PH.D. THESIS

Product Estimators for Hidden Markov Models

by Vahid Reza Ramezani
Advisor: Steven Marcus

PhD 2001-7

INSTITUTE FOR SYSTEMS RESEARCH

ABSTRACT

Title of Dissertation: PRODUCT ESTIMATORS FOR

HIDDEN MARKOV MODELS

Vahid Reza Ramezani, Doctor of Philosophy, 2001

Dissertation directed by: Professor S. I. Marcus
Department of Electrical and Computer Engineering

In this thesis, risk-sensitive estimation for Hidden Markov Models is studied from a dynamical systems point of view. We show that risk-sensitive estimators belong to a broader class of *product estimators* in which risk-sensitivity will be shown to be related to certain *scaling functions*. The product structure and the scaling functions perspective give us new insights into the underlying mechanism of risk-sensitive estimation. For the first time, in a series of theorems and examples, we relate risk-sensitivity to the dynamics of the underlying process and expose relations among the transition probabilities, risk-sensitivity and the *decision regions*. We introduce the risk-sensitive Maximum A Posterior Probability (MAP) criterion for HMM's with discrete range observation. This criterion is the discrete

time finite dimensional version of the classic risk-sensitive estimation problem for linear/quadratic partial observation case.

The risk-sensitive filters take into account the "higher order" moments of the of the estimation error. In the context of risk-sensitive MAP for HMM's, we clarify and quantify the influence of risk-sensitivity on the behavior of the sample paths of the estimator; the product structure representation will play an important role.

PRODUCT ESTIMATORS FOR

HIDDEN MARKOV MODELS

by

Vahid Reza Ramezani

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2001

Advisory Committee:

      Professor S. I. Marcus, Chairman
      Professor Eyad Abed
      Professor Rama Chellappa
      Professor Joel Cohen
      Professor William Levine

PREFACE

Beh Name Khodavande Jan o Kherad

DEDICATION

To my father

ACKNOWLEDGEMENTS

All my teachers from M. Zargar to S. Marcus.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The exponential (risk-sensitive) criterion of a quadratic function of the state and control for full state control was first proposed by Jacobson [5]. Whittle [15] produced the controller for the linear/quadratic partial observation case which required a state estimator *separated* from the control policy in a fashion analogous to the policy separation for the partial observation in the risk-neutral case. Speyer [9] treated the estimation problem for the linear/quadratic partial observation case and showed that a linear estimator is optimal among all non-linear and linear estimators. The non-linear discrete-time stochastic problem for the partially observed control problem was solved by a change of measure technique [7]. This so called *reference probability* approach was later used by Dey and Moore [10] to solve an estimation problem for a *partially observed Markov chain* or, as it is commonly referred to in signal processing literature, a *Hidden Markov Model* (HMM).

In its essence, our work can be traced back to the Speyer's paper [9] nearly a decade ago and is related to the Dey-Moore filter [10]. We are also interested in the estimation problem with the exponential criterion for HMM's. However, we have a different perspective; we view the estimator as a dynamical system whose

dynamics are inherited from the Markov chain through the partial observation and an optimization criterion. We are not only interested in the computation of the optimal estimator for an HMM and its properties for exponential criterion, but also in the *qualitative analysis* of its sample paths. Under *perfect observation* the dynamics of the estimator revert back to the Markov chain itself.

We show that risk-sensitive estimators belong to the broader class of *product estimators* in which risk-sensitivity will be shown to be related to certain *scaling functions*. The product structure and the scaling perspective will give us new insights into the underlying mechanism of risk-sensitive estimators. For the first time, in a series of theorems and examples, we relate risk-sensitivity to the dynamics of the underlying process and show that elegant relations exist among the transition probabilities, risk-sensitivity and the *decision regions*. Several problems for future research will be suggested.

In this chapter we will give a broad view of what this thesis is about and how it fits in the context of the existing literature. Next we will present a brief background of the problem starting with Speyer's linear/quadratic Gaussian problem.

1.1 Background

1.1.1 Speyer's problem

Consider the following linear stochastic discrete-time system

$$
\begin{cases}
X_{t+1} = AX_{t+1} + w_t; \\
Y_t = HX_t + v_t,
\end{cases}
$$

where all variables take their values in R^n, the initial distribution of the state

X_0 is zero-mean unit variance and $\{w_t, v_t\}$ is assumed to be zero-mean jointly Gaussian independent random variables. The objective is to find $\widehat{X_i}$ $i = 0, 1, ..., N$ (measurable functions of the observations $Y_0, ...Y_i$) such that the following criterion is minimized

$$J = E[exp(\theta \sum_{t=0}^{N} (X_t - \widehat{X_t})^T Q^T (X_t - \widehat{X_t})],$$

where θ is a constant scalar, Q is a positive semi-definite matrix and the expectation is over all underlying random variables.

