# EXPERIMENTAL DESIGN FOR PARTIALLY OBSERVED MARKOV DECISION PROCESSES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Leifur Thorbergsson

August 2014

EXPERIMENTAL DESIGN FOR PARTIALLY OBSERVED MARKOV DECISION PROCESSES

Leifur Thorbergsson, Ph.D.

Cornell University 2014

This thesis considers the question of how to most effectively conduct experiments in Partially Observed Markov Decision Processes so as to provide data that is most informative about a parameter of interest. Methods from Markov decision processes, especially dynamic programming, are introduced and then used in algorithms to maximize a relevant Fisher Information. These algorithms are then applied to two POMDP examples. The methods developed can also be applied to stochastic dynamical systems, by suitable discretization, and we consequently show what control policies look like in the Morris-Lecar Neuron model and the Rosenzweig MacArthur Model, and simulation results are presented. We discuss how parameter dependence within these methods can be dealt with by the use of priors, and develop tools to update control policies online. This is demonstrated in another stochastic dynamical system describing growth dynamics of DNA template in a PCR model.

## BIOGRAPHICAL SKETCH

Leifur grew up in Ísafjörður, Iceland. He is the son of Þorbergur Kjartansson and Frauke Eckhoff, and he has two sisters, Elisabeth Þorbergsdóttir and Oddný Þorbergsdóttir. He attended Menntaskólinn a Ísafirði high school, before majoring in Math at the University of Iceland in Reykjavík and graduating in 2008. He attended Cornell from 2008 till 2014, graduating with a Ph.D. in Statistics.

This thesis is dedicated to my parents, Frauke Eckhoff, and Thorbergur Kjartansson.

**ACKNOWLEDGEMENTS**

I would like to especially thank my advisor, Professor Giles Hooker for his invaluable guidance, and my committee, Professors James Booth and Bruce Turnbull for their helpful advice. Additionally I would like to thank Diana Drake and Beatrix Johnson, my teachers and fellow students in Statistics and ORIE, my housemates at Gamma Alpha, and my friends in Ithaca.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

CHAPTER 1

# INTRODUCTION

Hidden Markov Models have proven their usefulness across a wide variety of applications. In many of these applications, the user or the experimenter will have some way of influencing the transitions of the underlying Markov Chain, as in Markov Decision Processes, and such a process is called a Partially Observed Markov Decision Process (POMDP), see Monahan [6]. If we assume that the transition probability matrix is governed by some unknown parameters, an important problem is to understand how the process can be influenced to get data that is most informative about the parameters. We can think of this as experimental design for Partially Observed Markov Decision Processes.

We consider a POMDP $(x_t, y_t, u_t)_{t=0,\ldots,T}$. In this setting $x_t$ is an unobserved Markov Chain, where the transition probabilities depend in a parametric way on what control $u_t$ is chosen at time $t$ and an unknown parameter $\theta$. The process $y_t$ is observed and depends on which state $x_t$ is in.

Our goal is to find ways to use the controls $u_t$ to improve parameter estimates of $\theta$. Since the maximum likelihood estimates for $\theta$ will be asymptotically efficient, our general strategy will be to use the controls to try to minimize the sample variance of the maximum likelihood estimates of $\theta$. This will be achieved by maximizing a Fisher Information for $\theta$. The controls are calculated using dynamic programming, a popular maximization algorithm from Markov Decision Processes which outputs an adaptive control policy, i.e. the control chosen at time $t$ is based on observations up to time $t$.

In Chapter 2 we review the relevant theory from Hidden Markov Models

and modify it to allow for controls. We discuss forgetting properties of the filter and score function, which will be needed to prove convergence results in Chapter 3. Then we review relevant theory from Markov Decision Processes, especially dynamic programming and the Value Iteration Algorithm.

The first attempt at using dynamic controls to maximize a Fisher Information was by Hooker et al. [5] who proposed maximizing the Fisher Information that corresponds to direct observations of the underlying process $x_t$, labeled the Full Information Fisher Information (FOFI), and using a filter to compute $x_t$ if it is not observed directly. We extend their work by making use of the POMDP structure and we propose maximizing a Fisher Information that is based on the observations $y_t$, labeled the Partial Observation Fisher Information (POFI). We show that maximizing POFI directly using dynamic programming is computationally unfeasible, and in Chapter 3 we give two approximations to POFI, the truncated Partial Observation Fisher Information, and the Weighted Observation Fisher Information (WOFI) and bound the difference between them and POFI.

In Chapter 4 we discuss how these Fisher Information criteria are maximized using dynamic programming and discuss the computational complexity of running such algorithms. Then we describe parameter estimation techniques within POMDP's, review the asymptotic properties of the MLE and bound the difference between the asymptotic Fisher Information of our estimate and the theoretically best possible Fisher Information.

The methods developed have application value beyond Partially Observed Markov Decision Processes. In Chapter 6 we consider stochastic systems of the

form

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, \theta, u(t))dt + \Sigma^{1/2}d\mathbf{W}$$

where $\theta$ is the parameter of interest, to be estimated, $u(t)$ is a control that can be chosen by the user, $\mathbf{x}$ is the vector of state variables, $\mathbf{f}$ is a vector valued function, $\mathbf{W}$ a Wiener process, and additionally $\mathbf{x}(t)$ is only observed partially or noisily. By discretizing time, state and observation spaces the process can be approximated by a POMDP, allowing us to use the methods developed to devise a control policy that maximizes information about the parameter $\theta$.

In order to illustrate our methods we present five examples, with the first two being POMDP's and the latter three continuous stochastic systems. We use the unknown $\theta$ to calculate controls in all but the last example to highlight the differences between the different maximizing criteria, but in the last example we examine means to deal with the dependence of the Fisher Information criteria on the parameter of interest.

In Chapter 5 we consider 2 POMDP examples. First we hypothesize about the kind of systems in which policies based on POFI will lead to large improvement in parameter estimation over the FOFI policy. Following a discussion we construct a mock Partially Observed Markov Decision Process, in which this improvement is shown using a simulation study. To illustrate the real-world applicability of design in discrete POMDP's we consider a realistic POMDP from experimental economics. The model will consist of a simple adversarial game similar to the "rock - paper - scissor" game where one player tries to play in such a way that maximizes information about the other players' strategy.

In Chapter 7 we consider two diffusion processes. First a stochastic version of the Morris-Lecar Neuron model, a dynamical system which models voltage

in a single neural cell. This model is two dimensional, but only one dimension is observed. The model has multiple parameters and we investigate how the truncated POFI, WOFI and FOFI control policies perform in estimating them. Then we consider the Rosenzweig MacArthur Model, which describes a two species ecology, with a predator species consuming a prey species in a controlled environment, and we look into control policies towards estimating the rate of which prey is consumed.

The methods we use to calculate controls for maximizing Fisher Information will depend on the unknown parameter $\theta$. In Chapter 8 we illustrate how this problem can partially be overcome by assuming a prior for $\theta$ to calculate a control policy before running the experiment. Additionally we describe how, using data acquired as the experiment progresses, a posterior for $\theta$ can be used to calculate a more precise control policy. That is, parameter information from observations acquired at a time $t$ can be used to improve the policy used in what is left of the experiment. These methods will be based on the Value Iteration Algorithm (VIA), which is closely related to dynamic programming. This is illustrated in a fifth example, now from biology, a Polymerase chain reaction (PCR) experiment where DNA template is grown in liquid substrate. The population dynamics are modeled in a dynamical system with stochastic errors, and the aim is to estimate the half-saturation constant, a parameter which controls the saturation of the template. Here we compare using a prior for $\theta$ and using VIA to calculate a control policy.

CHAPTER 2

# FRAMEWORK, LITERATURE REVIEW AND THEORY

## 2.1 Framework

We consider a Markov decision process $(X_t, u_t)_{t=0,\dots,T}$. In this setting $X_t$ is a Markov chain, but the transition probabilities at time $t$ depend on a control $u_t$ chosen at that time. We assume a finite state space $\mathcal{X}$ for the state process $X_t$ and that the controls available belong to some finite set $\mathcal{U}$. We let $K$ denote the size of $\mathcal{X}$ and $l$ the size of $\mathcal{U}$. The transition probabilities are assumed to be parametric and we frequently write $p(x_{t+1}|x_t, u_t, \theta)$ short for $p(x_{t+1} = x^i|x_t = x^j, u_t = u^r, \theta)$ where $x^i, x^j \in \mathcal{X}$ and $u^r \in \mathcal{U}$.

In addition to this we assume that the process $X_t$ is latent and we only observe the related observations $Y_t \in \mathcal{Y}$ whose relation to the $X_t$ can also depend on $\theta$. We write $p(y_t|x_t, \theta)$ short for $p(y_t = y^i|x_t = x^j, \theta)$, where $x^j \in \mathcal{X}$ and $y^j \in \mathcal{Y}$, and let $L$ denote the size of $\mathcal{Y}$. This makes the system a Partially Observed Markov Decision Process (POMDP). It has a finite horizon $T$ in which we observe $y_0 \dots y_T$. We will use the short hand notation $y_{m:t}$ to denote $y_m, \dots, y_t$, i.e. the observations between time $m$ and $t$, and analogous notation for $u_t$ and $x_t$.

## 2.2 Hidden Markov Model theory, adjusted for controls

This section is devoted to expanding Hidden Markov Model Theory to Partially Observed Markov Decision Processes. We base it completely on Cappe et al. [1] and use their notation, only changing what is necessary. Reviewing

forward and backwards variables, see 2.2.2, will be useful to describe the EM algorithm in 4.4.1, but the main objective here is to prove Theorem 1 in Section 2.2.5 about the forgetting properties of the filter $p(x_t|y_{0:t}, u_{0:t-1}, \theta)$ and Theorem 3 in Section 2.2.7 about the forgetting properties of the corresponding score function. The latter is then used to prove Theorem 5 in Section 3.6 and Theorem 7 in Section 3.9. In most cases the changes will amount to adding controls and seeing that the theory follows through, although the proof of Theorem 3 has more substantial changes.

The forgetting properties of the filter $p(x_t|y_{0:t}, u_{0:t-1}, \theta)$ will describe the intuitive statement that the filter depends less on older observations than new, although showing this is somewhat subtle.

### 2.2.1 Framework

Cappe et al [1] allow for continuous state spaces, and thus use integrals instead of sums, etc. Since in this part we are only modifying their theory to allow for controls, we adopt their notation for all of Section 2.2.

Let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be the state space and the observations space respectively. Let

$$Q^u(x, A) = \int_A q^u(x, x')dx', \ A \in \mathcal{X}, u \in \mathcal{U}$$

be a transition kernel for our state space, where $u$ is a control, and $\mathcal{U}$ is finite. Also let

$$G(x, A) = \int_A g(x, y)dy, \ A \in \mathcal{Y}$$

be the transition kernel for moving from the state space to the observation space.

We generally assume that the Markov Chain is initialized with distribution $\nu$, and then runs for $n$ steps $x_{0:n} = x_0, \dots, x_n$ and that $n-1$ decisions are made on what controls $u$ to use. This results in $n$ observations $y_{0:n} = y_0, \dots, y_n$ and $n-1$ control $u_{0:n-1} = u_0, \dots, u_{n-1}$.

### 2.2.2 Forwards and Backwards variables

**Definition 1** (Definition 3.1.6 in [1])**.** *Conditional on $y_{0:k}$ and $u_{0:k-1}$ we define the forward variable*

$$\alpha_{\nu,k}(y_{0:k}, u_{0:k-1}, f) = \int \cdots \int f(x_k)\nu(dx_0)g(x_0, y_0) \prod_{l=1}^{k} Q^{u_{l-1}}(x_{l-1}, dx_l)g(x_l, y_l)$$

*and conditional on $y_{k+1:n}$ and $u_{k:n-1}$ we define the backward variable*

$$\beta_{k|n}(y_{k+1:n}, u_{k:n-1}, x) = \int \cdots \int Q^{u_k}(x, dx_{k+1})g(x_{k+1}, y_{k+1}) \prod_{l=k+2}^{n} Q^{u_{l-1}}(x_{l-1}, dx_l)g(x_l, y_l)$$

As in the classical case these satisfy recursion formulas

$$\alpha_{\nu,k}(y_{0:k}, u_{0:k-1}, f) = \int f(x_k) \int \alpha_{\nu,k-1}(y_{0:k-1}, u_{0:k-2}, dx_{k-1})Q^{u_{k-1}}(x_{k-1}, dx_k)g(x_k, y_k)$$

with initial condition

$$\alpha_{\nu,0}(f) = \int f(x_0)g(x_0, y_0)\nu(dx_0)$$

and similarly

$$\beta_{k|n}(y_{k+1:n}, u_{k:n-1}, x) = \int Q^{u_k}(x, dx_{k+1})g(x_{k+1}, y_{k+1})\beta_{k+1|n}(y_{k+2:n}, u_{k+1:n-1}, x_{k+1})$$

### 2.2.3 Forwards and Backwards Kernels

A standard result in HMM theory is that conditional on the observations $y_{0:n}$ the Process $\{X_k\}_{k\geq 0}$ still is a Markov Chain, although non-homogeneous, with a

transition kernel called the Forward Smoothing Kernel. We state the transition kernel here for our case, also conditional on the controls.

**Definition 2** (Definition 3.3.1 in [1]). *Forward Smoothing Kernels. Given $n \geq 0$ define for indices $k \in \{0, \ldots, n-1\}$ the transition kernels*

$$F_{k|n}(x, A, y_{k+1:n}, u_{k:n-1}) = \frac{\int_A Q^{u_k}(x, dx_{k+1}) g(x_{k+1}, y_{k+1}) \beta_{k+1|n}(x_{k+1})}{\beta_{k|n}(x)}$$

Note that the Forward Smoothing Kernels are defined in terms of the backward variables.

We are generally interested in calculating smoothers and filters for our POMDP.

**Definition 3** (Definition 3.1.3 in [1]). *We let $\phi_{\nu,k:l|n}$ denote the conditional distribution of $X_{k:l}$ given $Y_{0:n}$ and $u_{0:n-1}$.*

The Forward Smoothing Kernel allows us a convenient way of calculating the smoothing distributions. We first compute all the backward variables $\beta_{k|n}$ using the backward recursion given. We then note that $\phi_{\nu,0|n}$ can be calculated as

$$\phi_{\nu,0|n}(A) = \frac{\int_A \nu(dx_0) g(x_0, y_0) \beta_{0|n}(x_0)}{\int \nu(dx_0) g(x_0, y_0) \beta_{0|n}(x_0)}$$

and then we have the following recursion

$$\phi_{\nu,k+1|n}(x) = \int \phi_{\nu,k|n}(dx_k) F_{k|n}(x_k, x) = \phi_{\nu,k|n} F_{k|n}$$

where $F_{k|n}$ are the forward kernels, and the last equation is a short hand way of writing the integral.

Using this recursion repeatedly allows to express the smoother in the following way

$$\phi_{\nu,k|n}[y_{0:n}, u_{0:n-1}] = \phi_{\nu,0|n} \prod_{i=1}^{k} F_{i-1|n}[y_{i:n}, u_{i-1:n-1}]$$

### 2.2.4 Total Variation and Dobrushin Coefficient

To continue towards forgetting properties we introduce Total variation (see Definition 4.3.1 in [1]). Let $\xi$ be a signed measure, which can be negative, and let $\xi = \xi_+ - \xi_-$ where $\xi_+, \xi_-$ are (positive) measures. So if $X$ is the state space then

$$\|\xi\|_{TV} = \xi_+(X) + \xi_-(X)$$

Next, we let $K$ be a transition Kernel from $X$ to $Y$. The Dobrushin Coefficient (see Definition 4.3.7 in [1]) is defined as

$$\delta(K) = \frac{1}{2} \sup_{(x,x') \in X \times X} \|K(x, \cdot) - K(x', \cdot)\|_{TV}$$

The Dobrushin coefficient is sub-multiplicative (see Prop. 4.3.10 in [1]). If $K : X \to Y, R : Y \to Z$ are 2 transition kernels we have

$$\delta(KR) = \delta\left( \int K(\cdot, dx) R(x, \cdot) \right) \leq \delta(K)\delta(R)$$

It can be shown that $0 \leq \delta(K) \leq 1$, however to establish forgetting properties we often need $\delta(K) \leq 1 - \varepsilon$, where $\varepsilon > 0$.

The latter inequality holds if we assume the Doeblin Condition is satisfied:

**Assumption 1** (Assumption 4.3.12 in [1])**.** *There exist an integer $m \geq 1, \varepsilon \in (0, 1)$, and a probability measure $\nu$ on $(X, \mathcal{X})$ such that for any $x \in X$ and $A \in \mathcal{X}$,*

$$Q^m(x, A) \geq \varepsilon \nu(A)$$

Under these assumptions Lemma 4.3.13 in [1] gives $\delta(Q^m) \leq 1 - \varepsilon$.

We say that a filter $\phi_{\nu,k|n}$ has forgetting properties if it depends less and less on the initial distribution of $X_0 \sim \nu$, as $k$ increases. Specifically when comparing initial distributions $\nu$ and $\nu'$ we have

$$\phi_{\nu,k|n}(y_{0:n}, u_{0:n-1}, x_k) - \phi_{\nu',k|n}(y_{0:n}, u_{0:n-1}, x_k)$$
$$= \int \cdots \int (\phi_{\nu,0|n}(y_{0:n}, u_{0:n-1}, x_k) - \phi_{\nu',0|n}(y_{0:n}, u_{0:n-1}, x_k)) \prod_{i=1}^{k} F_{i-1|n}(x_{k-1}, x_k)$$

Now using Corollary 4.3.9 in [1] we have

$$\|\xi K - \xi' K\|_{TV} \leq \delta(K) \|\xi - \xi'\|_{TV}$$

where $\xi, \xi'$ are probability measures, $K$ a transition kernel.

Using this on our representation of the filters gives

$$\|\phi_{\nu,k|n} - \phi_{\nu',k|n}\|_{TV} \leq \delta\left(\prod_{i=1}^{k} F_{i-1|n}(y_{i:n}, \cdot)\right) \|\phi_{\nu,0|n} - \phi_{\nu',0|n}\|_{TV}$$

Now since the Dobrushin coefficient is sub-multiplicative

$$\leq \prod_{i=1}^{k} \delta\left(F_{i-1|n}(y_{i:n}, \cdot)\right) \|\phi_{\nu,0|n} - \phi_{\nu',0|n}\|_{TV}$$

and since the Dobrushin coefficient $\delta$ satisfies $0 \leq \delta \leq 1$ we at least have that the difference between the 2 filters is non-expanding.

Establishing forgetting properties thus amounts to showing $\delta(F_{i-1|n}(y_{i:n})) \leq 1 - \varepsilon$ for the forward smoothing kernels $F_{i|n}$. Note that so far no assumptions have been made on how quickly the Hidden Markov Model mixes. Those assumptions are made to get $\delta(F_{i|n}) \leq 1 - \varepsilon$.

### 2.2.5 Mixing Conditions and forgetting properties

Cappe et al. [1] establish contracting bounds on the Dobrushin coefficient by imposing Strong Mixing conditions on the transition probabilities of the Hidden Markov Model.

**Assumption 2** (Assumption 4.3.21 in [1])**.** *Strong Mixing Conditions in Hidden Markov Models. There exist a transition kernel $K : Y \to X$ and measurable functions $\varsigma^-$ and $\varsigma^+$ from $Y$ to $(0, \infty)$ such that for any $A \in \mathcal{X}$ and $y \in Y$,*

$$\varsigma^-(y)K(y, A) \leq \int_A Q(x, dx')g(x', y) \leq \varsigma^+(y)K(y, A)$$

In our case we have different transition kernels for each control. The weakest assumptions we can get away with is, if each transition kernel $Q^u$ has a corresponding transition kernel $K^u$ and measurable functions $\varsigma^-(y, u)$ and $\varsigma^+(y, u)$ satisfying the strong mixing condition. By letting $\varsigma^-(y) = \min_u \varsigma^-(y, u)$ and $\varsigma^+(y) = \max_u \varsigma^+(y, u)$ we see that we can consider the same $\varsigma$ functions for each transition kernel $Q^u$. We restate the Strong mixing conditions for POMDP's:

**Assumption 3.** *Modified Strong Mixing Conditions. For each control $u$ there exist a transition kernel $K^u : Y \to X$ and measurable functions $\varsigma^-$ and $\varsigma^+$ from $Y$ to $(0, \infty)$ such that for any $A \in \mathcal{X}$ and $y \in Y$,*

$$\varsigma^-(y)K^u(y, A) \leq \int_A Q^u(x, dx')g(x', y) \leq \varsigma^+(y)K^u(y, A)$$

Lemma 4.3.22 in Cappe et al. [1] uses the mixing conditions stated above to establish contracting bounds on the Dobrushin coefficient. We restate the lemma for the POMDP case, where we also condition on the controls, and use the modified mixing conditions.

