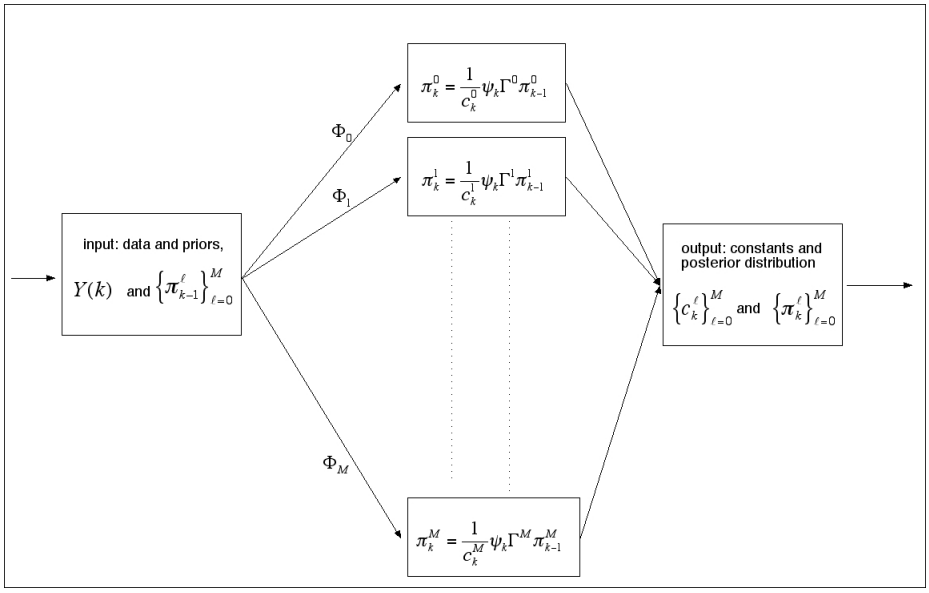

FIGURE 1. The flow diagram for a bank of nonlinear filters. At each time step the input is the previous bank of posterior distributions and the latest observation, from which a bank of updated posteriors is returned.

2.3. Banks of Filters for Parallel Computing. When the lowest states of the double process are static in time, there is the option to compute banks of filters for several hypotheses. For instance, recall from the tracking intentions example in Section 1.2 where θ_0 did not change over time and had initial distribution $\rho = \{\rho_0, ...\rho_M\}$. For the tracking intentions example, we define a set of hypotheses $\{\Phi_\ell\}_{\ell=0}^M$ where for each ℓ

$$\Phi_\ell : \{\theta_0 = \ell\}$$

then for each Φ we have a forward operator $\mathbf{\Gamma}^\ell$ from which we can compute a posterior given each hypothesis:

$$\pi_k^{\ell,i}(x) = \partial_x \mathbb{P}(\theta_k^1 = s_i, X_k \leq x | Y_1, \ldots, Y_k, \theta_0 = \ell)$$

$$= \frac{1}{c_k^\ell} \psi_{i,x}(Y_k) \left(\mathbf{\Gamma}^\ell \pi_{k-1}^\ell\right)^i (x)$$

where ψ is the likelihood function and c_k^ℓ is a normalizing constant,

$$c_k^\ell = \sum_i \int \psi_{i,x}(Y_k)(\mathbf{\Gamma}^\ell \pi_{k-1}^\ell)^i(x)dx = \mathbb{P}(Y_k | \theta_0 = \ell, Y_1, \ldots, Y_{k-1})$$

Reusing the sequence of normalizing constants we can compute the log-likelihood of the person's intentions:

$$L_k^\ell = \log \mathbb{P}(Y_1, \ldots, Y_k | \theta_0 = \ell) = \log c_1^\ell + \log c_2^\ell + ... + \log c_k^\ell$$

A posterior on θ_0 can then be computed using this likelihood, ρ, and Bayes rule:

$$\mathbb{P}(\theta_0 = \ell | Y_1, \ldots, Y_k) = \frac{\exp(L_k^\ell)\rho_\ell}{\sum_{\ell'=0}^M \exp(L_k^{\ell'})\rho_{\ell'}}$$

Figure 1 shows a flow chart of such an algorithm and how it divides the labor associated with each recursive step. The division of labor opens up the possibility of running the code in parallel on a cluster. These banks of filters will be used later in the example of Section 3, and in the next sub-section where we formulate a particle filtering algorithm.

2.4. Particle Filtering Algorithm. Particle filters were mentioned in Section 2.2 as a fast way of computing the posteriors. The particle filter is an approximation of the nonlinear filter and only returns an optimal estimate as the number of samples grows to infinity. However it is a consistent estimator because it does not make any assumptions which may introduce bias into the model. In this section we formulate a bank of particle filters for the tracking intentions example. The example can serve as a template for general particle filtering algorithms regardless of how many banks are used.

The intentions example fits the description given in Figure 1 where we split the algorithm into banks of filters based on θ_0 which is constant in time. When $\theta_0 \in \{0, 1, 2, ..., M\}$, we can split the filter into $(M+1)$-many banks, one filter for each hypothesis on θ_0. A bank of particle filters can then be constructed as follows:

For the ℓ^{th} hypothesis at time k, the n^{th} sample is a vector $(\theta^1_{0:k}, X_{0:k})^\ell(n)$ where

$$(\theta^1_{0:k}, X_{0:k})^\ell(n) = \left\{(\theta^1_0, X_0)^\ell(n), (\theta^1_1, X_1)^\ell(n), ..., (\theta^1_k, X_k)^\ell(n)\right\}$$

and has been independently sampled sequentially from $\mathbf{\Gamma}^\ell$. Letting the k^{th} likelihood of the n^{th} particle under $\mathbf{\Gamma}^\ell$ be written as

$$\psi_{\ell,n}(Y_k) = \mathbb{P}(Y_k|(\theta^1_k, X_k) = (\theta^1_k, X_k)^\ell(n)),$$

we can sequentially define an importance weight $\omega^\ell_k(n)$ which is proportional to the likelihood of the sequence of particles given the data:

$$\omega^\ell_k(n) = \frac{\psi_{\ell,n}(Y_k) \times ..... \times \psi_{\ell,n}(Y_1)}{\sum_n[\text{numerator}]} = \frac{\psi_{\ell,n}(Y_k)\omega^\ell_{k-1}(n)}{\sum_n[\text{numerator}]}$$

A particle filter produces an empirical distribution from the importance weights and invokes SIR to speed up the convergence of the empirical distribution to the true distribution (see [**2**] and [**9**]). Specifically, the particle filter can be used to approximate the log-likelihood function because it is a consistent estimator of the

TABLE 3. **Bank of Particle Filters**

step 1: input data, particles, particle weights and log-likelihoods, Y_{k+1}, $\{(\theta_k^1, X_k)^\ell(n)\}_{\ell,n}$, $\{\omega_k^\ell(n)\}_{\ell,n}$, and $\{L_k^\ell\}_\ell$ for $\ell \leq M$ and $n \leq$(sample size).
step 2: for each hypothesis, sequentially sample each particle and update the likelihood: $$(\theta_{k+1}^1, X_{k+1})^\ell(n) \sim \Gamma^\ell(\,\cdot\,|(\theta_k^1, X_k)^\ell(n))$$ $$L_{k+1}^\ell = \log\left(\sum_n \psi_{\ell,n}(Y_k)\omega_k^\ell(n)\right) + L_k^\ell$$
step 3: perform SIR for each hypothesis independently if at least one empirical distribution has too few important weights, i.e. if $\left(\sum_n(\omega_k^\ell(n))^2\right)^{-1} < \tau$ for at least one pair (i,j) where $0 < \tau <$(# of particles) is some threshold; otherwise for each hypothesis, update and normalize the importance weights: $$\omega_{k+1}^\ell(n) = \frac{\psi_{\ell,n}(Y_k)\omega_k^\ell(n)}{\sum_n[\text{numerator}]},$$
step 4: output the updated particles and weights, and the posterior on the hypotheses: $$\mathbb{P}(\theta_0 = \ell|Y_k, ..., Y_1) = \frac{\exp\left(L_k^\ell\right)\rho_\ell}{\sum_{\ell=0}^M[\text{numerator}]}$$

true log-likelihood. We see this in the limit as the number of samples approaches infinity:

$$\sum_{r=0}^{k-1}\log\left(\sum_n \psi_{\ell,n}(Y_k)\omega_r^\ell(n)\right) \xrightarrow{a.s.} \sum_{r=0}^{k-1}\log\mathbb{P}(Y_{r+1}|Y_1,\ldots Y_r,\theta_0=\ell) = L_k^\ell.$$

2.5. Sampling on a High-Dimensional State-Space. The iterative construction of DMCs allows us to sample from distributions of a very large dimension. In the

last section a particle filter was formulated, but it was assumed that sampling from $\mathbf{\Gamma}$ was always possible. This will not be the case, particularly when N (the number of components in the the DMC) is large. Recall in Section 1 where the kernel of $\mathbf{\Gamma}$ was written as the product of $N+1$ many transition probabilities. In such a situation $\mathbf{\Gamma}$ will be extremely small and computers will effectively consider it to be zero.

Recall the definitions and notations from Section 1, and consider a DMC $\vec{\theta}_k^N = (\theta_k^0, \ldots, \theta_k^N)$ with transition matrices $\{\Lambda^n\}_{n=0}^N$ where $\theta_k^n \in \mathcal{S}^n$ for all $n \leq N$. When $\vec{\theta}_k$ has very large dimension, the probability of any state-space element may be below the threshold of machine precision:

$$\Gamma(\ \vec{s}\ |\ \vec{r}\ ) < \epsilon_{machine}, \qquad\qquad \forall \vec{s}, \vec{r}$$

which means it will not be possible to sample particles the in the manner suggested in Section 2.4.

However, it is possible to sample from a high-dimensional DMC if the set of transition matrices $\{\Lambda^n\}_{n=0}^N$ individually lend themselves to sampling. In other-words, if it is possible to conditionally sample from each individual Λ^n, then there is an iterative procedure that samples the vector $\vec{\theta}_k$ component by component. Given $\vec{\theta}_{k-1} = \vec{r}$ there is the following iterative sampling algorithm for $\vec{\theta}_k$:

(1) sample s^0 from $\Lambda_0(\ \cdot\ |r^0)$ and set $\theta_k^0 = s^0 \in \mathcal{S}^0$

(2) for $n = 1, 2, ..., N$

sample s^n from $\Lambda^n(\ \cdot\ |r^n, s^{n-1})$ and set $\theta_k^n = s^n \in \mathcal{S}^n$.

This iterative scheme is analogous to the iterative form of the forward Baum-Welch equation given in proposition (2).

3. Numerical Simulation

In this section numerical simulations of the nonlinear filtering algorithm for tracking intentions are presented. The model and algorithm are those described in Section 1 and the example of Section 1.2. In the following sections the model is described in

detail, there is some discussion of the computational techniques for implementation, and then the simulated tracking results are presented.

3.1. Pinocchio Example. Pinocchio was overheard bragging that he has a gun and he will shoot out the only street lamp in the city. A hidden video camera and an ATR&D (automatic threat recognition and defense) system was installed by the concerned citizens in a hiding place with a good view of the lamp. The ATR&D is capable of analyzing the behavior of Pinocchio if he approaches the lamp and can neutralize him if the system decides that Pinocchio indeed has malicious intentions and the danger to the lamp becomes credible. The system is "all weather" in that it could function even in very bad weather with extremely low visibility.

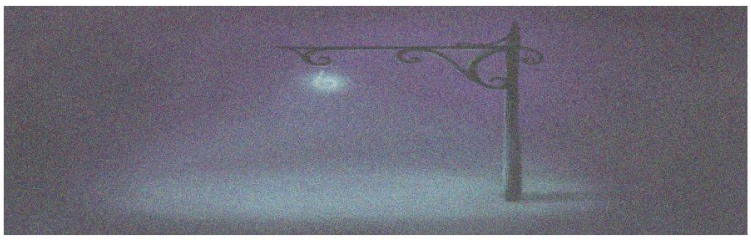

FIGURE 2. The only street lamp in the city which Pinocchio may or may not intend to shoot out.

The decision making algorithm is based on a 3-leveled DMC model. In our setting θ_0 is simply a two dimensional r.v. $(\theta_0^{int}, \theta_0^{gun})$. The states for θ_0^{int} are:

$$\{\theta_0^{int} = 0\} \quad = \quad \{\text{just kidding}\}$$

$$\{\theta_0^{int} = 1\} \quad = \quad \{\text{hostile}\}$$

The states for θ_0^{gun} are:

$$\{\theta_0^{gun} = 0\} \quad = \quad \{\text{doesn't have a gun}\}$$

$$\{\theta_0^{gun} = 1\} \quad = \quad \{\text{has a gun}\}$$

In addition, a statistical research firm has provided the city with the following a priori information on θ_0:

$$P\left(\theta_0^{int}=0|\theta_0^{gun}=0\right)=1 \qquad P\left(\theta_0^{int}=0|\theta_0^{gun}=1\right)<1/2$$

$$P\left(\theta_0^{gun}=0|\theta_0^{int}=0\right)>1/2 \qquad P\left(\theta_0^{gun}=0|\theta_0^{int}=1\right)=0$$

The second component of the DMC, $\theta_k^1 \in \{1,2,3,....,m\}$, is Pinocchio's pose at time k. Some poses are shooting poses and some are not. The set of poses that Pinocchio can take is shown in Figure 3.

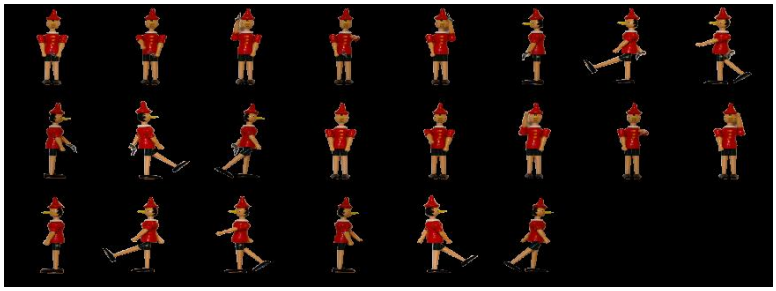

FIGURE 3. The alphabet of poses which Pinocchio can take.

The third component, X_k, is a vector containing a translation, rotation and dilation of Pinocchio's pose:

$$X_k = \begin{pmatrix} d_x(k) \\ d_y(k) \\ r_x(k) \\ r_y(k) \\ t_x(k) \\ t_y(k) \end{pmatrix}$$

Pinocchio's position relative to the center of the camera's view is given by the translation $(t_x(k), t_y(k))$, and his distance from the camera is given by the dilation $(d_x(k), d_y(k))$. If the camera perceives any shear or rotation in Pinocchio's pose then $(r_x(k), r_y(k))$ would account for it, but for simplicity we will assume that neither occurs and keep $r_x(k) = 0 = r_y(k)$ for $k \geq 0$.

If Pinocchio is in a standing pose then $X_{k+1} = X_k$; if he is a walking pose then the dynamics are given by a recursive structure:

$$
\begin{aligned}
d_x(k+1) &= 1.3^{t_y(k+1)} \\
d_y(k+1) &= 1.3^{t_y(k+1)} \\
r_x(k+1) &= 0 \\
r_y(k+1) &= 0 \\
t_x(k+1) &= d_x(k) + 1.3^{t_y(k)}\Delta X(i) + W^x_{k+1} \\
t_y(k+1) &= \kappa t_y(k) + W^y_{k+1},\ 0 < \kappa < 1
\end{aligned}
$$

where $\Delta X(i)$ is the distance Pinocchio travels on the x-axis during one time step if $\theta^1_k = i$, and $W_k = (W^x_k, W^y_k)^*$ are iid mean-zero Gaussian random vectors with covariance matrix

$$
Q = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}
$$

What is actually visible from the DMC is the function $H(i, x)$ which is a 2-D image of the street lamp with Pinocchio taking the i^{th} pose and translated by x from the center of the scene. Figure 4 shows some examples of some possible mappings which move Pinocchio around the street lamp.

FIGURE 4. We see here how the spatial location is a mapping of a pose from the center of the street lamp scene. The Pinocchio in the center illustrates $H(i, 0)$, and each arrow represents a possible value of X_k and how it will spatially relocate Pinocchio within the scene.

The observed visual data obtained by the ATR&D's camera is denoted by Y_k. The observation Y_k is an $n \times m$-pixeled snap shot of the street lamp at time k:

$$
Y_k = H(\theta^1_k, X_k) + \gamma Z_k,
$$

where Z_k is a frame of standard normal Gaussian noises with zero covariance between different pixels. The magnitude of γ depends on conditions such as lighting, fog, haze, etc. In Figure 5 there are some fairly low noise observation frames of Pinocchio walking around near the street lamp, and it does not appear that he has a gun, but it is not entirely clear.

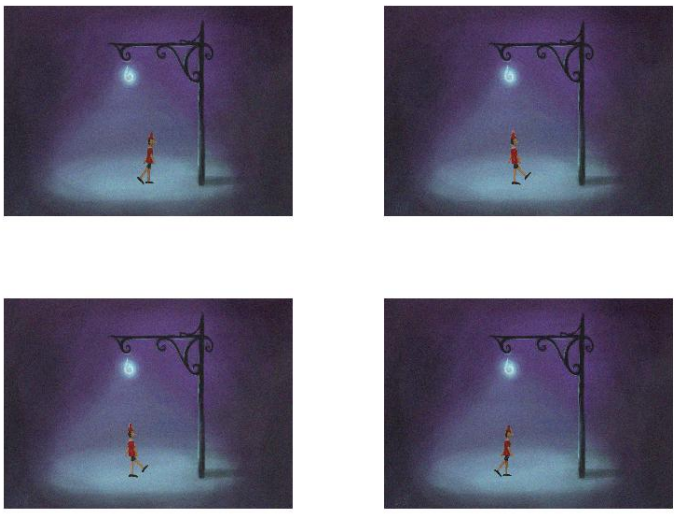

FIGURE 5. Low-noise observations of Pinocchio walking around near the street lamp.

3.2. Filtering Algorithm for Pinocchio. Pinocchio appears at the street lamp around midnight and hangs around while the ATR&D system attempts to determine whether his earlier statements are indeed true. To make matters worse there is heavy fog on this particular night and so there is an increase in γ.

The nonlinear filter computes the posterior probability for each of the following four hypotheses:

$$\begin{aligned}
\{00\} &= \{\theta_0^{int} = 0, \theta_0^{gun} = 0\} \\
\{01\} &= \{\theta_0^{int} = 0, \theta_0^{gun} = 1\} \\
\{10\} &= \{\theta_0^{int} = 1, \theta_0^{gun} = 0\} \\
\{11\} &= \{\theta_0^{int} = 1, \theta_0^{gun} = 1\}
\end{aligned}$$

The posterior distribution is then:

$$\pi_k^{\{\cdot\}}(i,x) = \mathbb{P}(X_k = x, \theta_k^1 = i | Y_1, \ldots, Y_k, \{\cdot\})$$

where $\{\cdot\}$ contains one of the four initial hypotheses. The Baum-Welch equation gives the following recursive form of the posterior distribution:

$$\pi_k^{\{\cdot\}}(i,x) = \frac{1}{C_k^{\{\cdot\}}} \psi_{i,x}(Y_k) \sum_j \int \Gamma^{\{\cdot\}}(i,x|j,x') \pi_{k-1}^{\{\cdot\}}(j,x') dx'$$

where $\Gamma^{\{\cdot\}}$ is the forward operator semi-group conditioned on the hypothesis $\{\cdot\}$ and $\psi_{i,x}(Y_k)$ is the likelihood of the i^{th} pose and spatial coordinates x given Y_k:

$$\Gamma^{\{\cdot\}}(i,x|j,x') = \mathbb{P}(X_k = x, \theta_k^1 = i | X_{k-1} = x', \theta_{k-1}^1 = j, \{\cdot\})$$

$$\psi_{i,x}(Y_k) = \mathbb{P}(Y_k | X_k = x, \theta_k^1 = i)$$

with $C_k^{\{\cdot\}}$ being a normalizing constant.

Given a set of observations $Y_1, \ldots, Y_k$, there is a log-likelihood function for any of the four initial hypotheses:

$$\begin{aligned} L_k^{00} &= \log \mathbb{P}(Y_1, \ldots, Y_k | \{00\}) \\ L_k^{01} &= \log \mathbb{P}(Y_1, \ldots, Y_k | \{01\}) \\ L_k^{10} &= \log \mathbb{P}(Y_1, \ldots, Y_k | \{10\}) \\ L_k^{11} &= \log \mathbb{P}(Y_1, \ldots, Y_k | \{11\}) \end{aligned}$$

for which there is a recursive formula:

$$L_k^{\{\cdot\}} = \log \mathbb{P}(Y_k | Y_1, \ldots, Y_{k-1}, \{\cdot\}) + L_{k-1}^{\{\cdot\}}$$

$$= \log \sum_i \int \mathbb{P}(Y_k, X_k = x, \theta_k^1 = i | Y_1, \ldots, Y_{k-1}, \{\cdot\}) dx + L_{k-1}^{\{\cdot\}}$$

$$= \log \sum_i \int \psi_{i,x}(Y_k) \left(\sum_j \int \Gamma^{\{\cdot\}}(i,x|j,x') \pi_{k-1}^{\{\cdot\}}(j,x') dx' \right) dx + L_{k-1}^{\{\cdot\}}$$

The posterior of each hypothesis is

$$p_k^{\{ij\}} = \mathbb{P}(\theta_k^{int} = i, \theta_k^{gun} = j | Y_1, ..., Y_k)$$

and can be computed through as

$$p_k^{\{ij\}} = \frac{\exp(L_k^{\{ij\}}) \times \mathbb{P}(\theta_0^{int} = i, \theta_0^{gun} = j)}{\sum_{i,j \in \{0,1\}} [\text{numerator}]}, \ \forall i, j \in \{0, 1\}$$

from which there is a maximum aposteriori (MAP) estimate:

$$(\hat{\theta}_0^{int}, \hat{\theta}_0^{gun}) \ = \arg\max_{i,j \in \{0,1\}} \exp(L_k^{\{ij\}}) \times \mathbb{P}(\theta_0^{int} = i, \theta_0^{gun} = j)$$

Finally, a decision whether or not to arrest Pinocchio must be made. Rather than arrest Pinocchio as soon as $p^{\{11\}}$ crosses some threshold, multiple observations are allowed to arrive in order to see how a series of lag-weighted posteriors grows or shrinks. Such a decision-making criterion will minimize false arrests. The decision-making criterion is based on a statistic S_k, defined as follows:

Let $S_0 = 0$, and then recursively update:

$$S_k = (1-a)p_k^{\{11\}} + aS_{k-1}$$

$$= (1-a)p_k^{\{11\}} + (1-a)ap_{k-1}^{\{11\}} + a^2 S_{k-2} = ... = (1-a)\sum_{\ell=1}^{k} a^{k-\ell} p_\ell^{\{11\}}$$

where $0 < a < 1$ is the lag coefficient. The decision to arrest Pinocchio is made if

$$S_k > 1 - \alpha$$

where $\alpha > 0$ is a small threshold; otherwise Pinocchio remains free but is still being watched and will be arrested at a later time if the lag posteriors exceed the threshold. The lag coefficient a and the threshold α should be chosen with respect to the set of transition laws $\{\Lambda^{\{ij\}}\}_{ij}$ and the "distance" between elements in the set. For instance, consider the setting where there are multiple transition laws that are very likely to have generated the observed data. Such a situation would occur if the "distance" between two or more transition laws was very small. In such a case it would be a

good idea to take a closer to 1 and α closer to zero in order to avoid false alarms. The example in Section 3.3.3 explores this idea further. For more on such decision making criteria (and change-point detection in general) see the book by Basseville and Nikiforov [**4**].