Speyer showed that the estimator which minimizes the above is linear but *not* a conditional mean estimator such as the Kalman filter, i.e., the *information state* of the estimator which condenses the sequence of past observations and determines the value of the estimates is not the conditional mean. Speyer showed that the *value functions,* i.e., the optimal value of the cost function $J(N)$ for all values of N, can be calculated by backward dynamic programming. From this backward dynamic programming, the optimal estimates were determined. Speyer used ideas developed earlier by Whittle [15] to simplify these calculations.

1.1.2 Reference probability method and Dey-Moore filter

The reference probability method or the *Girsanov change of measure* technique was first used to simplify the calculations of the non-linear filtering problem for continuous time stochastic processes. The basic idea, as Marcus puts it in [4], is that of defining a reference probability measure such that under this new measure the observation process is transformed into a "simpler" process. A version of Bayes formula shows that the conditional expectation under the original measure can be computed easily from the conditional expectation under the new measure.

These methods were developed for a finite dimensional state space in [6] and were used in [7][3][16][17]. In parallel development to the continuous case, a new measure was introduced under which the process could be "simplified".

In [10] this change of measure technique was applied to the risk-sensitive HMM's with continuous range observation. We will give a brief description of the problem formulation and its solution.

The objective is to find an estimator $\widehat{X_t}$ of the state X_t such that $\widehat{X_t}$ is a function of the observation up to t and minimizes

$$J_t = E[exp(\theta \Psi_{0,t})],$$

where

$$\Psi_{0,t} = \widehat{\Psi}_{0,t-1} + 1/2(X_t - \widehat{X_t})^T Q^T (X_t - \widehat{X_t})$$

and

$$\widehat{\Psi}_{m,n} := 1/2 \sum_{i=m}^{n} (X_i - \widehat{X_i})^T Q^T (X_i - \widehat{X_i}).$$

The difference with Speyer's problem definition (aside from the obvious difference of discrete vs. continuous state space) is that the minimization is performed over an *incremented horizon*. The estimates are defined recursively as we move forward in time. This resembles the standard *risk-neutral* filtering for the Markov chain where the filtering process does not involve dynamic programming and is carried out forward in time. Dey-Moore filtering methodology is claimed, therefore, to be more "natural" than the Speyer's approach [10]. We will see in this thesis that both these filters are special cases of a more general filter and that both are "natural" in different contexts. We will see that under certain conditions the two coincide.
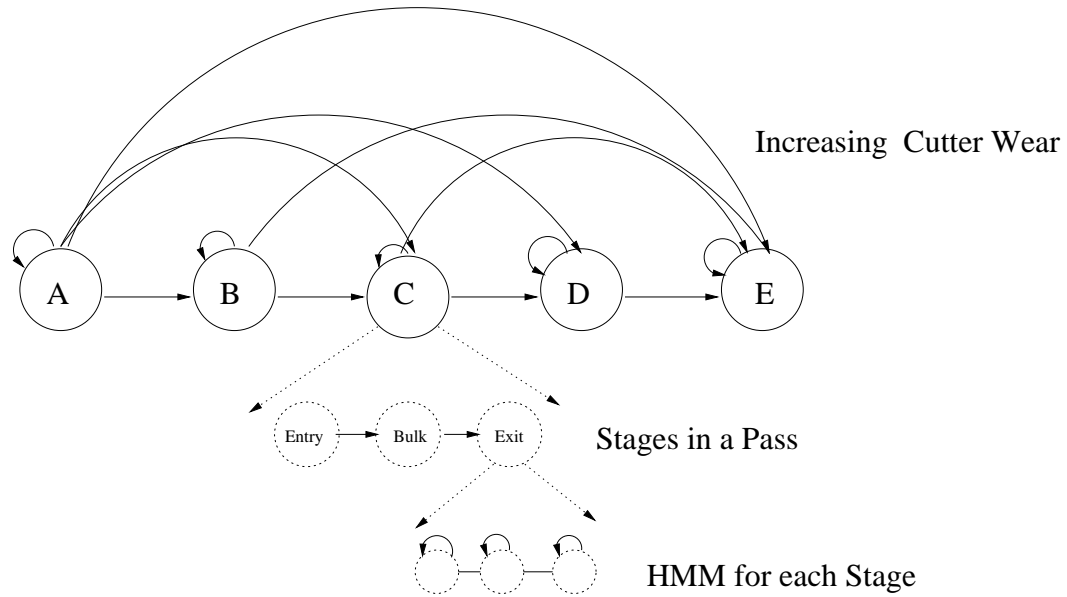
Figure 1.1: The progressive nature of the metal cutting process is modeled at several levels.