**Theorem 1** (Lemma 4.3.22 in [1])**.** *Under the strong mixing conditions the following holds*

*(i) For any non-negative integers $k$ and $n$ such that $k < n$ and $x \in X$,*

$$\prod_{j=k+1}^{n} \varsigma^-(y_j) \leq \beta_{k|n}[y_{k+1:n}, u_{k:n-1}](x) \leq \prod_{j=k+1}^{n} \varsigma^+(y_j)$$

(ii) *For any non-negative integers $k$ and $n$ such that $k < n$ and any probability measures $\nu$ and $\nu'$ on $(X, \mathcal{X})$,*

$$\frac{\varsigma^-(y_{k+1})}{\varsigma^+(y_{k+1})} \leq \frac{\int \nu(dx)\beta_{k|n}[y_{k+1:n}, u_{k:n-1}](x)}{\int \nu'(dx)\beta_{k|n}[y_{k+1:n}, u_{k:n-1}](x)} \leq \frac{\varsigma^+(y_{k+1})}{\varsigma^-(y_{k+1})}$$

(iii) *For any non-negative integers $k$ and $n$ such that $k < n$, there exists a transition kernel $\lambda_{k|n}$ from $(Y^{n-k}, \mathcal{Y}^{(n-k)})$ to $(X, \mathcal{X})$ such that for any $x \in X$, $A \in \mathcal{X}$, and $y_{k+1:n} \in Y^{n-k}$,*

$$\frac{\varsigma^-(y_{k+1})}{\varsigma^+(y_{k+1})}\lambda_{k,n}(y_{k+1:n}, u_{k:n-1}, A) \leq F_{k|n}[y_{k+1:n}, u_{k:n-1}](x, A)$$

$$\leq \frac{\varsigma^+(y_{k+1})}{\varsigma^-(y_{k+1})}\lambda_{k,n}(y_{k+1:n}, u_{k:n-1}, A)$$

(iv) *For any non-negative integers $k$ and $n$, the Dobrushin coefficient of the forward smoothing kernel $F_{k|n}[y_{k+1:n}, u_{k:n-1}]$ satisfies*

$$\delta(F_{k|n}[y_{k+1:n}, u_{k:n-1}]) \leq \rho_0(y_{k+1}) := 1 - \frac{\varsigma^-(y_{k+1})}{\varsigma^+(y_{k+1})}$$

*if $k < n$, and*

$$\delta(F_{k|n}[y_{k+1:n}, u_{k:n-1}]) \leq 1 - \int \varsigma^-(y)dy$$

*if $k \geq n$.*

*Proof.* The proof is the same as for the corresponding lemma in Cappe et al. [1], but with slight modifications to allow for conditioning on controls.

(i) Letting $A = X$ in the strong mixing conditions we find that for all $u$

$$\varsigma^-(y) \leq \int Q^u(x, dx')g(x', y) \leq \varsigma^+(y)$$

We also have

$$\begin{aligned}
\beta_{k|n}(x) &= \int_{x_{k+1}} \cdots \int_{x_n} Q^{u_k}(x, dx_{k+1}) g(x_{k+1}, y_{k+1}) \prod_{l=k+2}^{n} Q^{u_{l-1}}(x_{l-1}, dx_l) g(x_l, y_l) \\
&= \int_{x_{k+1}} Q^{u_k}(x, dx_{k+1}) g(x_{k+1}, y_{k+1}) \\
&\times \int_{x_{k+2}} \cdots \int_{x_n} Q^{u_{k+1}}(x_{k+1}, dx_{k+2}) g(x_{k+2}, y_{k+2}) \prod_{l=k+3}^{n} Q^{u_{l-1}}(x_{l-1}, dx_l) g(x_l, y_l) \\
&\leq \varsigma^+(y_{k+1}) \sup_{x_{k+1}} \int_{x_{k+2}} \cdots \int_{x_n} Q^{u_{k+1}}(x_{k+1}, dx_{k+2}) g(x_{k+2}, y_{k+2}) \\
&\quad \times \prod_{l=k+3}^{n} Q^{u_{l-1}}(x_{l-1}, dx_l) g(x_l, y_l) \\
&= \varsigma^+(y_{k+1}) \sup_x \beta_{k+1|n}(x) \leq \prod_{j=k+1}^{n} \varsigma^+(y_j)
\end{aligned}$$

The other inequality is similar.

(ii) Using the recursion for the backward variables we find

$$\begin{aligned}
&\int_x \nu(dx) \beta_{k|n}(y_{k+1:n}, u_{k:n-1}) \\
&= \int_x \int_{x_{k+1}} \nu(dx) Q^{u_k}(x, x_{k+1}) g(x_{k+1}, y_{k+1}) \beta_{k+1|n}(y_{k+2:n}, u_{k+1:n-1}, dx_{k+1}) \\
&= \int_{x_{k+1}} \left[ \int_x \nu(dx) Q^{u_k}(x, x_{k+1}) g(x_{k+1}, y_{k+1}) \right] \beta_{k+1|n}(y_{k+2:n}, u_{k+1:n-1}, dx_{k+1}) \\
&\leq \int_{x_{k+1}} \left[ \int_x \nu(dx) \varsigma^+(y_{k+1}) K^{u_k}(y_{k+1}, x_{k+1}) \right] \beta_{k+1|n}(y_{k+2:n}, u_{k+1:n-1}, dx_{k+1}) \\
&= \varsigma^+(y_{k+1}) \int_{x_{k+1}} K^{u_k}(y_{k+1}, x_{k+1}) \beta_{k+1|n}(y_{k+2:n}, u_{k+1:n-1}, dx_{k+1})
\end{aligned}$$

We get a similar inequality for $\varsigma^-$. Also note that the last integral doesn't depend on $\nu$, so it cancels when we take the ratio. The result follows.

(iii) We have that

$$\begin{aligned}
F_{k|n}[y_{k+1:n}, u_{k:n-1}](x, A) &= \frac{\int_A Q^{u_k}(x, dx_{k+1}) g(x_{k+1}, y_{k+1}) \beta_{k+1|n}(x_{k+1})}{\int Q^{u_k}(x, dx_{k+1}) g(x_{k+1}, y_{k+1}) \beta_{k+1|n}(x_{k+1})} \\
&\leq \frac{\varsigma^+(y_{k+1})}{\varsigma^-(y_{k+1})} \cdot \frac{\int_A K^{u_k}(y_{k+1}, dx_{k+1}) \beta_{k+1|n}(x_{k+1})}{\int K^{u_k}(y_{k+1}, dx_{k+1}) \beta_{k+1|n}(x_{k+1})}
\end{aligned}$$

and we can set

$$\lambda_{k|n}(y_{k+1:n}, u_{k:n-1}, A) = \frac{\int_A K^{u_k}(y_{k+1}, dx_{k+1})\beta_{k+1|n}(x_{k+1})}{\int K^{u_k}(y_{k+1}, dx_{k+1})\beta_{k+1|n}(x_{k+1})}$$

(iv) Using (*iii*) we find that

$$F_{k|n}[y_{k+1:n}, u_{k:n-1}](x, A) \geq \frac{\varsigma^-(y_{k+1})}{\varsigma^+(y_{k+1})}\lambda_{k|n}(y_{k+1:n}, u_{k:n-1}, A)$$

and thus Assumption 4.3.12 holds and Lemma 4.3.13 gives

$$\delta(F_{k|n}) \leq \rho_0(y_{k+1}) = 1 - \frac{\varsigma^-(y_{k+1})}{\varsigma^+(y_{k+1})}$$

□

**Theorem 2** (Proposition 4.3.23 in [1])**.** *Under the strong mixing conditions the following holds*

*(i) We let $\nu$ and $\nu'$ be two different initial distributions for $X_0$. Now for $k \leq n$*

$$\begin{aligned}
&\|\phi_{\nu,k|n}[y_{0:n}, u_{0:n-1}] - \phi_{\nu',k|n}[y_{0:n}, u_{0:n-1}]\|_{TV} \\
&\leq \left[\prod_{j=1}^{k} \rho_0(y_j)\right] \|\phi_{\nu,0|n}[y_{0:n}, u_{0:n-1}] - \phi_{\nu',0|n}[y_{0:n}, u_{0:n-1}]\|_{TV} \\
&\leq 2\left[\prod_{j=1}^{k} \rho_0(y_j)\right]
\end{aligned}$$

*(ii) For any non-negative integers $j, k, n$ such that $j \leq k \leq n$*

$$\begin{aligned}
&\|P_\nu(X_k \in \cdot\,|y_{0:n}, u_{0:n-1}) - P_\nu(X_k \in \cdot\,|Y_{j:n}, u_{j:n-1})\|_{TV} \\
&\leq 2\prod_{i=j}^{k} \rho_0(y_i)
\end{aligned}$$

*where $\nu$ is the initial distribution of $X_0$.*

*Proof.* (i) Earlier we had

$$\|\phi_{\nu,k|n} - \phi_{\nu',k|n}\|_{TV} \leq \prod_{i=1}^{k} \delta\left(F_{i-1|n}(y_{i:n}, \cdot)\right) \|\phi_{\nu,0|n} - \phi_{\nu',0|n}\|_{TV}$$

and the first inequality now follows from the Lemma 4.3.22 part (*iv*). The factor "2" follows from using the triangle inequality on the difference of two probability measures.

(ii) This is just like part (*i*) except we consider different initial distributions for $X_j$.

□

### 2.2.6 Fisher's identity

Fisher's identity (see Proposition 10.1.6 in [1]) gives an alternative way to calculate the score function $\frac{\partial}{\partial\theta} l(\theta)$. This is based on theory associated with the EM algorithm.

In general one can set $f(x;\theta) \equiv f(x,y;\theta)$, the joint pdf of $x, y$. The likelihood for $Y$ is $L(\theta) = \int f(x;\theta)dx$ and $l(\theta) = \log L(\theta)$ the loglikelihood. Set $p(x;\theta) = \frac{f(x;\theta)}{L(\theta)}$, the conditional of $X$ given $Y$.

Now set

$$Q(\theta,\theta') = \int \log f(x;\theta) p(x;\theta') dx = E[\log f(x;\theta)|Y]$$

and

$$H(\theta,\theta') = -\int \log p(x;\theta) p(x;\theta') dx$$

We find that

$$Q(\theta, \theta') = \int \log f(x;\theta) p(x;\theta') dx = \int \log(p(x;\theta)L(\theta)) p(x;\theta') dx$$
$$= l(\theta) + \int \log p(x;\theta) p(x;\theta') dx = l(\theta) - H(\theta, \theta')$$

It is easily seen that $H(\theta, \theta')$ is minimized as a function of $\theta$ at $\theta'$ and thus

$$\frac{\partial}{\partial\theta} l(\theta') = \frac{\partial}{\partial\theta} Q(\theta, \theta')|_{\theta=\theta'} + \frac{\partial}{\partial\theta} H(\theta, \theta')|_{\theta=\theta'} = \int \frac{\partial}{\partial\theta} \log f(x;\theta)|_{\theta=\theta'} p(x, \theta') dx$$

assuming we can exchange derivatives with integration. The last equation is called Fisher's identity.

In the POMDP case this translates to

$$f(x_{0:n}, y_{0:n}, u_{0:n-1}, \theta) = \nu(x_o) g(x_0, y_0; \theta) q^{u_0}(x_0, x_1; \theta) g(x_1, y_1; \theta)$$
$$\cdots q^{u_{n-1}}(x_{n-1}, x_n; \theta) g(x_n, y_n; \theta)$$

and then

$$\log f = \log \nu(x_0;\theta) + \log(x_0, y_0; \theta) + \sum_{k=0}^{n-1} \log(q^{u_k}(x_k, x_{k+1}; \theta) g(x_{k+1}, y_{k+1}; \theta))$$

and

$$Q(\theta, \theta') = E\left[\log f | Y_{0:n}, u_{0:n-1}\right]$$
$$= E_{\theta'}\left[\log \nu(x_0;\theta) | Y_{0:n}, u_{0:n-1}\right] + E_{\theta'}\left[\log g(x_0, y_0; \theta) | Y_{0:n}, u_{0:n-1}\right]$$
$$+ \sum_{k=0}^{n-1} E_{\theta'}\left[\log(q^{u_k}(x_k, x_{k+1}; \theta) g(x_{k+1}, y_{k+1}; \theta)) | Y_{0:n}, u_{0:n-1}\right]$$

We set $\phi(x, x', u, y) = \frac{\partial}{\partial\theta} \log(q^u(x, x';\theta) g(x', y';\theta))$ and get

$$\frac{\partial}{\partial\theta} l(\theta) = E_\theta\left[\frac{\partial}{\partial\theta} \log \nu(x_0;\theta) | Y_{0:n}, u_{0:n-1}\right] + E_\theta\left[\frac{\partial}{\partial\theta} \log g(x_0, y_0; \theta) | Y_{0:n}, u_{0:n-1}\right]$$
$$+ \sum_{k=0}^{n-1} E_\theta\left[\phi(x_k, x_{k+1}, u_k, y_{k+1};\theta) | Y_{0:n}, u_{0:n-1}\right]$$

This is a different expression of the score function from the usually considered

$$\frac{\partial}{\partial\theta}l(\theta) = \sum_{k=0}^{n-1}\frac{\partial}{\partial\theta}\log p(y_{k+1}|y_{0:k}, u_{0:k-1}, \theta)$$

### 2.2.7 Bounds on score function, with adjustments

Set $h_{k,x}(\theta) = \log\left[\int g(x_k, Y_k)P(X_k \in dx_k|Y_{0:k-1}, u_{0:k-1}, X_0 = x)\right]$. Then our usual log-likelihood is $l_{x,n}(\theta) = \sum_{k=0}^{n} h_{k,x}(\theta)$

We now wish to use the expression for $\frac{\partial}{\partial\theta}l(\theta)$ derived in the last section. We have that $\frac{\partial}{\partial\theta}l_{x,n}(\theta) = \sum_{k=0}^{n}\dot{h}_{k,x}(\theta)$ but also

$$\frac{\partial}{\partial\theta}l_{x,n}(\theta) = \frac{\partial}{\partial\theta}l_{x,0}(\theta) + \sum_{k=1}^{n}\left\{\frac{\partial}{\partial\theta}l_{x,k}(\theta) - \frac{\partial}{\partial\theta}l_{x,k-1}(\theta)\right\}$$

This gives an alternative expression of $\dot{h}_{k,x}$. We get $\dot{h}_{0,x}(\theta) = \frac{\partial}{\partial\theta}\log g(x_0, Y_0)$ and for $k \geq 1$

$$\begin{aligned}
\dot{h}_{k,x}(\theta) &= \frac{\partial}{\partial\theta}l_{x,k}(\theta) - \frac{\partial}{\partial\theta}l_{x,k-1}(\theta) \\
&= E\left[\sum_{i=1}^{k}\phi(X_{i-1}, X_i, Y_i)\middle|Y_{1:k}, u_{0:k-1}, X_0 = x\right] \\
&\quad - E\left[\sum_{i=1}^{k-1}\phi(X_{i-1}, X_i, Y_i)\middle|Y_{1:k-1}, u_{0:k-2}, X_0 = x\right]
\end{aligned}$$

This expression can be generalized to starting the process at other values than zero;

$$\begin{aligned}
\dot{h}_{k,m,x}(\theta) &= \log\left[\int g(x_k, Y_k)P(X_k \in dx_k|Y_{m:k-1}, u_{m:k-1}, X_m = x)\right] \\
&= E\left[\sum_{i=m+1}^{k}\phi(X_{i-1}, X_i, Y_i)\middle|Y_{m+1:k}, u_{m:k-1}, X_m = x\right] \\
&\quad - E\left[\sum_{i=m+1}^{k-1}\phi(X_{i-1}, X_i, Y_i)\middle|Y_{m+1:k-1}, u_{m:k-2}, X_m = x\right]
\end{aligned}$$

This is done in Cappe et al. [1] to extend the process to minus infinity ($m \to -\infty$). We don't extend the process to infinity, but rather think of $m$ as indicating lack of information, that is assuming that the process starts at $X_m$.

We now prove a modified Lemma 12.5.3 where we use the expression developed above.

**Theorem 3** (Lemma 12.5.3 in [1] modified)**.** *Assuming strong mixing conditions. Then for $k \geq 1$ Cappe et al. [1] prove the following inquality in the HMM case:*

$$\left(E|\dot{h}_{k,-m,x}(\theta) - \dot{h}_{k,\infty}(\theta)|^2\right)^{1/2} \leq 12\left(E\left[\sup_{x,x' \in X} |\phi_\theta(x, x', Y_1)|^2\right]\right)^{1/2} \frac{\rho^{(k+m)/2-1}}{1-\rho}$$

*We don't extend the process to $-\infty$, but rather starting at $X_0$ and we prove the following inequality, also for $k \geq 1$*

$$\left(E|\dot{h}_{k,0,x_0}(\theta) - \dot{h}_{k,m,x}(\theta)|^2\right)^{1/2} \leq 8 \sup_{x,x' \in X, u \in U, y \in Y} \|\phi_\theta(x, x', y, u)\| \frac{\rho^{(k-m)/2-1}}{1-\rho}$$

*where $\rho = \max_{y \in Y} \rho_0(y)$ (See Theorem 1).*

*Proof.* From the representation derived above for $\dot{h}$ we have

$$\dot{h}_{k,0,x_0}(\theta) = E\left[\sum_{i=1}^{k} \phi(X_{i-1}, X_i, Y_i, u_{i-1}) \middle| Y_{1:k}, u_{0:k-1}, X_0 = x_0\right] \tag{1}$$

$$- E\left[\sum_{i=1}^{k-1} \phi(X_{i-1}, X_i, Y_i, u_{i-1}) \middle| Y_{1:k-1}, u_{0:k-2}, X_0 = x_0\right] \tag{2}$$

and

$$\dot{h}_{k,m,x}(\theta) = E\left[\sum_{i=m+1}^{k} \phi(X_{i-1}, X_i, Y_i, u_{i-1}) \middle| Y_{m+1:k}, u_{m:k-1}, X_m = x\right] \tag{3}$$

$$- E\left[\sum_{i=m+1}^{k-1} \phi(X_{i-1}, X_i, Y_i, u_{i-1}) \middle| Y_{m+1:k-1}, u_{m:k-2}, X_m = x\right] \tag{4}$$

Just like in the proof of Lemma 12.5.3 in [1] we match together different pairs of terms within the sums, depending on their index $i$. More specifically for $i = k$

we match together the terms where $i = k$ in (1) and (3). For $\frac{k+m}{2} \leq i < k$ we match the terms in (1) with (3) and the terms in (2) with those in (4). For $m+1 \leq i < \frac{k+m}{2}$ we match terms in (1) with terms in (2) and terms in (3) with those in (4). That leaves $i \in 1, \ldots, m$ in $\dot{h}_{k,0,x_0}$ where we match (1) and (2).

If we look at the case where (1) is matched with (3) we have

$$
\begin{aligned}
&\|E[\phi_\theta(X_{i-1}, X_i, Y_i, u_{i-1})|Y_{m+1:k}, u_{m:k-1}, X_m = x] - E[\phi_\theta(X_{i-1}, X_i, Y_i, u_{i-1})|Y_{1:k}, u_{0:k-1}]\| \\
&= \Big| \int_{x_m} \int_{x_{i-1}} \int_{x_i} \phi_\theta(x_{i-1}, x_i, Y_i, u_i) F_{i-1}(x_{i-1}, dx_i) P_\theta(X_{i-1} \in dx_{i-1}|Y_{m+1:k}, u_{m:k-1}, X_m = x) \\
&\times [\delta_x(dx_m) - P_\theta(X_m \in dx_m|Y_{1:k}, u_{0:k-1})]\Big| \\
&\leq 2 \sup_{x,x' \in X, u \in U} \|\phi_\theta(x, x', Y_i, u)\| \rho^{(i-1)-m}
\end{aligned}
$$

where $F_{i-1} = F_{i-1;\theta}[y_{i:k}, u_{i-1:k}]$ is the Forward Smoothing Kernel, and the inequality stems from Proposition 4.3.23 ($i$) where the second line can be thought of as two different initial distributions for $X_m$, and the kernel $F$ is bounded by 1.

Matching (2) with (4) is similar. For matching (1) with (2) and (3) with (4) we need a "Backwards bound";

$$
\|P_\theta(X_i \in \cdot\,|Y_{m+1:k}, u_{m:k-1}, X_m = x) - P_\theta(X_i \in \cdot\,|Y_{m+1:k-1}, u_{m:k-2}, X_m = x)\|_{TV} \leq 2\rho^{k-1-i}
$$

that is established below, see Theorem 4. For matching (3) with (4) we get

$$
\begin{aligned}
&\|E_\theta[\phi_\theta(X_{i-1}, x_i, Y_i, u_{i-1})|Y_{m+1:k}, u_{m:k-1}, X_m = x] \\
&- E_\theta[\phi_\theta(X_{i-1}, x_i, Y_i, u_{i-1})|Y_{m+1:k-1}, u_{m:k-2}, X_m = x]\| \\
&= |\int_{x_{i-1}} \int_{x_i} \phi_\theta(x_{i-1}, x_i, Y_i, u_{i-1}) B_i(x_i, dx_{i-1}) \\
&\times [P_\theta(X_i \in dx_i|Y_{m+1:k}, u_{m:k-1}, X_m = x) - P_\theta(X_i \in dx_i|Y_{m+1:k-1}, u_{m:k-2}, X_m = x)]| \\
&\leq 2 \sup_{x,x' \in X, u \in U} \|\phi_\theta(x, x', Y_i, u)\| \rho^{(k-1)-i}
\end{aligned}
$$

where $B_i$ is the Backwards Smoothing Kernel described below. Matching (1) with (2) is a special case of the above.