3.3. Simulations and Results. If the ATR&D systems does not act fast, Pinocchio will indeed shoot out the light, but it will take a sufficiently long time for Pinocchio to get close enough to shoot out the lamp and so the ATR&D will have ample time to make a decision. When Pinocchio shoots out the light the ATR&D sees only darkness and the posterior probability of $(1, 1)$ is 1.

The algorithm is implemented as a particle filter in the manner described in Section 2.4, and for each hypothesis we have a sample size of 2000 particles. The algorithm used is described in detail in Table 2.4, and the scaling procedure from Section 2.1 is used to update the particle weights.

3.3.1. *Likelihood of θ_0 Under Zero Noise.* If it is the case that $\gamma = 0$ it simulates no noise, in which case the posterior of $(\theta_0^{gun}, \theta_k^1, X_k)$ becomes a point-mass function of the correct location, pose, and hypothesis on gun-possesion. However, θ_0^{int} is still not obvious because it is impossible to visually observe, so the filter's inference is still required to make a decision on whether or not Pinocchio has hostile intentions.

A zero-noise example serves to demonstrate why it was a good choice to use a double model. The example illustrates how inferences of θ_0 are made based on the probability of a possible sequence of (θ^1, X) given any one of the four hypotheses mentioned in Section 3.2. In general, regardless of the amount of noise in the observations, the path taken by (θ^1, X) through its state-space should be most probable under the transition laws of the correct hypothesis, otherwise the filter cannot accurately track θ_0.

With $\gamma = 0$ and assuming $H(\ell, x)$ is a one-to-one function in (ℓ, x), the likelihoods reveal the pose and location by themselves:

$$\psi_{\ell,x}(Y_k) = \mathbf{1}_{\{\theta_k^1 = \ell\}} \delta_{\{X_k = x\}}$$

(where δ is the Dirac-delta function) and therefore the posterior given $\theta_0 = (i,j)$ also reveals the pose and location:

$$\pi_k^{\{ij\}}(\ell, x) = \mathbf{1}_{\{\theta_k^1 = \ell\}} \delta_{\{X_k = x\}}$$

Furthermore, the log-likelihood of the hypothesis $\{ij\}$ given the data $Y_1, \ldots, Y_T$ is simply the log probability of $(\theta_1, X) = \{\theta_k^1, X_k\}_{k=0}^T$ under the forward operator $\Gamma^{\{ij\}}$:

$$L_T^{\{ij\}} = \log \mathbb{P}(Y_1, \ldots, Y_T | \theta_0 = (i,j)) =$$

$$= \log \left(\mathbb{P}(Y_T | Y_1, \ldots, Y_{T-1}, \theta_0 = (i,j)) \right) + L_{T-1}^{\{ij\}}$$

$$= \log \left(\sum_\ell \int \mathbb{P}(Y_T, X_T = x, \theta_T^1 = \ell | Y_1, \ldots, Y_{T-1}, \theta_0 = (i,j)) dx \right) + L_{T-1}^{\{ij\}}$$

$$= \log \left(\sum_\ell \int \underbrace{\psi_{\ell,x}(Y_T)}_{= \mathbf{1}_{\{\theta_T^1 = \ell\}} \delta_{\{X_T = x\}}} \sum_{\ell'} \int \Gamma^{\{ij\}}(x, \ell | x', \ell') \underbrace{\pi_{T-1}^{\{ij\}}(\ell', x')}_{= \mathbf{1}_{\{\theta_{T-1}^1 = \ell'\}} \delta_{\{X_{T-1} = x'\}}} dx' dx \right)$$

$$+ L_{T-1}^{\{ij\}}$$

$$= \log \left(\Gamma^{\{ij\}}(X_T, \theta_T^1 | X_{T-1}, \theta_{T-1}^1) \right) + L_{T-1}^{\{ij\}}$$

and by induction back to $k = 0$,

$$\ldots\ldots\ldots = \sum_{k=1}^T \log \left(\Gamma^{\{ij\}}(X_k, \theta_k^1 | X_{k-1}, \theta_{k-1}^1) \right) + \log \pi_0^{\{ij\}}(\theta_0^1, X_0)$$

$$= \log \mathbb{P}((\theta^1, X) | \theta_0 = (i,j))$$

When $\theta_0 = (1,1)$ and T is large enough, there should be a significant difference between $L_T^{\{11\}}$ and the other hypotheses:

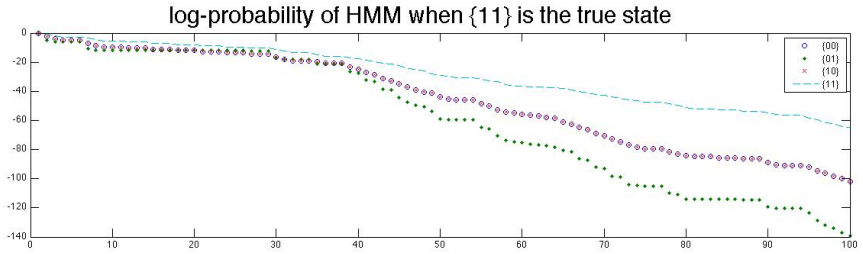

FIGURE 6. **A zero-noise example:** The log-likelihood of the data for a realization $\{\theta_k^1, X_k\}_{k=0}^{100}$ generated under $\Gamma^{\{11\}}$ when $\gamma = 0$ is simply the log-probability of the realization, $L_k^{\{ij\}} = \log \mathbb{P}((\theta_1, X)|\theta_0 = (i,j))$. The log-probability $L_{100}^{\{11\}}$ is significantly larger than $L_{100}^{\{ij\}}$ for $(i,j) \neq (1,1)$.

$$0 > L_T^{\{11\}} > L_T^{\{ij\}}, \ \forall (i,j) \neq (1,1)$$

In Figure 6 it is demonstrated that for $T = 100$ there is a clear separation of the log-probability of the realization of (θ_1, X) under the correct hypothesis from others. The intentions are not-directly observable even in the settings of lowest noise, but it is the inference of the filter that tracks the phenomenon in Figure 6 and returns a sequence of posteriors on θ_0 that suggest the correct hypothesis.

3.3.2. *High Noise Example.* This example uses almost the same realization of the DMC as the previous sub-section with $(\theta_0^{int}, \theta_0^{gun}) = (1,1)$, but now the observations have much more noise (i.e. the highest level of the DMC is different but the lower levels are all the same). Figure 7 shows some frames from this example. The figure illustrates how difficult it is to see Pinocchio, but the ATR&D is able to track him and has marked his location with an ellipse. Figure 8 shows the evolution of S_k when $a = .9$ and $\alpha = .95$. By time $k = 44$ the ATR&D has made the decision to arrest Pinnochio, but he will shoot out the lamp by time $k = 65$ if the simulation continues. Therefore, this example is a successful application of the nonlinear filtering algorithm because the ATR&D was confident that Pinocchio has a gun and hostile intentions, and has given the signal to make the arrest before he shot out the lamp.

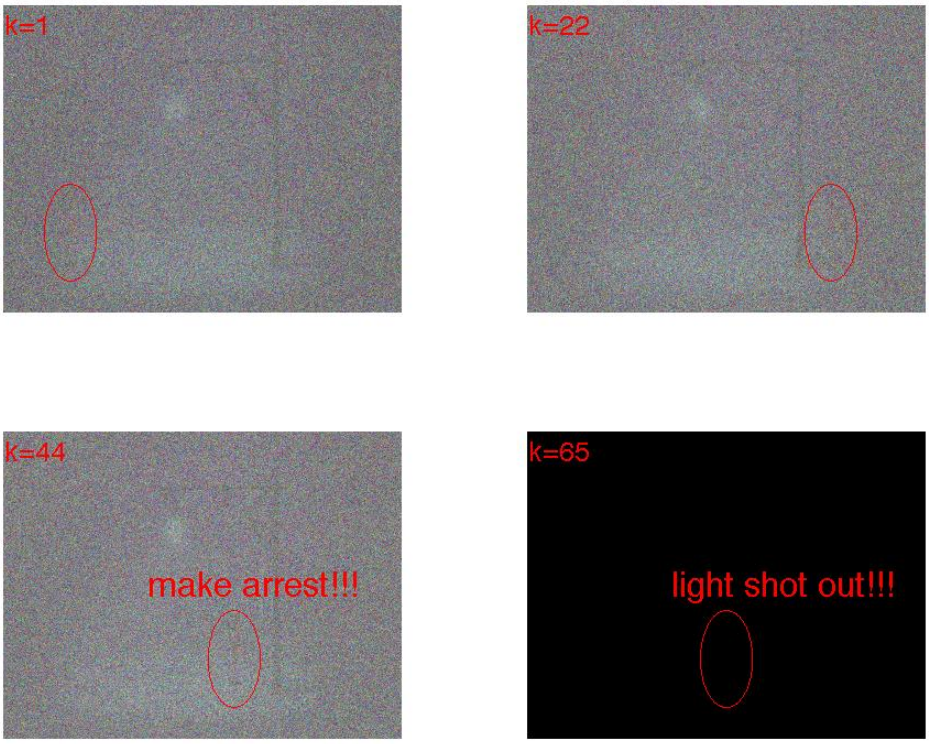

FIGURE 7. **High Noise Example.** This figure shows a high noise example of Pinocchio when he has a gun and his intentions are hostile. The pictures are the snapshots taken by the surveillance camera at times $k = 1, 22, 44, 65$, and although this realization of Pinocchio would have shot out the light, the decision to arrest Pinocchio was made with plenty of time to spare.

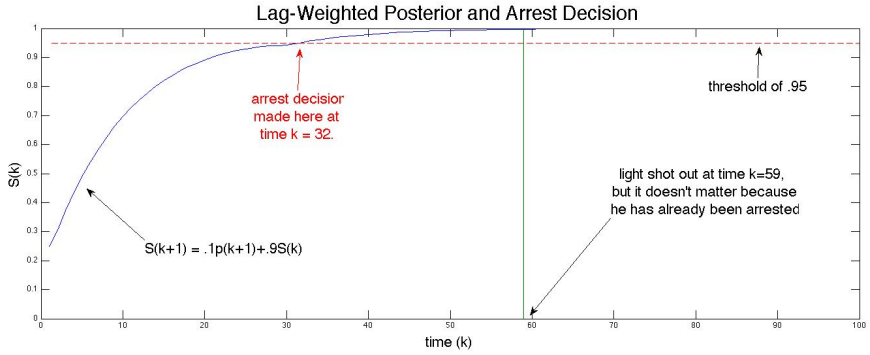

FIGURE 8. **High Noise Example.** This figure shows the lag-weighted posterior statistic S_k as described in Section 3.2. It is plotted against time with $a = .9$ and $\alpha = .95$. In this particular realization of Pinocchio, he shoots out the light at time $k = 59$, but when the ATR&D system is working it will arrest him earlier at time $k = 32$.

3.3.3. *Unsuccessful Tracking Example.* The ability of the filter to track intentions depends largely on the contrast between the different probability laws. For the previous example, in Figure 6 it was made clear that $\Lambda^{\{11\}}$ was the correct probability law, and then later in Figure 8 it was shown how S_k converged to 1, thereby causing the ATR&D to make the correct decision in a timely manner. That example served to show how well the filter could work when the set of transition laws $\{\Lambda^{\{ij\}}\}_{ij}$ has a significant amount of individuality among its elements. In this example, the filter will fail to make the decision to arrest because the transition probability laws are such that it is hard to tell which one generated the data.

The example is constructed in the same way as the example from the previous section, except for the following difference:

$$\Lambda^{\{01\}} = \Lambda^{\{11\}} + \epsilon$$

where ϵ is a perturbation which only makes minor changes in the transition probabilities. For the resulting process, regardless of whether $\theta_0 = (0,1)$ or $\theta_0 = (1,1)$, the probability of the outcome (θ_1, X) is roughly the same under both $\Lambda^{\{01\}}$ and $\Lambda^{\{11\}}$. In Figure 9 there is an example of a realization of (θ_1, X) under the law $\Lambda^{\{11\}}$ which has roughly the same probability under the law $\Lambda^{\{01\}}$. This type of characterization of the transition laws can be interpreted as a situation in which the agent being tracked has the ability to behave benignly even when their true intentions are hostile. Figure 9 shows the log-probability of a realization of (θ_1, X) generated by $\Lambda^{\{11\}}$ under all four hypotheses. In the figure it is seen that $\log P((\theta_1, X)|\theta_0 = (1,1)) \approx \log P((\theta_1, X)|\theta_0 = (0,1))$.

In such a situation, regardless of how much noise there is in the observations, it is hard to make an inference amount θ_0^{int} because there are multiple possibilities that are going to be equally likely. Indeed, as seen in Figure 10 the filter is indecisive and the statistic S_k never exceeds the threshold for making the arrest even though $\theta_0 = (1,1)$.

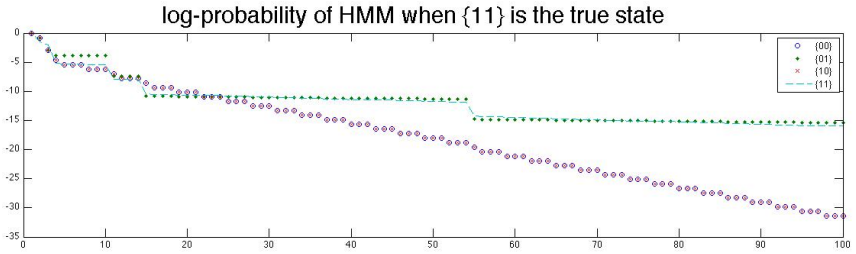

FIGURE 9. **Unsuccessful Tracking Example.** Suppose that we a have a realization $(\theta_1, X) = \{\theta_k^1, X_k\}_{k=1}^{100}$ generated under $\Lambda^{\{11\}}$. The figure plots the probability of the HMM (θ_1, X) under all four possible probability laws, from which we see that $\sum_k \log \Gamma^{\{01\}}(\theta_k^1, X_k|\theta_{k-1}^1, X_{k-1}) \approx \sum_k \log \Gamma^{\{11\}}(\theta_k^1, X_k|\theta_{k-1}^1, X_{k-1})$. This means that the ATR&D may be confident that Pinocchio has a gun, but it will have difficulty in deciding whether his behavior is hostile or benign.

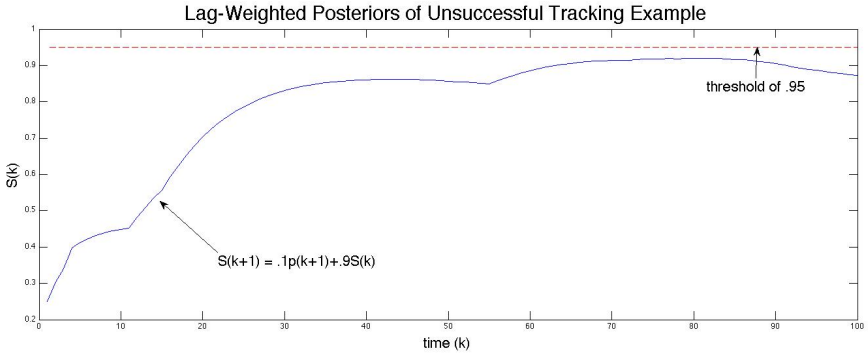

FIGURE 10. **Unsuccessful Tracking Example.** The lag-weighted posteriors, S_k. When $\Lambda^{\{11\}}$ and $\Lambda^{\{10\}}$ are very similar it makes it hard for the ATR&D to make a decision. We see in this figure how S_k never exceeds the 95% threshold.

4. Chapter Summary

Section 1 introduced double Markov processes as a type of Hidden Markov Model with which to model the intentions, poses and actions of a person or robotic agent, and Section 1.2 presented an example. Section 1 formulated a filtering algorithm for double Markov processes for which the necessary priors were available, and then the filter for a tracking intentions example was written explicitly. Section 2 addressed in detail the implementation issues for the filter and considered several techniques for improving algorithm performance. Finally, Section 3 presented an example where the filtering algorithm was carried out in detail and its performance was accessed. Section 3.3 showed that there needed to be significant differences in the transition

probabilities associated with each hypothesis on intentions in order for a filtering-based decision to be made with confidence.

The main result of this chapter is that for a large class of double Markov processes, given an accurate description of the state-space and priors, a filtering algorithm has the potential to track abstract attributes such as intentions and beliefs. The success of the algorithm depends largely on the contrast between transition rates associated with different intentions, otherwise the results of the filter will be ambiguous and the decision-making part of the algorithm will be hampered by uncertainty.

CHAPTER 4

A Kramers-Smoluchowski Approximation Applied to a Multi-Scale HMM

The essential tool in filtering problems is the recursive form of the Baum-Welch equations which are derived from Bayes rule and are used to obtain the posterior distribution of hidden states given observable data. Filtering for linear Gaussian models is done with Kalman filters, which are very fast while also yielding fully optimal estimates. More general filtering problems, including those that are highly nonlinear, rely solely on the Baum-Welch equations to obtain posterior distributions. The task of computing the posterior distribution for every element in the state space often takes enormous amounts of time and so many algorithms avoid optimal filtering and use approximations. However, when the hidden states of the model are constructed in a hierarchical and multi-scaled manner there may be a significant reduction in the computational complexity of the algorithms. In particular, there is potential for a fast filtering algorithm when one or more of the hidden states is fast mean-reverting (FMR) with respect to a local equilibrium state.

Given a vector-valued processe $\theta_t = (\theta_t^0, ..., \theta_t^N)$ where $t > 0$ is time and $N > 0$ is finite, the process is said to be FMR if the relaxation times to local equilibria of some components are fast compared to the mean time between transitions of the other components. For instance, there may be a relaxed set of states that buffers the front components from those in the rear:

$$(\underbrace{\theta_t^1, ..., \theta_t^n}_{\tilde{\theta}_t^0}, \overbrace{\underbrace{\theta_t^{n+1}, ..., \theta_t^{n+m}}_{\tilde{\theta}_t^1}}^{\text{FMR states}}, \underbrace{\theta_t^{n+m+1}, ..., \theta_t^N}_{\tilde{\theta}_t^2})$$

where $law(\tilde{\theta}_t^1) \to \mu_{i,j}$ relatively fast in t for constant $\tilde{\theta}_t^0 \equiv s_i$ and $\tilde{\theta}_t^2 \equiv s_j$. In this case the distribution of θ_t can be approximated by a combination of the invariant measure

of the FMR states and some average measure of the non-FMR states:

$$p_{\tilde{\theta}_t^0 \tilde{\theta}_t^1 \tilde{\theta}_t^2}(i, \ell, j) \approx \bar{p}_{\tilde{\theta}_t^0 \tilde{\theta}_t^2}(i, j) \mu_{i,j}(\ell)$$

where $\bar{p}$ is a joint distribution of $\tilde{\theta}_t^0$ and $\tilde{\theta}_t^2$ which can be computed without integrating over the FMR field. Depending on the system of hierarchical dependencies between components of θ_t, each model will have a different reduced form and so a general treatment of algorithms may be rather abstract. However, probability theory and stochastic differential equations can be used to derive a general form for the distributional approximations for FMR models.

This chapter considers target tracking problems where observations of an environment come in the form of images returned from a low resolution camera. The task is to design an algorithm that tracks a target that may or may not be present in the observed images. Given the observable data and the necessary priors for the model, a filter tracks the target by maintaining an up-to-date posterior probability distribution of some hidden states associated with the target. For instance, filters are often formulated to track some basic physical attribute, such as the target's position, velocity and acceleration, and more elaborate filters are capable of tracking hidden attributes such as a regime or state with which the target may be identified for a brief time period.

Given the necessary priors and model parameters, the recursive filtering algorithm has two steps: (1) *predict*, which is done by solving the forward Kolmogorov equation for the hidden states with the previous posterior as the initial conditioning; (2) *correct*, which is done by weighting the predicted probability of each state with its likelihood given the new data. The computational workload associated with nonlinear filtering algorithms is of the order of the number of hypotheses (usually somewhere between linear and quadratic order), which is practically impossible to compute in real-time. This potential curse of dimensionality is what inhibits the implementation of optimal algorithms in nonlinear problems of high-dimension.

If some hidden states have the FMR property then the computational complexity of filtering can be reduced. However, for this to be realized the time between observations used in the filtering algorithm must be chosen appropriately. If the time between observations is short compared to the relaxation time to equilibrium of the FMR states, then observations arrive too fast for a filter to take advantage of the FMR property. On the other-hand, if the time between observations is so long that there is relaxation of the distribution of the entire process, then it might be the case that potentially important transitions of the hidden states are not being tracked by the filter. Therefore, the observation times need to be well-picked to suit both the parameters of the FMR and non-FMR states. If these observation times are chosen correctly then there will be a reduction in the dimension of the filtering problem, and a reduction in computational cost will follow without reducing the overall performance of the filter.

The main result of this chapter is the derivation and analysis of a Baum-Welch equation that is of reduced dimension and thus faster to compute. The result is applicable to a particular multi-scaled target tracking model, as described in Section 1.4. The mathematical analysis involves the Kramers-Smoluchowski approximation for mean reverting processes, which is done using stochastic differential equations by Nelson [**28**] and Freidlin [**20**]. In the work presented in this chapter, a jump process is introduced among the hidden states as described and analyzed in Section 1.3. The analysis can also be carried out with asymptotic expansions, as is done with stochastic volatility models [**19**], but it is more effective and convenient to analyze the tracking models considered here by using directly the stochastic differential equations.

The conventional methods for filtering are derived in the book by Jazwinski [**25**], including Kalman filters for the basic classes of linear problems. Linear estimations of otherwise nonlinear models, so as to apply Kalman filters, have been tried in numerous filtering problems, as described in the book by Bar-Shalom and Li [**3**] and in the book by Anderson and Moore [**1**]. The nonlinear filtering algorithm has been successfully applied by Rozovksy [**6, 34**] to tracking problems, and by Tartakovsky

[**39, 8**] who has also explored its applications to change point detection algorithms in target tracking. Fast nonlinear filtering algorithms have been considered by many, including Rozovsky and Lototsky [**27**], but particle filters are more frequently used at present [**2, 9**], in particular the Rao-Blackwellized filters proposed by Gustafsson et al [**22, 36**] which look to marginally model nonlinearity with particles and then let the bulk of the filtering be carried out by multiple Kalman filters. Averaging over FMR states while still obtaining consistent approximations can be done with a version of the Kramers-Smoluchowski Theorem, which is presented in general in [**28**] and [**20**]. The book by Fouque, Papanicolaou, and Sircar [**19**] uses similar methods in financial mathematics to model the unobserved stochastic volatility in equities and elsewhere.