1.2 The objective and organization of the thesis

The theoretical development of this thesis was in part initiated by a basic question posed by our collaborating colleagues from the University of Washington in Seattle who were working on the problem of failure detection of certain cutting tools [1] (See Figure 1.1). They had developed a left to right *multi-scale* HMM representing the cutting process and were interested to know if a risk-sensitive estimator could be constructed for which risk-sensitivity was "scale dependent" [2]. (In Figure 1.1, the progressive nature of the metal cutting process is modeled at several levels. Progressive wear is modeled as a left to right Markov process constrained to only allow increasing levels of wear. The progress of a cutter through a single pass is also modeled as a left to right process, composed of sequences of HMM states.)

In other words, they were interested in risk-sensitive estimators for which risk-

sensitivity is not controlled by a parameter but is a dynamic function of the "scales". This question is a special case of the broader theoretical problem of constructing "state dependent" risk-sensitive estimators since the multi-scale representation is nothing but the aggregation of the states of the Markov chain. After constructing such state dependent risk-sensitive filters in the broader context of *product estimators* and in attempting to understand their behavior, we realized that an important problem had received little notice in the literature. How do the underlying dynamics of the process interact with risk-sensitivity to determine the dynamics of the estimated process? Perfect observation produces trivial results, but partial observation produces non-trivial results.

We deliberately postponed the connection between the *risk-sensitive* and the classic *risk-neutral* filters to this section. This connection is viewed differently in this thesis from the existing literature.

It is often said that risk-sensitive filters take into account the "higher order" moments of the of the estimation error. Roughly speaking, this follows from the analytic property of the exponential $e^x = \sum_{k=0}^{\infty} x^k/k!$ so that if Ψ stands for the sum of the error functions over some time interval then

$$E[exp(\theta\Psi] = E[1 + \theta\Psi + (\theta)^2(\Psi)^2/2 + ...].$$

Thus, higher order moments of the error are included in the minimization of the expected cost. At the expense of the "mean error cost", the variance and other higher order moments are considered and minimized, reducing the "risk" of large deviations and increasing our "confidence" in the estimator. The parameter $\theta > 0$ controls the extent to which the higher order moments are included. In particular, the first order approximation $\theta \to 0$ given by $E[exp(\theta\Psi] \cong 1 + \theta E\Psi$ indicates that

the minimization of the sum criterion or the risk neutral problem is recovered as the small risk limit of the exponential criterion. Thus, the exponential criterion is a more general problem and includes the the risk-neutral one as a limiting result.[1] There are also indications that risk-sensitive estimators have certain robustness properties when uncertainties exist in the description of the underlying model. The interested reader is referred to [20]-[23].

The exponential function, however, includes the higher order moments in a special way, according to the expansion $\sum_{k=0}^{\infty} x^k/k!$. This gives the exponential function a unique algebraic property: the product of exponential components is equal to the exponential of the sum of the exponents. What kind of estimators can we construct which are in some sense a generalization of the sum criterion if we begin with the algebraic property of the exponential?

In this thesis, we study these estimators in the context of HMM's, and show that once we begin with the product criterion, of the analytic property of the exponential, we only need the continuity to unity at zero and the strict monotonicity of the exponential function. Then, the risk-sensitivity parameter is replaced by certain scaling functions that scale the estimation metric along the sample paths of the estimator. Also, as a special case, we will obtain the Dey-Moore filter with the continuous range observation and quadratic criterion in Chapter 4.

In Chapter 3, we will consider discrete range observation with the Maximum A Posterior Probability (MAP) criterion. Later on, we will see that this is the "natural" criterion for the risk-sensitive HMM's because of its symmetric characteristics. In that chapter, the product estimators are studied in several examples.

---

[1]The case of $\theta < 0$ corresponding to the *risk-seeking* estimators will not be considered in this thesis.

We show that product estimators allow for *dynamic* risk-sensitivity and attempt to understand the implication of this property. We study how risk-sensitivity parameter and the transition probabilities (which determine the underlying dynamics) are coupled in the behavior of the estimator. We consider the *multi-scale* representation of the Markov chains and how this representation is dependent on the risk-sensitivity parameter. This, in turn, suggests a coupling of the risk-sensitivity parameter, the underlying dynamics and the availability of information. These examples will lead to several theorems and a proposition in Chapter 5.

In Chapter 5, only the exponential criterion is considered and the Speyer and Dey-Moore filtering ideas are unified by the introduction of *risk-sensitive filter banks.* The properties of the sample paths of these filter banks are described. Several problems in that direction are suggested in Chapter 6.

In Chapter 2 (preliminaries), our notation and the Girsanov's change of measure for HMM's for the *zero delay* model are established.