Going back to our original objective, we have

$$\left(E_\theta \|\dot{h}_{k,m,x}(\theta) - \dot{h}_{k,0,x_0}(\theta)\|^2\right)^{1/2} = \left(E\left\|\sum a_i\right\|^2\right)^{1/2}$$

where $\sum a_i$ is a sum over the pairs we considered above. Now by Minkowski's inequality we have

$$\leq \sum \left(E\,\|a_i\|^2\right)^{1/2}$$

Now we have that $\|a_i\| \leq 2\sup_{x,x'\in X,u\in U} \|\phi_\theta(x,x',Y_i,u)\|\rho^{b_i}$ where $b_i$ is the power of $\rho$ associated with $a_i$.

$$\leq \sum 2\left(E \sup_{x,x'\in X,u\in U} \|\phi_\theta(x,x',Y_i,u)\|^2\right)^{1/2} \rho^{b_i}$$

At this point Cappe et al. [1] argue that since in their case the process was started at infinity and the process is homogeneous the expected value over $Y_i$ is always the same by stationarity, and $Y_i$ can be exchanged by $Y_1$. Since arguing for stationarity is more of stretch for us, we also take the supremum over $Y$ an is also finite.

$$\begin{aligned}
&\leq 2\left(\sup_{x,x'\in X,u\in U,y\in Y} \|\phi_\theta(x,x',y,u)\|^2\right)^{1/2} \sum \rho^{b_i} \\
&= 2 \sup_{x,x'\in X,u\in U,y\in Y} \|\phi_\theta(x,x',y,u)\| \sum \rho^{b_i}
\end{aligned}$$

We now deal with the sum of $\rho$ to different powers.

From $i = k$ we have $\rho^{k-1-m}$ where we matched (1) with (3). For $\frac{k+m}{2} \leq i < k$ we have $2\rho^{i-1-m}$ where we matched (1) with (3) and (2) with (4). For $m+1 \leq i < \frac{k+m}{2}$ we have $2\rho^{k-1-i}$ from matching (1) with (2) and (3) with (4). Finally for $1 \leq i \leq m$

we have $\rho^{k-1-i}$ from matching (1) with (2). This gives

$$\begin{aligned}
\sum \rho^{b_i} &= \rho^{k-1-m} + \sum_{i=(k+m)/2}^{k-1} 2\rho^{i-1-m} + \sum_{i=m+1}^{(k+m)/2-1} 2\rho^{k-1-i} + \sum_{i=1}^{m} \rho^{k-1-i} \\
&\leq 2 \sum_{i=(k+m)/2}^{\infty} \rho^{i-1-m} + 2 \sum_{i=-\infty}^{(k+m)/2-1} \rho^{k-1-i} \\
&= 2\frac{\rho^{(k-m)/2-1}}{1-\rho} + 2\frac{\rho^{(k-m)/2}}{1-\rho} \leq 4\frac{\rho^{(k-m)/2-1}}{1-\rho}
\end{aligned}$$

Thus, finally we have

$$\left(E_\theta \|\dot{h}_{k,m,x}(\theta) - \dot{h}_{k,0,x_0}(\theta)\|^2\right)^{1/2} \leq 8 \sup_{x,x' \in X, u \in U, y \in Y} \|\phi_\theta(x, x', y, u)\| \frac{\rho^{(k-m)/2-1}}{1-\rho}$$

□

**Theorem 4** (Proposition 12.5.4 modified)**.**

$$\|P_\theta(X_i \in \cdot | Y_{m+1:k}, u_{m:k-1}, X_m = x) - P_\theta(X_i \in \cdot | Y_{m+1:k-1}, u_{m:k-2}, X_m = x)\|_{TV} \leq 2\rho^{k-1-i}$$

*Proof.* The idea behind this proof is to replicate all the results derived so far for the Backward Smoothing Kernel. That is, conditional on $Y_{m+1:k}$, $u_{m:k-1}$ and $X_m = x_m$ the time-reversed process $X$ is a non-homogeneous Markov Chain, where the conditional probability of moving from $X_{j+1}$ to $X_j$ given all the observations $Y_{m+1:k-1}$, controls $u_{m:k-2}$ and initial condition ends up only depending on $Y_{m+1:j}$, $u_{m:j}$ and the initial condition, and is governed by the Backwards Smoothing Kernel given by

$$\begin{aligned}
&B_{x_m,j}[u_{m+1:j}, u_{m:j}](x, f) \\
&= \frac{\int \cdots \int \prod_{r=m+1}^{j} Q^{u_{r-1}}(x_{r-1}, dx_r) g(x_r, y_r) f(x_j) Q^{u_j}(x_j, x)}{\int \cdots \int \prod_{r=m+1}^{j} Q^{u_{r-1}}(x_{r-1}, dx_r) g(x_r, y_r) Q^{u_j}(x_j, x)}
\end{aligned}$$

Just as we did in Lemma 4.3.22 we can show

$$\frac{\varsigma^-(y_j)}{\varsigma^+(y_j)} \nu_{x_m,j}[y_{m+1}, u_{m:j}] \leq B_{x_m,j}[y_{m+1:j}, u_{m:j}](x_j, \cdot) \leq \frac{\varsigma^+(y_j)}{\varsigma^-(y_j)} \nu_{x_m,j}[y_{m+1}, u_{m:j}]$$

where

$$\nu_{x_m,j}[y_{m+1}, u_{m:j}](f) = \frac{\int \cdots \int \prod_{r=m+1}^{j} Q^{u_{r-1}}(x_{r-1}, dx_r) g(x_r, y_r) f(x_j)}{\int \cdots \int \prod_{r=m+1}^{j} Q^{u_{r-1}}(x_{r-1}, dx_r) g(x_r, y_r)}$$

As we showed there this gives

$$\delta(B_{x_m,j}) \leq 1 - \frac{\varsigma^-(y_j)}{\varsigma^+(y_j)}$$

We now get that the 2 smoothers we are interested in can be thought of as smoothers of the reversed Markov Chain from $k-1$ to $m$ with 2 different initial distributions for $X_{k-1}$, the starting position. We get

$$\begin{aligned}
&\|P_\theta(X_i \in \cdot | Y_{m+1:k}, u_{m:k-1}, X_m = x) - P_\theta(X_i \in \cdot | Y_{m+1:k-1}, u_{m:k-2}, X_m = x)\|_{TV} \\
&\leq \|P_\theta(X_{k-1} \in \cdot | Y_{m+1:k}, u_{m:k-1}, X_m = x) - P_\theta(X_{k-1} \in \cdot | Y_{m+1:k-1}, u_{m:k-2}, X_m = x)\|_{TV} \\
&\times \prod_{j=i+1}^{k-1} \delta(B_{x_m,j}) \leq 2 \prod_{j=i+1}^{k-1} \rho_0(y_j) \leq 2\rho^{k-1-i}
\end{aligned}$$

(where $\rho = \max_{y \in Y} \rho_0(y)$) □

## 2.3 Markov Decision Processes theory

In this section we review relevant MDP theory, noting that methods like dynamic programming and the Value Iteration Algorithm will be useful in our pursuit to maximize various forms of Fisher Information.

We assume a Markov Decision Process $(X_t, u_t)_{t=0,\ldots,T}$ like we do in Section 2.1, but now without the observation process $(Y_t)_{t=0,\ldots,T}$. In the standard MDP problem we assume a reward function $C_t(X_t, u_t)$ and the objective is to maximize the total expected reward $W_1$

$$W_1 = E\left[\sum_{t=0}^{T} C_t(X_t, u_t)\right]$$

by use of the controls. The essence of dynamic programming is that by starting at time $T-1$ and working backwards, we can compute an optimal policy that maps a state $x_t$ to a control $u_t$ that accounts for the choices of $u_t$ that we will make in the future.

In a generic dynamic program we set $V_T = 0$ and and then going backwards from $t = T-1, \ldots, 0$ solve

$$V_t(x_t) = \max_{u_t}\{E_{x_{t+1}}[C_t(x_t, u_t) + V_{t+1}(x_{t+1})|x_t, u_t]\}$$

where $V_t$ is called the value function, and we get the associated control

$$u_t^*(x_t) = \operatorname*{argmax}_{u_t}\{E_{x_{t+1}}[C_t(x_t, u_t) + V_{t+1}(x_{t+1})|x_t, u_t]\}$$

for every state $x_t$. This will give us a policy of what control to use at a certain state $x_t$ at a certain time $t$. The use of these controls will maximize the expected total reward $E[\sum_t C_t(X_t, u_t)]$. We refer to Puterman [8] for a detailed description of dynamic programming.

It will be useful to consider other criteria then the expected total reward $W_1$. Assuming an infinite time horizon we consider the expected average reward $W_2$

$$W_2 = \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{t=0}^{n} C(x_t, u_t)\right]$$

(note that this limit doesn't always exist) and the expected total discounted reward $W_3$

$$W_3 = E\left[\sum_{t=0}^{\infty} \lambda^{t-1} C(x_t, u_t)\right]$$

that has a discounting factor $\lambda$ where $0 \leq \lambda < 1$, and exists if the reward $C$ is bounded. In both $W_2$ and $W_3$ we assume that the reward function $C$ is stationary (time independent).

Maximizing the expected total discounted reward $W_3$ if frequently done via the Value Iteration Algorithm (VIA), another popular MDP algorithm, again see [8]. In VIA we calculate

$$v^{n+1}(x_t, \theta) = \max_u \{E_{x_{t+1}}[C(x_t, u_t, \theta) + \lambda \cdot v^n(x_{t+1}, \theta)|x_t, u_t, \theta]\}$$

with the associated control

$$u^{n+1}(x_t, \theta) = \operatorname*{argmax}_u \{E_{x_{t+1}}[C(x_t, u_t, \theta) + \lambda \cdot v^n(x_{t+1}, \theta)|x_t, u_t, \theta]\}$$

in a while-loop until $v^n$ converges to some fixed point, within some tolerance. Convergence is guaranteed since each iteration of $v^n$ is a contraction mapping. We note that the output of VIA will be a stationary policy, i.e. a policy that only depends on the state $x_t$ and not the time $t$.

To analyze the expected average reward $W_2$ we need some additional assumptions on our MDP.

**Definition 4.** *We say that a MDP is unichain if for every deterministic stationary policy the transition probability matrix consists of a single recurrent class plus a possibly empty set of transient states.*

In section 8.4.2 in Puterman [8] it is shown that if a MDP is unichain, the reward $C$ is bounded and stationary (time independent) and the state space $\mathcal{X}$ and the action space $\mathcal{U}$ are finite then there exists a stationary policy that maximizes $W_2$. In section 8.5.1 they show that under these same assumptions, running VIA with $\lambda = 1$ converges in $W_2(u^n)$ (The expected average reward, if only using control $u^n$) to its maximum, even though the value function $v^n$ generally diverges. Also note that the operation of dynamic programming is analogous to the operation of VIA with $\lambda = 1$.

One of the reason we bring this up is that when we analyze a control policy for a MDP $(x_t, u_t)_{t=0,\dots,T}$ that is the output of a dynamic program, we will often only look at the policy $u_t^*$ when $t = 1$, and label it the long-term policy. By the above, we can argue that this is informative because it corresponds to a policy that is close maximizing the expected average reward, and we can expect the policy to converge in some sense as $t \to 0$ if $T$ is large enough (meaning that there generally will not be much difference between the policy at say time $t = 1$ and $t = 2$).

The last thing we mention from MDP theory is Blackwell optimality. It guarantees that a stationary policy that maximizes the expected total discounted reward $W_3$ also maximizes the expected average reward $W_2$ (or its lim sup if the limit doesn't exist), given that $\lambda$ is chosen close enough to one. This will be important when we consider algorithms based on VIA in Chapter 8. A Blackwell optimal control policy exists under the same assumptions as listed above. How small $1 - \lambda$ needs to be is generally hard to determine, and choosing $\lambda$ too high will cause VIA to converge slowly. See Puterman [8] chapter 10 for more on Blackwell optimality.

CHAPTER 3

# FISHER INFORMATION

## 3.1 Objectives

Consider again our framework stated in Section 2.1. Our goal is to use the controls $u_{0:T}$ to get an estimate of the parameter $\theta$ that is as accurate as possible. We will estimate $\theta$ by maximizing the likelihood, see section 4.4. MLE's are, under suitable regularity conditions, unbiased and asymptotically efficient, with asymptotic variance equal to the inverse Fisher Information

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, (FI(\theta))^{-1})$$

Thus our strategy will be to maximize a Fisher Information $FI(\theta, u_{0:T})$ by using the controls $u_t$ adaptively, that is at time $t$ the $u_t$ chosen can, and generally should, depend on the observations $y_1, \ldots, y_t$. We will now discuss various forms of Fisher Information and their properties.

## 3.2 FOFI

The first attempt at maximizing Fisher Information using controls was by Hooker et al. [5]. They considered constructing an optimal control policy for the Fisher Information that would apply if $(X_t)$ were observed directly;

$$FI = E \sum_{t=0}^{T-1} \left( \frac{\partial}{\partial \theta} \log p(x_{t+1}|x_t, u_t, \theta) \right)^2$$

We label this the Full Observation Fisher Information (FOFI). When considering continuous time stochastic systems, the state space is continuous, but we

use this Fisher Information as an approximation to the continuous state Fisher Information. An advantage of using FOFI is that when running the dynamic program the Markov property of the Markov Decision Process $(X_t, u_t)$ allows us to only consider a maximization over the state space $x_t \in \mathcal{X}$ but not past values $x_{0:t-1}$. The dynamic program for FOFI is given in Section 4.1.1.

However, maximizing FOFI can lead to suboptimal controls since it is not the correct Fisher Information for the data. Additionally, when the actual experiment is run we do not observe $X_t$. Instead we have to use the observed values to calculate a filter for the state $x_t$, $p(x_t|y_{0:t}, u_{0:t-1}, x_0, \theta)$, and use the control associated with the state that has the highest probability.

## 3.3 POFI

Since the objective is to use the controls to maximize the information about the parameter $\theta$ through the observed process $y_{0:t}$ it seems natural to maximize the Fisher Information associated with the observed process, in some sense the correct Fisher Information for the data,

$$FI(\theta) = E\left[\sum_{t=0}^{T-1}\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|y_{0:t}, u_{0:t}, x_0, \theta)\right)^2\right]$$

which we label as the Partial Observation Fisher Information (POFI), see Section 3.4 for details on its construction. When we consider continuous time dynamical systems the observation spaces will be continuous, but we will use this discretized Fisher Information as an approximation to the actual Fisher Information of the observations.

Maximizing POFI using dynamic programming or a similar algorithm is generally not feasible, due to the curse of dimensionality, see section 4.2. We thus try to approximate POFI, with Fisher Information like criteria that are easier to maximize, see Truncated POFI in section 3.5 and Weighted Observation Fisher Information in section 3.7. Also see Sections 4.2 and 4.3 for a description of the dynamic program for the respective criteria.

## 3.4 Expressing POFI

In this section we find useful expressions for POFI, the Fisher Information of a POMDP where we observe $y_{0:T}$ and use controls $u_{0:T-1}$, that are needed to derive convergence arguments and set up dynamical programs. We use the short hand notation

$$h_k(\theta) = \log p(y_k|y_{0:k-1}, u_{0:k-1}, x_0 = x)$$

where the dependence on $X_0 = x$ is frequently suppressed.

For data $Y_0, \ldots, Y_T$ the Fisher Information for $\theta$ can be expressed in one or two derivatives

$$FI = E\left[\sum_{t=0}^{T-1} -\ddot{h}_{t+1}\right] = E\left[\left(\sum_{t=0}^{T-1} \dot{h}_{t+1}\right)^2\right]$$

and we define the Fisher Information to Go at time $k$ to be

$$FI_k = E\left[\sum_{t=k}^{T-1} -\ddot{h}_{t+1}\middle| y_{0:k}, u_{0:k-1}\right] = E\left[\left(\sum_{t=k}^{T-1} \dot{h}_{t+1}\right)^2\middle| y_{0:k}, u_{0:k-1}\right]$$

where the equality is justified by both quantities being the Fisher Information for the same observations. We see that $FI_0 = FI$.

The Fisher Information to Go can be calculated recursively (in both one or two derivatives):

**Lemma 1.**

$$FI_k = E\left[-\ddot{h}_{k+1} + FI_{k+1}\middle| y_{0:k}, u_{0:k-1}\right] = E\left[\left(\dot{h}_{k+1}\right)^2 + FI_{k+1}\middle| y_{0:k}, u_{0:k-1}\right]$$

*Proof.* In the case of using two derivatives this follows from iterated expectation. In one derivative we have

$$\begin{aligned} FI_k &= E\left[\left(\sum_{t=k}^{T-1} \dot{h}_{t+1}\right)^2 \middle| y_{0:k}, u_{0:k-1}\right] \\ &= E\left[\left(\dot{h}_{k+1}\right)^2 + \left(\sum_{t=k+1}^{T-1} \dot{h}_{t+1}\right)^2 + 2\left(\dot{h}_{k+1}\right)\left(\sum_{t=k+1}^{T-1} \dot{h}_{t+1}\right) \middle| y_{0:k}, u_{0:k-1}\right] \end{aligned}$$

The cross term is

$$\begin{aligned} & E\left[2\left(\dot{h}_{k+1}\right)\left(\sum_{t=k+1}^{T-1} \dot{h}_{t+1}\right) \middle| y_{0:k}, u_{0:k-1}\right] \\ &= E\left[E\left[2\left(\dot{h}_{k+1}\right)\left(\sum_{t=k+1}^{T-1} \dot{h}_{t+1}\right) \middle| y_{0:k+1}, u_{0:k}\right] \middle| y_{0:k}, u_{0:k-1}\right] \\ &= E\left[2\left(\dot{h}_{k+1}\right) E\left[\left(\sum_{t=k+1}^{T-1} \dot{h}_{t+1}\right) \middle| y_{0:k+1}, u_{0:k}\right] \middle| y_{0:k}, u_{0:k-1}\right] \\ &= E\left[2\left(\dot{h}_{k+1}\right) \cdot 0 \middle| y_{0:k}, u_{0:k-1}\right] = 0 \end{aligned}$$

Thus

$$\begin{aligned} FI_k &= E\left[\left(\dot{h}_{k+1}\right)^2 + \left(\sum_{t=k+1}^{T-1} \dot{h}_{t+1}\right)^2 \middle| y_{0:k}, u_{0:k-1}\right] \\ &= E\left[\left(\dot{h}_{k+1}\right)^2 + E\left[\left(\sum_{t=k+1}^{T-1} \dot{h}_{t+1}\right)^2 \middle| y_{0:k+1}, u_{0:k}\right] \middle| y_{0:k}, u_{0:k-1}\right] \\ &= E\left[\left(\dot{h}_{k+1}\right)^2 + FI_{k+1} \middle| y_{0:k}, u_{0:k-1}\right] \end{aligned}$$

□

**Corollary 1.**

$$FI = E\left[\sum_{t=0}^{T-1}\left(\dot{h}_{t+1}\right)^2\right]$$

*and similarly*

$$FI_k = E\left[\sum_{t=k}^{T-1}\left(\dot{h}_{t+1}\right)^2\middle| y_{0:k}, u_{0:k-1}\right]$$

*Proof.* This follows from using induction and lemma 1. □

### 3.4.1 One or two derivates?

We note that we can both try to maximize the $FI$ expressed in one or in two derivates. We only used the former since it was slightly easier to calculate. There was no noticeable difference between the two in practice.

## 3.5 Truncated POFI

Running an exact dynamic program to maximize POFI is not feasible due to the curse of dimensionality, requiring us to do certain approximations. We set

$$h_{k,m,\nu_m}(\theta) = \begin{cases} \log p(y_k|y_{m:k-1}, u_{m:k-1}, \nu_m) & \text{if } m \geq 0 \\ \log p(y_k|y_{0:k-1}, u_{0:k-1}, \nu_0) & \text{if } m < 0 \end{cases}$$

where $\nu_m$ is the assumed distribution of $x_m$ and we will consider it to be fixed and known. Allowing $m$ to be negative will ease notation when $t - m < 0$. We

set

$$FI_{trunc} = E \sum_{k=0}^{T-1} \left( \dot{h}_{k+1,k-m,\nu_{k-m}}(\theta) \right)^2$$

and label it as the truncated Partially Observed Fisher Information, also see Section 4.2.

Similarly the truncated Fisher Information to go is

$$\begin{aligned} FI_{k,m} &= E \left[ \sum_{t=k}^{T-1} \left( \dot{h}_{t+1,t-m,\nu_{t-m}} \right)^2 \middle| y_{k-m:k}, u_{k-m:k-1} \right] \\ &= E \left[ \sum_{t=k}^{T-1} -\ddot{h}_{t+1,t-m,\nu_{t-m}} \middle| y_{k-m:k}, u_{k-m:k-1} \right] \end{aligned}$$

That the formulation in one derivative is equal to the one in two derivatives follows from the individual parts of each sum having a Fisher Information interpretation. In our notation we also have $FI_{0,m} = FI_{trunc}$.

### 3.5.1 Mixing conditions

Cappe et al. [1] establish forgetting properties of the filter by assuming mixing conditions for Hidden Markov Models. We use the same conditions, slightly modified to allow for controls, see Assumption 3, here restated for a discrete state space $\mathcal{X}$.

**Assumption 4.** *Modified Strong Mixing Conditions. For each control $u$ there exist a transition kernel $K^u : Y \to X$ and measurable functions $\varsigma^-$ and $\varsigma^+$ from $Y$ to $(0, \infty)$ such that for any $A \in \mathcal{X}$, $y \in Y$ and $x \in X$,*

$$\varsigma^-(y) K^u(y, A) \leq \sum_{x' \in A} p(y_{t+1} = y | x_{t+1} = x') p(x_{t+1} = x' | x_t = x, u_t = u) \leq \varsigma^+(y) K^u(y, A)$$

Cappe et al.'s [1] discussion on what models satisfy these conditions applies analogously to POMDP's. Given these conditions we prove the following bound in Theorem 3, which is a modification of Lemma 12.5.3 in [1];

$$(E|\dot{h}_{k,0,\nu_0}(\theta) - \dot{h}_{k,m,\nu_m}(\theta)|^2)^{1/2} \leq 8 \sup_{x,x' \in X, u \in U, y \in Y} ||\phi_\theta(x, x', y, u)|| \frac{\rho^{(k-m)/2-1}}{1-\rho}$$

where $\phi(x, x', u, y) = \frac{\partial}{\partial \theta} \log(p(x_{t+1} = x'|x_t = x, u_t = u, \theta)p(y_{t+1} = y'|x_{t+1} = x', \theta))$ and $\rho = \max_{y \in Y} 1 - \frac{\varsigma^-(y)}{\varsigma^+(y)}$

## 3.6 Truncated POFI convergence theorem

In this section we show that the truncated Fisher Information approaches the true Fisher Information exponentially as one conditions on more and more observations, while using the same controls.