Section 1 introduces the tracking model and explains which processes among the hidden states need to be inferred with filtering algorithms. In particular, Section 1.2 explains the traditional nonlinear filter algorithm, Section 1.3 introduces the Kramers-Smoluchowski approximation, and Section 1.4 introduces a filter of reduced dimension which is derived by exploiting the FMR property through the Kramers-Smoluchowski approximation. Section 2 proves the Kramers-Smoluchowski for the model introduced in Section 1, and provides an intuitive analysis and some numerical simulations in sections 2.2 and 2.3. Section 2.4 shows an example of a problem where the reduced filter performs well provided that the rate of mean reversion is fast enough.

1. Target Tracking Models with Fast Mean-Reversion

A basic hierarchical model in target tracking is a vector-valued process denoted by (θ_t, V_t^a, X_t^a). Suppose the target moves in space and observations are noisy pictures taken by a camera equipped with an algorithm to track the target. At time $t \geq 0$, the target velocity is a given vector-valued function $\overline{V}_{\theta_t} \in \mathbb{R}^d$ of a hidden state θ_t, where d is the dimension of the space and θ_t is a positive recurrent continuous-time Markov chain such that

$$\theta_t \in \{1, 2, ..., m\}, \text{ with matrix } \Lambda \text{ of transitions rates.}$$

The transition probabilities of θ_t are given by

$$\mathbb{P}(\theta_{t+\Delta t} = i | \theta_t = j) = P_{ji}(\Delta t) = \mathbf{1}_{\{i=j\}} + \Lambda_{ji}\Delta t + o(\Delta t)\ ,\ i, j = 1, 2 \ldots, m, \tag{30}$$

so that the forward and backward Kolomogorov equations are

$$\frac{dP_{ji}(t)}{dt} = \underbrace{\sum_{\ell} P_{j\ell}(t)\Lambda_{\ell i}}_{\text{forward}} = \underbrace{\sum_{\ell} \Lambda_{j\ell} P_{\ell i}(t)}_{\text{backward}}.$$

We assume that θ_t is a positive and recurrent Markov chain, which means that

$$0 < -\Lambda_{ii} < \infty, \qquad \forall i \leq m$$

$$\text{and} \qquad \mathbb{P}(\theta_t = i | \theta_0 = j) > 0, \qquad \forall i, j \leq m \text{ and } t > 0$$

For a pre-specified parameter $a > 0$, the camera does not track the target perfectly, but instead tracks using a biased velocity $V_t^a \in \mathbb{R}^d$, which after a short time interval becomes an unbiased estimate of the correct velocity. In addition, the camera is mounted on a platform that creates a lot of jitter, which is left unresolved in its pictures. This tracking system can be modeled with the following generalized Ornstein-Uhlenbeck (O-U) process for the velocity

$$dV_t^a = a(\overline{V}_{\theta_t} - V_t^a)dt + \sigma dB_t, \qquad \text{biased velocity (unobserved)}$$

where a is the rate mean reversion for fixed θ_t, and σ is the standard deviation of the jitter. For a sequence of observation times $\{t_k\}_{k=1}^{\infty}$, the observed data Y_k is given by a nonlinear function of the relative position error, which is denoted with a vector $X_t^a \in \mathbb{R}^d$ and is the integral over the interval $[t_{k-1}, t_k]$ of the difference between the true and the biased velocity scaled by a:

$$X_{t_k}^a = X_{t_{k-1}}^a + a\int_{t_{k-1}}^{t_k} (\overline{V}_{\theta_t} - V_t^a)dt, \qquad \text{scaled relative position error (unobserved)}$$

$$Y_k = H(X_{t_k}^a) + Z_k. \qquad \text{2-D pixel image from camera (observed)}$$

Here Z_k is an iid sequence of matrices of mean-zero Gaussian noise with spatial covariance R.

The vector of hidden states (θ_t, V_t^a, X_t^a) is an example of a multi-scale model with an ascending hierarchy of distributional dependencies (i.e. changes in the lower levels occur regardless of what has happened in the higher levels). Specifically, this model describes a jittery camera that is following a target based on the output of an algorithm that tracks the target's position reasonably well (e.g. a heat-seeking device), but does not have the capacity to track the more hidden process θ_t. The table on page 71 gives a detailed description of the model in a hierarchical format.

The joint distribution of the hidden state is denoted by u and the initial conditions of the distribution are considered as given:

$$u_t^a(x, v, i) = p_{X_t^a, V_t^a, \theta_t}(x, v, i)$$

$$u_0^a(x, v, i) = \text{given}$$

The joint distribution satisfies a forward Kolmogorov equation:

$$\frac{\partial}{\partial t} u_t^a = [M^a]^* u_t^a \tag{31}$$

where '$*$' denotes the operator adjoint, and M^a is

$$M^a = \begin{pmatrix} M_1^a & 0 & 0 & \dots & 0 \\ 0 & M_2^a & 0 & \dots & 0 \\ 0 & 0 & M_3^a & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \dots & M_m^a \end{pmatrix}$$

with the operators $\{M_i^a\}_{i=1}^m$ defined in terms of Λ given by equation (30), the gradient with respect to x, and the infinitesimal generator of V_t^a for fixed θ_t:

$$M_i^a = \mathcal{L}_i^a + \Lambda + a(\overline{V}_i - v) \cdot \nabla_x, \qquad \text{for all } i \leq m$$

where $\mathcal{L}_i^a$ is the infinitesimal generator of V_t^a for $\theta_t \equiv i$ given by

$$\mathcal{L}_i^a = \frac{1}{2}\sigma^2 \Delta_v + a(\overline{V}_i - v) \cdot \nabla_v$$

Solutions to equation (31) are generated by a forward operator semi-group $\{\Gamma_t^a\}_{t\geq 0}$. The kernel of this semi-group is the joint transition probability of the three states

$$\Gamma_t^a(x,v,i|x',v',j) = p_{X_t^a,V_t^a,\theta_t}(x,v,i|X_0^a = x', V_0^a = v', \theta_0 = j)$$

Given the initial condition u_0, for any time $t \geq 0$ the joint probability distribution of (X_t^a, V_t^a, θ_t) is given by

$$u_t^a(x,v,i) = \sum_j \int \Gamma_t^a(x,v,i|x',v',j) u_0^a(x',v',j) dx' dv' = \left(\Gamma_t^a u_0\right)(x,v,i)$$

For fixed $\theta_t \equiv i$, the generalized O-U process V_t^a converges in distribution to a normal random variable as $t \to \infty$:

$$V_t^a \Rightarrow N\left(\overline{V}_i, \frac{\sigma^2}{2a} I_{d\times d}\right), \qquad \text{as } t \to \infty, \text{ for } \theta_t \equiv i$$

where $I_{d\times d}$ denotes the d-dimensional identity matrix. For each $i \leq m$, we let μ_i^a denote the invariant distribution of V_t^a for $\theta_t \equiv i$, which we can write explicitly as

$$\mu_i^a(v) = \frac{1}{(\pi\sigma^2/a)^{d/2}} \exp\left(-a\frac{|v-\overline{V}_i|^2}{\sigma^2}\right)$$

1.1. Observations & Maximum Likelihood. For every observation Y_k, the likelihood of the event $\{X_{t_k}^a = x\}$ can be computed by considering the conditional mode of the distribution of Y_k given $X_{t_k}^a$. This is expressed by the function $\psi_k^a : \mathbb{R}^d \to \mathbb{R}^+$ given by

$$\psi_k^a(x) = p(Y_k|X_{t_k}^a = x) \propto \exp\left\{-\frac{1}{2}(\mathbf{Y}_k - \mathbf{H}(x))^* R^{-1}(\mathbf{Y}_k - \mathbf{H}(x))\right\} \tag{32}$$

where $\mathbf{Y}_k$ and $\mathbf{H}(x)$ denote the image data strung out into vectors, and '$*$' denotes vector transpose. The function ψ_k^a is the likelihood function, and the estimate returned from ψ_k^a is called the maximum likelihood estimate (MLE):

$$\widehat{X_{t_k}^a}^{mle} = \arg\max_x \psi_k^a(x)$$

Physical Model	**Mathematical Model**
Kinematic Mode (e.g. maneuvers)	θ_t, a finite state M.C., $\theta_t \in \{1, ..., m\}$ with matrix of transition rates Λ
Target Velocity	$\overline{V}_{\theta_t} \in R^d$
Camera Velocity (biased velocity)	$V_t^a \in \mathbb{R}^d$ where $dV_t^a = a(\overline{V}_{\theta_t} - V_t^a)dt + \sigma dB_t$
Position Error (scaled)	$dX_t^a = a(\overline{V}_{\theta_t} - V_t^a)dt$
Observed noisy image of the target taken by a camera with jitter	$Y_k = H(X_{t_k}^a) + Z_k$ where $\{t_k\}_{k=1}^T$ is a sequence of observation times.

This is a straight forward way to estimate $X_{t_k}^a$, but does not make a fully informed decision because it does not use all the available data. When there is high signal-to-noise ratio, the MLE starts to fail and inference needs to be used. More importantly, for many applications there is no MLE for estimating the hidden processes such as θ_{t_k}, but inferences of the hidden processes can still be made if a filtering algorithm is used.

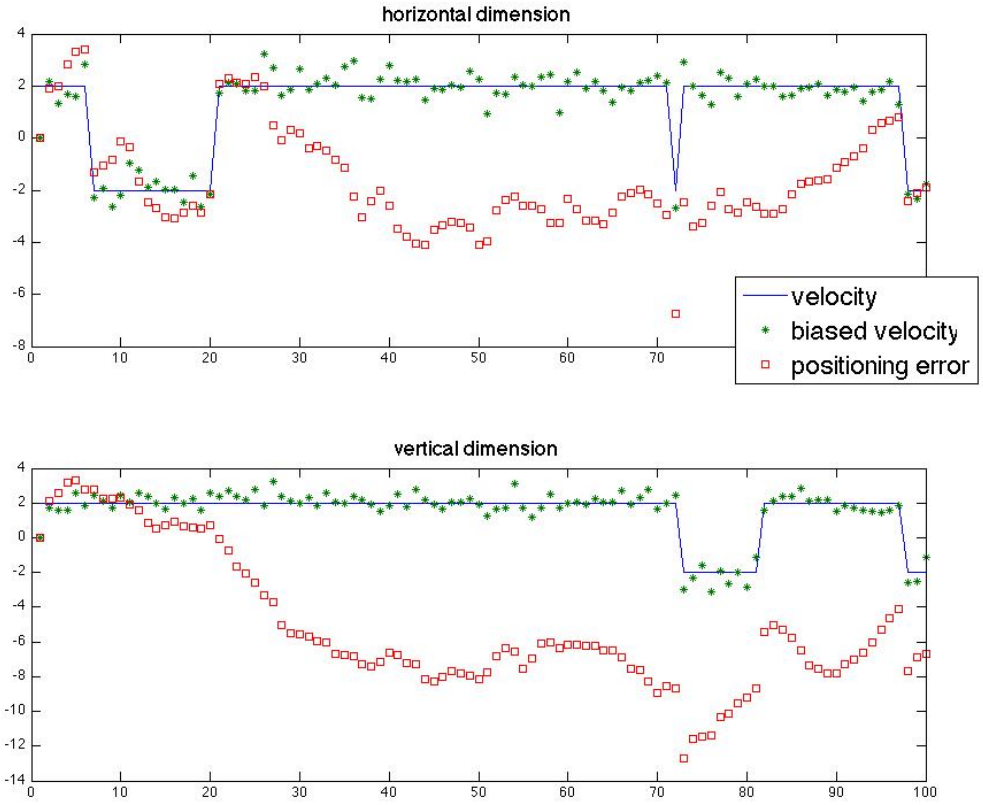

FIGURE 1. A sample of the process (θ_t, V_t^a, X_t^a) in two dimensions with the horizontal axis being time. The mean reversion rate a is such that V_t^a reverts to its mean fairly quickly with time.

1.2. The Non-Linear Filter. Given the observed data $Y_{1:k} = \{Y_1, Y_2, ..., Y_k\}$, the conventional way to track the target is by computing the posterior distribution of the hidden states, (θ_t, V_t^a, X_t^a). The posterior of the hidden states at time t_k is the distribution of $(\theta_{t_k}, V_{t_k}^a, X_{t_k}^a)$ given $Y_{1:k}$, denoted by

$$\pi_k^a(x, v, i) = p_{X_{t_k}^a, V_{t_k}^a, \theta_{t_k}}(x, v, i | Y_{1:k})$$

The forward Baum-Welch Equation recursively updates π_k^a given Y_k and π_{k-1}^a based on the likelihood function given in equation (32) and the priors given by equation (31). The forward Baum-Welch equation recursively computes the posterior distribution as follows:

$$\pi_k^a(x,v,i) = \frac{1}{c_k}\psi_k^a(x)\left(\Gamma_{\Delta t_k}\pi_{k-1}^a\right)(x,v,i), \qquad \text{forward Baum-Welch Equation} \tag{33}$$

where $\Delta t_k = t_k - t_{k-1}$ is the time between observation Y_k and Y_{k-1} $\forall k$, and c_k is a normalizing constant. This filter can be computed "exactly" with numerical integration or Monte Carlo methods.

Given the posterior distribution, there are several optimal estimates of the target's location, including the posterior mean:

$$\widehat{X_k^a} = \sum_i \int x\pi_k^a(x,v,i)dxdv = \mathbb{E}[X_{t_k}^a|Y_{1:k}]$$

and the maximum aposteriori (MAP) estimate:

$$\widehat{\theta_k^a} = \arg\max_i \int\int \pi_k^a(x,v,i)dxdv$$

However, by using the forward operator semi-group Γ, the conventional filter indiscriminately tracks every hidden state which often requires tremendous overhead in computing power. In the next section we identify a regime where the tracking model given in Section 1 is fast mean-reverting and allows us to forgo the tracking of the jittery state V_t^a, and focus on the relevant states, (θ_t, X_t^a).

1.3. Fast Mean Reversion. This is a direct approach to the analysis of the generalized O-U process for the biased velocity process V_t^a and the relative position error X_t^a. It generalizes the analysis by Nelson [**28**] to include the Markov chain θ_t. We write both X_t^a and V_t^a in integral form:

$$V_t^a = e^{-at}V_0^a + a\int_0^t e^{-a(t-s)}\overline{V}_{\theta_s}ds + \sigma\int_0^t e^{-a(t-s)}dB_s \tag{34}$$

$$X_t^a = X_0^a + a\int_0^t (\overline{V}_{\theta_s} - V_s^a)ds = X_0^a + \int_0^t dV_s^a - \sigma B_t = X_0^a + (V_t^a - V_0^a) - \sigma B_t \tag{35}$$

We will show that in general

$$X_t^a - X_0^a = \overline{V}_{\theta_t} - \overline{V}_{\theta_0} - \sigma B_t + E_t^a \tag{36}$$

where E_t^a is an error process that tends to zero point-wise (but not uniformly) almost-surely as $a \to \infty$, so that for any $t > 0$ fixed the relative position error will converge to a limit $\tilde{X}_t$ with probability 1:

$$\tilde{X}_t - \tilde{X}_0 = \lim_{a \to \infty} (X_t^a - X_0^a) = \overline{V}_{\theta_t} - \overline{V}_{\theta_0} - \sigma B_t \tag{37}$$

as $a \to \infty$. The limit in (37) is the theorem that covers the generalized Kramers-Smoluchowski approximation. Note that in this limit the error process makes changes in the hidden state θ_t more observable, through the given velocity function $\overline{V}_i$, and with additive Gaussian white noise.

1.4. FMR Models & Baum-Welch Equations of Reduced Dimension. The multi-scale process composed of (θ_t, V_t^a, X_t^a) as described in the previous sections is unobservable and there is an observation of the following form:

$$Y_k = H(X_{t_k}^a) + Z_k$$

where $H(\cdot)$ is a known nonlinear function of the error process, and Z is a noise process with a known distribution. We say that V_t^a is an FMR state if it satisfies **the FMR condition:**

$$a \gg \max_i(-\Lambda_{ii}) \tag{38}$$

Assuming that (θ_t, V_t^a, X_t^a) is in a regime that agrees with (54), the Kramers-Smoluchowski approximation from equation (37) can be used to bypass the dependence on V_t^a so that the algorithm can use an approximation of the Baum-Welch equations that is of reduced dimension. As noted in the introduction, the time interval between observation $\Delta t_k = t_k - t_{k-1}$ must be chosen so that

$$\frac{1}{a} \ll \Delta t_k \ll \min_i \frac{-1}{\Lambda_{ii}},$$

in which case V_t^a can be ignored and filtering can be carried out only with the limit process $(\theta_t, \tilde{X}_t)$. Such an algorithm is considered a 'reduced filter' provided that (54) is the valid, the filter will be computed faster with very little loss in overall performance.

In the limit as $a \to \infty$, there is the reduced posterior of $(\theta_{t_k}, \tilde{X}_{t_k})$, written as

$$\bar{\pi}_k(x,i) = p_{\tilde{X}_{t_k},\theta_{t_k}}(x,i|Y_{1:k}).$$

It is calculated recursively with the reduced Baum-Welch Equation:

$$\bar{\pi}_k(x,i) = \frac{1}{c_k}\tilde{\psi}_k(x)\sum_j \int \bar{\Gamma}_{\Delta t_k}(x,i|x',j)\bar{\pi}_{k-1}(x',j)dx' \tag{39}$$

where $\bar{\Gamma}$ is a the forward semi-group operator that generates solutions to the forward Kolmogorov equation for the joint distribution of (θ_t, X_t). From the Kramers-Smoluchowski theorem, the kernel of the semi-group operator $\bar{\Gamma}_{\Delta t_k}(x,i|x',j)$ is given by

$$\bar{\Gamma}_{\Delta t_k}(x,i|x',j) = \left(2\pi\sigma\sqrt{\Delta t_k}\right)^{-d/2} \exp\left\{-\frac{1}{2}\left\|\frac{x-x'-(\overline{V}_i - \overline{V}_j)}{\sigma\sqrt{\Delta t_k}}\right\|^2\right\} P_{ji}(\Delta t_k)$$

with $\tilde{\psi}_k(x)$ as the likelihood function as in (32):

$$\tilde{\psi}_k(x) = p(Y_k|\tilde{X}_{t_k} = x) = p_z(Y_k - H(x))$$

Notice that the biased velocity process V_t^a is absent in (39).

2. Analysis of the Generalized Kramers-Smoluchowski Approximation

This section provides some qualitative methods for understanding the Kramers-Smoluchowski approximation for the model proposed in Section 1. In Section 2.1 there is a derivation of the approximation from equation (36) and a proof of the limit from equation (37). In Sections 2.2 and 2.3 there is analysis in discrete time settings to give further intuition into the asymptotic behavior, and there is a detailed explanation of the numerical schemes used to simulate the SDEs and the Kramers-Smoluchowski limit.

2.1. The Generalized Kramers-Smoluchowski Approximation. The Kramers-Smoluchowski limit in equation (37) with a jump process is the main analytical result of this chapter. The first thing to notice is that the limiting process is discontinuous while for any $a < \infty$ the process X_t^a is continuous with probability 1. Therefore,

contrary to what occurs in the classical Kramers-Smoluchowski limit [**28, 20**], convergence in probability under the uniform norm does not hold:

$$\lim_{a\to\infty} \mathbb{P}\left(\max_{t\in[t_k,t_{k+1}]} |X_t^a - (X_{t_k}^a + \overline{V}_{\theta_t} - \overline{V}_{\theta_{t_k}} - \sigma(B_t - B_{t_k}))| > \beta \right) > 0,$$

for some $\beta > 0$. However, we can still show convergence still holds pt.-wise in some sense.

We now derive an expression for the error in equation (36). The analysis begins by calculating the time integral of the biased velocity process:

$$a\int_{t_k}^{t_{k+1}} V_s^a ds = (1 - e^{-a\Delta t_k})V_{t_k}^a$$

$$+a\int_{t_k}^{t_{k+1}} ds \int_{t_k}^{s'} e^{-a(s-s')}\overline{V}_{\theta_{s'}} ds' + a\sigma \int_{t_k}^{t_{k+1}} ds \int_{t_k}^{s'} e^{-a(s-s')} dB_{s'}$$

where $\Delta t_k = t_{k+1} - t_k$. The Brownian integral is rewritten by interchanging the two integrations:

$$\begin{aligned}
\sigma\int_{t_k}^{t_{k+1}} ds \int_{t_k}^{s'} e^{-a(s-s')} dB_{s'} &= \sigma \int_{t_k}^{t_{k+1}} dB_{s'} \int_{s'}^{t_{k+1}} e^{-a(s-s')} ds \\
&= \frac{\sigma}{a}\int_{t_k}^{t_{k+1}} (1 - e^{-a(t_{k+1}-s')}) dB_{s'} \\
&= \frac{\sigma}{a}\Delta B_{t_k} - \frac{\sigma}{a}\int_{t_k}^{t_{k+1}} e^{-a(t_{k+1}-s)} dB_s
\end{aligned}$$

where $\Delta B_{t_k} = B_{t_{k+1}} - B_{t_k}$. Then, with a similar calculation it becomes

$$a\int_{t_k}^{t_{k+1}} ds \int_{t_k}^{s'} e^{-a(s-s')}\overline{V}_{\theta_{s'}} ds' = \int_{t_k}^{t_{k+1}} \overline{V}_{\theta_s} ds - \int_{t_k}^{t_{k+1}} e^{-a(t_{k+1}-s)}\overline{V}_{\theta_s} ds$$

Putting these expressions together we have the following recursive formula for X_t^a:

$$\begin{aligned}
X_{t_{k+1}}^a &= X_{t_k}^a + a\int_{t_k}^{t_{k+1}} (\overline{V}_{\theta_s} - V_s^a) ds \\
&= X_{t_k}^a - (1 - e^{-a(\Delta t_k)})V_{t_k}^a + a\int_{t_k}^{t_{k+1}} e^{-a(t_{k+1}-s)}\overline{V}_{\theta_s} ds
\end{aligned}$$

$$-\sigma\Delta B_{t_k} + \sigma\int_{t_k}^{t_{k+1}} e^{-a(t_{k+1}-s)}dB_s$$

where $\Delta t_k = t_{k+1} - t_k$. This can be rewritten by adding and subtracting terms

$$X^a_{t_{k+1}} = X^a_{t_k} + a\int_{t_k}^{t_{k+1}} (\overline{V}_{\theta_s} - V^a_s)ds$$

$$= X^a_{t_k} + (1 - e^{-a\Delta t_k})(\overline{V}_{\theta_{t_{k+1}}} - \overline{V}_{\theta_{t_k}} + \overline{V}_{\theta_{t_k}} - V^a_{t_k}) + \int_{t_k}^{t_{k+1}} ae^{-a(t_{k+1}-s)}(\overline{V}_{\theta_s}$$

$$-\overline{V}_{\theta_{t_k}})ds - \sigma\Delta B_{t_k} + \sigma\int_{t_k}^{t_{k+1}} e^{-a(t_{k+1}-s)}dB_s$$

Comparing this with (36) we get the following expression for the error process over the interval Δt_k:

$$E^a_{\Delta t_k} = \underbrace{-e^{-a\Delta t_k}(\overline{V}_{\theta_{t_{k+1}}} - \overline{V}_{\theta_{t_k}})}_{E^{a1}_{\Delta t_k}} + \underbrace{(1 - e^{-a\Delta t_k})(\overline{V}_{\theta_{t_k}} - V^a_{t_k})}_{E^{a2}_{\Delta t_k}} \tag{40}$$

$$+ \underbrace{\int_{t_k}^{t_{k+1}} ae^{-a(t_{k+1}-s)}(\overline{V}_{\theta_s} - \overline{V}_{\theta_{t_k}})ds}_{E^{a3}_{\Delta t_k}} + \underbrace{\sigma\int_{t_k}^{t_{k+1}} e^{-a(t_{k+1}-s)}dB_s}_{E^{a4}_{\Delta t_k}} \tag{41}$$

Using the explicit expression of equations (40) and (41), the limit in equation (36) can now be proved by showing that each of the four terms on the right-hand side of (40) and (41) goes to zero in almost-surely.