Chapter 2

Preliminaries

In this chapter, our basic notation for the simplex representation of Markov chains and the *discrete-time Girsanov change of measure* are established. We present a *zero delay* model; i.e., as soon as a transition to a state is made, an observation of that state becomes available. The sources for this chapter are [6][7][25]. For a *non-zero delay* model, the reader is referred to [6].

In section 2.1, we begin with the *simplex representation of Markov chains* which leads to a pair of discrete state space equations in (2.15). This is accomplished by introducing *the martingale increments* (2.6) and (2.12) for the state and observation process of the Markov chain.

In section 2.2, the change of measure technique and the *conditional Bayes theorem* (Theorem 2.2.4) are developed. The chapter ends with Lemma 2.2.10 where we prove the martingale increment property (2.64) under the new measure. In proving Lemma 2.2.10, we show that the martingale property under the new measure is directly inherited from the martingale property under the old.

2.1 The simplex representation of Markov chains

All time series below are defined on a common probability space $(\mathbf{\Omega}, \mathcal{F}, \mathbf{P})$. Without loss of generality, we assume that the discrete-time Markov process X_t takes its values in the set of $N_{\mathbf{X}}$ elements

$$\mathbf{S}_{N_{\mathbf{X}}} = \{e_1, ..., e_{N_{\mathbf{X}}}\},$$

where $e_i's$ are the unit vectors in $R^{N_{\mathbf{X}}}$. (See Figure 2.1.)

Let $\mathcal{F}_t^0 = \sigma\{X_0, ..., X_t\}$ be the σ-algebra generated by the sequence $X_0, ..., X_t$ and $\mathcal{F}_t$ its completion. The Markov property implies

$$P(X_{t+1} = e_j | \mathcal{F}_t) = P(X_{t+1} = e_j | X_t). \tag{2.1}$$

Define the transition probability matrix $\mathcal{A}$ by

$$\mathcal{A} = (a_{ij}), \tag{2.2}$$

$$P(X_{t+1} = e_j | X_t = e_i) = a_{ij}. \tag{2.3}$$

Note that if X is a unit vector then

$$E[< X, e_i >] = \sum_{j=1}^{N_{\mathbf{X}}} < e_j, e_i > P(X = e_j) = P(X = e_i). \tag{2.4}$$

Therefore, we can write

$$E[X_{t+1} | \mathcal{F}_t] = E[X_{t+1} | X_t] = \mathcal{A}^T X_t, \tag{2.5}$$

where $\mathcal{A}^T$ denotes the transpose of $\mathcal{A}$.

Definition 2.1.1

$$Z_{t+1} := X_{t+1} - \mathcal{A}^T X_t. \tag{2.6}$$
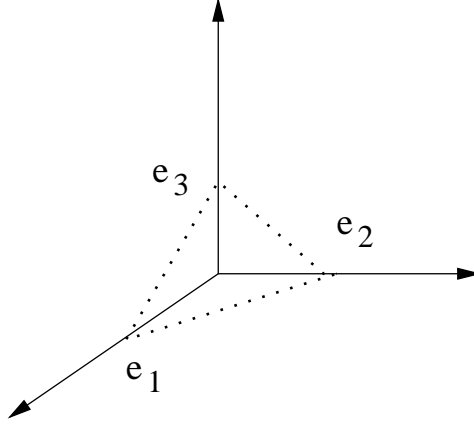
Figure 2.1: The simplex in three dimensions.

We have

$$X_{t+1} = \mathcal{A}^T X_t + Z_{t+1}.$$

Next, we show that $\{Z_t\}$ is a sequence of martingale increments. First, note that $E[\mathcal{A}^T X_t | X_t] = \mathcal{A}^T X_t$. It follows that

$$E[Z_{t+1}|\mathcal{F}_t] = E[X_{t+1} - \mathcal{A}^T X_t | X_t] = \mathcal{A}^T X_t - \mathcal{A}^T X_t = 0, \tag{2.7}$$

which means $\{Z_t\}$ is a sequence of martingale increments.

Suppose that the observation space is finite dimensional and the observation sequence is given by

$$Y_t = c \ (X_t, w_{t+1}),$$

where w_t are i.i.d. random variables, with Z_t and w_t mutually independent. $c(.,.)$ is a Borel measurable function. Let $\mathcal{Y}_t$ be the completion of $\sigma\{Y_0, ..., Y_t\}$, and $\mathcal{G}_t$ the completion of $\sigma(X_0, ..., X_t, Y_1, ...Y_{t-1})$, the σ-algebra generated by the state and observation processes up to time t and t-1 respectively. Clearly $\mathcal{G}_t \subset \mathcal{G}_{t+1}...$, $\mathcal{Y}_t \subset \mathcal{Y}_{t+1}...$ and so each σ-field forms an increasing family or a *filteration*. We say a function f is $\mathcal{G}_t$-measurable if it is a function of $X_0, ..., X_t, Y_1, ...Y_{t-1}$.