By Corollary 1 the true Fisher Information (POFI) is

$$FI(\theta, u_{0:T-1}) = E \sum_{k=0}^{T-1} \left(\dot{h}_{k+1,0,\nu_0}(\theta)\right)^2$$

but since that is hard to optimize we consider

$$FI_{trunc} = FI_{0,m}(\theta, u_{0:T-1}) = E \sum_{k=0}^{T-1} \left(\dot{h}_{k+1,k-m,\nu_{k-m}}(\theta)\right)^2$$

see definitions for $\dot{h}$ above. Here we use Fisher Information in one derivative, but as noted above it is equivalent to using the formulation in two derivatives. Also note that where $k - m < 0$ we just set it to 0 and use the initial distribution of $x_0$.

**Lemma 2.** *Assume the mixing conditions in Assumption 4 hold. Then*

$$\left(E(\dot{h}_{k+1,0,\nu_0} + \dot{h}_{k+1,k-m,\nu_{k-m}})^2\right)^{1/2}$$
$$\leq 16 \sup_{x,x' \in X, u \in U, y \in Y} |\phi_\theta(x, x', y, u)| \frac{\rho^{1/2}}{1-\rho} + 2 \sup_{u_0} \left(E(\dot{h}_{1,0,\nu_0})^2\right)^{1/2}$$

*Proof.* Set

$$A(m') = \sup_{u_1,\ldots,u_{m'-1}} \left(E(\dot{h}_{m',0,\nu_0})^2\right)^{1/2}$$

which sets an upper bound on the length of $\dot{h}_m$. Note that $A(m')$ also bounds $(E(\dot{h}_{k+1,k-m,\nu_{k-m}})^2)^{1/2}$ since $\nu_{k-m} = \nu_0$. Now

$$\begin{aligned}
&\left(E(\dot{h}_{k+1,0,\nu_0} + \dot{h}_{k+1,k-m,\nu_{k-m}})^2\right)^{1/2} \\
&\leq \left(E(\dot{h}_{k+1,0,\nu_0} - \dot{h}_{k+1,k-m',\nu_{k-m'}})^2\right)^{1/2} + \left(E(\dot{h}_{k+1,k-m,\nu_{k-m}} - \dot{h}_{k+1,k-m',\nu_{k-m'}})^2\right)^{1/2} \\
&+ 2\left(E(\dot{h}_{k+1,k-m',\nu_{k-m'}})^2\right)^{1/2} \\
&\leq 16 \sup|\phi_\theta| \frac{\rho^{(1+\min(m,m'))/2}}{1-\rho} + 2A(m'+1)
\end{aligned}$$

using Theorem 3. Setting $m' = 0$ gives the result, although that might not be the best bound. □

**Lemma 3.** *Assume the conditions in Assumption 4 hold. Then, for any control policy and any k such that $k - m \geq 0$, we have*

$$\left|E(\dot{h}^2_{k+1,0,\nu_0} - \dot{h}^2_{k+1,k-m,\nu_{k-m}})\right| \leq 8M(\theta) \sup_{x,x\in X,u\in U,y\in Y} |\phi_\theta(x,x',y,u)| \frac{\rho^{(m+1)/2-1}}{1-\rho}$$

*where $M(\theta) = 16 \sup|\phi_\theta| \frac{\rho^{1/2}}{1-\rho} + 2\sup_{u_0}\left(E(\dot{h}_{1,0,\nu_0})^2\right)^{1/2}$ is the bound from lemma 2.*

*Proof.*

$$\left|E(\dot{h}^2_{k+1,0,\nu_0} - \dot{h}^2_{k+1,k-m,\nu_{k-m}})\right| \leq \left|E(\dot{h}_{k+1,0,\nu_0} - \dot{h}_{k+1,k-m,\nu_{k-m}}) \cdot (\dot{h}_{k,0,\nu_0} + \dot{h}_{k,k-m,\nu_{k-m}})\right|$$

and by Cauchy Schwarz

$$\leq \left(E\left|\dot{h}_{k+1,0,\nu_0} - \dot{h}_{k+1,k-m,\nu_{k-m}}\right|^2\right)^{1/2} \cdot \left(E\left|\dot{h}_{k,0,\nu_0} + \dot{h}_{k,k-m,\nu_{k-m}}\right|^2\right)^{1/2}$$

For the first parenthesis we use Theorem 3 to get

$$\left(E\left|\dot{h}_{k+1,0,v_0} - \dot{h}_{k+1,k-m,v_{k-m}}\right|^2\right)^{1/2} \leq 8 \sup_{x,x\in X,u\in U,y\in Y} |\phi_\theta(x,x,y,u)| \frac{\rho^{(m+1)/2-1}}{1-\rho}$$

and the second one is bounded by the Lemma 2. □

**Theorem 5.** *Assume the conditions in Assumption 4 hold. Then, for $m < T$ and any control policy, we have*

$$\left|FI - FI_{0,m}\right| \leq c_1(T-1-m)\rho^{m/2}$$

*where $c_1 = 8M(\theta)\sup_{x,x'\in X,u\in U,y\in Y}|\phi_\theta(x,x',y,u)|\frac{1}{\rho^{1/2}(1-\rho)}$ and $M(\theta)$ the bound from lemma 2; $M(\theta) = 16\sup|\phi_\theta|\frac{\rho^{1/2}}{1-\rho} + 2\sup_{u_0}\left(E(\dot{h}_{1,0,v_0})^2\right)^{1/2}$.*

*Proof.*

$$\begin{aligned}
\left|FI - FI_{0,m}\right| &= \left|E\sum_{k=0}^{T-1}\left(\dot{h}_{k+1,0,v_0}(\theta)\right)^2 - E\sum_{k=0}^{T-1}\left(\dot{h}_{k+1,k-m,v_{k-m}}(\theta)\right)^2\right| \\
&= \left|\sum_{k=m+1}^{T-1} E\left(\dot{h}^2_{k+1,0,v_0} - \dot{h}^2_{k+1,k-m,v_{k-m}}\right)\right| \\
&\leq \sum_{k=m+1}^{T-1}\left|E\left(\dot{h}^2_{k+1,0,v_0} - \dot{h}^2_{k+1,k-m,v_{k-m}}\right)\right| \\
&\leq (T-1-m)8M(\theta)\sup_{x,x\in X,u\in U,y\in Y}|\phi_\theta(x,x,y,u)|\,\frac{\rho^{(m+1)/2-1}}{1-\rho}
\end{aligned}$$

by Lemma 3. □

Exactly the same arguments can be used to show that the truncated Fisher Information to Go $FI_{k,m}$ approaches the true Fisher Information to Go as $m$ increases.

## 3.7 WOFI

We now consider a different way to approximate the Partial Observation Fisher Information. POFI is expressed as

$$POFI = E\left[\sum_{t=0}^{T-1}\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|y_{0:t}, u_{0:t}, x_0, \theta)\right)^2\right]$$

and examining the score function gives

$$\begin{aligned}&\frac{\partial}{\partial\theta}\log p(y_{t+1}|y_{0:t}, u_{0:t}, x_0, \theta)\\ &= \frac{\sum_{x_t}\left(\frac{\partial}{\partial\theta}p(y_{t+1}|x_t, u_t, \theta)p(x_t|y_{0:t}, u_{0:t-1}, \theta) + p(y_{t+1}|x_t, u_t, \theta)\frac{\partial}{\partial\theta}p(x_t|y_{0:t}, u_{0:t-1}, \theta)\right)}{\sum_{x_t} p(y_{t+1}|x_t, u_t, \theta)p(x_t|y_{0:t}, u_{0:t-1}, \theta)}\end{aligned}$$

where $p(y_{t+1}|x_t, u_t, \theta) = \sum_{x_{t+1}} p(y_{t+1}|x_{t+1}, \theta)p(x_{t+1}|x_t, u_t, \theta)$. If we now assume that the filter $p(x_t|y_{0:t}, u_{0:t-1}, \theta)$ is fairly accurate at determining $x_t$ and that it is not very dependent on $\theta$ ($\frac{\partial}{\partial\theta}p(x_t|y_{0:t}, u_{0:t-1}, \theta) \approx 0$ ) we can motivate the reward function

$$C_t(x_t, u_t, y_{t+1}, \theta) = \left(\frac{\frac{\partial}{\partial\theta}p(y_{t+1}|x_t, u_t, \theta)}{p(y_{t+1}|x_t, u_t, \theta)}\right)^2 = \left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|x_t, u_t, \theta)\right)^2$$

as an approximation to the POFI reward function. See Section 3.8 for details. The corresponding “Fisher Information" for the whole system is now given as

$$FI = E\left[\sum_{t=0}^{T-1}\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|x_t, u_t, \theta)\right)^2\right]$$

and we label it as the Weighted Observation Fisher Information (WOFI). Running a dynamic program with WOFI, see section 4.3, has the same computational cost as FOFI, that is $O(TK^2l)$, but as with FOFI the state $X_t$ needs to be estimated at runtime with cost $O(K^2)$ at each time point $t$.

## 3.8 WOFI approximates POFI theorem

In this section we show how the WOFI criteria approximates the POFI criteria given that the following assumption holds for all $t = 1, \ldots, T$.

**Assumption 5.** *For any history of observations and controls* $(y_{0:t}, u_{0:t-1})$ *at time t and a* $\varepsilon > 0$ *there exists a state* $x^* \in \mathcal{X}$ *such that* $1 - p(x^*|y_{0:t}, u_{0:t-1}, \theta) < \varepsilon$. *Additionally we assume that for some* $M > 0$

$$\left|\frac{\partial}{\partial\theta} p(x_i|y_{0:t}, u_{0:t-1}, \theta)\right| \leq M p(x_i|y_{0:t}, u_{0:t-1}, \theta) \text{ for all } i \neq *$$

Assumption 5 states that the filter $p(x_t|y_{0:t}, u_{0:t}, \theta)$ at time $t$ is close to having a point mass at some state $x^*$. Note that since $\sum_{x_t} \frac{\partial}{\partial\theta} p(x_t|y_{0:t}, u_{0:t-1}, \theta) = 0$ we also have $|\frac{\partial}{\partial\theta} p(x^*|y_{0:t}, u_{0:t-1}, \theta)| \leq M(1 - p(x^*|y_{0:t}, u_{0:t-1}, \theta))$ in Assumption 5.

We label the $t'$th element of WOFI as

$$W_t = E\left[\left(\frac{\partial}{\partial\theta} \log p(y_{t+1}|x_t, u_t, \theta)\right)^2\right]$$

and the $t'$th element of POFI is

$$P_t = E\left[\left(\frac{\partial}{\partial\theta} \log p(y_{t+1}|y_{0:t}, u_{0:t}, \theta)\right)^2\right]$$

where we drop the dependence on $x_0$ in notation.

Let $\mathcal{W} = \{(y_{t+1}, x_t) \in (\mathcal{Y}, \mathcal{X}) \text{ such that } p(y_{t+1}|x_t) > 0\}$ and let

$$v_{min} = \min\{p(y_{t+1}|x_t); (y_{t+1}, x_t) \in \mathcal{W}\}$$
$$u_{max} = \max\left\{\left|\frac{\partial}{\partial\theta} p(y_{t+1}|x_t)\right|; (y_{t+1}, x_t) \in \mathcal{W}\right\}$$

We now get

**Theorem 6.** *Assuming that Assumption 5 holds for the filter $p(x_t|y_{0:t}, u_{0:t-1}, \theta)$ at time $t$, we have*

$$|W_t - P_t| \leq 12L\left(\frac{u_{max}}{v_{min}}\right)^2 \frac{\varepsilon}{1-\varepsilon} + \frac{4LM^2}{v_{min}}\frac{\varepsilon^2}{1-\varepsilon} + \left(\frac{(2L+1)Mu_{max}}{v_{min}} + M^2\right)\varepsilon$$

*where $L$ is the dimension of $\mathcal{Y}$ and $M$ and $\varepsilon$ are from Assumption 5.*

*Proof.* Set

$$W_{t,cond} = E\left[\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|x_t, u_t, \theta)\right)^2 \middle| y_{0:t}, u_{0:t-1}\right]$$

and

$$P_{t,cond} = E\left[\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|y_{0:t}, u_{0:t}, \theta)\right)^2 \middle| y_{0:t}, u_{0:t-1}\right]$$

We have that

$$|W_t - P_t| = \left|E\left[W_{t,cond} - P_{t,cond}\right]\right| \leq E\left|W_{t,cond} - P_{t,cond}\right|$$

We now show that the bound holds for $\left|W_{t,cond} - P_{t,cond}\right|$ irrespective of the history $y_{0:t}, u_{0:t-1}$, which suffices to prove this Theorem. We suppress the control $u_t$ and the parameter $\theta$ in notation to save space. Let $x^*$ be as defined in Assumption 5 and set $p^* = p(x_t = x^*|y_{0:t})$ and and $p_i = p(x_t = x_i|y_{0:t})$, where $x_i \neq x^*$.

We set $u = u(y_{t+1}) = \frac{\partial}{\partial\theta}p(y_{t+1}|x_t = x^*)$, $v = v(y_{t+1}) = p(y_{t+1}|x_t = x^*)$, $u_0 = u_0(y_{t+1}) = E_{X_t}\left[\frac{\partial}{\partial\theta}p(y_{t+1}|x_t)\middle|y_{0:t}\right]$ and $v_0 = v_0(y_{t+1}) = E_{X_t}\left[p(y_{t+1}|x_t)|y_{0:t}\right]$. Also let $\tilde{\mathcal{Y}}(x) = \{y \in \mathcal{Y} : p(y_{t+1} = y|x_t = x) > 0\}$.

This allows us to write

$$W_{t,cond} = E\left[\left(\frac{\frac{\partial}{\partial\theta}p(y_{t+1}|x_t, u_t, \theta)}{p(y_{t+1}|x_t, u_t, \theta)}\right)^2 \middle| y_{0:t}\right] = E\left[\left(\frac{u}{v}\right)^2 \middle| y_{0:t}\right]$$

and

$$
\begin{aligned}
P_{t,cond} &= E\left[\left(\frac{E_{X_t}[\frac{\partial}{\partial\theta}p(y_{t+1}|x_t)|y_{0:t}] + \sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}]}\right)^2\middle| y_{0:t}\right] \\
&= E\left[\left(\frac{u_0}{v_0}\right)^2\middle| y_{0:t}\right] + E\left[\left(\frac{\sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}]}\right)^2\middle| y_{0:t}\right] \\
&+ 2E\left[\frac{u_0\sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{(E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}])^2}\middle| y_{0:t}\right]
\end{aligned}
$$

A bit of algebra gives

$$
\begin{aligned}
\left(\frac{u}{v}\right)^2 &= (u_0^2 + (u^2 - u_0^2))\frac{1}{v_0^2}\left(1 + \frac{v_0^2 - v^2}{v^2}\right) \\
&= \left(\frac{u_0}{v_0}\right)^2 + \frac{u^2 - u_0^2}{v_0^2} + \frac{u_0^2(v_0^2 - v^2)}{v_0^2 v^2} + \frac{(u^2 - u_0^2)(v_0^2 - v^2)}{v_0^2 v^2}
\end{aligned}
$$

and we get

$$
\begin{aligned}
W_{t,cond} &= E\left[\left(\frac{u}{v}\right)^2\middle| y_{0:t}\right] = E\left[\left(\frac{u}{v}\right)^2\middle| x_t = x^*\right]p^* + \sum_{i\neq *} E\left[\left(\frac{u}{v}\right)^2\middle| x_t = x_i\right]p_i \\
&= E\left[\left(\frac{u_0}{v_0}\right)^2\middle| x_t = x^*\right]p^* + E\left[\frac{u^2 - u_0^2}{v_0^2} + \frac{u_0^2(v_0^2 - v^2)}{v_0^2 v^2} + \frac{(u^2 - u_0^2)(v_0^2 - v^2)}{v_0^2 v^2}\middle| x_t = x^*\right]p^* \\
&+ \sum_{i\neq *} E\left[\left(\frac{u}{v}\right)^2\middle| x_t = x_i\right]p_i \\
&= P_{t,cond} + E\left[\frac{u^2 - u_0^2}{v_0^2} + \frac{u_0^2(v_0^2 - v^2)}{v_0^2 v^2} + \frac{(u^2 - u_0^2)(v_0^2 - v^2)}{v_0^2 v^2}\middle| x_t = x^*\right]p^* \\
&+ \sum_{i\neq *} E\left[\left(\frac{u}{v}\right)^2\middle| x_t = x_i\right]p_i - \sum_{i\neq *} E\left[\left(\frac{u_0}{v_0}\right)^2\middle| x_t = x_i\right]p_i \\
&- E\left[\left(\frac{\sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}]}\right)^2\middle| y_{0:t}\right] - 2E\left[\frac{u_0\sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{(E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}])^2}\middle| y_{0:t}\right]
\end{aligned}
$$

The superfluous expectations are bounded in Lemmas 5, 6, 7, 8, 9, 10 and 11 below. □

**Corollary 2.** *Assuming that Assumption 5 holds for the filter $p(x_t|y_{0:t}, u_{0:t-1}, \theta)$ at times*

$t = 1, \ldots, T$ *we have*

$$\left| E\left[\sum_{t=0}^{T-1}\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|x_t, u_t, \theta)\right)^2\right] - E\left[\sum_{t=0}^{T-1}\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|y_{0:t}, u_{0:t}, x_0, \theta)\right)^2\right]\right|$$
$$\leq T\left(12L\left(\frac{u_{max}}{v_{min}}\right)^2\frac{\varepsilon}{1-\varepsilon} + \frac{4LM^2}{v_{min}}\frac{\varepsilon^2}{1-\varepsilon} + \left(\frac{(2L+1)Mu_{max}}{v_{min}} + M^2\right)\varepsilon\right)$$

**Lemma 4.** *Under Assumption 5*

$$|v(y_{t+1}, x^*) - v_0(y_{t+1})| \leq 1 - p^*$$
$$|u(y_{t+1}, x^*) - u_0(y_{t+1})| \leq 2u_{max}(1 - p^*)$$

*Proof.*

$$\begin{aligned}|v(y_{t+1}, x^*) - v_0(y_{t+1})| &= \left|p(y_{t+1}|x^*)(1-p^*) - \sum_{i\neq *} p(y_{t+1}|x_i)p_i\right| \\ &\leq \max\left(p(y_{t+1}|x^*)(1-p^*), \sum_{i\neq *} p(y_{t+1}|x_i)p_i\right) \\ &\leq \max\left((1-p^*), \sum_{i\neq *} p_i\right) = 1 - p^*\end{aligned}$$

The $u$ case;

$$|u(y_{t+1}, x^*) - u_0(y_{t+1})| \leq \left|\frac{\partial}{\partial\theta}p(y_{t+1}|x^*)\right|(1-p^*) + \sum_{i\neq *}\left|\frac{\partial}{\partial\theta}p(y_{t+1}|x_i)\right|p_i \leq 2u_{max}(1-p^*)$$

□

**Lemma 5.** *Under Assumption 5*

$$\left|E\left[\frac{u^2 - u_0^2}{v_0^2}\middle| x_t = x^*\right]\right| p^* \leq \frac{4u_{max}^2(1-p^*)}{p^*}\sum_{y_{t+1}\in\tilde{\mathcal{Y}}}\frac{1}{p(y_{t+1}|x^*)}$$

*Proof.*

$$\left|E\left[\frac{u^2 - u_0^2}{v_0^2}\middle| x_t = x^*\right]\right| \leq E\left[\frac{|u + u_0||u - u_0|}{v_0^2}\middle| x_t = x^*\right] \leq 2u_{max}E\left[\frac{|u - u_0|}{v_0^2}\middle| x_t = x^*\right]$$

Now only summing over $y_{t+1} \in \tilde{\mathcal{Y}}(x^*)$

$$
\begin{aligned}
E\left[\frac{|u-u_0|}{v_0^2}\middle| x_t = x^*\right] &= \sum_{y_{t+1}\in\tilde{\mathcal{Y}}} \frac{\left|\frac{\partial}{\partial\theta}p(y_{t+1}|x^*)(1-p^*) - \sum_{i\neq *}\frac{\partial}{\partial\theta}p(y_{t+1}|x_i)p_i\right|}{\left(p(y_{t+1}|x^*)p^* + \sum_{i\neq *}p(y_{t+1}|x_i)p_i\right)^2} p(y_{t+1}|x^*) \\
&\leq \sum_{y_{t+1}\in\tilde{\mathcal{Y}}} \frac{2u_{max}(1-p^*)}{\left(p(y_{t+1}|x^*)p^*\right)^2} p(y_{t+1}|x^*) \\
&= \frac{2u_{max}(1-p^*)}{(p^*)^2} \sum_{y_{t+1}\in\tilde{\mathcal{Y}}} \frac{1}{p(y_{t+1}|x^*)}
\end{aligned}
$$

□

**Lemma 6.** *Under Assumption 5*

$$
\left|E\left[\frac{u_0^2(v_0^2-v^2)}{v_0^2v^2}\middle| x_t = x^*\right]\right| p^* \leq \frac{2u_{max}^2(1-p^*)}{p^*} \sum_{y_{t+1}\in\tilde{\mathcal{Y}}} \frac{1}{p(y_{t+1}|x^*)^2}
$$