- The first term goes to zero almost-surely because $\overline{V}_{\theta_t}$ is bounded:

$$|E^{a1}_{\Delta t_k}| \leq 2e^{-a\Delta t_k}\|\overline{V}\| \overset{a\to\infty}{\longrightarrow} 0$$

- The Brownian term, $E^{a4}_{\Delta t_k}$, can be shown to go to zero almost-surely with an application of stochastic integration by parts:

$$\int_{t_k}^{t_{k+1}} e^{-a(t_{k+1}-s)}dB_s$$

$$= B_{t_{k+1}} - e^{-a(t_{k+1}-t_k)}B_{t_k} - a\int_{t_k}^{t_{k+1}} e^{-a(t_{k+1}-s)}B_s ds$$

$$\xrightarrow{a.s.} B_{t_{k+1}} - B_{t_{k+1}} = 0, \qquad \text{as } a \to \infty$$

where the continuity of B_t was applied in order to extract $B_{t_{k+1}}$ from the integral as $a \nearrow \infty$. From this it follows that $E^{a4}_{\Delta t_k} \xrightarrow{a.s.} 0$ as $a \to \infty$.

- For the second error term, $E^{a2}_{\Delta t_k}$, we can apply the fact that $E^{a4}_{\Delta t_k} \xrightarrow{a.s.} 0$ as $a \to \infty$ to the following general formula for $V^a_{t_k}$,

$$V^a_{t_k} = e^{-at_k}V^a_0 + (1 - e^{-at_k})\overline{V}_{\theta_{t_k^-}} + \sigma \int_0^{t_k} e^{-a(t_k - s)} dB_s \xrightarrow{a.s} \overline{V}_{\theta_{t_k^-}}, \qquad \text{as } a \nearrow \infty$$

but $\mathbb{P}(\overline{V}_{\theta_{t_k}} \neq \overline{V}_{\theta_{t_k^-}}) = 0$, and so

$$V^a_t \xrightarrow{a.s.} \overline{V}_{\theta_{t_k}}, \qquad \text{as } a \to \infty$$

which is sufficient to prove $E^{a2}_{\Delta t_k} \xrightarrow{a.s.} 0$ as $a \to \infty$.

- For the third term, $E^{a3}_{\Delta t_k}$, let $\tau = \max\{t \in [t_k, t_{k+1}] : \theta_t \neq \theta_{t^-}\}$. With probability 1 the last jump time happens before the end of the interval, $\mathbb{P}(\tau < t_{k+1}) = 1$, and so by dividing the integral at τ on sets of positive measure we have:

$$\left|E^{a2}_{\Delta t_k}\right| = \left| \int_{t_k}^{t_{k+1}} a(\overline{V}_{\theta_s} - \overline{V}_{\theta_{t_{k+1}}}) e^{-a(t_{k+1} - s)} ds \right|$$

$$= \left| \int_{t_k}^{\tau} a(\overline{V}_{\theta_s} - \overline{V}_{\theta_{t_{k+1}}}) e^{-a(t_{k+1} - s)} ds + \int_{\tau}^{t_{k+1}} a(\underbrace{\overline{V}_s - \overline{V}_{\theta_{t_{k+1}}}}_{=0 \text{ a.s.}}) e^{-a(t_{k+1} - s)} ds \right|$$

$$\overset{a.s.}{=} e^{-a(t_{k+1} - \tau)} \left| \int_{t_k}^{\tau} a(\overline{V}_{\theta_s} - \overline{V}_{\theta_{t_{k+1}}}) e^{-a(\tau - s)} ds \right| \leq 2e^{-a(t_{k+1} - \tau)} \|\overline{V}\| \int_{t_k}^{\tau} a e^{-a(\tau - s)} ds$$

$$= 2e^{-a(t_{k+1} - \tau)} \|\overline{V}\| \left(1 - e^{-a(\tau - t_k)}\right) \xrightarrow{a \to \infty} 0$$

This is sufficient for $E^{a3}_{\Delta t_k} \xrightarrow{a.s.} 0$ as $a \to \infty$ and completes the proof of equation (37).

2.2. Kramers-Smoluchowski Approximation in Discrete Time. In this section the Kramers-Smoluchowski approximation is shown to be the limit of a discrete-time approximation of the multi-scaled system. The integral solutions of the SDEs are approximated with discrete sums with time-step Δt, and the analysis shows that the discrete approximations converge almost surely to the Kramers-Smoluchowski limit as the rate of mean reversion increases. This section is not an alternative proof to equations (36) or (37), but rather serves as an analytical alternative in the understanding of how the limiting behavior arises.

In Section 2.1 it was shown that the integral form of the SDE for X_t^a is

$$X_t^a = X_0^a - (1-e^{-at})V_0^a + a\int_0^t \overline{V}_{\theta_s} e^{-a(t-s)} ds - \sigma B_t + \sigma \int_0^t e^{-a(t-s)} dB_s$$

$$= X_0^a - (1-e^{-at})V_0^a + \int_0^t \overline{V}_{\theta_s} d\left(e^{-a(t-s)}\right) - \sigma B_t + \sigma \int_0^t e^{-a(t-s)} dB_s$$

The integrals can be approximated by taking a very small time step Δt such that

$$0 = t_0 < t_1 < ..... < t_n = t$$

where $t_\ell = \ell \Delta t$ for $\ell = 0, 1, 2, ..., n$. Letting

$$g = e^{-a\Delta t}$$

define the discrete-time process $[X]_{t_n}^a \approx X_{t_n}^a$ to be a Riemann approximation of the integrals:

$$(42)\quad [X]_{t_n}^a = X_0^a - (1-g^n)V_0^a + \sum_{\ell=0}^{n-1} \overline{V}_{\theta_{t_{\ell+1}}} \left(g^{n-(\ell+1)} - g^{n-\ell}\right) - \sigma B_{t_n} + \sigma \sum_{\ell=0}^{n-1} g^{n-\ell} \Delta B_{t_\ell}$$

where $\Delta B_{t_\ell} = B_{t_{\ell+1}} - B_{t_\ell}$ for all ℓ and the middle integral is a forward sum. It is important to approximate the Itô integral with a backward Riemann sum to avoid the bias introduced by Stratanovich-type integrals (for more on Riemann-type approximations of SDEs see the book by Øxendale [**30**].) Rearranging terms, a step-wise recursive formula can be written,

$$[X]_{t_n}^a + V_0^a + \sigma B_{t_n} - X_0^a = g^n V_0^a + (1-g) \sum_{\ell=0}^{n-1} \overline{V}_{\theta_{t_{\ell+1}}} g^{n-(\ell+1)} + \sigma \sum_{\ell=0}^{n-1} g^{n-\ell} \Delta B_{t_\ell}$$

$$= g\left([X]^a_{t_{n-1}} + V^a_0 + \sigma B_{t_{n-1}} - X^a_0\right) + (1-g)\overline{V}_{\theta_{t_n}} + \sigma g \Delta B_{t_{n-1}}$$

and then grouping terms correctly, we have a step-wise recursion for $[X]^a_{t_n}$:

$$[X]^a_{t_n} = g[X]^a_{t_{n-1}} + (1-g)X^a_0 - \sigma(1-g)B_{t_n} + (1-g)\left(\overline{V}_{\theta_{t_n}} - V^a_0\right) \tag{43}$$

which will converge in probability as Δt shrinks:

$$[X]^a_{t_n} \overset{\Delta t \to 0}{\longrightarrow} X^a_t$$

In addition the recursive form of equation (43) reveals the Kramers-Smoluchowski approximation for fixed Δt as the rate of mean reversion grows:

$$[X]^a_{t_n} - [X]^a_{t_{n-1}} \xrightarrow{a.s.} \overline{V}_{\theta_{t_n}} - \overline{V}_{\theta_{t_{n-1}}} - \sigma \Delta B_{t_{n-1}}, \qquad \text{as } a \to \infty$$

where $\Delta B_{t_{n-1}} = B_{t_n} - B_{t_{n-1}}$.

2.3. Numerical Simulations for the Kramers-Smoluchowski Approximation. The analysis of the SDEs so far has been theoretical and now needs to be put in an algorithmic form for numerical simlations. In general, an Euler-Maruyama or Milstein scheme can be used to generate approximate solutions to an SDE provided that the SDE's coefficients allow for certain stability conditions to hold (see [**23, 35**] for more on numerical schemes for SDEs). However, the SDE for the biased velocity has a coefficient with the factor a with $a \to \infty$, and so Euler-Maruyama, Milstein and similar schemes will have several places where division and multiplication by 'a' can create numerical error that is not accounted for by the theoretical analysis. The goal in this section is to find a discrete approximation scheme for the tracking model (θ_t, V^a_t, X^a_t) that is well-behaved as $a \to \infty$.

Generating the process θ_t is straight forward. Given $\theta_0 = j$ and any time $t > 0$, a sample from $\mathbb{P}(\theta_t | \theta_0 = j)$ is taken from the exponential of the matrix Λ of transition rates. Assuming that Λ is diagonalizable, its canonical decomposition is written as $\Lambda = QDQ^{-1}$ where the D is a diagonal matrix with the eigenvalues of Λ on the

diagonal, and Q is a matrix whose columns are the eigenvectors of Λ. The sampling distribution is then

$$P(\theta_t = i|\theta_0 = j) = (e^{\Lambda t})_{ji} = (Qe^{Dt}Q^{-1})_{ji} \tag{44}$$

Time-homogeneity of θ_t allows this sampling distribution to be used sequentially for any initial time and any length transition period.

The process V_t^a on the other-hand is more complicated. A forward Euler approximation for this scheme is not a good choice because for obvious reasons there will be numerically unstable behavior for $a \gg 1$. For instance, for a fixed time step Δt and $a < \infty$, the Euler-Maruyama scheme is

$$V_{t+\Delta t}^a = V_t^a + a(\overline{V}_{\theta_t} - V_t^a)\Delta t + \sigma \Delta B_t$$

where $\Delta B_t = B_{t+\Delta t} - B_t$. This scheme is clearly not a good discretization of the processes in an asymptotic setting because the amplification factor is $a \nearrow \infty$.

An alternative scheme can be found by looking at V_t^a in integrated form. Letting $g = e^{-a\Delta t}$ it is easily seen that the process $V_{t+\Delta t}^a$ conditioned on V_t^a is given in integral form as

$$V_{t+\Delta t}^a = e^{-a\Delta t}V_t^a + a\int_t^{t+\Delta t} e^{-a(t+\Delta t-s)}\overline{V}_{\theta_s}ds + \sigma\int_t^{t+\Delta t} e^{-a(t+\Delta t-s)}dB_s$$

$$\approx gV_t^a + g\overline{V}_{\theta_{t+\Delta t}}\int_t^{t+\Delta t} d\left(e^{-a(t-s)}\right) + \sigma\int_t^{t+\Delta t} e^{-a(t+\Delta t-s)}dB_s$$

$$= gV_t^a + g(g^{-1}-1)\overline{V}_{\theta_{t+\Delta t}} + \sigma\int_t^{t+\Delta t} e^{-a(t+\Delta t-s)}dB_s$$

$$=_d gV_t^a + (1-g)\overline{V}_{\theta_{t+\Delta t}} + \underbrace{\sigma\sqrt{\frac{1-g^2}{2a\Delta t}}}_{=\tilde{\sigma}^a}\Delta B_t$$

which yields a discrete scheme which is asymptotically stable in the rate of mean reversion.

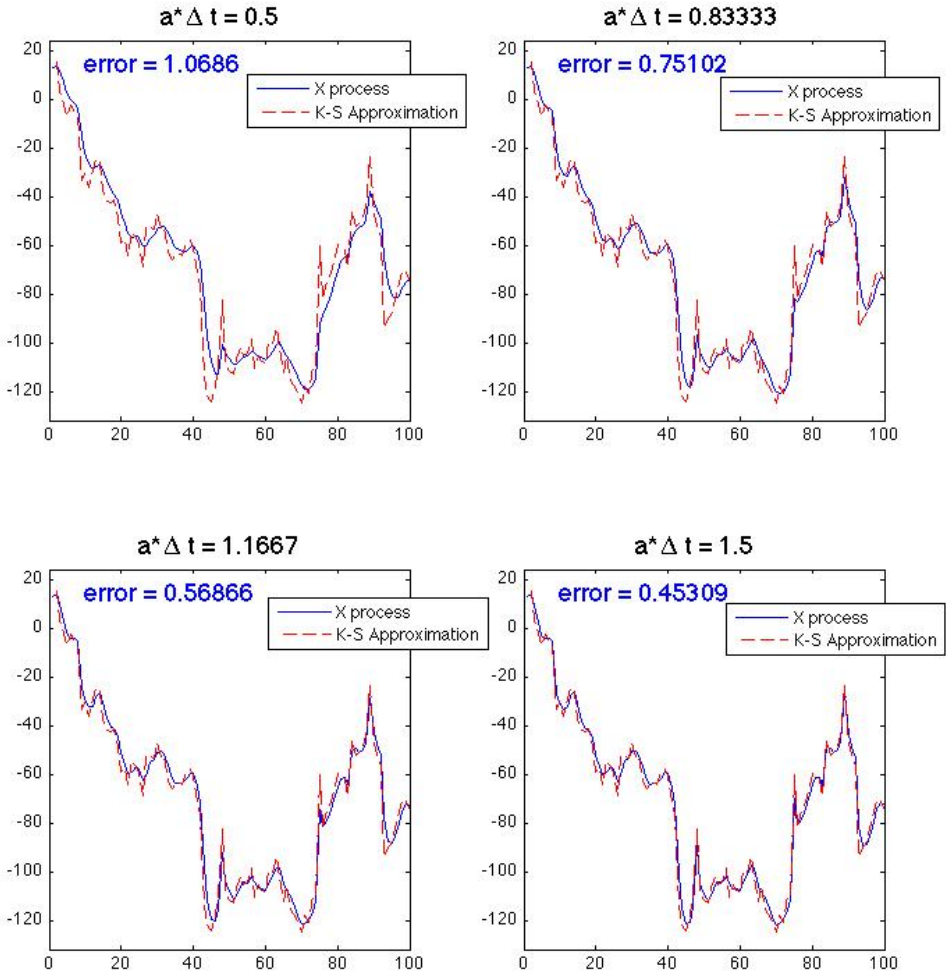

FIGURE 2. This figure displays the convergence of the process X_k^a to the Kramers-Smoluchowski approximation as $a \to \infty$. The sample processes are generated using equations (44), (45) and (46). For several realizations of (θ_k, B_k) the exact solution is compared to the approximation. The error is the empirical average of MSE, $MSE = \sqrt{\mathbb{E}\sum_k |X_k^a - \tilde{X}_k|^2}$ where $\tilde{X}_k$ denotes the Kramers-Smolushowski approximation.

Finally, to generate the process X_t^a, we consider the form of X_t^a as was shown in equation (35):

$$X_{t+\Delta t}^a = X_t^a + V_{t+\Delta t}^a - V_t^a - \sigma \Delta B_t$$

This equation is a scheme for generating realizations of X_t^a from realizations of V_t^a, and it also illustrates how X_t^a is indeed a velocity. Therefore, for a discretization of

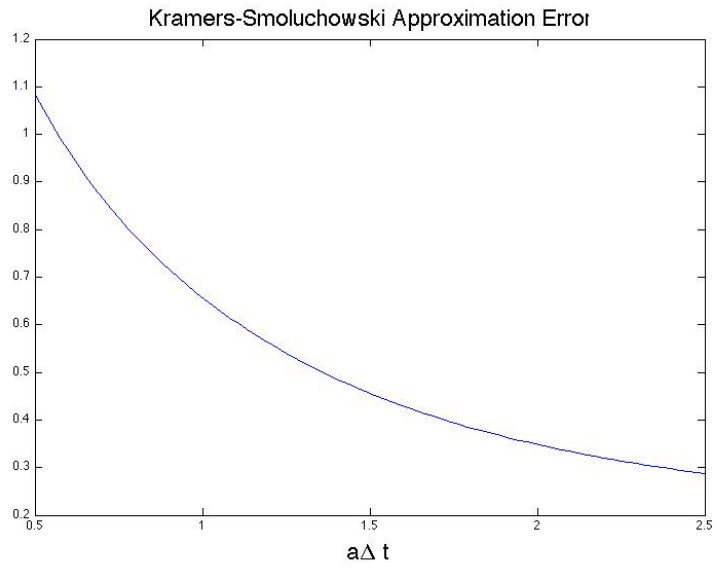

FIGURE 3. In this figure the log-error of X_k^a compared with its Kramers-Smoluchowski approximation is computed as $a \to \infty$. The sample processes are generated using equations (44), (45) and (46). The vertical axis in the plot is an empirical approximation of $\log\left(\sqrt{\mathbb{E}\sum_k |X_k^a - \tilde{X}_k|^2}\right)$ where $\tilde{X}_k$ denotes the Kramers-Smolushowski approximation.

the time-domain $\{t_k\}_{k=0}^n$ where

$$0 = t_0 < t_1 < ..... < t_n = t$$

with $t_{k+1} - t_k = \Delta t$ for all $k < n$, there is the following discrete scheme:

$$V_{t_{k+1}}^a = gV_{t_k}^a + (1-g)\overline{V}_{\theta_{t_{k+1}}} + \underbrace{\sigma\sqrt{\frac{1-g^2}{2a\Delta t}}}_{=\tilde{\sigma}^a}\Delta B_{t_k} \tag{45}$$

$$X_{t_{k+1}}^a = X_{t_k}^a + V_{t_{k+1}}^a - V_{t_k}^a - \sigma\Delta B_{t_k} \tag{46}$$

where $\Delta B_{t_k} = B_{t_{k+1}} - B_{t_k}$.

Approximate realizations of (θ_t, V_t^a, X_t^a) can be generated with equations (44), (45) and (46) for any $a > 0$, and the distribution of the these approximate realizations will converge to the correct distribution as $\Delta t \searrow 0$.

The sub-scripts in the time points will be dropped for brevity in the notation, $t_k \to k$, whenever it is appropriate. Given $a > 0$, (θ_k, V_k^a, X_k^a), the process B_k and a

time step Δt, the process is recursively updated as follows:

(1) Generate θ_{k+1} from the distribution given by:

$$\mathbb{P}(\theta_{k+1} = i | \theta_k = j) = (e^{\Lambda \Delta t})_{ji}$$

(2) Generate V^a_{k+1} given θ_{k+1} and V^a_k:

$$V^a_{k+1} = gV^a_k + (1-g)\overline{V}_{\theta_{k+1}} + \tilde{\sigma}^a \Delta B_k$$

where $g = e^{-a\Delta t}$, $\tilde{\sigma}^a = \sigma\sqrt{(1-g^2)/(2a\Delta t)}$, and $\Delta B_k = B_{k+1} - B_k \sim iidN(0, \sqrt{\Delta t})$.

(3) Generate X^a_{k+1} given V^a_{k+1}, V^a_k and X^a_k:

$$X^a_{k+1} = X^a_k + V^a_{k+1} - V^a_k - \sigma \Delta B_k$$

Provided that $-1/\Lambda_{ii} \gg 1/a$, taking the limit as $a \to \infty$ causes the approximated process to converge almost-surely to the Kramers-Smoluchowski approximation, which can be written recursively in terms of θ_{k+1}, θ_k and B_k:

$$\tilde{X}_{k+1} = \tilde{X}_k + (\overline{V}_{\theta_{k+1}} - \overline{V}_{\theta_k}) - \sigma B_k, \qquad \text{Kramers-Smoluchowski approximation}$$

Figures 2 and 3 compare $\tilde{X}_k$ to the process X^a_k when steps (1), (2) and (3) are used to generate samples. The error is defined as

$$err(a, \Delta t) = \sqrt{\mathbb{E}\sum_k \|X^a_k - \tilde{X}_k\|^2}$$

which is seen in the figure to approach zero as $a\Delta t$ gets large.

2.4. Simulation of FMR Filtering. In this section a simulated example is used to demonstrate the performance of the reduced filter in equation (39). The filter is effective when the parameter a is large enough to make the (54) a valid assumption. In contrast, the results are not so good when the reduced filter is applied to the same model but with a smaller so that (54) does not apply.

Consider the noisy observations shown in figure 4, where the target makes a

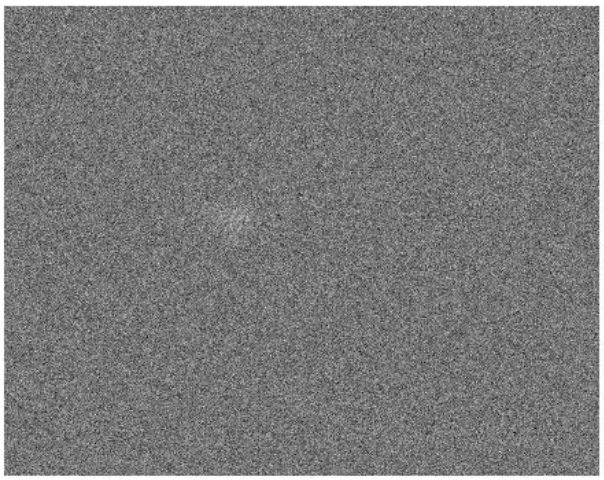

FIGURE 4. An example of an observed image from which the filter must infer the location and state of the target. The dim glow is the true location of the target.

very dim glow in the frame from which the filter must track. Other elements such as obstructions, correlated noise, or multiple targets can be incorporated into the model, but for this example the filter is tested in a simple scenario in order to see that it works. In figure 5 the results are shown for a particular simulation where the target was well-tracked by the filter.