11

The pair (X_t, Y_t) is our basic example of a Hidden Markov Model or, as in the control literature, a partially observed Markov chain. Both terminologies refer to the indirect observation of the state. The observation is given by a zero delay model, i.e., the observation Y_t is gathered from the state X_t without delay. In this thesis, we only consider the homogeneous HMM's, i.e., the matrix $\mathcal{A}$ and the function $c(.,.)$ do not depend on the time parameter t.

Suppose, without loss of generality, that $c(.,.) \in \mathbf{S}_{N_{\mathbf{Y}}}$ where $\mathbf{S}_{N_{\mathbf{Y}}} = \{e_1, ..., e_{N_{\mathbf{Y}}}\}$ and $e_i's$ are the unit vectors in $R^{N_{\mathbf{Y}}}$. We have

$$P(Y_t = e_j | X_0, X_1, ... X_t, Y_0, ... Y_{t-1}) = P(Y_t = e_j | X_t). \tag{2.8}$$

Write

$$\mathcal{C} = (c_{ij}) \tag{2.9}$$

$$P(Y_t = e_j | X_t = e_i) = c_{ij}, \qquad \sum_{i=1}^{N_{\mathbf{Y}}} c_{ij} = 1. \tag{2.10}$$

Thus, we have

$$E[Y_t | X_t] = \mathcal{C}^T X_t. \tag{2.11}$$

Define

$$W_{t+1} := Y_t - \mathcal{C}^T X_t. \tag{2.12}$$

First, note that $E[\mathcal{C}^T X_t | X_t] = \mathcal{C}^T X_t$. It follows that

$$E[W_{t+1} | \mathcal{G}_t] = E[Y_t - \mathcal{C}^T X_t | X_t] = \mathcal{C}^T X_t - \mathcal{C}^T X_t = 0, \tag{2.13}$$

so W_t is a $\mathcal{G}_t$ martingale increment and

$$Y_t = \mathcal{C}^T X_t - W_{t+1}.$$

Since w_t are i.i.d. and mutually independent of Z_t, the W_t are conditionally independent of Z_t given $\mathcal{G}_t$.

Definition 2.1.2 $Y_t^i := <Y_t, e_i>$ and $c_t := (c_t^1, ..., c_t^{\mathbf{N_Y}})^T$ where $c_t^i := E[Y_t^i|\mathcal{G}_t] = \sum_{j=1}^{N_{\mathbf{X}}} cji <e_j, X_t>$. Then

$$c_t = E[Y_t|\mathcal{G}_t] = \mathcal{C}X_t. \tag{2.14}$$

$$\sum_{i=1}^{N_{\mathbf{Y}}} c_t^i = 1 \quad \sum_{i=1}^{N_{\mathbf{Y}}} Y_t^i = 1$$

Lemma 2.1.3 Let $diag(V)$ be the diagonal matrix whose non-zero elements along its diagonal are given by the elements of V, i.e., $(diag(V))_{ii} = Z_{ii}$. Then

$$\begin{aligned} Z_{t+1}Z_{t+1}^T = \quad & diag(\mathcal{A}^T X_t) + diag(Z_{t+1}) - \mathcal{A}^T diag(X_t \mathcal{A}) \\ & - \mathcal{A}^T X_t Z_{t+1}^T - Z_{t+1}(\mathcal{A}^T X_t)^T \end{aligned}$$

and

$$\begin{aligned} <Z_{t+1}> \quad := \quad & E[Z_{t+1}(Z_{t+1})^T|\mathcal{F}_t] \\ = \quad & E[Z_{t+1}(Z_{t+1})^T|X_t] \\ = \quad & diag(\mathcal{A}^T X_t) - \mathcal{A}^T diag(X_t)\mathcal{A}. \end{aligned}$$

Proof:

By definition of Z_{t+1}, we have

$$X_t X_t^T = \mathcal{A}^T X_t (\mathcal{A}^T X_t)^T + \mathcal{A}^T X_t Z_{t+1}^T + Z_{t+1}(\mathcal{A}^T X_t)^T + Z_{t+1}(Z_{t+1})^T,$$

but

$$X_t X_t^T = diag(X_t) = diag(\mathcal{A}^T X_t) + diag(Z_{t+1}),$$

and the first equality follows. By the martingale increment property and conditioning on X_t, we have

$$<Z_{t+1}> = E[Z_{t+1}(Z_{t+1})^T|X_t] = diag(\mathcal{A}^T X_t) - \mathcal{A}^T diag(X_t)\mathcal{A}.$$

This finishes the proof of the lemma.