*Proof.*

$$
\begin{aligned}
\left|E\left[\frac{u_0^2(v_0^2-v^2)}{v_0^2v^2}\middle| x_t = x^*\right]\right| &\leq E\left[\frac{u_0^2(v_0+v)|v_0-v|}{v_0^2v^2}\middle| x_t = x^*\right] \\
&\leq u_{max}^2 E\left[\left(\frac{1}{v_0v^2} + \frac{1}{v_0^2v}\right)|v_0-v|\middle| x_t = x^*\right] \\
&\leq \frac{2u_{max}^2}{(p^*)^2} E\left[\frac{|v_0-v|}{v^3}\middle| x_t = x^*\right]
\end{aligned}
$$

since $v_0(y_{t+1}) = p(y_{t+1}|x^*)p^* + \sum_{i\neq *}p(y_{t+1}|x_i)p_i \geq p(y_{t+1}|x^*)p^* = v(y_{t+1},x^*)p^*$. Only summing over $y_{t+1} \in \tilde{\mathcal{Y}}(x^*)$ we get

$$
E\left[\frac{|v_0-v|}{v^3}\middle| x_t = x^*\right] = \sum_{y_{t+1}\in\tilde{\mathcal{Y}}} \frac{|E_{X_t}[p(y_{t+1}|x_t)] - p(y_{t+1}|x^*)|}{p(y_{t+1}|x^*)^3} p(y_{t+1}|x^*) \leq \sum_{y_{t+1}\in\tilde{\mathcal{Y}}} \frac{1-p^*}{p(y_{t+1}|x^*)^2}
$$

□

**Lemma 7.** *Under Assumption 5*

$$
\left|E\left[\frac{(u^2-u_0^2)(v_0^2-v^2)}{v_0^2v^2}\middle| x_t = x^*\right]\right| p^* \leq \frac{4u_{max}^2(1-p^*)^2}{p^*} \sum_{y_{t+1}\in\tilde{\mathcal{Y}}} \frac{1}{p(y_{t+1}|x^*)^2}
$$

*Proof.*

$$\begin{aligned}\left|E\left[\frac{(u^2-u_0^2)(v_0^2-v^2)}{v_0^2v^2}\middle|x_t=x^*\right]\right| &\le E\left[(u+u_0)\left(\frac{1}{v_0^2v}+\frac{1}{v^2v_0}\right)|u-u_0||v-v_0|\middle|x_t=x^*\right]\\ &\le \frac{2u_{max}}{(p^*)^2}E\left[\frac{|u-u_0||v-v_0|}{v^3}\middle|x_t=x^*\right]\\ &\le \frac{2u_{max}}{(p^*)^2}\sum_{y_{t+1}\in\tilde{\mathcal{Y}}}\frac{2u_{max}(1-p^*)^2}{p(y_{t+1}|x^*)^2}\\ &= \frac{4u_{max}^2(1-p^*)^2}{(p^*)^2}\sum_{y_{t+1}\in\tilde{\mathcal{Y}}}\frac{1}{p(y_{t+1}|x^*)^2}\end{aligned}$$

□

**Lemma 8.** *Under Assumption 5*

$$\sum_{i\neq *}E\left[\left(\frac{u}{v}\right)^2\middle|x_t=x_i\right]p_i\le\frac{Lu_{max}^2}{v_{min}}(1-p^*)$$

*where L is the dimension of* $\mathcal{Y}$.

*Proof.* Only summing over $y_{t+1}\in\tilde{\mathcal{Y}}(x_i)$ for each $i$, we get

$$\begin{aligned}\sum_{i\neq *}E\left[\left(\frac{u}{v}\right)^2\middle|x_t=x_i\right]p_i &= \sum_{i\neq *}\sum_{y_{t+1}\in\tilde{\mathcal{Y}}}\left(\frac{\frac{\partial}{\partial\theta}p(y_{t+1}|x_i)}{p(y_{t+1}|x_i)}\right)^2p(y_{t+1}|x_i)p_i\\ &= \sum_{i\neq *}\sum_{y_{t+1}\in\tilde{\mathcal{Y}}}\frac{\left(\frac{\partial}{\partial\theta}p(y_{t+1}|x_i)\right)^2}{p(y_{t+1}|x_i)}p_i\\ &\le \frac{Lu_{max}^2}{v_{min}}(1-p^*)\end{aligned}$$

□

**Lemma 9.** *Under Assumption 5*

$$\sum_{i\neq *}E\left[\left(\frac{u_0}{v_0}\right)^2\middle|x_t=x_i\right]p_i\le\frac{u_{max}^2(1-p^*)}{v_{min}^2}$$

*Proof.* We assume that if $p(y_{t+1}|x_t) = 0$ then $\frac{\partial}{\partial\theta}p(y_{t+1}|x_t) = 0$ as well. This gives us that

$$\left(\frac{\frac{\partial}{\partial\theta}p(y_{t+1}|x^*)p^* + \sum_{j\neq *}\frac{\partial}{\partial\theta}p(y_{t+1}|x_j)p_j}{p(y_{t+1}|x^*)p^* + \sum_{j\neq *}p(y_{t+1}|x_j)p_j}\right)^2 \leq \left(\frac{u_{max}}{v_{min}}\right)^2$$

This gives us that

$$\begin{aligned}\sum_{i\neq *}E\left[\left(\frac{u_0}{v_0}\right)^2\middle| x_t = x_i\right]p_i &= \sum_{i\neq *}\sum_{y_{t+1}\in\tilde{\mathcal{Y}}}\left(\frac{\frac{\partial}{\partial\theta}p(y_{t+1}|x^*)p^* + \sum_{j\neq *}\frac{\partial}{\partial\theta}p(y_{t+1}|x_j)p_j}{p(y_{t+1}|x^*)p^* + \sum_{j\neq *}p(y_{t+1}|x_j)p_j}\right)^2 p(y_{t+1}|x_i)p_i \\ &\leq \left(\frac{u_{max}}{v_{min}}\right)^2\sum_{i\neq *}\sum_{y_{t+1}\in\tilde{\mathcal{Y}}}p(y_{t+1}|x_i)p_i \\ &= \left(\frac{u_{max}}{v_{min}}\right)^2(1-p^*)\end{aligned}$$

□

**Lemma 10.** *Under Assumption 5*

$$E\left[\left(\frac{\sum_{x_t}p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}]}\right)^2\middle| y_{0:t}\right] \leq \frac{4LM^2(1-p^*)^2}{v_{min}p^*} + M^2(1-p^*)$$

.

*Proof.*

$$\left(\frac{\sum_{x_t}p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}]}\right)^2 \leq \left(\frac{p(y_{t+1}|x^*)M(1-p^*) + \sum_{i\neq *}p(y_{t+1}|x_i)Mp_i}{p(y_{t+1}|x^*)p^* + \sum_{i\neq *}p(y_{t+1}|x_i)p_i}\right)^2 \leq M^2$$

If $p(y_{t+1}|x^*) > 0$ we also have

$$\begin{aligned}\left(\frac{p(y_{t+1}|x^*)M(1-p^*) + \sum_{i\neq *}p(y_{t+1}|x_i)Mp_i}{p(y_{t+1}|x^*)p^* + \sum_{i\neq *}p(y_{t+1}|x_i)p_i}\right)^2 &\leq M^2\left(\frac{(1-p^*) + \sum_{i\neq *}p_i}{p(y_{t+1}|x^*)p^*}\right)^2 \\ &= 4M^2\left(\frac{(1-p^*)}{p(y_{t+1}|x^*)p^*}\right)^2\end{aligned}$$

We now have

$$
\begin{aligned}
& E\left[\left(\frac{\sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}]}\right)^2\middle| y_{0:t}\right] \\
&\leq \sum_{y_{t+1}\in\tilde{\mathcal{Y}}} 4M^2\left(\frac{(1-p^*)}{p(y_{t+1}|x^*)p^*}\right)^2 p(y_{t+1}|x^*)p^* + \sum_{i\neq *}\sum_{y_{t+1}\in\tilde{\mathcal{Y}}} M^2 p(y_{t+1}|x_i)p_i \\
&\leq \frac{4LM^2(1-p^*)^2}{v_{min}p^*} + M^2(1-p^*)
\end{aligned}
$$

□

**Lemma 11.** *Under Assumption 5*

$$
E\left[\frac{u_0\sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{(E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}])^2}\middle| y_{0:t}\right] \leq \frac{u_{max}}{v_{min}}(2L+1)M(1-p^*)
$$

.

*Proof.* Using $\left|\frac{u_0}{v_0}\right| \leq \frac{u_{max}}{v_{min}}$ from Lemma 9 we get

$$
E\left[\frac{u_0\sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{(E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}])^2}\middle| y_{0:t}\right] \leq \frac{u_{max}}{v_{min}} E\left[\frac{\sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}]}\middle| y_{0:t}\right]
$$

and like in Lemma 10 we get

$$
\begin{aligned}
& E\left[\frac{\sum_{x_t} p(y_{t+1}|x_t)\frac{\partial}{\partial\theta}p(x_t|y_{0:t})}{E_{X_t}[p(y_{t+1}|x_t)|y_{0:t}]}\middle| y_{0:t}\right] \\
&\leq \sum_{y_{t+1}\in\tilde{\mathcal{Y}}} 2M\frac{(1-p^*)}{p(y_{t+1}|x^*)p^*}p(y_{t+1}|x^*)p^* + \sum_{i\neq *}\sum_{y_{t+1}\in\tilde{\mathcal{Y}}} Mp(y_{t+1}|x_i)p_i \\
&\leq 2LM(1-p^*) + M(1-p^*) = (2L+1)M(1-p^*)
\end{aligned}
$$

□

## 3.9 Best POFI convergence theorem

Remembering that POFI is the true Fisher Information of our data, which we can influence by the choice of our control policy, we are interested in how well controls that arise from an approximated POFI criteria maximize the original POFI criteria, compared with a theoretical best policy. This is also discussed in Section 4.4.

We assume that the Fisher Information to Go for POFI;

$$FI_k = E\left[\sum_{t=k}^{T-1}\left(\dot{h}_{t+1,0,v_0}\right)^2\right]$$

is approximated by

$$\widetilde{FI}_k = E\left[\sum_{t=k}^{T-1}\left(\dot{h}_{t+1,t-m,v_{k-m}}\right)^2\right] \text{ or } \widetilde{FI}_k = E\left[\sum_{t=k}^{T-1}\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|x_t,u_t,\theta)\right)^2\right]$$

that is either the truncated POFI or WOFI.

Given that our controls are obtained by dynamic programming, we have that the optimal control at time $t$ is dependent on the optimal control obtained at time $t+1$. Let $u_1^*,\ldots,u_{T-1}^*$ denote the set of optimal controls obtained in this manner, i.e. $u_k^*,\ldots,u_{T-1}^*$ maximize $FI_k$ and let $u_{0,m}^*,\ldots,u_{T-1,m}^*$ denote the approximated control policy where $u_{k,m}^*,\ldots,u_{T-1,m}^*$ maximize the approximated criteria $\widetilde{FI}_k$.

The following theorem quantifies the loss in Fisher Information from using approximate controls instead of exact ones, in an experiment of length $T$.

**Theorem 7.** *Given that the mixing conditions in Assumption 4 hold and we calculate a control policy $u_{0,m}^*,\ldots,u_{T-1,m}^*$ using the truncated POFI, we have*

$$0 \leq FI(u_0^*,\ldots,u_{T-1}^*) - FI(u_{0,m}^*,\ldots,u_{T-1,m}^*) \leq c_2 T(T+1)\rho^{m/2}$$

*where* $c_2 = 8M(\theta)\sup|\phi_\theta(x,x',y,u)|\frac{1}{\rho^{1/2}(1-\rho)}$*, and* $M(\theta)$ *is the bound from lemma 2.*

*Alternatively, given that the filter assumptions in Assumption 5 hold for every* $t = 1,\dots,T$*, and we calculate a control policy* $u^*_{0,m},\dots,u^*_{T-1,m}$ *using WOFI we have*

$$\begin{aligned}0 &\le FI(u^*_0,\dots,u^*_{T-1}) - FI(u^*_{0,m},\dots,u^*_{T-1,m})\\ &\le T(T+1)\left(12L\left(\frac{u_{max}}{v_{min}}\right)^2\frac{\varepsilon}{1-\varepsilon} + \frac{4LM^2}{v_{min}}\frac{\varepsilon^2}{1-\varepsilon} + \left(\frac{(2L+1)Mu_{max}}{v_{min}} + M^2\right)\varepsilon\right)\end{aligned}$$

*where the constants are given in Section 3.8*

*Proof.* We analyze the difference by bounding errors in each step of the dynamic program inductively, starting at time $t = T-1$ and going backwards. If we use truncated POFI to calculate a control policy, we set

$$\gamma = 8M(\theta)\sup|\phi_\theta(x,x',y,u)|\frac{\rho^{(m+1)/2-1}}{1-\rho}$$

, see Lemma 3, while if we use WOFI then we set

$$\gamma = \left(\left(\frac{u_{max}}{v_{min}}\right)^2\frac{\varepsilon}{1-\varepsilon} + \frac{4LM^2}{v_{min}}\frac{\varepsilon^2}{1-\varepsilon} + \left(\frac{(2L+1)Mu_{max}}{v_{min}} + M^2\right)\varepsilon\right)$$

and refer to Theorem 6.

We find that

$$\begin{aligned}0 &\le FI_{T-1}(u^*_{T-1}) - FI_{T-1}(u^*_{T-1,m})\\ &\le FI_{T-1}(u^*_{T-1}) - FI_{T-1}(u^*_{T-1,m}) + (\widetilde{FI}_{T-1}(u^*_{T-1,m}) - \widetilde{FI}_{T-1}(u^*_{T-1}))\end{aligned}$$

so far only using that $u^*_{T-1}$ maximizes $FI_{T-1}$ and $u^*_{T-1,m}$ maximizes $\widetilde{FI}_{T-1}$.

$$\begin{aligned}&\le \left|FI_{T-1}(u^*_{T-1}) - \widetilde{FI}_{T-1}(u^*_{T-1})\right| + \left|FI_{T-1}(u^*_{T-1,m}) - \widetilde{FI}_{T-1}(u^*_{T-1,m})\right|\\ &\le 2\gamma\end{aligned}$$

by either lemma 3 or Theorem 6.

We now inductively assume

$$\left|FI_{T-s}(u^*_{T-s:T-1}) - FI_{T-s}(u^*_{T-s:T-1,m})\right| \leq s(s+1)\gamma$$

where $u^*_{T-s:T-1,m} = u^*_{T-s,m}, \ldots, u^*_{T-1,m}$ and then get

$$\begin{aligned}
&\left|FI_{T-s}(u^*_{T-s:T-1}) - \widetilde{FI}_{T-s}(u^*_{T-s:T-1,m})\right| \\
&\leq \left|FI_{T-s}(u^*_{T-s:T-1}) - FI_{T-s}(u^*_{T-s:T-1,m})\right| + \left|FI_{T-s}(u^*_{T-s:T-1,m}) - \widetilde{FI}_{T-s}(u^*_{T-s:T-1,m})\right| \\
&\leq s(s+1)\gamma + s\gamma = s(s+2)\gamma \qquad (1)
\end{aligned}$$

Now moving from $s$ to $s+1$ we have

$$\widetilde{FI}_{T-(s+1)}(u^*_{T-(s+1):T-1,m}) \geq \widetilde{FI}_{T-(s+1)}(u^*_{T-(s+1)}, u^*_{T-s:T-1,m})$$

since $u^*_{T-(s+1):T-1,m}$ are the controls that maximize $\widetilde{FI}_{T-(s+1)}$. By adding and subtracting the same quantity we get the following equivalent inequality

$$\left(\widetilde{FI}_{T-(s+1)}(u^*_{T-(s+1):T-1,m}) - FI_{T-(s+1)}(u^*_{T-(s+1):T-1,m})\right) \qquad (2)$$

$$-\left(\widetilde{FI}_{T-(s+1)}(u^*_{T-(s+1)}, u^*_{T-s:T-1,m}) - FI_{T-(s+1)}(u^*_{T-(s+1):T-1})\right) \qquad (3)$$

$$\geq FI_{T-(s+1)}(u^*_{T-(s+1):T-1}) - FI_{T-(s+1)}(u^*_{T-(s+1):T-1,m}) \geq 0$$

Line (2) is bounded by $(s+1)\gamma$ by lemma 3/Theorem 6 and line (3) by $\gamma + s(s+2)\gamma$ using (1) and lemma 3/Theorem 6. Therefore

$$\begin{aligned}
&\left|FI_{T-(s+1)}(u^*_{T-(s+1):T-1}) - FI_{T-(s+1)}(u^*_{T-(s+1):T-1,m})\right| \\
&\leq (s+1)\gamma + \gamma + s(s+2)\gamma = (s+1)(s+2)\gamma
\end{aligned}$$

and for the whole experiment we find

$$\left|FI(u^*_{0:T-1}) - FI(u^*_{0:T-1,m})\right| \leq T(T+1)\gamma$$

□

CHAPTER 4

# CONTROL THEORETIC ALGORITHMS APPLIED TO FISHER INFORMATION PROBLEMS

In this chapter we show how the various forms of Fisher Information considered, FOFI, truncated POFI and WOFI, can be maximized using dynamic programming. We provide pseudocodes and analyze their computational complexities. The computations required can be split into computations done prior to the experiment and computations that are required while running the experiment. A direct comparison is not completely feasible since FOFI and WOFI require computations at runtime while truncated POFI does not as discussed below.

## 4.1 FOFI Dynamic Program

Hooker et al. [5] considered constructing an optimal control policy for the Fisher Information that would apply if $(X_t)$ were observed directly, that is the Full Observation Fisher Information (FOFI)

$$FI = E \sum_{t=0}^{T-1} \left( \frac{\partial}{\partial \theta} \log p(x_{t+1}|x_t, u_t, \theta) \right)^2$$

When considering continuous time stochastic systems, the state space is continuous, but we use this Fisher Information as an approximation to the continuous state Fisher Information. An advantage of using FOFI is that when running the dynamic program the Markov property of the Markov Decision Process $(X_t, u_t)$ allows us to only consider a maximization over the state space $x_t \in \mathcal{X}$ but not past values $x_{0:t-1}$.

However, maximizing FOFI can lead to suboptimal controls since it is not the correct Fisher Information for the data. Additionally, when the actual experiment is run we do not observe $X_t$. Instead we have to use the observed values to get a probability distribution (a filter) on the state $x_t$, $p(x_t|y_{0:t}, u_{0:t-1}, x_0, \theta)$ and use the control associated with the state that has the highest probability.

### 4.1.1 Pseudocode for FOFI and computational complexity

We set the reward function as $C(x_t, u_t, \theta) = \left(\frac{\partial}{\partial\theta} \log p(x_{t+1}|x_t, u_t, \theta)\right)^2$. The pseudocode for this dynamic program is:

$FI_T = 0$

**for** $t = (T-1) \to 0$ **do**

$\forall\ x_t$ and calculate and store

$FI_t(x_t, \theta) = \max_{u_t} \{E_{x_{t+1}}[C(x_t, u_t, \theta) + FI_{t+1}(x_{t+1}, \theta)|x_t, u_t, \theta]\}$

$u_t^*(x_t, \theta) = \text{argmax}_{u_t} \{E_{x_{t+1}}[C(x_t, u_t, \theta) + FI_{t+1}(x_{t+1}, \theta)|x_t, u_t, \theta]\}$

**end for**

We assume that the transition probability matrix $p(x_{t+1}|x_t, u_t, \theta)$ is given. Calculating $\left(\frac{\partial}{\partial\theta} \log p(x_{t+1}|x_t, u_t, \theta)\right)^2$ is negligible compared to the calculations required for the dynamic program; If we set

$$g_t(x_t, x_{t+1}, u_t, \theta) = \left(\frac{\partial}{\partial\theta} \log p(x_{t+1}|x_t, u_t, \theta)\right)^2$$

then for a given time $t$ in the dynamic program we need to maximize

$$E\left[g_t(x_t, x_{t+1}, u_t, \theta) + V_{t+1}(x_{t+1}, \theta)|x_t\right]$$

over $u_t \in \mathcal{U}$ for each $x_t \in \mathcal{X}$, where $V_{t+1}$ is the value function from the previous step $t+1$. This calculation requires adding $g_t$ and $V_{t+1}$ which are two $K^{\times 2} \times l$ tensors with cost $K^2 l$. Next we need a dot product between $g_t + V_{t+1}$ and $p(x_{t+1}|x_t, u_t)$ over the $x_{t+1}$ dimension which has cost $O(K^2 l)$. Finally maximizing over $u_t$ for each $x_t$ has cost $O(Kl)$. Thus each step $t$ has cost $O(K^2 l)$ and the dynamic program in total has cost $O(TK^2 l)$

In runtime a filter is required to estimate the state $x_t$. The filter for time $t+1$ can be calculated via the following recursive formula

$$p(x_{t+1}|y_{0:t+1}, u_{0:t}) \propto \sum_{x_t} p(y_{t+1}|x_{t+1}) p(x_{t+1}|x_t, u_t) p(x_t|y_{0:t}, u_{0:t-1})$$

and then normalizing. This requires $2K$ dot products of vectors of length $K$, with cost $O(K^2)$ and the normalization has cost $O(K)$. Thus we have $O(K^2)$ computations at each time step $t$ during runtime.

## 4.2 Truncated POFI dynamic program

The most natural Fisher Information to maximize is the Fisher Information of our observed process, the Partial Observation Fisher Information (POFI) which we can express as

$$FI(\theta) = E\left[\sum_{t=0}^{T-1} \left(\frac{\partial}{\partial \theta} \log p(y_{t+1}|y_{0:t}, u_{0:t}, x_0, \theta)\right)^2\right]$$

also see Sections 3.3, 3.4 and 3.5.