In this example, there are four modes of velocity that the target can take, which we model with the process $\theta_t \in \{1, 2, 3, 4\}$ which makes transitions with the following intensities,

$$\Lambda = \begin{bmatrix} -4 & 4/3 & 4/3 & 4/3 \\ 8/3 & -8 & 8/3 & 8/3 \\ 4 & 4 & -12 & 4 \\ 8 & 8 & 8 & -24 \end{bmatrix}$$

Indeed, this is a diagonalizable matrix and so the exponential can be easily computed to form a matrix of transition probabilities for any time-step $\Delta t > 0$ as was shown in equation (44) of Section 2.3. The actual velocity of the target given θ_t is given by

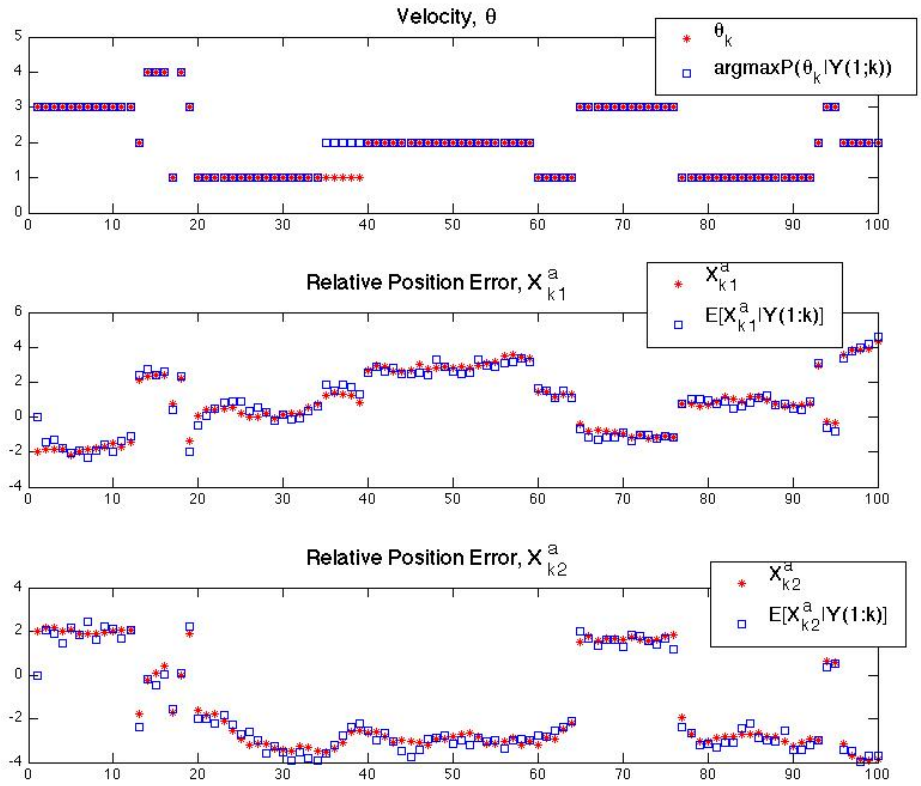

FIGURE 5. The FMR reduced filter from equation (39) with $a = 250$ and $\Delta t = .01$. Both the processes θ_k and X_k^a are tracked based on the information provided by a sequence of frames that look like the one given in figure 4. The significant spots are the changes in the kinematic mode. The filter is able to catch the changes without tracking the process V_k^a.

the discrete function $\overline{V}$,

$$\overline{V} = [\overline{V}_1, \overline{V}_2, \overline{V}_3, \overline{V}_4] = \left[\begin{pmatrix} -.2 \\ -.2 \end{pmatrix}, \begin{pmatrix} -.2 \\ .2 \end{pmatrix}, \begin{pmatrix} .2 \\ -.2 \end{pmatrix}, \begin{pmatrix} .2 \\ .2 \end{pmatrix} \right]$$

The model parameters for the SDEs for V_t^a and X_t^a are taken as follows: $a = 250$, and $\sigma = 4$. The discrete schemes from Section 2.3 are used with time-step $\Delta t = .01$ to generate $V_k^a = (V_{k1}^a, V_{k2}^a) \in \mathbb{R}^2$ and $X_k^a = (X_{k1}^a, X_{k2}^a) \in \mathbb{R}^2$. For such a choice of parameters, the coefficient g is very small,

$$g = e^{-a\Delta t} = e^{-10} \approx 4 \times 10^{-5} \ll 1,$$

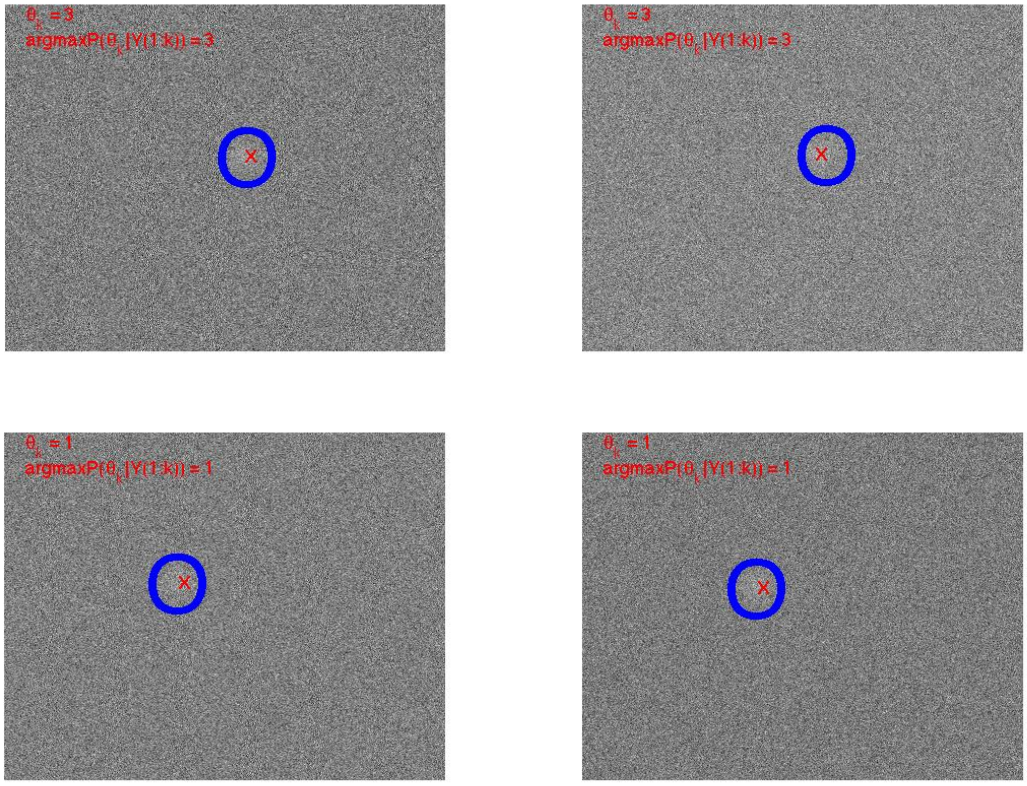

FIGURE 6. The filtering of the processes as seen from the camera's view-finder. The red 'X' marks the target's true location, and the blue 'O' marks the filtering estimate of the location.

and so this example agrees with equation (54) and can safely be considered in the FMR regime.

To approximate the domain of $X_k^a \in \mathbb{R}^2$, the real line is approximated by a pair of 1000-point lattices $\{x_i\}_{i=1}^{1000}$ so that

$$x_i = -20 + (i-1)\frac{40}{999}, \qquad \text{for } i = 1, 2, 3, ..., 1000$$

to which the posterior assigns mass to each pair $\{(x_i, x_j)\}_{i,j=1}^{1000}$. Letting π_k denote the posterior probability, the probability mass across the discrete fields is labeled as

$$\pi_k^{i,j,\ell} = \mathbb{P}\left(X_k^a = (x_i, x_j), \theta_k = \ell | Y_{1:k}\right),$$

for $i, j = 1, 2, 3...., 1000$.

The observation model generates a 1000×1000 pixel image $Y_k = [Y_k^{ij}]_{i,j=1}^{1000}$ where for each (i, j) the pixel intensity is given by:

$$Y_k^{ij} = \exp\left(-.5\|(x_i, x_j) - X_k^a\|^2\right) + \gamma Z_k^{ij}$$

where $\gamma = 1$, $Z_k^{ij} \sim iidN(0, 1)$, and (x_i, x_j) is an element of the 1000×1000 lattice which divides a 2-D sub-space of $\mathbb{R}^2$. For any (i, j), the likelihood function is then

$$\psi_k^{ij} = \exp\left(-\frac{1}{2\gamma^2} \sum_{(i',j') \in nbh_\gamma(i,j)} \left|\exp\left(-.5\|(x_{i'}, x_{j'}) - (x_i, x_j)\|^2\right) - Y_k^{i'j'}\right|^2\right)$$

where $nbh_\gamma(i, j) = \{(i', j') : \|(x_i, x_j) - (x_{i'}, x_{j'})\| \leq 4\gamma\}$, and the averaged Baum-Welch filtering equation as given by equation (39) yields the following recursive update formula for the posterior:

$$\pi_k^{i,j,\ell}$$

$$= \frac{1}{c_k}\psi_k^{ij} \sum_{\ell'=1}^{m} \left\{ \sum_{(i',j') \in nbh_\sigma^{\ell\ell'}(i,j)} \exp\left(-\frac{1}{2\sigma^2\Delta t}\left\|(x_{i'}, x_{j'}) - (\overline{V}_\ell - \overline{V}_{\ell'})/\sigma + (x_i, x_j)\right\|^2\right)\right\}$$

$$\times \left(e^{\Lambda\Delta t}\right)_{\ell'\ell} \pi_{k-1}^{i',j',\ell'}$$

where $nbh_\sigma^{\ell\ell'}(i, j) = \left\{(i', j') : \|(x_i, x_j) - (\overline{V}_\ell - \overline{V}_{\ell'})/\sigma - (x_{i'}, x_{j'})\| \leq 4\sigma\sqrt{\Delta t}\right\}$. Figures 5 and 6 show a successful tracking of the pair (θ_k, X_k^a) in this parameter regime.

Since the number of lattice points in $nbh_\gamma(\ \cdot\ )$ and $nbh_\sigma(\ \cdot\ )$ is much smaller than the total number of lattice points, the number of flops required for the recursive update formula is of linear order:

$$\mathcal{O}(m \times (\text{the number of lattice points}))$$

This is significantly less work than is required to recursively update with the full Baum-Welch equations in (33), which call for an additional discretization of a 2-D

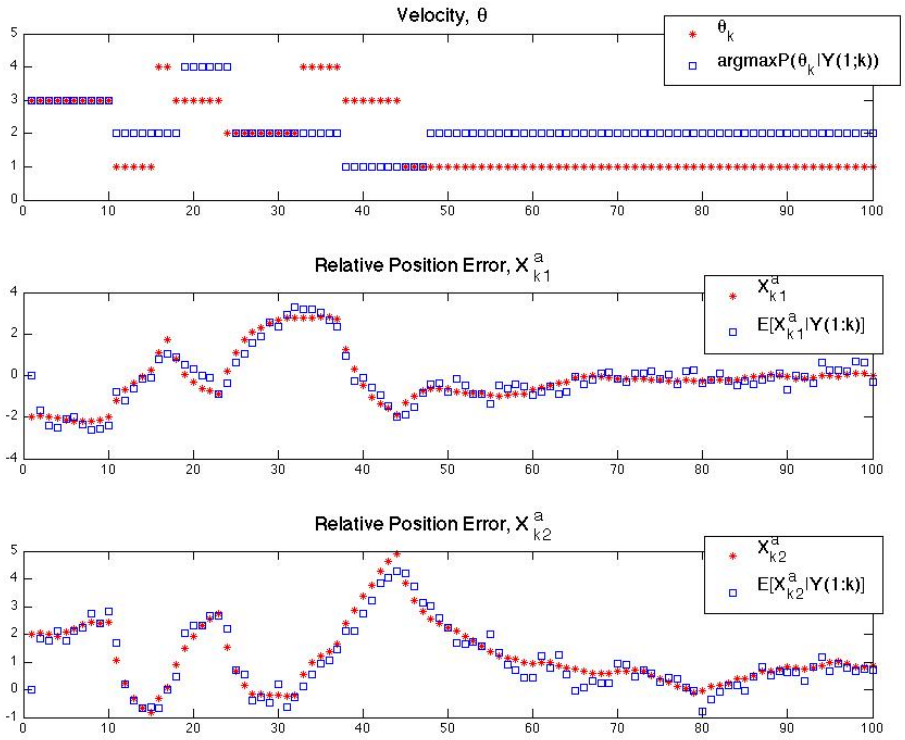

FIGURE 7. The FMR reduced filter from equation (39) with $a = 40$ and $\Delta t = .01$. There is significant error in tracking Θ_k because the (54) is not valid for this parameterization of the model.

sub-space of $\mathbb{R}^2$ for V_k^a. This second 2-D lattice will be approximately the same size as the lattice for X_k^a, and therefore the workload will be increased to quadratic order:

$$\mathcal{O}(m \times (\text{the number of lattice points})^2)$$

This will require significantly more memory and will take a lot more time to compute even in problems of moderate dimensions. For instance, in this section's example, the number of states is $m = 4$ and the number of 2-D lattice points is 1000^2, which means the savings in operations for each recursive update is a factor of one million:

$$4 \times 1000^2 \ll 4 \times 1000^2 \times 1000^2$$

Finally, it is important to examine cases when the reduced filter fails. The simplest way to show the limitations of the reduced filter is to adjust a so that the model

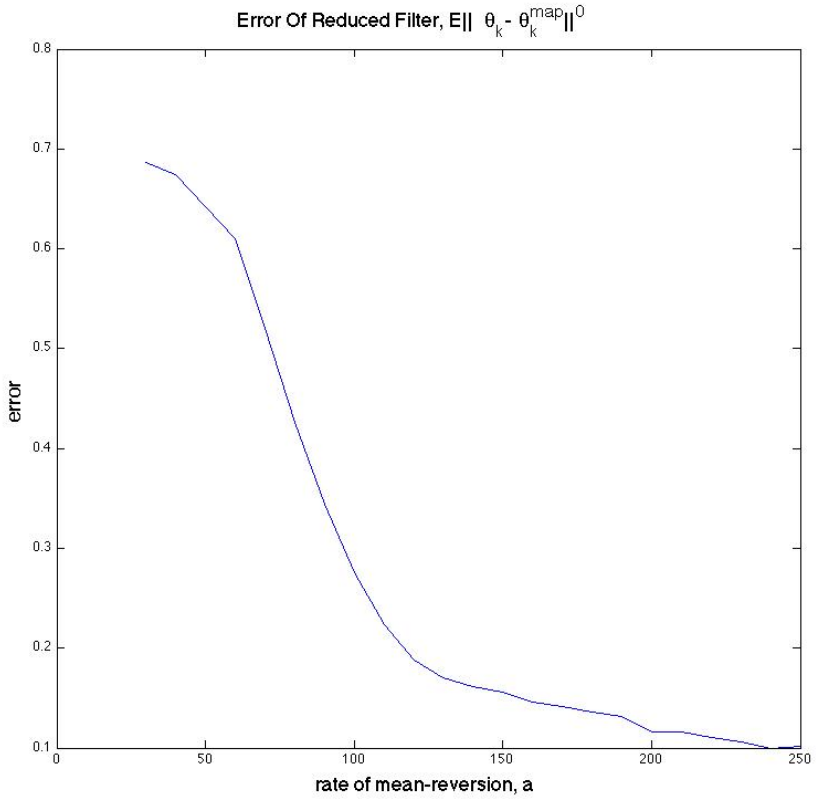

FIGURE 8. As a decreases, the error increases. This plot shows the proportion of time that the MAP estimator of the reduced filter is wrong as a function of a. The vertical axis is the error, $err = \mathbb{E}|\theta_k - \hat{\theta}_k^{map}|^0 = \mathbb{E}\mathbf{1}_{\hat{\theta}_k^{map} \neq \theta_k}$.

disagrees with equation (54). In other-words, choose a so that $g = e^{-a\Delta t}$ is not close to zero, making significant the time it takes for V_t^a to revert back to its conditional mean. In the example, if a is made smaller the error increases, as is seen in figure 7 where $a = 40$. The increase in error as a decreases is further analyzed in figure 8, where the plot shows the expected proportion of time that the MAP estimator is wrong, which can be written as the expectation of the 'zero-norm' or the 0-1 loss-function,

$$err(a) = \mathbb{E}|\theta_k - \hat{\theta}_k^{map}|^0 = \mathbb{E}\mathbf{1}_{\theta_k \neq \hat{\theta}_k^{map}}$$

For the parameter values $40 \leq a \leq 250$, the plot illustrates how the reduced filter cannot track the hidden states for smaller a because it has been applied in setting where the FMR assumption of (54) is not valid.

3. Chapter Summary

To summarize, this chapter has proposed a multi-scale tracking model with states that are fast mean-reverting in such a way that allows for distributional relaxation over time intervals of a certain length. In particular, the main results was the derivation of a nonlinear filtering algorithm of reduced dimension with discrete-time observations which exploits these fast mean-reverting states through a Kramers-Smoluchowski limit. This filter is faster to compute but will not sacrifice too much in terms of accuracy and will be optimal as the rate of mean-reversion approaches infinity. Presented in the chapter were some numerical schemes which can be used to approximate and simulate processes both for filtering and to get sense of how much error is associated with the Kramers-Smoluchowski approximation. Finally, some data was simulated to demonstrate the capacity of the reduced filter to track, and also the inability of the reduced filter to track when misapplied.

CHAPTER 5

Nonlinear Filters of Reduced Dimension for Multi-Scale Stochastic Models

Nonlinear Filtering problems for multi-scale models are optimally solved with the recursive form of the Baum-Welch equations. In practice, such filtering algorithms may be near impossible to implement because of the costly computations associated with updating the hypotheses of the hidden states. However, the class of multi-scale models where one or more states is fast-mean reverting (FMR) may permit filters that are computed on a state-space of reduced dimension. Such filters can be shown to be asymptotically optimal as the rate of mean-reversion goes to infinity.

A multi-scale nonlinear filtering problem with discrete data is formulated as follows:

$$\frac{\partial}{\partial t}\mathbb{P}(\Theta^\epsilon(t)=s_i)=\sum_j Q_{ji}(X^\epsilon(t))\mathbb{P}(\Theta^\epsilon(t)=s_j), \qquad \Theta^\epsilon(t) \text{ is hidden,}$$

$$dX^\epsilon(t)=\frac{1}{\epsilon}\left(\Theta^\epsilon(t)-X^\epsilon(t)\right)dt+\frac{1}{\sqrt{\epsilon}}dW(t) \qquad \text{hidden,}$$

$$\frac{\partial}{\partial t}I^\epsilon(t)=h(X^\epsilon(t)) \qquad \text{hidden,}$$

$$Y_k^\epsilon=I^\epsilon(t_{k+1})-I^\epsilon(t_k)+Z_k \qquad \text{observed}$$

where $\Theta^\epsilon(t)$ is positive recurrent on a finite space, $\epsilon>0$ and $Z_k\sim iidN(0,1)$. The posterior distribution of the finite states is

$$\pi_k^\epsilon(s_i,x,v)=\frac{\partial}{\partial x}\frac{\partial}{\partial v}\mathbb{P}(\Theta^\epsilon(t_k)=s_i,X^\epsilon(t_k)\le x,I^\epsilon(t_k)\le v|Y_{1:k}^\epsilon)$$

for which it is well-known (see the book by Jazwinski or the 1989 paper by Rabiner, [**25, 32**]) that the **forward Baum-Welch** equation recursively updates the posterior

distribution given $(Y_1^\epsilon, ..., Y_{k+1}^\epsilon) = (y_1, ..., y_{k+1})$:

$$\pi_{k+1}^\epsilon(i, x, v) = \frac{1}{c_{k+1}} \frac{\partial}{\partial y} \mathbb{P}(Y_{k+1}^\epsilon \leq y_{k+1} | I^\epsilon(t_{k+1}) = v)$$

$$\times \sum_j \int \int \Gamma_{\Delta t_k}^\epsilon(i, x, v | j, x', v') \pi_k^\epsilon(j, x', v') dx' dv' \tag{47}$$

where Γ^ϵ is the forward operator semi-group which generates the joint distribution of $(\Theta^\epsilon(t), X^\epsilon(t), I^\epsilon(t))$.

Analysis of multi-scale models has been done extensively by Yin and Zhang in [**41, 42**], in which they also analyzed the asymptotics of the Kalman filter and the Wonham filter. Filtering algorithms for systems of this type are important in target tracking [**3, 31, 6**], and can be implemented with particle filters [**2, 9**], or for with Rao-Blackwellized filters in order to exploit the potential for linear filtering [**22, 36**]. Multi-scale models are also of interest in stochastic volatility models [**19**].

Of interest in this thesis is the potential for fast nonfiltering algorithms of reduced dimension when the parameter ϵ is very small. In this chapter, we address the nonlinear filtering problem when data arrives discretely in time and is modeled as a nonlinear function of the hidden states $(\Theta^\epsilon(t), X^\epsilon(t), I^\epsilon(t))$ where $X^\epsilon(t)$ is fast-mean-reverting. The main result is that the nonlinear filter converges point-wise as a function of the data for any test-function,

$$\mathbb{E}[g(\Theta^\epsilon(t_k)) | Y_{1:k}^\epsilon = y_{1:k}] \to \mathbb{E}[g(\Theta^0(t_k)) | Y_{1:k}^0 = y_{1:k}] \qquad \text{as } \epsilon \searrow 0, \forall y_{1:k} \in \mathbb{R}^k$$

where $\Theta^0(t_k)$ and Y_k^0 are the weak limits of $\Theta^\epsilon(t_k)$ and Y_k^ϵ as $\epsilon \searrow 0$, respectively. This chapter contains a proof of this result for processes of the OU type, but can easily be generalized to the class of fast-mean-reverting stochastic differential equations.

1. Definitions, Parameters & the MainResult

Formally, consider the Markov process $(\Theta^\epsilon(t), X^\epsilon(t), I^\epsilon(t))$ for $\epsilon > 0$ and $t \in [0, T]$, on the probability space $(\Omega, \mathcal{F}_t, \mathbb{P})$. Let $\Theta^\epsilon(t)$ be a jump process taking value in the

space of finite values $\mathcal{S} = \{s_1, ..., s_m\}$. The process $X^\epsilon(t)$ is an Ornstein-Uhlenbeck (OU) process with a shifting-mean,

$$dX^\epsilon(t) = \frac{1}{\epsilon}\left(\Theta^\epsilon(t) - X^\epsilon(t)\right)dt + \frac{1}{\sqrt{\epsilon}}dW_t \tag{48}$$

where W_t is Brownian motion and $\epsilon > 0$ is an arbitrarily small parameter. We say that '$1/\epsilon$' is the rate of mean-reversion for $X^\epsilon(t)$. From this, the third process is defined as

$$I^\epsilon(t) = I_0 + \int_0^t h(X^\epsilon(\tau))d\tau$$

where $h : \mathbb{R} \to \mathbb{R}$ is bounded.