Similarly, we can show

$$< W_{t+1} >:= E[W_{t+1}(W_{t+1})^T|\mathcal{G}_t] = diag(\mathcal{C}^T X_t) - \mathcal{C}^T diag(X_t)\mathcal{C}.$$

In summary then, we have developed the following state space representation of the Hidden Markov Models

$$\begin{cases} X_{t+1} = \mathcal{A}^T X_t + Z_{t+1} \\ Y_t = \mathcal{C}^T X_t + W_{t+1} \end{cases} \tag{2.15}$$

$$X_t \in \mathbf{S}_{N_{\mathbf{X}}}, \qquad Y_t \in \mathbf{S}_{\mathbf{Y}}, \qquad t = 0, 1, ...,$$

which we will refer to as the *discrete state space representation*. The matrix entries satisfy

$$\sum_{i=1}^{N_{\mathbf{X}}} a_{ij} = 1, \quad \sum_{i=1}^{N_{\mathbf{Y}}} c_{ij} = 1.$$

Z_t and W_t are martingale increments satisfying

$$E[Z_{t+1}|\mathcal{F}_t] = 0, \qquad E[W_{t+1}|\mathcal{G}_t] = 0,$$

$$< Z_{t+1} >:= E[Z_{t+1}(Z_{t+1})^T|X_t] = diag(\mathcal{A}^T X_t) - \mathcal{A}^T diag(X_t)\mathcal{A}.$$

$$< W_{t+1} >:= E[W_{t+1}(W_{t+1})^T|\mathcal{G}_t] = diag(\mathcal{C}^T X_t) - \mathcal{C}^T diag(X_t)\mathcal{C}.$$

Note that Usually a Hidden Markov Model is defined as a four tuple $< \mathbf{X}, \mathbf{Y}, \mathcal{A}, \mathcal{C} >$ where $\mathcal{A}$ is the transition matrix, $\mathbf{Y} = \{1, 2, ..., N_{\mathbf{Y}}\}$ is the set of observations, $\mathbf{X} = \{1, 2, ..., N_{\mathbf{X}}\}$ is the finite set of the states and $\mathcal{C} := [c_{i,j}]$ is the $N_{\mathbf{X}} \times N_{\mathbf{Y}}$ state/observation matrix with $P(Y_t = j|X_t = i) = c_{i,j}$. We can project $\mathbf{X} = \{1, 2, ..., N_{\mathbf{X}}\}$ onto $\mathbf{S}_{N_{\mathbf{X}}} = \{e_1, ..., e_{N_{\mathbf{X}}}\}$ and $\mathbf{Y} = \{1, 2, ..., N_{\mathbf{Y}}\}$ onto $\mathbf{S}_{N_{\mathbf{Y}}} = \{e_1, ..., e_{N_{\mathbf{Y}}}\}$. Thus, without loss of generality, we can work in $R^{N_{\mathbf{X}}}$ and $R^{N_{\mathbf{Y}}}$ and utilize the results of this chapter to carry out our computations.

2.2 Change of measure.

Throughout this dissertation, we will use a discrete-time version of Girsanov's theorem. For an accessible treatment of the continuous time Girsanov's theorem, the reader is referred to [4] and [6]. A detailed development of the discrete-time version will be given below. The basic idea is to introduce a new measure $P^\dagger$ such that under $P^\dagger$ the observations have certain i.i.d. properties. The computations of the conditional expectations and their optimization will be greatly simplified under the new measure. The results are then projected back to the old measure by an inverse change of measure technique.

Define

$$\lambda_l = \prod_{i=1}^{N_\mathbf{Y}} (\frac{1}{N_\mathbf{Y} c_l^i})^{Y_l^i}, \tag{2.16}$$

$$\Lambda_t = \prod_{l=0}^{t} \lambda_l. \tag{2.17}$$

Note that in general, with U one of the unit vectors in R^N, any real valued function $f(U) = (f_1, ..., f_N)$ can be written as

$$f(U) = \sum_{i=1}^{N} f(e_i) X^i = \sum_{i=1}^{N} f_i U^i \quad f(e_i) = f_i :=< f, e_i > . \tag{2.18}$$

Furthermore, $Y_l^i = 1$ for only one i at each l, and $Y_l^i = 0$ otherwise, so that λ_t is the product of unity terms and non-unity terms, so that by (2.18) for $f = \lambda_t$, we can write

$$\lambda_t = \lambda_t(Y_t) = \sum_{i=1}^{N_\mathbf{Y}} Y_t^i / (N_\mathbf{Y} c_t^i).$$

Lemma 2.2.1 With the above definitions,

$$E[\lambda_t | \mathcal{G}_t] = 1. \tag{2.19}$$

15

Proof:

By the above,

$$E[\lambda_t|\mathcal{G}_t] \;=\; E[\;\prod_{i=1}^{N_\mathbf{Y}} (\frac{1}{N_\mathbf{Y} c_t^i})^{Y_t^i} \;|\mathcal{G}_t] \tag{2.20}$$

$$=\; E[\;\sum_{i=1}^{N_\mathbf{Y}} \frac{1}{N_\mathbf{Y} c_t^i} Y_t^i \;|\mathcal{G}_t] \tag{2.21}$$

$$=\; \sum_{i=1}^{N_\mathbf{Y}} \frac{1}{N_\mathbf{Y} c_t^i} E[\;Y_t^i \;|\mathcal{G}_t] \tag{2.22}$$

$$=\; 1/N_\mathbf{Y} \sum_{i=1}^{N_\mathbf{Y}} \frac{1}{c_t^i} \; c_t^i = 1. \tag{2.23}$$

Definition 2.2.2 For a sequence of σ-algebras $\mathcal{G}_0, \mathcal{G}_1, ...,$ denote by $\bigvee_{l=1}^{\infty} \mathcal{G}_l$ the σ-algebra generated by their union.

Define a new measure $P^\dagger$ on $(\Omega, \bigvee_{l=1}^{\infty} \mathcal{G}_t)$ by setting the restriction of the Radon-Nikodym derivative $\frac{dP^\dagger}{dP}$ to the sigma algebra $\mathcal{G}_{t+1}$ equal to Λ_t

$$\frac{dP^\dagger}{dP}|\;\mathcal{G}_{t+1} = \Lambda_t. \tag{2.24}$$

In the integral representation, this means

$$P^\dagger(B) = \int_B \Lambda_t dP \quad B \in \mathcal{G}_{t+1}. \tag{2.25}$$

That $P^\dagger$ is a measure on $(\Omega, \bigvee_{l=1}^{\infty} \mathcal{G}_l)$ follows from Kolmogorov's extension theorem (Appendix A). First, recall that $\{\mathcal{G}_l\}$ is a nested family. To apply the theorem, we must show that the family of distributions defined by the above integral is *compatible* [25] or equivalently

Lemma 2.2.3

$$\int_B \Lambda_t dP = \int_B \Lambda_{t-1} dP \quad B \in \mathcal{G}_t. \tag{2.26}$$

Proof:

16

Let I_S be the indicator function of a set S. Then

$$\int_B \Lambda_t dP \quad = \quad E[I_B \Lambda_t] \qquad (2.27)$$

$$= \quad E\{E[I_B \Lambda_t | \mathcal{G}_t]\} \qquad (2.28)$$

$$= \quad E\{I_B \Lambda_{t-1} E[\lambda_t | \mathcal{G}_t]\} \qquad (2.29)$$

$$= \quad E[I_B \Lambda_{t-1}] = \int_B \Lambda_{t-1} dP. \qquad (2.30)$$

Also note that

$$\int_\Omega \Lambda_t dP \quad = \quad E[\Lambda_t]$$

$$= \quad E[\Lambda_0] = E[\lambda_0] = 1$$

which implies $P^\dagger$ is a probability measure. The other conditions for applying Kolmogorov's extension theorem are easily verified.

Note that for any $\mathcal{G}_t$ measurable function f

$$E^\dagger[f] = \int f dP^\dagger = \int f \frac{dP^\dagger}{dP} dP = \int f \Lambda_{t-1} dP = E[\Lambda_{t-1} f]. \qquad (2.31)$$

An important observation is that the change of measure defined by

$$\Lambda_{t+l} = \frac{dP^\dagger}{dP} \Big| \ \mathcal{G}_{t+1} \subset \mathcal{G}_{t+1+l} \qquad (2.32)$$

(for $\forall l \geq 1$) is equivalent to the above change of measure ($l = 0$) because of the property $\int_B \Lambda_t dP = \int_B \Lambda_{t-1} dP \ B \in \mathcal{G}_t$. This does not violate the uniqueness clause of the Radon-Nikodym theorem since we are viewing $\mathcal{G}_{t+1}$ as a subset of $\mathcal{G}_{t+l}$ and as such the Radon-Nikodym derivative on $\mathcal{G}_{t+1}$ is Λ_{t+l}. The uniqueness clause requires that the Radon-Nikodym derivative be measurable; and, for Λ_{t+l} to be measurable with respect to $\mathcal{G}_{t+1}$, we must set $l = 0$. This instructive result demonstrates that the uniqueness clause of the theorem should be applied with care. In chapter 5, for computational reasons, we will let l=1.