To maximize POFI with a dynamic program we set

$$C_t(y_{0:t}, u_{0:t}, \theta) = \left(\frac{\partial}{\partial \theta} \log p(y_{t+1}|y_{0:t}, u_{0:t}, x_0, \theta)\right)^2$$

and we try to maximize the total reward $FI(\theta) = E[\sum_t C_t(y_{0:t}, u_{0:t}, \theta)]$. Note that in this instance the reward function depends on the entire history of observations and controls up to time $t$.

The Value function in the corresponding dynamic program is

$$FI_t(y_{0:t}, u_{0:t-1}, \theta) = \max_{u_t} \left\{ E_{y_{t+1}}[C_t(y_{0:t}, u_{0:t}, \theta) + FI_{t+1}(y_{0:t+1}, u_{0:t}, \theta) | y_{0:t}, u_{0:t}, \theta] \right\}$$

and we denote it the Fisher Information to Go .

A problem here is that just in the first step of the dynamic program ($t = T-1$) we would have to calculate the Fisher Information to Go for $L^{T-1}l^{T-2}$ many combinations of $y_{0:t}$ and $u_{0:t-1}$. This is formidable for even modest dimensions. We therefore approximate the process by conditioning only on the last $m+1$ observations in the Fisher Information;

$$FI_{trunc} = E \sum_{t=0}^{T-1} \left( \frac{\partial}{\partial \theta} \log p(y_{t+1} | y_{t-m:t}, u_{t-m:t}, \nu_{t-m}, \theta) \right)^2$$

where $\nu_{t-m}$ is some prior that we assume for $x_{t-m}$, although we generally suppress it in notation since we assume it is fixed. As before, if $t - m < 0$ we set $t - m : t$ to mean $0 : t$ to ease notation.

### 4.2.1 Pseudocode for POFI

The reward becomes $C(y_{t-m:t}, u_{t-m:t}, \theta) = \left( \frac{\partial}{\partial \theta} \log p(y_{t+1} | y_{t-m:t}, u_{t-m:t}, \theta) \right)^2$ and

$$FI_{t,m}(y_{t-m:t}, u_{t-m:t-1}, \theta) = \max_{u_t} \left\{ E_{y_{t+1}}[C + FI_{t+1,m} | y_{t-m:t}, u_{t-m:t}, \theta)] \right\}$$

the Fisher Information To Go. The pseudocode for the corresponding dynamic program is:

$FI_{T,m} = 0$

**for** $t = (T-1) \rightarrow 0$ **do**

$\forall\ y_{t-m:t}, u_{t-m:t-1}$ and calculate and store

$FI_{t,m}(y_{t-m:t}, u_{t-m:t-1}, \theta) = \max_{u_t} \left\{ E_{y_{t+1}}[C + FI_{t+1,m} | y_{t-m:t}, u_{t-m:t}, \theta] \right\}$

$u_t^*(y_{t-m:t}, u_{t-m:t-1}, \theta) = \text{argmax}_{u_t} \left\{ E_{y_{t+1}}[C + FI_{t+1,m} | y_{t-m:t}, u_{t-m:t}, \theta] \right\}$

**end for**

For this approximate dynamic program to be sensible we want the truncated Fisher Information to approach the true Fisher Information as $m$ increases. In Theorem 5 we show that $|FI - FI_{trunc}| \leq c_1(T-1-m)\rho^{m/2}$ where the constants $c_1$ and $\rho$ do not depend on $m$ or $T$, assuming certain technical mixing conditions which we have stated in detail in Assumption 4.

Theorem 5 states that $FI_{trunc}$ approaches the true Fisher Information exponentially as $m$ increases, and is thus a viable approximation for the Fisher Information in the dynamic program.

The runtime of the dynamic program however also grows exponentially in $m$ and we found that while setting $m = 0$, i.e. conditioning on one observation, gave poor results in some of our simulations, conditioning on two observations, i.e. $m = 1$, generally gave good results when compared to other control policies. Setting $m = 2$ increased runtime greatly and was in some applications infeasible without making more approximations to how the dynamic program is run. The exact effect of increasing $m$ is quite problem specific.

### 4.2.2 Truncated POFI, computational complexity

Here the dynamic program maximizes the truncated Partial observation Fisher Information,

$$FI_{trunc} = E\sum_{t=0}^{T-1}\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|y_{t-m:t}, u_{t-m:t}, \nu_{t-m}, \theta)\right)^2$$

We note that $p(y_{t+1}|y_{t-m:t}, u_{t-m:t}, \theta)$ is a $L^{\times m+2} \times l^{\times m+1}$ tensor, and it can be calculated using Bayes rule at the cost $O(K^2L^{m+2}l^{m+1})$. Calculating $\frac{\partial}{\partial\theta}\log p(y_{t+1}|y_{t-m:t}, u_{t-m:t}, \theta)$ can also be done at the cost $O(K^2L^{m+2}l^{m+1})$, but can also be effectively approximated using the finite difference approximation to the derivative.

The cost analysis of the truncated POFI dynamic program is just like the analysis of FOFI. At a given time $t$ adding $g_t$ and $V_{t+1}$ has cost $O(L^{m+2}l^{m+1})$, the dot product between $g_t + V_{t+1}$ and $p(y_{t+1}|y_{t-m:t}, u_{t-m:t})$ has cost $O(L^{m+2}l^{m+1})$ and the maximization has cost $O(L^{m+1}l^{m+1})$.

The dynamic program thus has cost $O(TL^{m+2}l^{m+1})$, which in some cases could be constrained by choosing $L$ lower than $K$, and $m = 1$.

## 4.3 WOFI dynamic program

The Weighted Observation Fisher Information

$$FI = E\left[\sum_{t=0}^{T-1}\left(\frac{\partial}{\partial\theta}\log p(y_{t+1}|x_t, u_t, \theta)\right)^2\right]$$

is motivated as an approximation to POFI, while preserving the Markov property of FOFI, also see Sections 3.7 and 3.8. In Theorem 6 we show how WOFI approximates POFI, given that the filter $p(x_t|y_{0:t}, u_{0:t-1}, \theta)$ is precise, see Assumption 5.

We set the reward function to $C_t(x_t, u_t, y_{t+1}, \theta) = \left(\frac{\partial}{\partial\theta} \log p(y_{t+1}|x_t, u_t, \theta)\right)^2$ in order to run a dynamic program with WOFI which will be similar to the FOFI dynamic program. It has the same computation cost as the one for FOFI, that is $O(TK^2l)$, and also needs a filter to estimate the state $x_t$ at time $t$, at the same cost $O(K^2)$.

## 4.4 Parameter estimation

After running an experiment, using one of the control policies, the parameter $\theta$ is estimated either via an EM algorithm or by directly maximizing the loglikelihood, see Sections 4.4.1 and 4.4.2. For the asymptotic properties of the MLE we refer to Cappe et al. [1] as well, where conditions for consistency and asymptotic normality in Hidden Markov Models are given. The central elements of their proof are the stationarity of the process $(X_t, Y_t)$ along with forgetting properties of the filter (see Theorem 2 in Section 2.2.5). We note that if we employ a time-independent control policy (as we do in in Chapter 8), we obtain to a Hidden Markov Model and can rely on [1] if we assume stationarity. That the forgetting properties of Hidden Markov Models can be extended to POMDP's points to a more general asymptotic theory for the MLE in POMDP's, but this is not pursued further.

Theorem 5 shows that using the truncated POFI is a good approximation to the Partial Observation Fisher Information for running a dynamic program and Theorem 6 shows the same for WOFI, under respective assumptions. This provides a control policy that is an approximation to the optimal control policy. Now consider using this approximate policy to run an experiment and then

estimating $\theta$ by evaluating the MLE. The asymptotic variance of this MLE will be the inverse of the Partial Observation Fisher Information, with controls from the approximate policy. It is therefore of interest to compare POFI, evaluated with an optimal policy, and POFI, evaluated with one of the approximate policies. In Section 3.9, Theorem 7 we show that, conditional on the mixing conditions in Assumption 4, that

$$0 \leq FI(u_0^*, \ldots, u_{T-1}^*) - FI(u_{0,m}^*, \ldots, u_{T-1,m}^*) \leq c_2 T(T+1)\rho^{m/2}$$

where $u_0^*, \ldots, u_{T-1}^*$ are the optimal controls, $u_{0,m}^*, \ldots, u_{T-1,m}^*$ the truncated POFI optimal controls and FI is the Partial Observation Fisher Information. The constants $c_2$ and $\rho$ do not depend on $m$ or $T$. Alternatively, conditional on the filter assumptions in Assumption 5, we have that

$$\begin{aligned} 0 &\leq FI(u_0^*, \ldots, u_{T-1}^*) - FI(u_{0,m}^*, \ldots, u_{T-1,m}^*) \\ &\leq T(T+1)\left(12L\left(\frac{u_{max}}{v_{min}}\right)^2 \frac{\varepsilon}{1-\varepsilon} + \frac{4LM^2}{v_{min}}\frac{\varepsilon^2}{1-\varepsilon} + \left(\frac{(2L+1)Mu_{max}}{v_{min}} + M^2\right)\varepsilon\right) \end{aligned}$$

where $u_{0,m}^*, \ldots, u_{T-1,m}^*$ are the WOFI optimal controls and FI is the Partial Observation Fisher Information. The constants are given in section 3.8.

That the asymptotic variance of the MLE converges to the lowest possible variance, as either $m \to \infty$ or $\varepsilon \to 0$, further supports our approximations.

### 4.4.1 EM algorithm

Given the data $y_0, \ldots, y_T$ one can estimate the parameter $\theta$ with the EM algorithm. This is well documented in the Hidden Markov Models literature, see [1] for example, so we will only describe it briefly.

The forward variable is defined as $\alpha_t(x) = P(x_t = x, y_{0:t}, u_{0:t-1}|\theta)$, and the backward variable as $\beta_{t|T}(x) = P(y_{t+1:T}, u_{t:T-1}|x_t = x, \theta)$. See section 2.2.2 for details on how they are calculated. We set

$$\gamma_{t|T}(x) = P(x_t = x|y_{0:T}, u_{0:T-1}, \theta) = \frac{\alpha_t(x)\beta_{t|T}(x)}{\sum_{x\in\mathcal{X}} \alpha_t(x)\beta_{t|T}(x)}$$

and

$$\begin{aligned}\xi_t(x_1, x_2) &= P(x_t = x_1, x_{t+1} = x_2|y_{0:T}, u_{0:T-1}, \theta) \\ &= \frac{\alpha_t(x_1)p(x_{t+1} = x_2|x_t = x_1, u_t, \theta)p(y_{t+1}|x_{t+1} = x_2)\beta_{t+1|T}(x_2)}{\sum_{x\in\mathcal{X}} \alpha_t(x)\beta_{t|T}(x)}\end{aligned}$$

The complete data log-likelihood is

$$l_{comp}(\theta) = \sum_{t=0}^{T-1} \log p(x_{t+1}|x_t, u_t, \theta) + \sum_{t=0}^{T} \log p(y_t|x_t, \theta)$$

With $\gamma_{t|T}$ and $\xi_t$ for a fixed $\theta^*$ we can now define the function $Q(\theta|\theta^*)$, which performs the EM expectation step

$$Q(\theta|\theta^*) = E\left[l_{comp}(\theta)\middle|y_{0:T}, u_{0:T-1}, \theta^*\right]$$

Maximizing the function $Q$ over $\theta$ gives an update to $\theta^*$. Alternating between expectation and maximization is the EM algorithm, and $\theta^*$ converges to the MLE, see [1] again. Convergence of the EM algorithm is discussed in Cappe et al. [1] for Hidden Markov Models and extends naturally to POMDP's.

### 4.4.2 Direct Maximum Likelihood

In some cases it is possible to directly maximize the log-likelihood

$$l(\theta) = \sum_{t=0}^{T-1} \log p(y_{t+1}|y_{0:t}, u_{0:t-1}, \theta)$$

In practice we implemented this by calculating $l(\theta)$ on a grid of values $\{\theta_1, \ldots, \theta_m\}$. If the maximizing value over this grid was at the endpoints, then the estimate was set to be that endpoint, otherwise a quadratic polynomial was fit to the maximizing value along with its two adjacent grid values. The value maximizing that polynomial was then taken to be the estimate.

CHAPTER 5

# DISCRETE EXAMPLES

## 5.1 6 state example

While the FOFI strategy has been shown to be effective in Hooker et al. [5] it is possible to define systems in which the strategy is not optimal and may in fact be worse than just using fixed or random controls. Usually certain parts of state space will give more information about a parameter than others, given that the state space is perfectly observed. In these cases optimal controls would try to move the process to these states. However, if the state space is only partially observed, most information might be obtained in different parts of state space and the FOFI controls become suboptimal. In cases like this the truncated POFI and WOFI often do better than FOFI, since they take advantage of the observation process. In this example, we demonstrate a system where using FOFI, WOFI and truncated POFI leads to different control policies, and using a simulation study, we show that using a truncated POFI or WOFI policy produces less variable parameter estimates than then using a FOFI policy.

Consider a discrete time Markov chain $x_t$ with state space $S_x = \{1, 2, 3\}$ and a transition probability matrix

$$P = \begin{bmatrix} \frac{1}{2} - \frac{p}{4} + \frac{u}{4} & \frac{1}{3} & .4 - \frac{u}{4} \\ \frac{p}{2} & \frac{1}{3} & .15 \\ \frac{1}{2} - \frac{p}{4} - \frac{u}{4} & \frac{1}{3} & .45 + \frac{u}{4} \end{bmatrix}$$

where the parameter of interest is $p \in [0, .5]$ and the control is $u \in \{-1, 1\}$. For $x_t = 1$ or $x_t = 3$, choosing the control $u = 1$ will increase the probability of the

Markov chain staying in its current state while choosing $u = -1$ will increase the probability of it leaving its state.

Now assume this process isn't observed directly but through a related process $y_t$ with state space $S_y = \{1, 2\}$ whose transition probabilities depend on which state $x_t$ is in. We denote the transition probability matrices with $(P_k)_{\{i,j\}} = p(y_{t+1} = j | y_t = i, x_t = k)$ given by

$$P_1 = \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix}, P_2 = \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix}, P_3 = \begin{bmatrix} 1 - \frac{p}{2} & \frac{p}{2} \\ \frac{p}{2} & 1 - \frac{p}{2} \end{bmatrix}$$

If $x_t$ were observed we would get information about the parameter $p$ when $x_t$ leaves state 1 and from $y_t$ when $x_t = 3$. The idea here is that since the FOFI controls assume the whole state space is observed they might encourage $x_t$ to be in state 1, while the truncated POFI controls and the WOFI controls take into account what is actually observed and might choose the controls more intelligently. Indeed when calculating the controls according to FOFI the long run control is to "leave one's state" if $x_t = 3$ and "stay in one's state" if $x_t = 1$. The WOFI policy takes observations into account and does the reverse as FOFI. It is harder to predict and interpret the controls that result from using the truncated POFI, but the control policy is given in Table 5.1. We set the truncation factor to $m = 1$, that is the policy at time $t$ depends on $(y_t, y_{t-1}, u_{t-1})$.

To illustrate this difference, a simulation study was carried out to test what method performed best: The process $x_t$ was run for 1000 steps with $p = .37$, using controls chosen by truncated POFI, WOFI and FOFI. Additionally we ran a simulation of the same length, but where the control was chosen randomly, with $u = -1$ and $u = 1$ having equal probability. Then the parameter $p$ was

| $u_t$ | 1 | -1 | -1 | 1 | -1 | 1 | 1 | -1 |
|---|---|---|---|---|---|---|---|---|
| $y_t$ | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| $y_{t-1}$ | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 |
| $u_{t-1}$ | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 |

Table 5.1: Long run control policy that results from using a truncated POFI in the 6 state example. The first column describes which control to use for a given history $(y_t, y_{t-1}, u_{t-1})$ of observations and control.

| | bias | st. dev. | MSE |
|---|---|---|---|
| FOFI | .0009 | .0823 | .0068 |
| WOFI | .0009 | .0506 | .0026 |
| tr. POFI | .0040 | .0526 | .0028 |
| Random | .0017 | .0702 | .0049 |

Table 5.2: Simulation results for the 6 state example. We see that the controls chosen by truncated POFI or WOFI make for more accurate estimates of $p$. The FOFI policy does worse than a random policy.

estimated using an EM algorithm. This was done 500 times to get an empirical distribution for the estimates of $p$. The results are given in Table 5.2. Estimates of $p$ using the truncated POFI or WOFI policy had the lowest MSE and variability. Estimates using a FOFI policy were comparatively worse than using a random policy.

## 5.2 Gamble Safe example

The following example describes an application of the above methods in the context of experimental economics. The problem is derived from Sachat et. al. [9], in which we wish to model how humans change their game-playing strategies over time.

We set up a game with two players: a Row player and a Column player.

| | Left | Right |
|---|---|---|
| Left | 2,0 | 0,1 |
| Right | 1,2 | 1,1 |

Table 5.3: Rewards in the Gamble Safe game. The first number is the reward for the Row player and the second number the reward for the Column player, given a certain outcome.

They repeatedly play a game where both simultaneously choose either left or right, and they get rewards depending on the outcome according to Table 5.3; the Row player would for example get 2 and the Column player 0 if both chose left. We follow [9] and assume that at any given play the Column player follows one of two strategies: the Nash-equilibrium strategy of choosing either left or right with 50% probability or the Gamble-safe strategy, where they only choose right. The player will pick either strategy based on a multinomial logistic model, where the probabilities depend on the last two plays of the Row player, and the last strategy chosen by the Column player. This results in a Partially Observed Markov Decision Process with the strategy employed being a hidden state giving rise to observed plays.

Let $S_t$ denote the strategy chosen by the Column player at time $t$, $U_t$ denote the action played by the Row player at time $t$. Let $S_t = -1$ if the Nash-equlibrium is chosen, $S_t = 1$ if the Gamble-safe strategy is chosen. Also let $U_t = 1$ if the Row player plays right, $U_t = -1$ if he plays left. Similarly $Y_t$ will denote the plays of the Column player. The strategy $S_{t+1}$ chosen at time $t + 1$ will then be chosen according to

$$P(S_{t+1} = -1) = \frac{e^x}{1 + e^x} \text{ and } P(S_{t+1} = 1) = \frac{1}{1 + e^x}$$

where we let

$$x = 1.2U_t + U_{t-1} + \theta S_t$$

The experiment is set up with two natural strategies for the Column player and we can think of $\theta$ as the persistence of strategies. The purpose of this experiment is to elicit information about how humans persist in strategy choice, and we therefore investigate how the plays of the Row player can be used to obtain an estimate of $\theta$ that is as precise as possible.

To cast this into our usual setting we think of $S_t$ being the unobserved underlying Markov Chain, $U_t$ as the control and $Y_t$ as the observed process. Since the transition probabilities from $S_t$ depend on $U_{t-1}$ (a part of the history at time $t-1$) we augment the state space to include $U_{t-1}$, i.e. $R_t = (S_t, U_{t-1})$ will be our underlying Markov Chain. At this point we could run the dynamic programs for both FOFI, truncated POFI and WOFI, but controls calculated that way will depend deterministically on the plays of the Column player. Seeing that realistically deterministic plays can often easily be countered in adversarial games, it is better to follow a strategy that includes some randomness in the plays. So we let $W_t \in \{-1, 1\}$ be the strategy of the Row player in such a way that

$$\left.\begin{array}{ll} U_t = 1 & \text{w.p. } .8 \\ U_t = -1 & \text{w.p. } .2 \end{array}\right\} \text{ if } W_t = 1, \quad \text{and} \quad \left.\begin{array}{ll} U_t = 1 & \text{w.p. } .2 \\ U_t = -1 & \text{w.p. } .8 \end{array}\right\} \text{ if } W_t = -1$$

These kind of changes are easily incorporated in the dynamic program for both FOFI, truncated POFI and FOFI, by adding an expectation over $W_t$ at every step $t$.

We set $\theta = .7$ and calculated the FOFI, truncated POFI and WOFI policies. We also consider a random policy, where the probability of choosing either control was set to $1/2$.

To compare the two policies we ran a simulation study with $T = 500$, and

| Adversarial Game | | | |
|---|---|---|---|
| | bias | st. dev. | MSE |
| FOFI | 0.00 | 0.33 | 0.11 |
| WOFI | 0.02 | 0.27 | 0.07 |
| tr. POFI | 0.02 | 0.30 | 0.09 |
| Random | 0.01 | 0.33 | 0.11 |

Table 5.4: Simulation results for Adversarial Game. The FOFI policy is similar to the random policy. Truncated POFI does slightly better that FOFI and WOFI does slightly better than truncated POFI.

1000 simulations for every control policy. The parameter $\theta$ was estimated using an EM algorithm. The results of this estimation under each policy are given in Table 5.4 where the WOFI controls produce least variance and the most accurate estimates.