The space Ω is then defined as

$$\Omega = D\left([0,T];\mathcal{S}\right) \times C\left([0,T];\mathbb{R}^2\right)$$

were $D\left([0,T];\mathcal{S}\right)$ is the set of right-hand continuous piece-wise constant functions in $\mathcal{S}$. There is a matrix-operator Q where

$$Q(x) = (Q_{ji}(x)) \in \mathbb{R}^{m\times m},$$

which is a transition rate matrix for all $x \in \mathbb{R}$ such that for any bounded function $g(\theta) : \mathcal{S} \to \mathbb{R}$,

$$Q(x)g(\theta)\Big|_{\theta=s_i} = \sum_j Q_{ij}(x)g(s_j),$$

The dynamics of the distribution of $\Theta^\epsilon(t)$ are determined the Martingale,

$$\mathbb{E}\left[g(\Theta^\epsilon(t)) - \int_0^t Q(X^\epsilon(\tau))g(\Theta^\epsilon(\tau))d\tau \Big| \mathcal{F}_s\right] = g(\Theta^\epsilon(s)) - \int_0^s Q(X^\epsilon(\tau))g(\Theta^\epsilon(\tau))d\tau$$

for $s < t$ and any test function g. For simplicity, we assume that $\Theta^\epsilon(t)$ communicates on all possible states of $\mathcal{S}$, and that there are constants $0 < \alpha \leq \beta < \infty$ which insure that jumps cannot occur at an infinite rate and there are no cemetery states,

$$\sup_x(-Q_{ii}(x)) \leq \beta < \infty \tag{49}$$

$$\inf_x(-Q_{ii}(x)) \geq \alpha > 0 \tag{50}$$

for all $i \leq m$. We also impose the initial condition that is independent of ϵ,

$$\mathbb{P}(\Theta^\epsilon(0) = s_i) = \rho_0(s_i), \qquad \text{for any } \epsilon > 0 \text{ and } \forall i \leq m. \tag{51}$$

It is well-known that the infinitesimal generator of $(\Theta^\epsilon(t), X^\epsilon(t), I^\epsilon(t))$ is the operator $\frac{1}{\epsilon}\mathcal{L} + Q(x) + h(x)\frac{\partial}{\partial t}$ where

$$\mathcal{L} = \begin{pmatrix} \mathcal{L}_1 & 0 & \dots & 0 \\ 0 & \mathcal{L}_2 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & \mathcal{L}_m \end{pmatrix}$$

and

$$\mathcal{L}_i = (s_i - x)\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial^2}{\partial x^2}$$

so that for any scalar-valued function $g(\Theta^\epsilon(t), X^\epsilon(t), I^\epsilon(t))$ with compact support, two derivatives in x and one in I (both bounded), the dynamics are determined by the Martingale,

$$\mathbb{E}\left[g(\Theta^\epsilon(t), X^\epsilon(t), I^\epsilon(t)) - \int_0^t \left(\frac{1}{\epsilon}\mathcal{L} + Q + h\frac{\partial}{\partial I}\right) g(\Theta^\epsilon(\tau), X^\epsilon(\tau), I^\epsilon(\tau))d\tau \Big| \mathcal{F}_s\right]$$

$$= g(\Theta^\epsilon(s), X^\epsilon(s), I^\epsilon(s)) - \int_0^s \left(\frac{1}{\epsilon}\mathcal{L} + Q + h\frac{\partial}{\partial I}\right) g(\Theta^\epsilon(\tau), X^\epsilon(\tau), I^\epsilon(\tau))d\tau \tag{52}$$

for some test function g. Equation (52) is the weak form of the backward equation for the Markov process $(\Theta^\epsilon(t), X^\epsilon(t), I^\epsilon(t))$.

A filtering problem arises when $(\Theta^\epsilon(t), X^\epsilon(t), I^\epsilon(t))$ is only observable through data that is a noisy function of the pair. Such hidden variables are referred to as a *Hidden Markov Model* (HMM). Of particular interest are models in which data is a function only of $X^\epsilon(t)$ so that prior knowledge of the HMM's dynamics is a necessary tool with which to make inferences about $\Theta^\epsilon(t)$.

Let the observed data be collected at time-points $(t_k)_{k=1,2,3,\dots}$ such that

$$t_{k+1} - t_k = \Delta t_k, \qquad \forall k \geq 0$$

An observation Y_k^ϵ is a vector-valued function of the unobserved,

$$(53) \quad Y_k^\epsilon = \int_{t_k}^{t_{k+1}} h\left(X^\epsilon(\tau)\right) d\tau + Z_k = I^\epsilon(t_{k+1}) - I^\epsilon(t_k) + Z_k \qquad k = 1, 2, 3, 4., ...$$

where h is a known and bounded function and Z_k is unobserved but known to be an iid random variable with mean zero and finite variance. For convenience, take

$$Z_k \sim iidN(0, R(\Delta t_k)), \qquad \text{where } R(\Delta t_k) = \mathbb{E} Z_k Z_k^*$$

with $\|R(\Delta t_k)\| \sim \sqrt{\Delta t}$. Given the model parameters and some data $Y_{1:k}^\epsilon = (Y_\ell^\epsilon)_{\ell \leq k}$, the posterior distribution of the HMM is denoted as π_k^ϵ and is given recursively by the forward Baum-Welch equation shown in equation (47). The posterior distribution is sometimes referred to as the 'optimal filter' because it is essential for computing optimal estimators of the hidden processes. For instance, letting $\mathcal{F}_k^{Y^\epsilon} = \sigma(Y_{1:k}^\epsilon)$ (i.e. the σ-algebra generated by the data up to time t_k), the following are optimal estimators for their respective optimization problems:

$$\widehat{X}_k^\epsilon = \mathbb{E}[X^\epsilon(t_k) | \mathcal{F}_k^{Y^\epsilon}] = \sum_i \int x \pi_k^\epsilon(x, s_i) dx = \arg \min_{x \in \mathcal{F}_k^{Y^\epsilon}} \mathbb{E} \|X^\epsilon(t_k) - x\|_2^2$$

$$\widehat{\Theta}_k^\epsilon = \arg\max_i \int \pi_k^\epsilon(x, s_i) dx = \arg \min_{\theta \in \mathcal{F}_k^{Y^\epsilon}} \mathbb{E} \mathbf{1}_{\{\Theta^\epsilon(t_k) \neq \theta\}}$$

1.1. Fast Mean-Reversion and a Fast Filter. In an asymptotic sense, we say that $X^\epsilon(t)$ is *fast mean-reverting* (FMR) relative to $\Theta^\epsilon(t)$ if

$$(54) \qquad \sup_x \left(-Q_{ii}(x)\right) \ll 1/\epsilon, \qquad \forall i \leq m$$

The condition set forth in (54) leads to the construction of fast filtering algorithms which closely approximate the optimal algorithm in (47) provided that the ratio of $\Delta t_k / \epsilon$ is large. For all $i, j \leq m$ define the following,

$$\mu_i(x) = \frac{1}{\sqrt{\pi}} e^{-(x - s_i)^2}, \qquad \bar{Q}_{ji} = \int Q_{ji}(x) \mu_j(x) dx \qquad \text{and,} \qquad \bar{h}_{s_i} = \int h(x) \mu_i(x) dx$$

With this notation, the following filter of reduced dimension is asymptotical optimal for small ϵ:

THEOREM 4. *Assume* (49) *is true. Then for any bounded function* $g : \mathcal{S} \to \mathbb{R}$*, point-wise for any* $y_{1:k} = \{y_1, y_2, ..., y_k\} \in \mathbb{R}^k$ *the optimal filter has the following limit as* $\epsilon \searrow 0$*,*

$$\mathbb{E}\left[g(\Theta^\epsilon(t_k))\middle| Y^\epsilon_{1:k} = y_{1:k}\right] \to \sum_i g(s_i)\bar{\pi}_k(s_i) \qquad \text{as } \epsilon \searrow 0$$

where

$$\bar{\pi}_k(s_i) = \frac{1}{c_k}\bar{\psi}_i(y_k)\sum_j \left(e^{\bar{Q}^*\Delta t_k}\right)_{ji}\bar{\pi}_{k-1}(s_j) \tag{55}$$

with $\bar{\psi}_i(y_k) = \mathbb{E}\left[\frac{1}{\sqrt{2\pi}}e^{-.5\left(y_k - \int_{t_k}^{t_{k+1}}\bar{h}_{\Theta^0(\tau)}d\tau\right)^2}\middle|\Theta^0(t_k) = s_i\right]$ *and* c_k *being a normalizing constant.*

One main component of the proof of theorem 4 is the path-wise weak limit of $(\Theta^\epsilon(\cdot), I^\epsilon(\cdot))$,

$$(\Theta^\epsilon(\cdot), I^\epsilon(\cdot)) \Rightarrow (\Theta^0(\cdot), I^0(\cdot)) \tag{56}$$

where $(\Theta^0(\cdot), I^0(\cdot))$ is a Markov process with generator $\bar{Q} + \bar{h}_{s_i}\frac{\partial}{\partial I}$. Section 2 goes through the steps to prove that the limit in (56) holds given the model construction of Section 1.

The other component of the proof of theorem 4 is the use of the exponential Martingale to change measure in calculating the posterior. This approach is used in deriving the Zakai equation (in the case of continuous-time observations) as was done by Zakai in 1969 [**40**] , and is also a part of Rozovsky's proof of the uniqueness of solutions to the Zakai equation [**33**].

Proof of Theorem 4: Consider a change of measure $\tilde{\mathbb{P}}^\epsilon$ under which Y^ϵ_k is an iid standard normal Gaussian process. For any vector $(y_1, ..., y_k) \in \mathbb{R}^k$ define the process

$$M_k(y_{1:k}; I^\epsilon(\cdot)) = \prod_{i=1}^k \frac{\exp\left\{-\frac{1}{2}\left(y_i - (I^\epsilon(t_{i+1}) - I^\epsilon(t_i))\right)^2\right\}}{\exp\left\{-\frac{1}{2}y_i^2\right\}}$$

Letting $M_0 \equiv 1$, it is easily checked that $M_k(Y^\epsilon_{1:k}; I^\epsilon(\cdot))^{-1}$ is a $\mathbb{P}^\epsilon$-martingale and provides the change of measure between $\tilde{\mathbb{P}}$ and $\mathbb{P}$,

$$\frac{d\mathbb{P}^\epsilon_{t_k}}{d\tilde{\mathbb{P}}^\epsilon_{t_k}} = M_k(Y^\epsilon_{1:k}; I^\epsilon(\cdot))$$

From the Kalianpour-Striebel formula, we have

$$\mathbb{E}\left[g(\Theta^\epsilon(t_k))\middle|Y^\epsilon_{1:k}=y_{1:k}\right]$$

$$=\frac{\tilde{\mathbb{E}}\left[M_k(Y^\epsilon_{1:k};I^\epsilon(\cdot))g(\Theta^\epsilon(t_k))\middle|Y^\epsilon_{1:k}=y_{1:k}\right]}{\tilde{\mathbb{E}}\left[M_k(Y^\epsilon_{1:k};I^\epsilon(\cdot))\middle|Y^\epsilon_{1:k}=y_{1:k}\right]}$$

From the weak limit of $(\Theta^\epsilon(\cdot),I^\epsilon(\cdot))$ in (56) and by independence of $\Theta^\epsilon(\cdot)$ and $I^\epsilon(\cdot)$ from $(Y^\epsilon_i)_{i\leq k}$ under $\tilde{\mathbb{P}}^\epsilon$, this is equal to a function of $y_{1:k}$ which converges point-wise,

$$=\frac{\tilde{\mathbb{E}}\left[M_k(y_{1:k};I^\epsilon(\cdot))g(\Theta^\epsilon(t_k))\right]}{\tilde{\mathbb{E}}\left[M_k(y_{1:k};I^\epsilon(\cdot))\right]}\rightarrow\frac{\tilde{\mathbb{E}}\left[M_k(y_{1:k};I^0(\cdot))g(\Theta^0(t_k))\right]}{\tilde{\mathbb{E}}\left[M_k(y_{1:k};I^0(\cdot))\right]}$$

Letting Y^0_k denote the process

$$Y^0_k=\int_{t_k}^{t_{k+1}}\bar{h}_{\Theta^0(\tau)}d\tau+Z_k$$

it is easily seen that $M_k(Y^0_{1:k};I^0(\cdot))$ is the Martingale change of measure from $\tilde{\mathbb{P}}$ where $(Y^0_i)_{i\leq k}$ is independent,

$$\frac{d\mathbb{P}^0_{t_k}}{d\tilde{\mathbb{P}}^0_{t_k}}=M_k(Y^0_{1:k};I^0(\cdot))$$

Therefore,

$$\mathbb{E}\left[g(\Theta^\epsilon(t_k))\middle|Y^\epsilon_{1:k}=y_{1:k}\right]\rightarrow\frac{\tilde{\mathbb{E}}\left[M_k(y_{1:k};I^0(\cdot))g(\Theta^0(t_k))\right]}{\tilde{\mathbb{E}}\left[M_k(y_{1:k};I^0(\cdot))\right]}=\mathbb{E}\left[g(\Theta^0(t_k))\middle|Y^0_{1:k}=y_{1:k}\right].$$

Therefore, in connection with the Baum-Welch equation, the posterior expectation converges to

$$\mathbb{E}\left[g(\Theta^0(t_k))\middle|Y^0_{1:k}=y_{1:k}\right]=\sum_i g(s_i)\bar{\pi}_k(s_i)$$

where the posterior distribution $\bar{\pi}_k$ is given recursively as

$$\bar{\pi}_k(s_i)=\frac{1}{c_k}\frac{\partial}{\partial y}\mathbb{P}(Y^0_k\leq y_k|\Theta^0(t_k)=s_i)\sum_j\left(e^{\bar{Q}^*\Delta t_k}\right)_{ji}\bar{\pi}_{k-1}(s_j),$$

with $\frac{\partial}{\partial y}\mathbb{P}(Y^0_k\leq y|\Theta^0(t_k)=s_i)=\bar{\psi}_i(y)$, which proves the theorem.

■

2. Tightness of Marginal Measures and Weak Convergence to a Unique Limit

This section contains the lemmas and theorems necessary to prove (56). In doing so, it is useful to describe a marginal probability space $(\bar{\Omega}, \bar{\mathcal{F}}_t, \bar{\mathbb{P}})$ for the non-Markov process $(\Theta^\epsilon(t), I^\epsilon(t))$, where

$$\bar{\Omega} = D\left([0,T];\mathcal{S}\right) \times C\left([0,T];\mathbb{R}\right)$$

and define the measures

$$\bar{\mathbb{P}}^\epsilon(\theta) = \bar{\mathbb{P}}(\Theta^\epsilon(\cdot) = \theta(\cdot))$$

for all $\epsilon > 0$. It can be shown that the process $\Theta^\epsilon(\cdot)$ is tight in the space $D([0,T];\mathcal{S})$ and $I^\epsilon(\cdot)$ is tight in the space $C([0,T];\mathbb{R})$, which therefore makes the family $(\bar{\mathbb{P}}^\epsilon)_{\epsilon>0}$ tight in $\bar{\Omega}$. Furthermore, for any bounded and differentiable $g : \mathcal{S} \times \mathbb{R} \to \mathbb{R}$, we can define the following process for any path $(\Theta(\cdot), I(\cdot)) \in \bar{\Omega}$,

$$\bar{M}_g(t) = g\left(\Theta(t), I(t)\right) - g\left(\Theta(0), I(0)\right) - \int_0^t \left(\bar{Q} + \bar{h}_{\Theta(\tau)} \frac{\partial}{\partial I}\right) g(\Theta(\tau), I(\tau)) d\tau$$

We will show that

$$\lim_{\epsilon \searrow 0} \mathbb{E}^{\bar{\mathbb{P}}^\epsilon}\left[\bar{M}_g(t) - \bar{M}_g(s) \middle| \mathcal{F}_s\right] = 0$$

when $s = 0$ which is sufficient to conclude that any convergent subsequence of $(\bar{\mathbb{P}}^\epsilon)_{\epsilon>0}$ must have a limit $\bar{\mathbb{P}}$ for which $\bar{M}_g$ is a $\bar{\mathbb{P}}$-Martingale. Therefore, from the tightness of the marginal measures $(\bar{\mathbb{P}}^\epsilon)_{\epsilon>0}$ it follows that

$$(\Theta^\epsilon(\cdot), I^\epsilon(\cdot)) \Rightarrow (\Theta^0(\cdot), I^0(\cdot))$$

where $(\Theta^0(\cdot), I^0(\cdot))$ uniquely solves the $\bar{\mathbb{P}}$-Martingale problem associated with $\bar{M}_g$, and is a Markov process with generator $\bar{Q} + \bar{h}_{s_i} \frac{\partial}{\partial I}$.

To summarize, the goal of this section is to prove (56) by showing that the marginal measures $(\bar{\mathbb{P}}^\epsilon)_{\epsilon>0}$ are tight in $\bar{\Omega}$ and that every convergent subsequence must converge to the measure of a process $(\Theta^0(\cdot), I^0(\cdot))$ which is the unique solution to a specific Martingale problem. This method of proving weak convergence is pointed out in the

book by Yin and Zhang [**41**], where in chapter 7 on page 186 they outline the method in four steps.

2.1. Tightness. General result regarding tightness of measures and weak convergence can be found in the book by Ethier and Kurtz [**18**], (chapter 3). In fact the proof of the tightness of $\Theta^\epsilon(\cdot)$ that is shown in this section is a specific case of lemma 6.1 on page 122 of their book.

Tightness of a family of measures means that for all $\delta > 0$ there exists a compact set $K_\delta \subset \bar{\Omega}$ such that

$$\sup_\epsilon \bar{\mathbb{P}}^\epsilon(K_\delta^c) < \delta$$

It turns out that because of the bound in (49), the family of measures is tight in $\bar{\Omega}$.

LEMMA 5. *Assuming that the initial condition I_0 is almost-surely bounded, the bound in (49) insures that $(\bar{\mathbb{P}}^\epsilon)_{\epsilon>0}$ is tight in $\bar{\Omega}$.*

Proof of Lemma 5: First, let's prove that $I^\epsilon(\cdot)$ is tight in $C([0,T];\mathbb{R})$. Since there exists a positive constant $\kappa < \infty$ such that $\mathbb{P}(|I_0| < \kappa) = 1$ and because h is bounded, the function $I^\epsilon(t)$ is uniformly bounded,

$$|I^\epsilon(t)| \leq \kappa + T\|h\|_\infty \qquad \forall t \in [0,T], \forall \epsilon > 0$$

Furthermore, $I^\epsilon(t)$ is uniformly equicontinuous with modulus of continuity $\|h\|_\infty$,

$$|I^\epsilon(t) - I^\epsilon(s)| \leq \|h\|_\infty |t-s| \qquad \forall t,s \in [0,T], \forall \epsilon > 0$$

Therefore, by the Ascoli theorem, any subsequence of $I^\epsilon(\cdot)$ has a further subsequence that converges uniformly, and therefore the support of all measures $\bar{\mathbb{P}}^\epsilon$ on I^ϵ is compact. Therefore, there is a set $K_\delta(I)$ such that

$$\bar{\mathbb{P}}^\epsilon(K_\delta(I)) > 1 - \delta/2 \qquad \forall \epsilon > 0.$$

Next, to prove compactness of $\Theta^\epsilon(\cdot)$, consider any $f \in D([0,T],\mathcal{S})$ and define the switching times $\{\tau_k(f)\}_{k=0}^\infty$ for $k = 1,2,3,...$

$$\tau_k(f) = \begin{cases} \inf\{s \in (\tau_{k-1}(f), T] : f(s) \neq f(s^-) \text{ or } s = T\} & \text{if } \tau_{k-1}(f) < T \\ T & \text{otherwise} \end{cases}$$

with $\tau_0(f) = 0$. For any $\delta > 0$ and any compact set $U \subset \mathbb{R}$, let the set $A(U, \delta)$ be defined as

$$A(U,\delta) = \{f : f(s) \in U \text{ for all } s \leq T, \tau_k(f) - \tau_{k-1}(f) > \delta, \forall k \text{ with } \tau_k(f) < T\}$$

The closure of $A(U, \delta)$ is compact:

Given any sequence $f_n \in A(U, \delta)$, by a diagonal argument there exists a subsequence f_{n_ℓ} s.t.

$$\tau_k(f_{n_\ell}) \to t_k \in [0, T]$$

and

$$f(\tau_k(f_{n_\ell})) \to a_k \in U$$

Therefore,

$$f_{n_\ell}(s) \to a_k, \qquad \text{for } s \in [t_k, t_{k+1})$$

which proves that $\overline{A(U,\delta)}$ is sequentially compact, and since $D([0, t]; \mathcal{S})$ is equipped with the point-wise metric, the subset is therefore compact.

For the Markov process $\Theta^\epsilon(\cdot)$, take U s.t. $\{s_1, ..., s_m\} \subset U$ and recall that

$$\mathbb{P}^\epsilon(A(U,\delta)^c) \leq 1 - e^{-\beta\delta} \leq \beta\delta \qquad \forall \epsilon > 0$$

which shows that $\Theta^\epsilon(\cdot)$ is compact in $D([0, T]; \mathbb{R})$.

Finally, tightness of $(\bar{\mathbb{P}}^\epsilon)_{\epsilon>0}$ follows by considering the compact set $K_\delta(I) \times \overline{A(U, \delta/2\beta)}$,

$$\bar{\mathbb{P}}^\epsilon\left((K_\delta(I) \times \overline{A(U,\delta/2\beta)})^c\right) \leq \bar{\mathbb{P}}^\epsilon\left(K_\delta(I)^c\right) + \bar{\mathbb{P}}^\epsilon\left(\overline{A(U,\delta/2\beta)}^c\right) < \delta/2 + \delta/2 = \delta$$

■

In this section it will be show (intheorem 7) that

$$\mathbb{E}^{\mathbb{P}^\epsilon}\left[\bar{M}_g(t) - \bar{M}_g(s)|\mathcal{F}_s\right] \to 0 \qquad \text{as } \epsilon \searrow 0 \qquad \forall t \in [0, T]$$

for $s = 0$. As a consequence of this and lemma 5 it follows that any convergent sequence of $(\bar{\mathbb{P}}^\epsilon)_{\epsilon>0}$ must converge to to a measure $\bar{\mathbb{P}}$ for which $\bar{M}_g$ is a $\bar{\mathbb{P}}$-Martingale. Therefore, the pair $(\Theta^\epsilon(\cdot), I^\epsilon(\cdot))$ must converge weakly to the Markov process $(\Theta^0(\cdot), I^0(\cdot))$ with generator $\bar{Q} + \bar{h}_{s_i}$.