CHAPTER 6

# DIFFUSION PROCESSES

In order to apply the methods described above to dynamical systems, we need to approximate them by a suitable Partially Observed Markov Decision Process. We achieve this by discretizing time, state and observation spaces. Here we consider continuous stochastic dynamical systems of the form

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, \theta, u(t))dt + \Sigma_1^{1/2} d\mathbf{W}$$

where $\theta$ is the parameter of interest, to be estimated, $u(t)$ is a control that can be chosen by user, $\mathbf{x}$ is the vector of state variables, $\mathbf{f}$ is a vector valued function and $\mathbf{W}$ a Wiener process. The dynamical system is approximated on a fine grid of times $(t\delta)_{t=0,\ldots,T}$ and we obtain a discrete-time model

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \delta \mathbf{f}(\mathbf{x}_t, \theta, u_t) + \sqrt{\delta}\boldsymbol{\epsilon}_{1t}$$

where $\boldsymbol{\epsilon}_{1t} \sim N(0, \Sigma_1)$ are independent normal random variables. We assume the underlying state variables $x_t$ are only observed partially or noisily.

$$\mathbf{y}_t = \mathbf{g}(\mathbf{x}_t) + \boldsymbol{\epsilon}_{2t}$$

where $\boldsymbol{\epsilon}_{2t} \sim N(0, \Sigma_2^2)$.

## 6.1 Discretizing a Diffusion Process

In order to approximate this as a Markov Chain, the state space is discretized in each dimension and the model is then thought of as moving between the different boxes. The probability of moving from box to box is approximated

using the normal p.d.f. at the midpoints of the boxes. In the examples covered in Chapter 7, only equidistant discretization is considered, but this restriction can be readily removed. If we label the two midpoints as $i_1$ and $i_2$ and the area of the second box as $A_x$ this probability is given as $p(x_{t+1} = i_2 | x_t = i_1, u_t, \theta)$

$$\approx \frac{\exp\left(-\frac{1}{2}(i_2 - (i_1 + \delta\mathbf{f}(i_1, \theta, u_t)))^T \Sigma_1^{-1} (i_2 - (i_1 + \delta\mathbf{f}(i_1, \theta, u_t)))\right) \cdot A_x}{(2\pi)^{k/2} \det(\Sigma_1)^{1/2}}$$

where $k$ is the dimension of $\mathbf{x}$. The probabilities are then normalized to make sure they sum to 1. If the controls $u_t$ can be chosen on a continuous scale then this scale has to be discretized as well. $(x_t, u_t)$ is then a Markov Decision Process, and one can run the FOFI dynamic program.

For the truncated POFI and the WOFI dynamic program the observation space needs to be discretized as well. The probability of what observation box is observed depends on in which box the underlying Markov Chain is in. If we label the midpoint of the underlying Markov chain midpoint as $i$ and the midpoint of the observed process box midpoint as $j$, and the area of the latter box as $A_y$ this probability is given as

$$p(y_t = j | x_t = i) \approx \frac{1}{(2\pi)^{k/2} \det(\Sigma_2)^{1/2}} \exp\left(-\frac{1}{2}(j - g(i))^T \Sigma_2^{-1} (j - g(i))\right) \cdot A_y$$

These probabilities are also normalized to sum to 1. The process $(x_t, y_t, u_t)$ is now a Partially Observed Markov Decision Process and one can run an appropriate POMDP dynamic program.

## 6.2 FOFI and WOFI in Diffusion Processes

Hooker et al. [5] came up with experimental design for Diffusion Processes;

$$d\mathbf{x_t} = \mathbf{f}(\mathbf{x_t}, \theta, u_t)dt + \Sigma(\mathbf{x_t})^{1/2} dW_t$$

that is similarly based on using dynamic programming to maximize the FOFI likelihood. However since their treatment is within the framework of diffusion processes, the reward functions are given in terms of $f$ and $\Sigma$ instead of transition probabilities of the approximating POMDP. We review their calculations for the FOFI criteria and show how they can be extended for the WOFI criteria. The truncated POFI criteria doesn't simplify the way the WOFI and FOFI criteria do.

The diffusion process likelihood for $\theta$ is

$$l(\theta|\mathbf{x}) = \frac{1}{2}\int_0^T \mathbf{f}(\mathbf{x_t}, \theta, u_t)^T \cdot \Sigma^{-1}(\mathbf{x_t}) \cdot \mathbf{f}(\mathbf{x_t}, \theta, u_t)dt - \int_0^T \mathbf{f}(\mathbf{x_t}, \theta, u_t) \cdot \Sigma^{-1}(\mathbf{x_t}) \cdot d\mathbf{x_t}$$

with the associated Fisher Information is

$$I(\theta, u) = E\int_0^T \left\| \frac{\partial}{\partial\theta}\mathbf{f}(\mathbf{x_t}, \theta, u_t)\right\|^2_{\Sigma(\mathbf{X_t})} dt$$

where $\|\mathbf{z}\|_\Sigma = \mathbf{z}^T\Sigma^{-1}\mathbf{z}$.

Hooker et al. [5] approximate this Fisher information by discretizing time. For a diffusion process discretized at time $t_i = i\Delta t, i = 1, \ldots, T$ we get

$$\mathbf{x_{i+1}} = \mathbf{x_i} + \mathbf{f}(\mathbf{x_i}, \theta, u_i)\Delta t + \sqrt{\Delta t}\Sigma(\mathbf{x_i})^{1/2}\epsilon_\mathbf{i}$$

where $\epsilon_\mathbf{i}$ are independent vectors of independent standard normal random variables. We can now discretize the continuous Fisher Information above, or derive it from the discretized diffusion process, either way we get the Full Observation Fisher Information (FOFI) as

$$\widehat{FI}(\theta) = \sum_{t=1}^{T} E\left[\left\| \frac{\partial}{\partial\theta}\mathbf{f}(\mathbf{X_i}, \theta, u_i)\right\|^2_{\Delta t\Sigma(\mathbf{X_i})}\right](\Delta t)^2$$

and this is in correspondence with what you would get in the POMDP framework, i.e.

$$E\left[\left(\frac{\partial}{\partial\theta}\log p(\mathbf{X_{t+1}}|\mathbf{X_t}, \theta, u_t)\right)^2\right] = E\left[\left\| \frac{\partial}{\partial\theta}\mathbf{f}(\mathbf{X_t}, \theta, u_t)\right\|^2_{\Delta t\Sigma(\mathbf{X_t})}\right](\Delta t)^2$$

Commonly the state space is observed noisily, or not completely, where the amount of noise could depend on location in state space. We assume for now that

$$\mathbf{y(t)} = A\mathbf{x(t)} + b + \Sigma_2^{1/2}\epsilon_{\mathbf{t}}$$

and note that this encompasses common situations such as not observing a state altogether, observing the sum of multiple states, but doesn't allow the observational variance to depend on the state $\mathbf{x(t)}$.

The WOFI criteria within POMDP's is

$$E\left[\left(\frac{\partial}{\partial\theta}\log p(\mathbf{Y_{t+1}}|\mathbf{X_t},\theta,u_t)\right)^2\right]$$

and within the discretized diffusion process

$$\mathbf{y_{i+1}} = A\left(\mathbf{x_i} + \mathbf{f}(\mathbf{x_i},\theta,u_i)\Delta t\right) + b + \sqrt{\Delta t}A\Sigma(\mathbf{x_i})^{1/2}\epsilon_{\mathbf{1,i}} + \Sigma_2^{1/2}\epsilon_{\mathbf{2,i}}$$

where $\epsilon_{1,i}, \epsilon_{2,i}$ are independent standard normal vectors. We see how, with WOFI, the Fisher Information criteria changes naturally to

$$E\left[\left(\frac{\partial}{\partial\theta}\log p(\mathbf{Y_{t+1}}|\mathbf{X_t},\theta,u_t)\right)^2\right] = E\left[\left\|A\frac{\partial}{\partial\theta}\mathbf{f}(\mathbf{X_t},\theta,u_t)\right\|^2_{\Sigma^*(\mathbf{X_t})}\right](\Delta t)^2$$

where $\Sigma^*(\mathbf{x_t}) = \Delta t A\Sigma(\mathbf{x_t})A^T + \Sigma_2$. Note that by adjusting $\Delta t$ we can adjust process variance relative to observation variance.

More generally we would be interested in $\mathbf{y(t)} = \mathbf{g(x(t))} + \Sigma_2^{1/2}\epsilon_{\mathbf{t}}$ where $g$ is not necessarily a linear mapping, or allowing the observational error to depend on $\mathbf{x(t)}$, that is $\mathbf{y(t)} = \mathbf{x(t)} + \Sigma_2^{1/2}(\mathbf{x_t})\epsilon_{\mathbf{t}}$. The WOFI criteria

$$E\left[\left(\frac{\partial}{\partial\theta}\log p(\mathbf{Y_{t+1}}|\mathbf{X_t},\theta,u_t)\right)^2\right]$$

is harder to write out directly in this case, since $\mathbf{Y_{t+1}}$ conditional on $\mathbf{X_t}$ is not necessarily normal anymore.

We don't pursue these issues in detail but note that the first case could be approximated with a linear mapping $\mathbf{g}(\mathbf{x_{t+1}}) \approx g(x_t) + J_g(x_t)(x_{t+1} - x_t)$, which above would amount to setting $b = g(x_t)$ and $A = J_g(x_t)$

CHAPTER 7

# CONTINUOUS EXAMPLES

## 7.1 Morris Lecar model

The Morris Lecar Model [11] describes oscillatory electric behavior in a single neural cell, as regulated by flow of Potassium and Calcium ions across the cell membrane. These models are defined in terms of state variables $v_t$ and $n_t$ representing the voltage across the membrane and the flux of the Potassium channel respectively.

$$C_m \dot{v}_t = I_t - g_l \cdot (v_t - E_l) - g_K \cdot n_t \cdot (v_t - E_K) - g_{Ca} \cdot m_\infty(v_t) \cdot (v_t - E_{Ca}) \tag{7.1}$$

$$\dot{n}_t = -\phi \cdot (n_t - n_\infty(v_t))/\tau_n(v_t) \tag{7.2}$$

where $m_\infty(v) = \frac{1}{2}(1 + \tanh((v - v_1)/v_2))$, $\tau_n(v) = \operatorname{sech}((v - v_3)/(2v_4))$ and $n_\infty(v) = \frac{1}{2}(1+\tanh((v-v_3)/v_4))$. We will write $C_m \dot{v}_t = F_1(v_t, n_t)$ and $\dot{n}_t = F_2(v_t, n_t)$ as shortcuts equations (7.1) and (7.2). The voltage between cells depends on Potassium and Calcium concentrations, and on the amount of leakage. The further these factors are away from their equilibriums $E_l, E_K, E_{Ca}$ the greater the rate of change in voltage. The multiplicative value $n_t$ changes the conductance of the potassium channel and is modeled through the second differential equation in which $n_t$ is driven towards a voltage-dependent equilibrium level defined by $n_\infty(v_t)$ but converges to this at a much slower rate then the dynamics of $v_t$. The neuron is stimulated by an external applied current, $I_t$ (our control), and $v_t$ is measured. Our goal is to maximize information about the parameters $C_m, g_{Ca}$ and $\phi$, considered separately.

We consider a stochastic version of this neural firing model, derived from [10], by adding $\sigma dW_1$ and $\tilde{\sigma} dW_2$ to equations (7.1) and (7.2) respectively, where $W_1$ and $W_2$ are independent Wiener processes. Stochastic models are important in this context in order to accommodate observable variation in the inter-spike interval where a deterministic model will require a fixed period; see [3], for example.

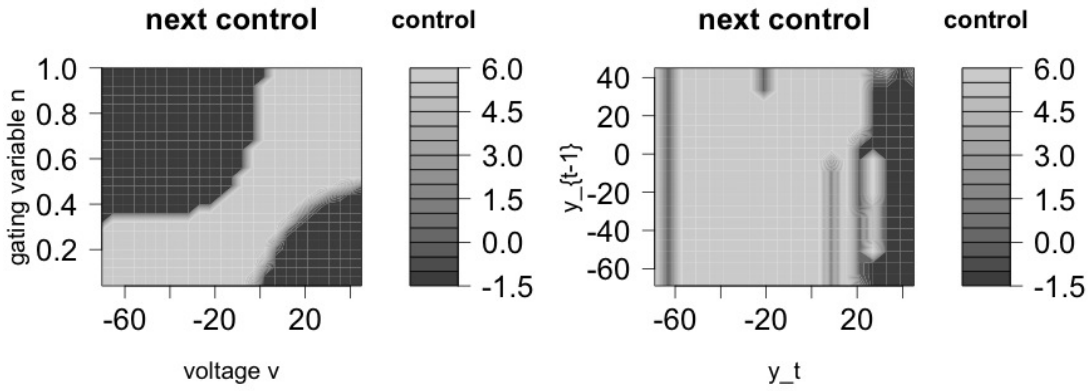
The first step is to discretize these equations with respect to time. We get that $v_t(t + dt) = v(t) + dt \cdot F_1(v(t), n(t))/C_m + \sigma \sqrt{dt} \cdot \varepsilon_1$ and $n_t(t + dt) = n(t) + dt \cdot F_2(v(t), n(t)) + \tilde{\sigma} \sqrt{dt} \cdot \varepsilon_2$ where $\varepsilon_1, \varepsilon_2 \sim N(0, 1)$.

We discretized $v_t$ onto the range $[-75, 45]$ and $n_t$ onto $[0, 1]$, after running a few trial versions of the model. Both ranges where discretized into 25 intervals. Only $v_t$ is measured and it is measured noisily,

$$y_t = v_t + \varepsilon_t$$

where $\varepsilon_t \sim N(0, 1)$. The observation space was discretized to the same range as $v_t$ but into 20 intervals. These approximations give rise to a Partially Observed Markov Decision Process to which our methods can be applied. The values for the parameters were set to be $C_m = 20$, $g_{Ca} = 4.4$, $g_l = 2.0$, $E_k = -84.0$, $E_l = -60$, $E_{Ca} = 120.0$, $\phi = .04$, $v_1 = -1.2$, $v_2 = 18.0$, $v_3 = 2.0$, $v_4 = 30.0$, $\sigma = \tilde{\sigma} = 1$ and $dt = 1$. The controls range was set to be $[-1.5, 6.0]$ and discretized to the set $I_t \in \{-1.5, 0.0, 1.5, 3.0, 4.5, 6.0\}$. We considered experimental design for the parameters $C_m$, $g_{Ca}$ and $\phi$, considered separately. FOFI, WOFI and truncated POFI controls were calculated using dynamic programming, where the truncation factor $m = 1$ was chosen.

When calculating a Fisher Information reward to use in a dynamic program, we generally use the estimated transition probabilities of the POMDP, for exam-




(a) FOFI control

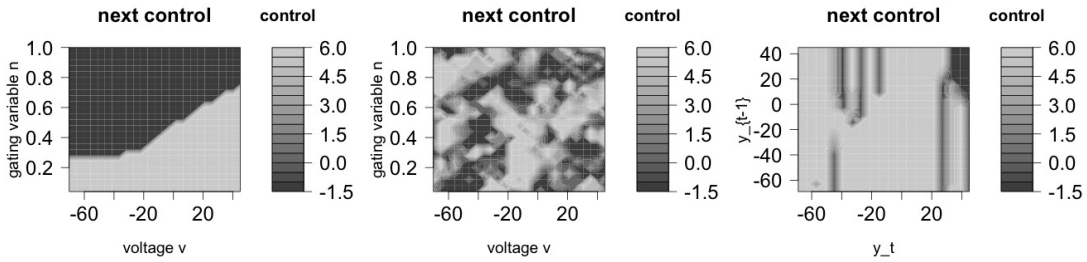
(b) truncated POFI control

Figure 7.1: Long term controls of FOFI and truncated POFI for the parameter $g_{Ca}$. The FOFI plot gives the control to use, given a certain position in state space. The truncated POFI control will depend on the last two observations and the last control, but fixing the last control as, for example, $I_{t-1} = 6$ one can plot which control to use given combinations of the last two observations.

ple the WOFI reward is $\left(\frac{\partial}{\partial\theta}p(y_{t+1}|x_t,u_t,\theta)\right)^2$. As discussed in Section 6.2 we can frequently calculate the FOFI and the WOFI reward using the function $f$ and the covariance matrix $\Sigma$. If we look at $\theta = g_{Ca}$ for example, we see that it only appears in the $v$ dimension, $C_m\dot{v}_t = I_t - g_l\cdot(v_t - E_l) - g_K\cdot n_t\cdot(v_t - E_K) - g_{Ca}\cdot m_\infty(v_t)\cdot(v_t - E_{Ca})$ and we get that

$$\left(\frac{\partial}{\partial\theta}p(x_{t+1}|x_t,u_t,\theta)\right)^2 \propto \left(m_\infty(v_t)\cdot(v_t - E_{Ca})\right)^2$$

Since the observations process assumes that we only observe the $v(t)$ dimension with some normal noise we have that $A = (1,0)$ in Section 6.2. This shows that the WOFI reward $\left(\frac{\partial}{\partial\theta}p(y_{t+1}|x_t,u_t,\theta)\right)^2$ is proportionally the same as the FOFI reward, and we shouldn't expect any difference between the corresponding policies. The FOFI and truncated POFI long term policies for $g_{Ca}$ are given in Figure 7.1. The WOFI policy is not shown since it coincides with the FOFI policy.




(a) FOFI control (b) WOFI control (c) truncated POFI control

Figure 7.2: Long term policy of FOFI, WOFI and truncated POFI for the parameter $\phi$. The FOFI policy is clear cut while the WOFI policy is only picking up on numerical noise. In the truncated POFI policy we fix $I_{t-1} = 6$ to get a plot of which control to use given combinations of the last two observations.

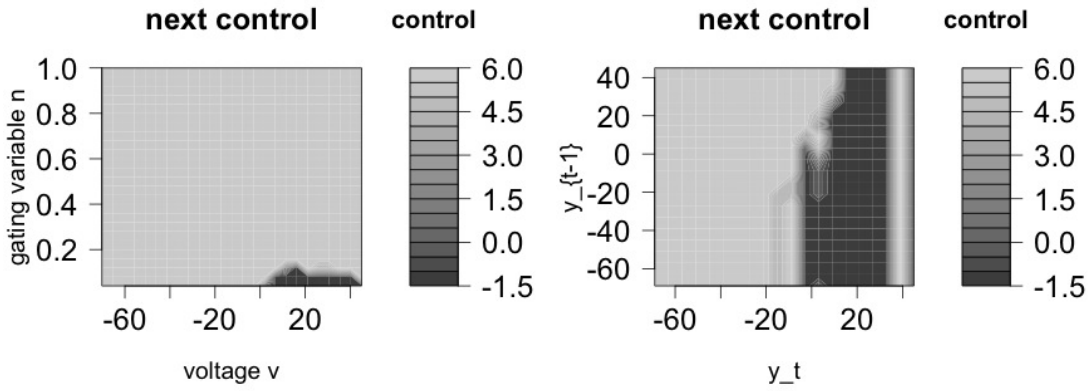
Experimental design for $\phi$ is trickier since it only appears in the second dimension $\dot{n}_t = -\phi \cdot (n_t - n_\infty(v_t))/\tau_n(v_t)$. Since the WOFI reward re-weights the FOFI reward depending on how the states are observed, we get that the WOFI reward breaks down in this case;

$$\left(\frac{\partial}{\partial\theta} p(y_{t+1}|x_t, u_t, \theta)\right)^2 = 0$$

See figure 7.2 for long term policies for $\phi$. We see that the truncated POFI policy seems rather unclear, while the WOFI policy just picks up on numerical noise.

The parameter $C_m$ only appears in the $v$ dimension, and the WOFI and FOFI policies coincide again, see figure 7.3 for FOFI and truncated POFI long term policies. The longterm FOFI policy seems to almost only choose the highest possible control, while the truncated POFI policy varies more.

A simulation study was run for each of the three parameters $g_{Ca}, \phi, C_m$ using FOFI and truncated POFI policies (skipping WOFI since it was either the same as FOFI or non sensible). The system was simulated within the discretized Markov Chain framework with 100 time steps and all schemes had 100 simulations. The parameter in question was estimated for each simulation using an




(a) FOFI control

(b) truncated POFI control

Figure 7.3: Long term policy of FOFI and truncated POFI for the parameter $C_m$. In the truncated POFI policy we fix $I_{t-1} = 6$ to get a plot of which control to use given combinations of the last two observations.

| parameter | | bias | st. dev. | MSE |
|---|---|---|---|---|
| $C_m$ | FOFI | .4234 | 2.4722 | 6.2913 |
| | tr. POFI | .4129 | 2.4068 | 5.9632 |
| | Fixed | .9098 | 3.4240 | 12.551 |
| $g_{Ca}$ | FOFI | .0613 | .3671 | .1385 |
| | tr. POFI | .0158 | .3706 | .1376 |
| | Fixed | .0249 | .6193 | .3841 |
| $\phi$ | FOFI | .00485 | .01085 | .00014 |
| | tr. POFI | .00257 | .01037 | .00011 |
| | Fixed | .01357 | .02643 | .00088 |

Table 7.1: Simulation results for the Morris-Lecar model, consider the parameters $C_m, g_{Ca}, \phi$ separately. We see that the truncated POFI and FOFI policies outperform the fixed policy $I_t = 1.5$ in all cases, and the truncated POFI policy seems to perform slightly better than the FOFI policy for the three parameters considered.

EM algorithm. As a baseline comparison we also ran a simulation study using a fixed control ($I_t = 1.5$). The results are given in Table 7.1. The difference between the truncated POFI and FOFI turns out to be not very dramatic, likely due to the observations providing a great deal of information about the underlying state variables, which is when FOFI performs well.

## 7.2 Rosenzweig MacArthur model

The Rosenzweig MacArthur model describes the population dynamics of a two species ecology, a prey species $C$ (generally a type of algae in chemostat experiments) and a predator species $B$ (rotifers in chemostat, a microscopic animal). The chemostat experiment consist of a tank filled with a nutrient rich medium which the prey species consumes, and the predator consumes the prey. See Hooker [4] for details.

The model can be expressed in various approximately equivalent ways, but we focus on the diffusion model formation of the model, $d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \Sigma(\mathbf{x})^{1/2}dW$ with

$$\mathbf{x} = \begin{pmatrix} C \\ B \end{pmatrix}, \mathbf{f} = \begin{pmatrix} \rho C(\kappa_C - C) - \frac{\gamma\beta CB}{\kappa_B + C} \\ \frac{\beta CB}{\kappa_B + C} - mB \end{pmatrix}$$

and $dW = (dW_1, dW_2)$, a two dimensional independent Wiener process with

$$\Sigma(\mathbf{x}) = \begin{pmatrix} \rho C(\kappa_C - C) + \frac{\gamma^2\beta BC}{\kappa_B + C} & -\frac{\gamma\beta BC}{\kappa_B + C} \\ -\frac{\gamma\beta BC}{\kappa_B + C} & \frac{\beta BC}{\kappa_B + C} + mB \end{pmatrix}$$

The algae grows logistically according to $\rho C(\kappa_C - C)$ where $\kappa_C$ is an upper bound on the population, and $\rho$ controls the growth speed. Rotifers reproduce proportionally to the number of algae according to $\beta BC/(\kappa_B + C)$, and $\gamma$ controls how many algae are needed to create one new rotifer. The rotifers die proportionally to their population according to $-mB$.

We assume a controllable dilution rate $\delta$, which affects both the algae population limit $\kappa_C$ and the rotifer death rate $m$ in the following way; $\kappa_C = \kappa^+/(\kappa^- + \delta)$

and $m = m_0 + \delta$. We assume the following parameter values; $\rho = 4.17 * 10^{-7}$, $\beta = .75$, $\gamma = 30$, $\kappa^+ = 180000$, $\kappa^- = .4$ and $m_0 = .04$.

Discretizing the state space is challenging in this form since the algae $C$ can become very large. Instead we take logarithms; $x_1 = \log(C)$ and $x_2 = \log(B)$, and consider the derived diffusion model $d\tilde{x} = \tilde{\mathbf{f}}(\exp(\tilde{\mathbf{x}}))dt + \tilde{\Sigma}(\exp(\tilde{\mathbf{x}}))^{1/2}dW$

We get

$$\tilde{\mathbf{f}}(\mathbf{x}) = \frac{\mathbf{f}(\mathbf{x})}{\mathbf{x}} = \begin{pmatrix} \rho(\kappa_C - C) - \frac{\gamma\beta B}{\kappa_B + C} \\ \frac{\beta C}{\kappa_B + C} - m \end{pmatrix}$$

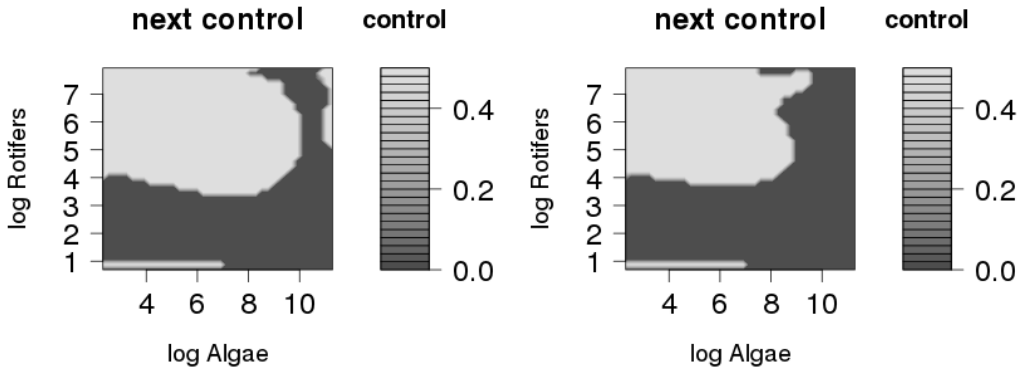
and

$$\begin{aligned} \tilde{\Sigma}(\mathbf{x}) &= diag(1/\mathbf{x})\Sigma(\mathbf{x})diag(1/\mathbf{x}) \\ &= \begin{pmatrix} \frac{\rho(\kappa_C - C)}{C} + \frac{\gamma^2\beta B}{C(\kappa_B + C)} & -\frac{\gamma\beta}{\kappa_B + C} \\ -\frac{\gamma\beta}{\kappa_B + C} & \frac{\beta C}{B(\kappa_B + C)} + \frac{m}{B} \end{pmatrix} \end{aligned}$$

After considering various sample paths of the system, we discretize the $x_1$ dimension onto the range $[2.3, 11.3]$, and the $x_2$ dimension to the range $[.7, 8]$. Both ranges were discretized evenly into 40 intervals. To add stability diagonal noise was added to $\tilde{\Sigma}$, which was proportional to the squared bin size in the discretization of $(x_1, x_2)$. Three possible control values were provided; $\delta \in \{0, .2, .5\}$.

We assumed that the observations were binomial samples of the algae, $y_t = Bin(C_t, p) = Bin(\exp(x_{1,t}), p)$ where the sampling coefficient was set at $p = .1$. This was approximated by a normal distribution $N\Big(C_t p, C_t p(1-p)\Big)$.

We considered the problem of estimating the parameter $\beta$ with maximal precision and to do that we ran a dynamic program with $T = 300$ steps for the WOFI and FOFI criteria and an example of the controls can be seen in figure 7.4. As in the Morris Lecar model, we should expect the WOFI and FOFI controls to




(a) FOFI control (b) WOFI control

Figure 7.4: Long term controls in the Rosenzweig MacArthur model, FOFI left, WOFI right.

| Rosenzweig MacArthur Model | | | |
|---|---|---|---|
| | bias | st. dev. | MSE |
| FOFI | .0016 | .0341 | .0012 |
| WOFI | .0075 | .0270 | .0008 |

Table 7.2: Simulation results for the Rosenzweig MacArthur Model.

be similar, due to the parameter $\beta$ having the same form in both dimensions of the system, although some variation might be due to $\Sigma$ also contribution parameter information. The truncated POFI turned out to be computationally harder to handle in this example, and was thus not considered.

We ran a simulation study of 200 simulations for both FOFI and WOFI, and estimated the parameter $\beta$ by maximizing the relevant likelihood. The results are given in Table 7.2. The performance between WOFI and FOFI seems similar.

CHAPTER 8

# PARAMETER DEPENDENCE OF DYNAMIC PROGRAM

In the examples above we calculated the dynamic program assuming knowledge of the parameter $\theta$, the very thing we wish to estimate with maximal precision. Since the dynamic programs we have considered are run before the experiment is started we generally won't have data to estimate $\theta$. Additionally, for the FOFI simulations we have used $\theta$ directly to estimate $x_t$ within the filter to get the appropriate control, but this will not be possible in practice. There are a few ways of dealing with this.

Assuming some prior information one can use a prior for $\theta$ to run the dynamic program. To do this, we add one more expectation for $\theta$ at every time step $t$, and then maximize the expected Fisher Information to get the best control. In the FOFI case this means maximizing

$$E_\theta\left[E\sum_{t=0}^{n}\left(\frac{\partial}{\partial\theta}\log p(x_{t+1}|x_t,u_t,\theta)\right)^2\right]$$

This strategy was employed in Hooker et al. [5].

The rather obvious deficiency here, for all our Fisher Information criteria, is that as the experiment runs, we get observations that can be used to improve our prior for $\theta$, and could be used to get better controls, if we could brake the experiment and rerun the dynamic program.

## 8.1 Online updating

In some systems the time spent in each state is very short, too short to perform many calculations, making it valuable to have a "look-up table" of con-

trols. Here the truncated POFI controls have an advantage over the FOFI and WOFI controls, in the sense that they are of the “look-up" kind, as FOFI and WOFI require estimation of the underlying $x_t$ process, before the control can be looked up.

In other systems, there is time to do some calculations between transitions. Note, for example, that at time $t$ we have observed $y_0, \ldots, y_t$ and this will allow us to calculate a posterior distribution $\pi(\theta|y_{0:t}, u_{0:t-1})$ for our parameter of interest. This posterior could then be used to run the dynamic program again, as described above, from time $T-1$ to time $t$. This can be quite time consuming if done at each time step $t$, so we propose a method that relies on the Value Iteration Algorithm (VIA), see Section 2.3 for a description of VIA.

### 8.1.1 Value Iteration Algorithm

As discussed in Section 2.3, in VIA we calculate

$$v^{n+1}(x_t, \theta) = \max_u \{E_{x_{t+1}}[C(x_t, u_t, \theta) + \lambda \cdot v^n(x_{t+1}, \theta)|x_t, u_t, \theta]\}$$

where $0 \leq \lambda < 1$, and this maximizes the expected total discounted reward $W_3 = E\left[\sum_{t=0}^{\infty} \lambda^{t-1} C(x_t, u_t)\right]$. Also covered in Section 2.3 is that if $\lambda$ is close enough to one, Blackwell optimality guarantees that controls that maximize $W_3$ also maximize the expected average reward $W_2 = \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{t=0}^{n} C(x_t, u_t)\right]$, or its lim sup if the limit doesn't exist.

We can therefore say that our aim with VIA is to maximize what we in the truncated POFI case label, the average truncated Partial Observation Fisher In-

formation

$$\lim_{n\to\infty} \frac{1}{n} E_\theta E_{\mathbf{y}|\theta} \sum_{t=0}^{n} \left( \frac{\partial}{\partial \theta} \log p(y_{t+1}|y_{t-m:t}, u_{t-m:t}, x_0, \theta) \right)^2$$

or in the FOFI case, the average Full Observation Fisher Information and similar for WOFI. This is a reasonable quantity to maximize in order to obtain a time-invariant policy, see Section 2.3 for conditions on the existence of a average criteria.

We propose running VIA at every time step $t$, but to use the posterior for $\theta$, $\pi(\theta|y_{0:t}, u_{0:t-1})$, which is conditioned on all the data observed so far, instead of using the prior for $\theta$. This will give a control that maximizes the average Fisher Information, using all the parameter information that is available at time $t$. Instead of starting VIA at each time $t$ with $v^1 = 0$, considerable time can be saved by using the last value vector $v^n$ from the previous run of VIA at time $t-1$. This is because the posterior for $\theta$ often doesn't change much between time steps, and the last $v^n$ from time $t-1$ thus being relatively close to the fixed point at time $t$.

Let $v_t^n$ denote the value vector at time $t$ at the $n$'th iteration of the $t$'th VIA and let $\pi(\theta|y_{0:t}, u_{0:t-1})$ denote the posterior for $\theta$ given observations up till time $t$. Also, to ease notation, let $\mathbf{z_t} = y_{t-m:t}, u_{t-m:t-1}$. The pseudocode for this modified VIA using the truncated POFI is:

Set $v_1^0 = 0$ and $n = 0$

**for** $t = 0 \to T$ **do**

  **while** $\|v^n - v^{n-1}\| > \varepsilon$ **do**

    $\forall$ $\mathbf{z_t}$ and calculate and store

$v_t^{n+1}(\mathbf{z_t}) =$

$$\max_{u_t} \sum_{\theta} \sum_{y_{t+1}} \left[ \left( \frac{\partial}{\partial \theta} \log p(y_{t+1}|\mathbf{z_t}, u_t, \theta) \right)^2 + \lambda v_t^n(\mathbf{z_{t+1}}) \right) p(y_{t+1}|\mathbf{z_t}, u_t, \theta) \pi(\theta|y_{0:t}, u_{0:t-1}) \Bigg]$$

n=n+1

**end while**

Set $v_{t+1}^0 = v_t^n$

Now let

$u_t(\mathbf{z_t}) =$

$$\operatorname*{argmax}_{u_t} \sum_{\theta} \sum_{y_{t+1}} \left[ \left( \frac{\partial}{\partial \theta} \log p(y_{t+1}|\mathbf{z_t}, u_t, \theta) \right)^2 + \lambda v_t^n(\mathbf{z_{t+1}}) \right) p(y_{t+1}|\mathbf{z_t}, u_t, \theta) \pi(\theta|y_{0:t}, u_{0:t-1}) \Bigg]$$

Use control $u_t$, and observe $y_{t+1}$ and then update the posterior for $\theta$,

$$\pi(\theta|y_{0:t+1}, u_{0:t}) = \frac{p(y_{t+1}|y_{0:t}, u_{0:t}, \theta)\pi(\theta|y_{0:t}, u_{0:t-1})}{\sum_\theta p(y_{t+1}|y_{0:t}, u_{0:t}, \theta)\pi(\theta|y_{0:t}, u_{0:t-1})}$$

**end for**

Updating FOFI and WOFI policies online using VIA can be done in a similar way. In the next example we compare fixed policies with policies that are updated in run-time.

### 8.1.2 PCR model

Polymerase chain reaction is a well established method to copy and multiply DNA. We are interested in modeling the growth dynamics of DNA template ($x_t$), for a fixed amount of substrate. The model we use is

$$x_{t+1} = (1 - u_t)x_t + dt \frac{a(1 - u_t)x_t}{(b + (1 - u_t)x_t)^2} + \sqrt{dt} \cdot \varepsilon_1$$

where $\varepsilon_1 \sim N(0, \sigma_1^2)$. Here $x_t$ is the amount of DNA template, $a$ and $b$ the parameters of the model and $u_t$ the control, the percentage of template removed at each time point. We are interested in estimating the parameter $b$, labeled the half-saturation constant. A good reference for PCR models is [2].

We measure the amount of DNA template at each time point, but with an error. Our observations are

$$y_t = x_t + \varepsilon_2 \text{ where } \varepsilon_2 \sim N(0, \sigma_2^2)$$

and thus we have a dynamical system which when discretized becomes a Partially Observed Markov Decision Process.

The range for $x_t$ was set to be $[0, 15]$ and then discretized into 200 intervals, and $y_t$ was discretized to the same range, but only into 50 intervals. The parameter values were set to be $a = 2.0$, $b = 4.2$, $\sigma_1 = \sigma_2 = 1$, $dt = 1$ and the possible values of the control $u_t \in \{0, .2, .4, .6, .8, 1\}$.

Still with the objective of maximizing Fisher Information, we more realistically assume priors for the parameters of the system, as discussed above. We conducted a simulation study using controls based on these priors for the truncated POFI and FOFI, and then compared their performance to controls that are updated online using VIA, also both for truncated POFI and FOFI. WOFI was left out in this example, as the focus was more on the effect of updating the parameter priors. As a baseline comparison we also ran simulations using fixed controls and simulations where the true parameter is used (unrealistically) to calculate the control policy via dynamic programming as in the previous examples. For fixed controls we report the simulation with the lowest MSE, which was when $u_t = .2$.

| uniform prior, without VIA | | | |
|---|---|---|---|
| | bias | st. dev. | MSE |
| FOFI | 0.1059 | 0.6598 | 0.4465 |
| tr. POFI | 0.0053 | 0.6189 | 0.3831 |

| inaccurate prior, without VIA | | | |
|---|---|---|---|
| | bias | st. dev. | MSE |
| FOFI | 0.0755 | 0.6374 | 0.4120 |
| tr. POFI | 0.0516 | 0.7051 | 0.4998 |

| fixed control ($u_t = .2$ for all $t$) | | | |
|---|---|---|---|
| fixed | .1264 | .7466 | 0.5734 |

| uniform prior, with VIA | | | |
|---|---|---|---|
| | bias | st. dev. | MSE |
| FOFI | 0.0388 | 0.6180 | 0.3834 |
| tr. POFI | 0.0766 | 0.5999 | 0.3658 |

| inaccurate prior, with VIA | | | |
|---|---|---|---|
| | bias | st. dev. | MSE |
| FOFI | 0.0713 | 0.6787 | 0.4657 |
| tr. POFI | 0.0954 | 0.6750 | 0.4648 |

| True parameter, without VIA | | | |
|---|---|---|---|
| | bias | st. dev. | MSE |
| FOFI | 0.0659 | 0.6235 | 0.3932 |
| tr. POFI | 0.0323 | 0.6249 | 0.3916 |

Table 8.1: Simulation results for the PCR Model using two kinds of priors, truncated POFI and FOFI, with and without VIA.

The range for $b$ was set to be $b \in [1.7, 8.0]$ and then we discretized that interval into 10 points $\{1.7, 2.4, 3.1, 3.8, 4.5, 5.2, 5.9, 6.6, 7.3, 8.0\}$. We then considered a uniform prior on these points and a prior that is somewhat inaccurate, and puts the weight .9 on the point 7.3 and gives the others equal weight. The discounting factor for VIA was set to be $\lambda = .9$.

Our simulation study had the time length $T = 200$ and there were 600 simulations for each case. The parameter $b$ was estimated using an EM algorithm. The simulation results are given in Table 8.1.

We note that when we calculate the controls prior to the experiment (No online updating), both the truncated POFI and FOFI controls are significantly better than using a fixed control, and truncated POFI seems to do better than FOFI when we use an uniform prior. Interestingly in the FOFI case, calculating the controls using the inaccurate prior does better then using the uniform prior, likely due to a reduction in prior variance, in spite of additional bias.

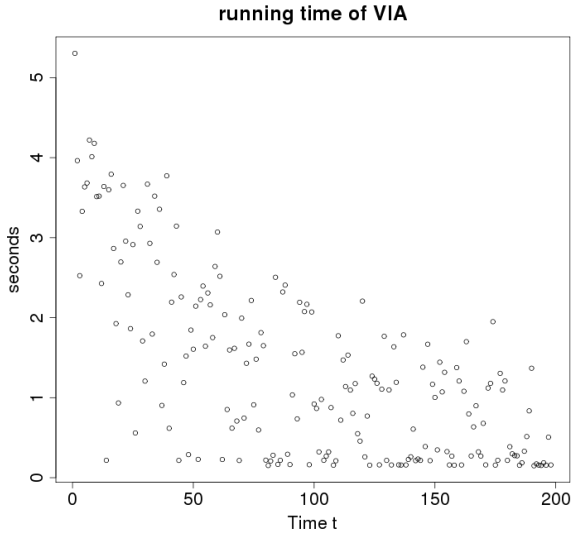



Figure 8.1: Running time of VIA at each time step $t$, for POFI using a uniform prior for the PCR model.

Accuracy increases in most cases when we allow for online updating using the VIA algorithm. Starting the VIA with an uniform prior does better than starting with the inaccurate one, which is probably due to the VIA having to spend more time “repairing" the prior. Also, we note that VIA controls with uniform prior have a similar performance to a control policy using the true (unknown) parameter.

Additionally, in Figure 8.1, we see that using the previous final value vector as the starting value vector of VIA when going from time point $t$ to $t + 1$, does save considerable time, and more so as $t$ grows and the posterior for the parameter starts to change less.

# CHAPTER 9

# CONCLUSION

We have compared three ways to conduct experimental design in parametric POMDP's, based on using dynamic programming to maximize the truncated Partial Observation Fisher Information, the Weighted Observation Fisher Information and the Full Observation Fisher Information. We have proven how the prior two criteria approximate POFI, the true Fisher Information of the data, under suitable assumptions.

Settings can arise where controls chosen by FOFI are not optimal, due to focusing on the underlying process rather than the observed process, and in these cases controls chosen with a POFI approximating criteria often perform better, as in the six state example and the adversarial game. In some of the examples analyzed they performed similarly.

In recent years, there has been growing interest in statistical procedures within dynamical systems, such as parameter estimation and hypothesis testing, and many of these procedures could be performed more efficiently given good experimental design. In the examples covered we fully discretized the state and observational spaces to transform dynamical systems with stochastic errors into partially observed Markov decision processes, allowing us to use the methods developed for POMDP's to our advantage.

We also noted how the problem of parameter dependence can be overcome by averaging over a prior. Additionally given that there is enough time between consecutive time steps, we showed how the controls can be efficiently updated online using observations gathered so far, by using a variant of the Value Itera-

tion Algorithm. This was demonstrated in the PCR example.

Finding controls that maximize information about parameters is a computationally challenging task. We have successfully demonstrated techniques for up to two dimensional systems, for a one dimensional parameter. Adding dimensions in state, parameter or observation space quickly make the methods considered computationally intractable. Considering a longer lag of past observations for the truncated POFI might also increase accuracy, but again at the cost of computation time. The biggest challenge of these methods that remains is to extend them to higher dimensional systems.

## BIBLIOGRAPHY


[1] O. Cappe, Moulines E., and Ryden T. *Inference in Hidden Markov Models*. Springer, 2005.

[2] P. Haccou, P. Jagers, and V.A. Vatutin. *Branching processes: Variation, growth, and extinction of populations*, volume 5. Cambridge Univ Pr, 2005.

[3] G. Hooker. Forcing function diagnostics for nonlinear dynamics. *Biometrics*, 65:613–620, 2009.

[4] G. Hooker. Incarnations of the rosensweig macarthur model. 2014.

[5] G. Hooker, K. K. Lin, and B. Rogers. Control theory and experimental design in diffusion processes. Unpublished, Department of Biological Statistics and Computational Biology, Cornell University. 2012.

[6] G.E. Monahan. A survey of partially observable markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982.

[7] W. Powell. *Approximate Dynamic Programming: solving the curses of dimensionality*. Wiley, 2007.

[8] M.L. Puterman. *Markov Decision Processes - Discrete Stochastic Dynamic Programming*. Wiley, Hoboken, NJ, 2005.

[9] J. Shachat, J. T. Swarthouty, and L. Wei. Man versus nash: An experiment on the self-enforcing nature of mixed strategy equilibrium. Unpublished, Wang Yanan Institute for Studies in Economics, Xiamen University. 2011.

[10] Gregory Smith. Modeling the stochastic gating of ion channels. In *Computational Cell Biology*, volume 20 (II) of *Interdisciplinary Applied Mathematics*. 2002.

[11] D. Terman and B. Ermentrout. *Mathematical Foundations of Neuroscience*. Springer, 2010.