To prove that the family of marginal measures will converge to the unique solution to the Martingale problem, the following lemma will be necessary:

LEMMA 6. *For any positive $\Delta t > 0$ and any function g such that $\|g\|_\infty + \left\|\frac{\partial}{\partial I} g\right\|_\infty \leq C$, the expectation of the following terms converge to a term of $o(\Delta t)$ as $\epsilon \searrow 0$*

$$\lim_{\epsilon \searrow 0} \left| \mathbb{E}\left[\int_0^{\Delta t} \left(Q(X^\epsilon(s)) - \bar{Q}\right) g(\Theta^\epsilon(s), I^\epsilon(s)) ds \middle| \mathcal{F}_0 \right] \right| \leq 2\beta^2 \Delta t^2$$

$$\lim_{\epsilon \searrow 0} \left| \mathbb{E}\left[\int_0^{\Delta t} \left(h(X^\epsilon(s)) - \bar{h}_{\Theta^\epsilon(s)} \frac{\partial}{\partial I}\right) g(\Theta^\epsilon(s), I^\epsilon(s)) ds \middle| \mathcal{F}_0 \right] \right| \leq 2\beta^2 \Delta t^2$$

Proof of lemma 6: Start by proving the first statement. By iteratively conditioning the σ-algebras, the expectation can be separated into two parts,

$$\left| \mathbb{E}\left[\int_0^{\Delta t} \left(Q(X^\epsilon(s)) - \bar{Q}\right) g(\Theta^\epsilon(s), I^\epsilon(s)) ds \middle| \mathcal{F}_0 \right] \right|$$

$$\leq \left| \mathbb{E}\left[\int_0^{\Delta t} \left(Q(X^\epsilon(s)) - \bar{Q}\right) g(\Theta^\epsilon(s), I^\epsilon(s)) ds \middle| \mathcal{F}_0 \cap \{\Theta^\epsilon(s) \equiv \Theta^\epsilon(0) \ \forall s \in [0, \Delta t]\} \right] \right|$$

$$\times \mathbb{P}(\Theta^\epsilon(s) \equiv \Theta^\epsilon(0) \text{ for } s \in [0, \Delta t])$$

$$+ \left| \mathbb{E}\left[\int_0^{\Delta t} \left(Q(X^\epsilon(s)) - \bar{Q}\right) g(\Theta^\epsilon(s), I^\epsilon(s)) ds \middle| \mathcal{F}_0 \cap \{\max_{s \leq \Delta t} |\Theta^\epsilon(s) - \Theta^\epsilon(0)| > 0\} \right] \right|$$

$$\times \left(1 - \mathbb{P}(\Theta^\epsilon(s) \equiv \Theta^\epsilon(0) \text{ for } s \in [0, \Delta t])\right)$$

Now, it obvious that $\mathbb{P}(\Theta^\epsilon(s) \equiv \Theta^\epsilon(0) \text{ for } s \in [0, \Delta t]) \leq 1$, and from the bound in (49) it follows that

$$1 - \mathbb{P}(\Theta^\epsilon(s) \equiv \Theta^\epsilon(0) \text{ for } s \in [0, \Delta t]) \leq 1 - e^{-\beta \Delta t}$$

On intervals where $\Theta^\epsilon(t)$ is constant, we have

$$X^\epsilon(t) =_d \Theta^\epsilon(0) + e^{-t/\epsilon}(x - \Theta^\epsilon(0)) + N\left(0, \frac{1}{2}\left(1 - e^{-2t/\epsilon}\right)\right)$$

and on these intervals the densities of $X^\epsilon(t)$ given $X(0) = x$ can be written as $\mu_{t/\epsilon}(y|x) \in \mathbb{R}^{m\times m}$ as

$$\mu_{t/\epsilon}(y|x) = \frac{1}{\sqrt{\pi(1 - e^{-2t/\epsilon})}} \begin{bmatrix} e^{-\frac{(y-s_1-e^{-2t/\epsilon}(x-s_1))^2}{1-e^{-2t/\epsilon}}} & \dots & 0 \\ & \ddots & \\ 0 & \dots & e^{-\frac{(y-s_m-e^{-2t/\epsilon}(x-s_m))^2}{1-e^{-2t/\epsilon}}} \end{bmatrix}$$

From here, there is the following upper-bound,

$$\left|\mathbb{E}\left[\int_0^{\Delta t} \left(Q(X^\epsilon(s)) - \bar{Q}\right) g(\Theta^\epsilon(s), I^\epsilon(s))ds \middle| \mathcal{F}_0\right]\right|$$

$$\leq \left|\mathbb{E}\left[\int_0^{\Delta t} \left(Q(X^\epsilon(s)) - \bar{Q}\right) g(\Theta^\epsilon(s), I^\epsilon(s))ds \middle| \mathcal{F}_0 \cap \{\Theta^\epsilon(s) \equiv \Theta^\epsilon(0)\ \forall s \in [0, \Delta t]\}\right]\right|$$

$$+2\beta\Delta t(1 - e^{-\beta\Delta t})$$

$$= \left|\int_0^{\Delta t} \left(\int \mu_{s/\epsilon}(y|X^\epsilon(0))Q(y)dy - \bar{Q}\right) g(\Theta^\epsilon(0), I^\epsilon(0))ds\right| + 2\beta\Delta t(1 - e^{-\beta\Delta t})$$

$$\leq \|g\|_\infty \left\| \int_{\epsilon^\alpha}^{\Delta t} \underbrace{\left(\int \mu_{s/\epsilon}(y|X^\epsilon(0))Q(y)dy - \bar{Q}\right)}_{\text{goes to zero as } \epsilon \searrow 0} ds \right\| + \underbrace{2\beta\Delta t(1 - e^{-\beta\Delta t})}_{\leq 2\beta^2\Delta t^2}$$

$$+ \underbrace{\|g\|_\infty \sup_x \|Q(x) - \bar{Q}\|}_{=C'} \epsilon^\alpha$$

where $0 < \alpha < 1$. Now let $\gamma(\Delta t, \epsilon)$ denote a bound on the under the norm. It is clear the $\gamma(\Delta t, \epsilon) \to 0$ as $\epsilon \searrow 0$ for all $\Delta t > 0$. Therefore, we have

$$\left|\mathbb{E}\left[\int_0^{\Delta t} \left(Q(X^\epsilon(s)) - \bar{Q}\right) g(\Theta^\epsilon(s), I^\epsilon(s))ds \middle| \mathcal{F}_0\right]\right|$$

$$\leq \Delta t\gamma(\Delta t,\epsilon) + 2\beta^2\Delta t^2 + C'\epsilon^\alpha \to 2\beta^2\Delta t^2$$

which proves the first statement of the lemma.

Repeating these steps with $Q(X^\epsilon(s))$ replaced by $h(X^\epsilon(s))$ and $\bar{Q}$ replace by $\bar{h}_{\Theta^\epsilon(s)}$, the second statement in the lemma is proven provided that there is a constant such that

$$\|g\|_\infty + \left\|\frac{\partial}{\partial I}g\right\|_\infty \leq C.$$

■

With lemma 6, the key calculation for proving weak convergence of $(\Theta^\epsilon(\cdot), I^\epsilon(\cdot))$ can be shown. Namely, from the time-homogeneity and Markov property of the solution to the Martingale problem associated with $\bar{M}_g$, it is sufficient to show that the expectation under $\mathbb{P}^\epsilon$ of the function $\bar{M}_g(t)$ conditioned on $\mathcal{F}_0$ converges to zeros as $\epsilon \searrow 0$.

THEOREM 7. *For any function $g(\Theta(t), I(t))$ with $\|g\|_\infty + \left\|\frac{\partial}{\partial I}g\right\|_\infty \leq C$, condition (49) and the boundedness of h insure that $\mathbb{E}^{\mathbb{P}^\epsilon}[\bar{M}_g(t)|\mathcal{F}_0] \to 0$ as $\epsilon \searrow 0$.*

Proof of Theorem 7: For any positive $t \leq T$, fix a step-size $\Delta t > 0$ (not the same as Δt_k that was described in section 1), and consider the following discretization of the interval $[0, t]$,

$$0 < \Delta t < 2\Delta t < \ldots \cdots < N\Delta t = t$$

For any $n < N$, let $t_n = n\Delta t$. We then have a collapsing sum

$$\mathbb{E}^{\mathbb{P}^\epsilon}[\bar{M}_g(t)|\mathcal{F}_0]$$

$$= \mathbb{E}\left[g(\Theta^\epsilon(t), I^\epsilon(t)) - g(\Theta^\epsilon(0), I^\epsilon(0))|\mathcal{F}_0\right]$$

$$-\mathbb{E}\left[\int_0^t \left(\bar{Q} + \bar{h}_{\Theta^\epsilon(s)}\frac{\partial}{\partial I}\right) g(\Theta^\epsilon(s), I^\epsilon(s))ds \middle| \mathcal{F}_0\right]$$

$$= \mathbb{E}\left[\sum_{n=0}^{N-1} g(\Theta^\epsilon(t_{n+1}), I^\epsilon(t_{n+1})) - g(\Theta^\epsilon(t_n), I^\epsilon(t_n))|\mathcal{F}_0\right]$$

$$-\mathbb{E}\left[\sum_{n=0}^{N-1}\int_{t_n}^{t_{n+1}} \left(\bar{Q} + \bar{h}_{\Theta^\epsilon(s)}\frac{\partial}{\partial I}\right) g(\Theta^\epsilon(s), I^\epsilon(s))ds\middle|\mathcal{F}_0\right]$$

$$= \underbrace{\mathbb{E}\left[\sum_{n=0}^{N-1} \mathbb{E}[g(\Theta^\epsilon(t_{n+1}), I^\epsilon(t_{n+1}))|\mathcal{F}_{t_n}] - g(\Theta^\epsilon(t_n), I^\epsilon(t_n))\middle|\mathcal{F}_0\right]}_{(*)}$$

$$\underbrace{-\mathbb{E}\left[\sum_{n=0}^{N-1}\int_{t_n}^{t_{n+1}} \mathbb{E}\left[\left(\bar{Q} + \bar{h}_{\Theta^\epsilon(s)}\frac{\partial}{\partial I}\right) g(\Theta^\epsilon(s), I^\epsilon(s))\middle|\mathcal{F}_{t_n}\right] ds\middle|\mathcal{F}_0\right]}_{(**)}$$

The summand in $(**)$ can be shown to be equal to the summand in $(*)$ plus a correction term,

$$\int_{t_n}^{t_{n+1}} \mathbb{E}\left[\left(\bar{Q} + \bar{h}_{\Theta^\epsilon(s)}\frac{\partial}{\partial I}\right) g(\Theta^\epsilon(s), I^\epsilon(s))\middle|\mathcal{F}_{t_n}\right] ds$$

$$= \underbrace{\int_{t_n}^{t_{n+1}} \mathbb{E}\left[\left(\bar{Q} + \bar{h}_{\Theta^\epsilon(s)}\frac{\partial}{\partial I} - Q(X^\epsilon(s)) - h(X^\epsilon(s))\frac{\partial}{\partial I}\right) g(\Theta^\epsilon(s), I^\epsilon(s))\middle|\mathcal{F}_{t_n}\right] ds}_{=e_n(\epsilon,\Delta t))}$$

$$+\int_{t_n}^{t_{n+1}} \mathbb{E}\left[\left(Q(X^\epsilon(s)) - h(X^\epsilon(s))\frac{\partial}{\partial I}\right) g(\Theta^\epsilon(s), I^\epsilon(s))\middle|\mathcal{F}_{t_n}\right] ds$$

$$= \mathbb{E}[g(\Theta^\epsilon(t_{n+1}), I^\epsilon(t_{n+1}))|\mathcal{F}_{t_n}] - g(\Theta^\epsilon(t_n), I^\epsilon(t_n)) + e_n(\epsilon, \Delta t)$$

By lemma 6, the correction term $e_n(\epsilon, \Delta t)$ can be controlled so that the limit as $\epsilon \searrow 0$ is $o(\Delta t)$. Therefore,

$$\left|\mathbb{E}^{\mathbb{P}^\epsilon}[\bar{M}_g(t)|\mathcal{F}_0]\right| \leq \sum_{n=0}^{N-1} |e_n(\epsilon, \Delta t)| \to \sum_{n=0}^{N-1} o(\Delta t) = O(\Delta t) \qquad \text{as } \epsilon \searrow 0$$

and the entire sum converges to something of order $O(\Delta t)$ as $\epsilon \searrow 0$, which proves the theorem since Δt is arbitrarily small.

■

Theorem 7 confirms that any convergent sequence in the family of tight measure $(\bar{\mathbb{P}}^\epsilon)_{\epsilon>0}$ must have a limit $\bar{\mathbb{P}}$ which is the measure for which paths in the space $\bar{\Omega}$ uniquely solve the Martingale problem associated with $\bar{M}_g$. Therefore, this proves (56) and the reduced filter of theorem 4 stands.

3. Numerical Simulations

3.1. Discrete Time Schemes. For the example in this section let Θ_t be independent of $X^\epsilon(t)$. Suppose both the jumps of $\Theta(t)$ and the observations of Y_k^ϵ occur at set of equidistant time-points $\{t_k\}_{k=0}^N$ with

$$0 = t_0 < t_1 < t_2 < .... < t_N = T$$

and $t_{i+1} = t_i + \Delta t$ for any $0 \leq i < N$. We use the convention that $\Theta(t)$ is constant on the half-open intervals,

$$\Theta(t) = \Theta(t_k), \qquad \forall t \in [t_k, t_{k+1})$$

which makes $\Theta(t_k)$ a Markov chain with discrete jump times, and so for each time $k < N$ there are the following transition probabilities for $\Theta(t_k)$,

$$\mathbb{P}(\Theta(t_{k+1}) = s_i | \Theta(t_k) = s_j) = p_{ji}$$

Now, because the process $\Theta(t)$ is piece-wise constant on the intervals $[t_k, t_{k+1})$, the solution for an OU process can be applied on each interval. Furthermore, the solution $X^\epsilon(t_k)$ can be thought of as a distributional equivalent to the process $\widetilde{X}^\epsilon(t_k)$ given by,

$$\widetilde{X}^\epsilon(t_{k+1}) = \widetilde{X}^\epsilon(t_k)e^{-\Delta t/\epsilon} + \Theta(t_k)(1 - e^{-\Delta t/\epsilon}) + \sqrt{\frac{1 - e^{-2\Delta t/\epsilon}}{2}}\mathcal{W}_k$$

where $\mathcal{W}_k \sim iidN(0,1)$. Then, assuming that the function h is sufficiently well-behaved (for instance Lipshitz) the observed process can be approximated with an Euler scheme with high probability,

$$Y_k^\epsilon = \int_{t_k}^{t_{k+1}} h(X^\epsilon(\tau))d\tau + Z_k \approx h(X_{t_k}^\epsilon)\Delta t + Z_k \tag{57}$$

where $Z_k \sim N(0, \Delta t)$.

Based on the model we've described so far and the approximated observation process, it is well-known that the Baum-Welch equation for this system is

$$\pi_{k+1}^\epsilon(x, s_i) = \frac{1}{c_{k+1}}\mathbb{P}(Y_{k+1}^\epsilon | X^\epsilon(t_{k+1}) = x) \sum_j \int \Gamma_{\Delta t}^\epsilon(x, i|x', j)\pi_k^\epsilon(x', s_j)dx'$$

where c_{k+1} is a normalizing constant, and Γ is the joint-transition kernel:

$$\Gamma_{\Delta t}^\epsilon(x, i|x', j) = \frac{\partial}{\partial x} p(X^\epsilon(t_{k+1}) \leq x, \Theta(t_k) = s_i | X^\epsilon(t_k) = x', \Theta(t_{k-1}) = s_j)$$

$$\propto G\left(\frac{x - x'e^{-\Delta t/\epsilon} - s_i(1 - e^{-\Delta t/\epsilon})}{\sqrt{(1 - e^{-2\Delta t/\epsilon})/2}}\right) p_{ji}$$

where $G(\cdot)$ is a standard normal Gaussian kernel.

An essential condition for an averaged filter to work in discrete time is a *fast mean-reverting* condition,

$$\frac{1}{\Delta t} \ll \frac{1}{\epsilon}$$

which insures the relaxation of X_t^ϵ in between observations and allows for the following weak approximation,

$$Y_k^\epsilon \approx_d \bar{h}_{\Theta(t_k)}\Delta t + \Delta B_k$$

In fast mean-reverting regimes, the optimal filter is approximated by putting the data into the reduced filter,

$$\bar{\pi}_{k+1}(s_i) \approx \frac{1}{c_k}\bar{\psi}_i(Y_k^\epsilon) \sum_j p_{ji}\bar{\pi}_k(s_j)$$

where $\bar{\psi}_i(Y_k^\epsilon) \propto G\left(\frac{Y_k^\epsilon - \bar{h}_{\Theta_{t_k}}\Delta t}{2\sqrt{\Delta t}}\right)$.

3.2. Error Graphs and Plots. There are two posterior estimates of Θ_{t_k}, an optimal and an averaged:

$$\widehat{\Theta}_k^{opt} = \arg\max_{s_i} \int \pi_k^\epsilon(x, s_i) = \arg\min_{\theta \in \mathcal{F}_k^{Y^\epsilon}} \mathbb{E}^\epsilon \mathbf{1}_{\Theta(t_k) \neq \theta}$$

$$\widehat{\Theta}_k^{avg} = \arg\max_{s_i} \bar{\pi}_k(s_i)$$

for which the 0-1 error of the averaged estimate should approach the error of the optimal estimate as $\epsilon \searrow 0$,

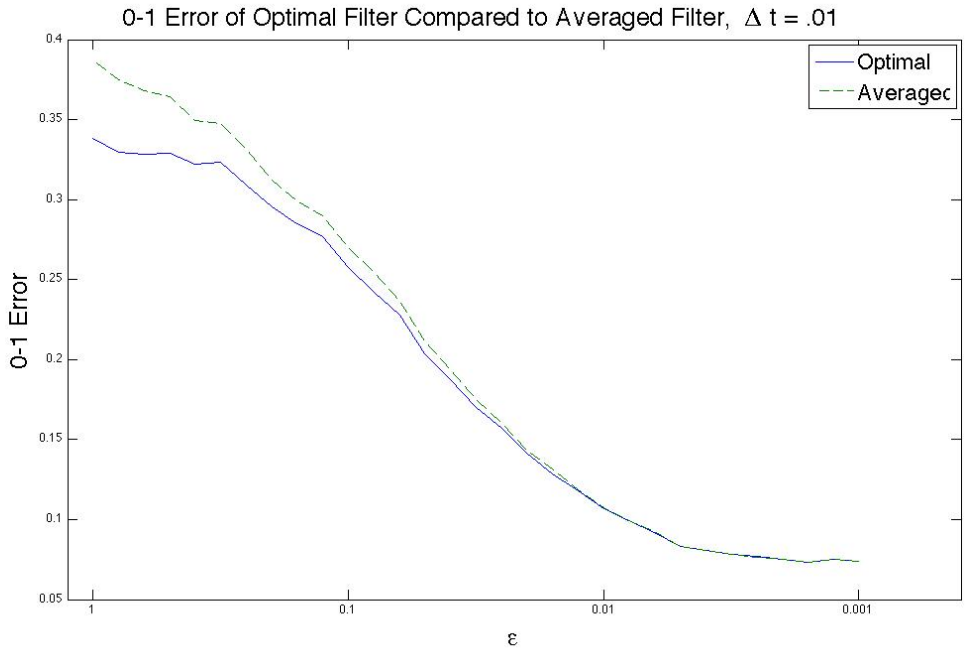

FIGURE 1. The error of the averaged filter decreases to the that of the optimal filter as $\epsilon \searrow 0$. Also, the signal-to-noise ratio decreases with ϵ, and so the error of the optimal filters decreases too.

$$0 \leq \mathbb{E}^\epsilon \left(\mathbf{1}_{\Theta_{t_k} \neq \widehat{\Theta}_k^{avg}} - \mathbf{1}_{\Theta_{t_k} \neq \widehat{\Theta}_k^{opt}} \right) \xrightarrow{\epsilon \searrow 0} 0$$

where $\mathbb{E}^\epsilon$ is the expectation operator of the measure for the specified rate of mean-reversion. Convergence of the averaged error to the optimal error is shown in Figure 1, where the plots are an approximation of the 0-1 error given by

$$\mathbb{E}^{\epsilon}\mathbf{1}_{\Theta(t_k)\neq\widehat{\Theta}_k^{\{\cdot\}}} \approx \frac{1}{T}\sum_{\ell=1}^{T}\mathbf{1}_{\Theta(t_\ell)\neq\widehat{\Theta}_\ell^{\{\cdot\}}}$$

for some k greater than the memory length of Θ, and T very large. It should also be noted that since the signal-to-noise ratio increases as ϵ decreases, the optimal error also decreases with ϵ.

3.3. Empirical Estimates of Optimal Sampling Rate. In this section we analyze the optimal rate at which observations should be sampled relative to a increasing rate of mean reversion. In other-words, we seek Δt which optimizes the expected loss with a non-negative penalty term added when sampling is too sparse,

$$\min_{\Delta t}\left(\mathbb{E}^{\epsilon}\mathbf{1}_{\Theta(t_k)\neq\widehat{\Theta}_k^{avg}} + \alpha(\Delta t)\right) \tag{58}$$

In other-words, sampling too fast will eliminate any hypothesis regarding a fast mean-reverting regime, while sampling too slow will be penalized because the target has not been tracked often enough. For instance, suppose we set,

$$\alpha(\Delta t) = \frac{1}{2}\Delta t^2$$

Then, since neither $\Delta t = 0$ or $\Delta t = \infty$ are optimal, we differentiate and set it equal to zero:

$$\frac{\partial}{\partial\Delta t}\left(\mathbb{E}^{\epsilon}\mathbf{1}_{\Theta(t_k)\neq\widehat{\Theta}_k^{avg}} + \frac{\Delta t^2}{2}\right) = \frac{\partial}{\partial\Delta t}\mathbb{E}^{\epsilon}\mathbf{1}_{\Theta(t_k)\neq\widehat{\Theta}_k^{avg}} + \Delta t = 0$$

Using empirical data we can approximate the derivatives of the sequence of errors, and then construct the optimal sequence of time steps. In this case the optimal is a fixed-point of the expected loss derivative,

$$\Delta t(\epsilon) = -\frac{\partial}{\partial\Delta t}\mathbb{E}^{\epsilon}\mathbf{1}_{\Theta(t_k)\neq\widehat{\Theta}_k^{avg}}\Big|_{\Delta t=\Delta t(\epsilon)}$$

Similarly, for $\alpha(\Delta t) = \log(1+\Delta t)$, differentiating leads to an optimal sampling rate that is the fixed-point of the inverse of expected loss derivative,

$$\Delta t(\epsilon) = \left(\frac{\partial}{\partial\Delta t}\mathbb{E}^{\epsilon}\mathbf{1}_{\Theta(t_k)\neq\widehat{\Theta}_k^{avg}}\right)^{-1}\Big|_{\Delta t=\Delta t(\epsilon)} - 1$$

Figure on page 111 shows some empirical analysis of the optimal Δt.

4. Chapter Summary

In summary, this chapter has proven a series of theorems and lemmas which confirm that the non-fast-mean-reverting processes marginally converge to a weak limit and that the nonlinear filter converges point-wise as a function of the data. The results depended largely on the properties of the OU process, but the proofs could be generalized to the class of fast-mean-reverting diffusion processes. Displayed in the later sections were some numerics to show how the filter behaves as a function of the rate of mean-reversion, and finally presents a more elaborate criterion for evaluating filter-performance. It is clear that the averaged filter is a good approximation for ϵ small, but possible future work needs to be done to understand the optimization problem of Section 3.3.

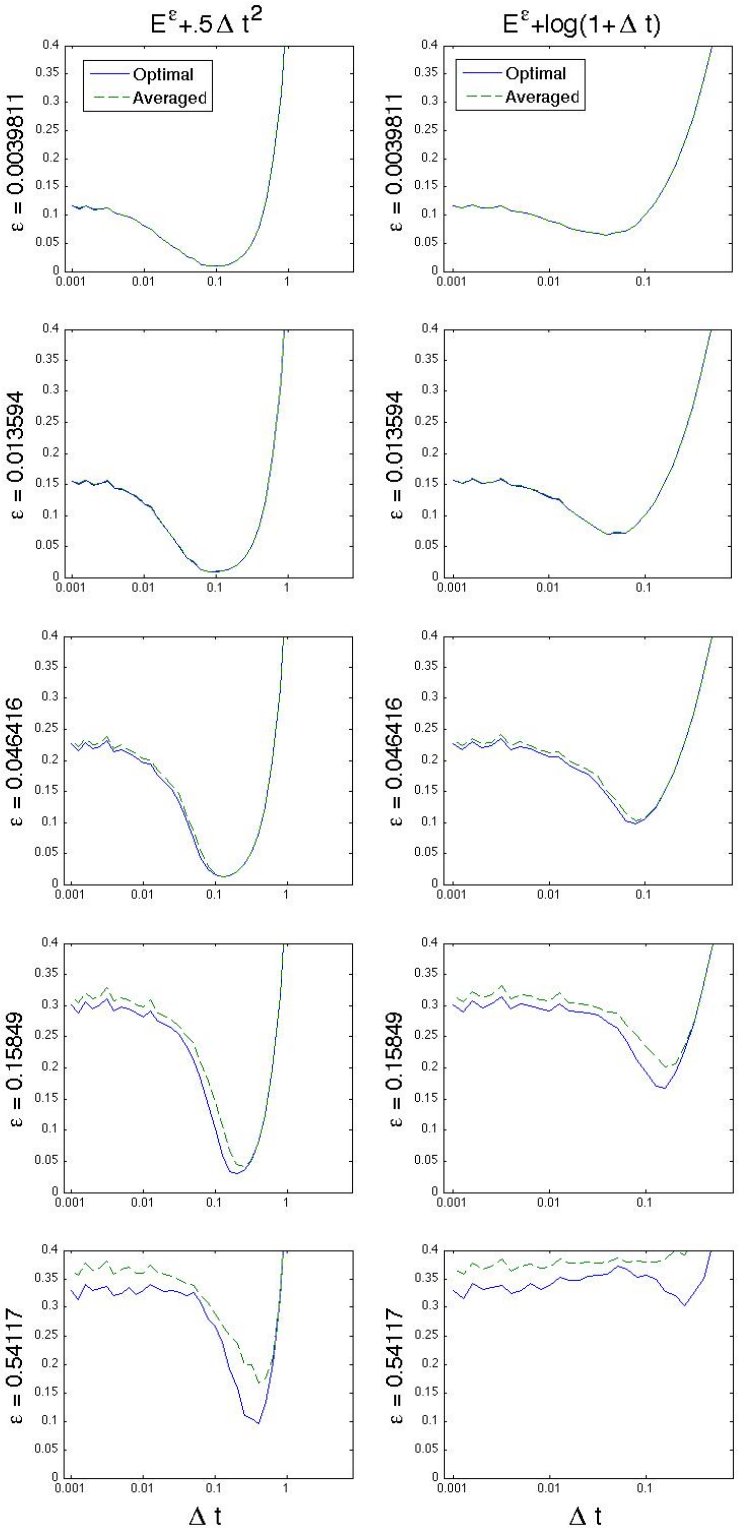

FIGURE 2. The penalized error term, $E^\epsilon + \frac{1}{2}\Delta t^2$ and $E^\epsilon + \log(1+\Delta t)$, for different Δt and ϵ. Notice how the optimal Δt increases with ϵ.

CHAPTER 6

Conclusion

The formulation of a filtering algorithm was the common theme of all four chapters of this dissertation. In some chapters it was challenging to formulate but easy to implement, and in other chapters the challenges were reversed.

Chapter 2 introduced the basic tools of nonlinear filtering by demonstrating their novel application to a hidden Markov model for stochastic volatility. Chapter 3 displayed a nonlinear filtering algorithm for human-tracking from video data, which is currently a popular way to model targets and human motion in computer vision. Chapter 4 had a novel use of a Kramers-Smoluchowski approximation to derive a fast filter of reduced dimension for a tracking model where the input-data consists of noisy video frames with considerable jitter. Finally, Chapter 5 rigorously derived and implemented a nonlinear filtering algorithm for a multi-scale model, and thoroughly exploited the ergodic averaging that happens between discrete observations in order to show that functions that are parameterized by the fast time-scales can be replaced by their invariant averages in the approximate nonlinear filter.

Relevant work that will succeed this thesis is the further analysis and applications of the algorithms to financial data, and further development of the software for human-tracking. Multi-scale modeling is also a topic with broad applications and there will undoubtedly be new and improved filtering algorithms that need to be analyzed.

APPENDIX A

Double Markov Chains

PROOF. *(proposition 2)* Suppose that $u_{k-1}^N(\vec{s})$ has been computed for all $\vec{s}$. For $n = 0$, we have

$$u_k^0(\vec{s}) = \sum_r \mathbb{P}(\theta_k^0 = s^0, \theta_{k-1}^0 = r, \theta_{k-1}^1 = s^1, ..., \theta_{k-1}^N = s^N)$$

$$= \sum_r \mathbb{P}(\theta_k^0 = s^0 | \theta_{k-1}^0 = r, \theta_{k-1}^1 = s^1, ..., \theta_{k-1}^N = s^N)$$
$$\times \mathbb{P}(\theta_{k-1}^0 = r, \theta_{k-1}^1 = s^1, ..., \theta_{k-1}^N = s^N)$$

$$= \sum_r \Lambda^0(s^0|r) u_{k-1}^N \left((r, s^1, ..., s^N)\right)$$

which proves the first part of the proposition. Now, for any $n > 0$, given $u_k^{n-1}(\vec{s})$, we have a similar breakdown:

$$u_k^n(\vec{s}) = \sum_r \mathbb{P}(\theta_k^0 = s^0, ...\theta_k^{n-1} = s^{n-1}, \theta_k^n = s^n, \theta_{k-1}^n = r, ..., \theta_{k-1}^N = s^N)$$

$$= \sum_r \mathbb{P}(\theta_k^n = s^n | \theta_k^0 = s^0, ...\theta_k^{n-1} = s^{n-1}, \theta_{k-1}^n = r, ..., \theta_{k-1}^N = s^N)$$

$$\times \mathbb{P}(\theta_k^0 = s^0, ...\theta_k^{n-1} = s^{n-1}, \theta_{k-1}^n = r, ..., \theta_{k-1}^N = s^N)$$

$$= \sum_r \mathbb{P}(\theta_k^n = s^n | \theta_k^{n-1} = s^{n-1}, \theta_{k-1}^n = r) u_k^{n-1} \left((s^0, ..., s^{n-1}, r, s^{n+1}, ..., s^N)\right)$$

$$= \sum_r \Lambda^n(s^n | r, s^{n-1}) u_k^{n-1} \left((s^0, ..., s^{n-1}, r, s^{n+1}, ..., s^N)\right)$$

□

PROPOSITION 8. *A DMC is a vector-valued Markov chain in time.*

PROOF. A Markov chain is a process whose future does not depend on its history but depends only on the present. The distribution of $\vec{\theta}_k$ given its entire history can be broken down as follows:

$$\mathbb{P}(\vec{\theta}_k = \vec{s}|\vec{\theta}_{k-1}, ..., \vec{\theta}_0) = \Pi_{n=0}^{N}\mathbb{P}(\theta_k^n = s^n|\theta_k^0 = s^0, ..., \theta_k^{n-1} = s^{n-1}, \vec{\theta}_{k-1}, ..., \vec{\theta}_0)$$

$$= \Lambda^0(s^0|\theta_{k-1}^0)\Pi_{n=1}^{N}\Lambda^n(s^n|\theta_{k-1}^n, s^{n-1})$$

On the other-hand, the distribution of $\vec{\theta}_k$ given only the knowledge of $\vec{\theta}_{k-1}$ can be broken down in a similar fashion,

$$\mathbb{P}(\vec{\theta}_k = \vec{s}|\vec{\theta}_{k-1}) = \Pi_{n=0}^{N}\mathbb{P}(\theta_k^n = s^n|\theta_k^0 = s^0, ..., \theta_k^{n-1} = s^n, \vec{\theta}_{k-1})$$

$$= \Lambda^0(s^0|\theta_{k-1}^0)\Pi_{n=1}^{N}\Lambda^n(s^n|\theta_{k-1}^n, s^{n-1})$$

The two of these conditional probabilities are equal, which therefore implies that $\vec{\theta}_k$ satisfies a Markov property over the time index k. □

If the state-space of the observation process is discrete, the *likelihood* is defined by the probability

$$L\left(y_0, ..., y_k\right) := \mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k\right) \qquad \text{where } y_\ell \in F.$$

If the state-space of the observation process is $\mathbb{R}^1$ the *likelihood* is defined by the probability density function

$$L\left(y_0, ..., y_k\right) := p\left(y_0, ..., y_k\right) := \partial\mathbb{P}\left(Y_0 \leq y_0, ..., Y_k \leq y_k\right)/\partial y_1 ... \partial y_k \qquad \text{where } y_\ell \in \mathbb{R}^d.$$

PROPOSITION 9.

$$L\left(y_0, ..., y_k\right)$$
$$= \sum_{\vec{s}_{i_0}, ..., \vec{s}_{i_k}} \pi_0\left(\vec{s}_{i_0}\right)\Gamma(\vec{s}_{i_1}|\vec{s}_{i_0}).....\Gamma(\vec{s}_{i_k}|\vec{s}_{i_{k-1}})\psi_{\vec{s}_{i_0}}\left(y_0\right)...\psi_{\vec{s}_{i_k}}\left(y_k\right)$$

PROOF. We will prove the proposition in the case of the observation with discrete state-space. By the "memoryess channel" assumption and formula (26) we have

$$\mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k, \vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right) =$$

$$\mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k | \vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right) \mathbb{P}\left(\vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_k = \vec{s}_{i_k}\right)$$

$$= \pi_0\left(\vec{s}_{i_0}\right) \Gamma(\vec{s}_{i_1} | \vec{s}_{i_0}).....\Gamma(\vec{s}_{i_k} | \vec{s}_{i_{k-1}}) \psi_{\vec{s}_{i_0}}\left(y_0\right) ... \psi_{\vec{s}_{i_k}}\left(y_k\right)$$

To conclude the proof it suffices to sum up both sides of the equation over all sets $\vec{s}_{i_0}, ..., \vec{s}_{i_k}$. □

PROOF. *(proposition 3)* Let us consider the discrete case (i.e. the case when the state-space of the white noise is F). The continuous case can be dealt with in a completely analogous way. We start by considering the following:

$$\phi_k\left(\vec{s}\right) = \mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k, \vec{\theta}_k = \vec{s}\right) =$$

$$\sum_{\vec{s}_{i_0}, ..., \vec{s}_{i_{k-1}}} \mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k, \vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_{k-1} = \vec{s}_{i_{k-1}}, \vec{\theta}_k = \vec{s}\right)$$

$$\sum_{\vec{s}_{i_0}, ..., \vec{s}_{i_{k-1}}} \mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k | \vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_{k-1} = \vec{s}_{i_{k-1}}, \vec{\theta}_k = \vec{s}\right) \times$$

$$\mathbb{P}\left(\vec{\theta}_0 = \vec{s}_{i_0}, ..., \vec{\theta}_{k-1} = \vec{s}_{i_{k-1}}, \vec{\theta}_k = \vec{s}\right).$$

By (27), the conditional term becomes

$$\mathbb{P}\left(Y_0 = y_0, ..., Y_k = y_k | \theta_0 = \vec{s}_{i_0}, ..., \theta_{k-1} = \vec{s}_{i_{k-1}}, \vec{\theta}_k = \vec{s}\right) =$$

$$\psi_{\vec{s}_{i_0}}\left(y_0\right) ... \psi_{\vec{s}_{i_{k-1}}}\left(y_{k-1}\right) \psi_{\vec{s}}\left(y_k\right).$$

Formula (26) yields the following for the unconditional term:

(59)

$$\mathbb{P}\left(\theta_0 = \vec{s}_{i_0}, ..., \theta_{k-1} = \vec{s}_{i_{k-1}}, \theta_k = \vec{s}\right) = \pi_0\left(\vec{s}_{i_0}\right) \Gamma(\vec{s}_{i_1} | \vec{s}_{i_0}).....\Gamma(\vec{s}_{i_{k-1}} | \vec{s}_{i_{k-2}}) \Gamma(\vec{s} | \vec{s}_{i_{k-1}})$$

Putting these two parts together, we have the result:

(60)

$$\phi_k(\vec{s}) =$$

$$= \psi_{\vec{s}}(y_k) \sum_{\vec{s}_{i_0},...,\vec{s}_{i_{k-1}}} \psi_{\vec{s}_{i_0}}(y_0) ... \psi_{\vec{s}_{i_{k-1}}}(y_{k-1}) \pi_0(\vec{s}_{i_0}) \Gamma(\vec{s}_{i_1}|\vec{s}_{i_0}) ..... \Gamma(\vec{s}_{i_{k-1}}|\vec{s}_{i_{k-2}}) \Gamma(\vec{s}|\vec{s}_{i_{k-1}})$$

$$= \psi_{\vec{s}}(y_k) \sum_{\vec{r}} \phi_{k-1}(\vec{r}) \Gamma(\vec{s}|\vec{r})$$

□

Mini-Corollaries:

(1) The likelihood function $L(y_0, ..., y_n) = \sum_{\vec{s}} \phi_k(\vec{s})$.

(2) The posterior distribution of the current state given past observations

$$\pi_k(\vec{s}) = \mathbb{P}\left\{\vec{\theta}_k = \vec{s} | Y_0 = y_0, ..., Y_k = y_k\right\} = \phi_k(\vec{s}) / \sum_{\vec{r}} \phi_k(\vec{r}).$$

The function $\phi_k = (\phi_k(\vec{s}_1), ..., \phi_k(\vec{s}_{\mathbf{n}}))_{k \geq 0}$ is a vector-valued stochastic process. By definition, $\phi_k(\vec{s}) \geq 0$ for all $\vec{s}$. Since, $\sum_{\vec{s}} \phi_k(\vec{s}) \neq 1$, ϕ_k is often referred to as the *unnormalized posterior distribution* (resp. *density*) function of the state process $\vec{\theta}_k$ given observation $\mathbf{Y}^k = (Y_1, ..., Y_k)$ in discrete (resp. continuous) case. In both cases we will use the same abbreviation: UPDF. The vector-valued stochastic process $\pi_k = (\pi_k(\vec{s}_1), ..., \pi_k(\vec{s}_{\mathbf{n}}))_{k \geq 0}$ in the discrete (resp.continuous) case will be called *filtering distribution* (resp. *density*) function (FDF).

After the posterior distribution has been computed, an estimate is chosen based on the optimization criterion deemed suitable for the problem. Essentially, a suitable criterion comes down to the choice of a loss function $\mathcal{M}(\cdot)$. For any $n \leq N$, we seek an estimate of θ_k^n which minimizes the expected posterior loss:

$$\widehat{\theta}_k^n = \arg\min_{s \in \mathcal{S}^n} \mathbb{E}\left[\mathcal{M}(\theta_k^n, s) \middle| \mathcal{Y}_k\right]$$

$$= \arg\min_{r \in \mathcal{S}^n} \sum_{\vec{s}} \mathcal{M}(s^n, r)\pi_k(\vec{s})$$

When θ_k^n takes values in a discrete state-space, we may take $\mathcal{M}$ to be a 0-1 loss function, $\mathcal{M}(a,b) = \mathbf{1}_{\{a \neq b\}}$, which leads to the maximum aposteriori (MAP) estimator being optimal:

$$\widehat{\theta}_k^n = \arg\min_{s \in \mathcal{S}^n} \mathbb{E}\left[\mathbf{1}_{\{\theta_k^n \neq s\}} \middle| \mathcal{Y}_k\right]$$

$$= \arg\max_{r \in \mathcal{S}^n} \left(\sum_{s^0 \in \mathcal{S}^0} ..... \sum_{s^{n-1} \in \mathcal{S}^{n-1}} \sum_{s^{n+1} \in \mathcal{S}^{n+1}} .... \sum_{s^N \in \mathcal{S}^N} \pi_k\left(s^0, \ldots, s^{n-1}, r, s^{n+1}, \ldots, S^N\right) \right)$$

$$= \arg\max_{r \in \mathcal{S}^n} \mathbb{P}(\theta_k^n = r | \mathbf{Y}^k)$$

When θ_k^n takes values in a continuum, a possible estimate is the posterior expectation

$$\widehat{\theta}_k^n = \sum_{\vec{s}} s^n \pi_k(\vec{s})$$

which is the optimal estimate when $\mathcal{M}(a,b) = \|a-b\|_2^2$. In linear Gaussian models the posterior is known apriori to be Gaussian and so for such models the Kalman filter can quickly return the posterior mean and posterior covariance matrix. Another special case is when the posterior distribution of θ_k^n is known to be unimodal and symmetric, in which case a 0-1 loss function is the same as any other symmetric loss function. In general, the choice of a loss function may be a difficult decision, but regardless of the choice, an optimal estimate will always be returned provided we are able to compute the posterior distribution.

Bibliography

[1] B. Anderson, J. Moore, Optimal Filtering Dover Publications, 1979.

[2] S. Asmussen, P. Glynn, Stochastic Simulation: Algorithms and Analysis. Springer, 2007.

[3] Y. Bar-Shalom, X. R. LI, Estimation and Tracking: Principles, Techniques and Software. Boston: Artech House, 1993.

[4] M. Basseville, I. Nikiforov, Detection of Abrupt Changes: Theory and Application. Prentice-Hall, Inc., April 1993.

[5] T. Björk, Arbitrage Theory in Continuous Time, 2nd Edition, Oxford University Press, 2004.

[6] R. B. Blazek, A. Petrov, B. L. Rozovskii,"Interacting Banks of Bayesian Matched Filters." SPIE Proceedings: Signal and Data Processing of Small Targets, Vol. 4048, Orlando, FL, 2000.

[7] M. Broadie, A. Jain, "The Effects of Jumps and Discrete Sampling on Volatility and Variaince Swaps," International Journal of Theoretical and Applied Finance, Vol. 11, No. 8, 761-797, 2008.

[8] J. Brown, A. Tartakovsky, "Adaptive spatial-temporal filtering methods for clutter removal and target tracking" Aerospace and Electronic Systems, IEEE Transactions, Oct. 2008 Volume: 44, Issue: 4, page(s): 1522-1537.

[9] O. Cappe, E. Moulines, T. Ryden, Inference in Hidden Markov Models. Springer 2005.

[10] B. Carlin, T. Louis, Bayes and Empirical Bayes Methods for Data Analysis. 2nd edition, Boco Raton: Chapman and Hall/ CRC Press, 2000

[11] P. Carr, R. Lee, "Realized Volatility and Variance: Options via Swaps" Risk, May 2007.

[12] P. Carr, L. Wu, "A Tale of Two Indices," The Journal of Derivatives, Spring 2006.

[13] R. Cont, T. Kokholm, "A Consistent Pricing Model for Index Options and Volatility Derivatives," 19th Annual Derivatives, Securities and Risk Management Conference, 2009.

[14] D. Cremers. "Dynamical Statistical Shape Priors for Level Set-Based Tracking." IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 28, NO. 8, August 2006

[15] J. Cvitanic, B. Rozovsky,I. Zaliapan, "Numerical Estimation of Volatility Values From Discretely Observed Data." 2005.

[16] A.P. Dempster, N.M Laird, D.B. Rubin, . (1977)."Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society. Series B (Methodological) 39 (1): pp. 1-38

[17] E. Derman, K. Demeterfi, M. Kamal, J. Zou, "More Than You Ever Wanted to Know About Volatility Swaps" Journal of Derivatives 6(4), 1999, pages 9-32.

[18] S. Ethier, T. Kurtz, Markov Processes: Characterization and Convergence. Wiley, 1986.

[19] J.-P. Fouque, G. Papanicolaou, and R. Sircar "Derivatives in Financial Markets with Stochastic Volatility", Cambridge University Press, 2000.

[20] M. Freidlin, "Some Remarks on the Kramers-Smoluchowski Approximation", Journal of Statistical Physics, 117, pages 617-634, 2004.

[21] P. Guan, A. Weiss, A. O. Bălan, M. J. Black, "Estimating Human Shape and Pose from a Single Image." ICCV 2009.

[22] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, P. Nordlund, "Particle Filters for Positioning, Navigation and Tracking," 2002.

[23] D. Higham, "An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations," Siam Review, Vol. 43, No. 3, pp. 525-546.

[24] S. Howisn, A. Rafailidis, H. Rasmussen, "On the pricing and hedging of volatility derivatives" Applied Mathematical Finance, Volume 11, Issue 4 December 2004 , pages 317 - 346.

[25] A.H. Jazwinski, Stochastic Processes and Filtering Theory. Academic Press, New York, 1970.

[26] V.R Lesser, F. Hayes-Roth, M. Birnbaum, and R. Cronk, "Selection of Word Islands in the HEARSAY-II Speech Understanding System." Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 2 , pp. 791-794. 1977.

[27] S. Lototsky, B. Rozovsky, "Fast Nonlinear Filter for Continuous-Discrete Time Multiple Models." Proceedings of the 35th Conference on Decision and Control. Kobe, Japan. December 1996.

[28] E. Nelson, "Dynamical Theories of Brownian Motion", Princeton University Press, 1967.

[29] D. Ormoneit, M. J. Black, T. Hastie, H. Kjellstöm, "Representing cyclic human motion using functional analysis." Image and Vision Computing, 23(14):1264-1276, 12 Dec. 2005."

[30] B. Øxendale Stochastic Differential Equations: An Introduction with Applications (2003) Berlin: Springer.

[31] A. Petrov, B. Rozovsky, "Optimal Nonlinear Filtering for Track-Before-Detect in IR Image Sequences." SPIE proceedings: Signal and Data Processing of Small Targets, vol. 3809, Denver, Co, 1999.

[32] L. R. Rabiner "A tutorial on Hidden Markov Models and selected applications in speech recognition". Proceedings of the IEEE 77 (2), (February 1989). pp. 257-286.

[33] B. Rozovsky, "A simple proof of uniqueness for Kushner and Zakai equations." In Stochastic analysis, ed. E. Mayer-Wolf, 449-458. Boston: Academic Press, 1991.

[34] B. Rozovsky, A. Tartakovsky, G. Yaralov, "An Adaptive Bayesian Approach to Fusion of Imaging and Kinematic Data." Proceedings of the 2nd International Conference on Information Fusion (Fusion99). Sunnyvale, CA, 6-8 July, 1999, pp. 1029-1036.

[35] Y. Saito, T. Mitsui, "Stability Analysis of Numerical Schemes for Stochastic Differential Equations," SIAM Journal of Numerical Analysis, Vol.33 (1996) pp. 2254-2267.

[36] T. Schön, G. Gustafsson, P. Norlund, "Marginalized Particle Filters for Mixed Linear/Nonlinear State-Space Models," Signal Processing, IEEE Transactions on Volume 53, Issue 7, July 2005 Page(s): 2279 - 2289

[37] J. Shi and S. Osher, "A Nonlinear Inverse Scale Space Method for a Convex Multiplicative Noise Model", SIAM J. Imaging Sciences, Vol 1, 2008, pp. 294-321.

[38] H. Sidenbladh, M. Black, "Learning Image Statisitics for Bayesian Tracking." International Conference on Computer Vision, Vancouver Canada, July 2001.

[39] A. Tartakovsky, "Multihypothesis Invariant Sequential Tests with Applications to the Quickest Detection of Signals in Multiple-Resolution Element Systems." Proceedings of the 1998 Conference on Information Sciences and Systems. Princeton University, Princeton, NJ., March 1998, vol. 2, pp. 906-911.

[40] Zakai, M. "On the optimal filtering of diffusion processes." Zeit. Wahrsch. 11 230-243. 1969

[41] G. Yin, Q. Zhang, Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach. Springer, 1998.

[42] G. Yin, Q. Zhang, Discrete-Time Markov Chains: Two-Time-Scale Methods and Applications. Springer, 2005.