Next, we state and prove the Conditional Bayes Theorem:

Theorem 2.2.4 Suppose $P^\dagger$ and P are probability measures on a measurable space $(\Omega, \mathcal{M})$ with $\frac{dP^\dagger}{dP} = \Lambda$ and $P^\dagger$ absolutely continuous with respect to P. Suppose $\mathcal{G} \subset \mathcal{M}$ is a sub-sigma algebra. Then, if f is is any $dP^\dagger$ integrable random variable

$$E^\dagger[f|\mathcal{G}_t] = \psi$$

where

$$\psi = \begin{cases} \frac{E[\Lambda f|\mathcal{G}]}{E[\Lambda|\mathcal{G}]} & \text{if } E[\Lambda|\mathcal{G}] > 0; \\ 0 & otherwise. \end{cases}$$

Proof:

Define $\psi = E[\Lambda f|\mathcal{G}]/E[\Lambda|\mathcal{G}]$ if $E[\Lambda|\mathcal{G}] > 0$ and $\psi = 0$ otherwise. We will show $E^\dagger[f|\mathcal{G}_t] = \psi$. This is equivalent to proving $\int_A E^\dagger[f|\mathcal{G}_t]dP^\dagger = \int_A \psi dP^\dagger$ for every set $A \in \mathcal{G}$. Define $G = \{\omega : E[\Lambda|\mathcal{G}] = 0\}$, so $G \in \mathcal{G}$. Then $\int_G E[\Lambda|\mathcal{G}]dP = 0 = \int_G \Lambda dP$ and $\Lambda \geq 0$ a.s. So, either P(G)=0 or the restriction of Λ to G is 0 a.s. In either case, $\Lambda = 0$ a.s. on G. Let $G^c = \{\omega : E[\Lambda|\mathcal{G}] > 0\}$. Suppose $A \in \mathcal{G}$, then $A = B \cup C$ where $B = A \cap G^c$ and $C = A \cap G$. Furthermore,

$$\begin{aligned} \int_A E^\dagger[f|\mathcal{G}_t]dP^\dagger &= \int_A f dP^\dagger = \int_A f\Lambda dP \\ &= \int_B f\Lambda dP + \int_C f\Lambda dP. \end{aligned}$$

But, $\Lambda = 0$ a.s. on $C \subset G$, so

$$\int_C f\Lambda dP = 0 = \int_C \psi dP^\dagger$$

by definition.

We have

$$\begin{aligned} \int_B \psi dP^\dagger &= \int_B \frac{E[\Lambda f|\mathcal{G}]}{E[\Lambda|\mathcal{G}]}dP^\dagger & (2.33) \\ &= E^\dagger\{I_B \frac{E[\Lambda f|\mathcal{G}]}{E[\Lambda|\mathcal{G}]}\} & (2.34) \end{aligned}$$

18

$$= E\{\Lambda I_B \frac{E[\Lambda f|\mathcal{G}]}{E[\Lambda|\mathcal{G}]}\} \tag{2.35}$$

$$= E\{E\{\Lambda I_B \frac{E[\Lambda f|\mathcal{G}]}{E[\Lambda|\mathcal{G}]}|\mathcal{G}\}\} \tag{2.36}$$

$$= E\{I_B E[\Lambda|\mathcal{G}] \frac{E[\Lambda f|\mathcal{G}]}{E[\Lambda|\mathcal{G}]}\} \tag{2.37}$$

$$= E[I_B E[\Lambda f|\mathcal{G}]] \tag{2.38}$$

$$= E[I_B \Lambda f]. \tag{2.39}$$

We have proved

$$\int_B \Lambda f dP = \int_B \psi dP^\dagger$$

Thus,

$$\int_C \Lambda f dP + \int_B \Lambda f dP = \int_A \Lambda f dP$$
$$= \int_A E^\dagger[f|\mathcal{G}]dP^\dagger = \int_A \psi dP^\dagger$$

and the theorem follows.

Definition 2.2.5 A sequence $\{f_t\}$ is said to be $\mathcal{G}$ adapted if each f_t is $\mathcal{G}_t$ measurable. We have the following result:

Lemma 2.2.6 If $\{f_t\}$ is a $\mathcal{G}$ adapted sequence of random variables, then

$$E^\dagger[f_t|\mathcal{Y}_{t-1}] = \frac{E[\Lambda_{t-1} f|\mathcal{Y}_{t-1}]}{E[\Lambda_{t-1}|\mathcal{Y}_{t-1}]} \tag{2.40}$$

We now state and prove an important independence property of the sequence of the observations under the new measure.

Theorem 2.2.7 Under $P^\dagger$

$$P^\dagger(Y_t^j = 1|\mathcal{G}_t) = 1/N_\mathbf{Y}. \tag{2.41}$$

This implies that Y_t is a sequence of i.i.d. random variables, each with a uniform distribution $1/N_\mathbf{Y}$.

Proof:

$$
\begin{aligned}
P^\dagger(Y_t^j = 1|\mathcal{G}_t) &= E^\dagger[< Y_t, e_j > |\mathcal{G}_t] &\text{(2.42)} \\
&= \frac{E[\Lambda_t < Y_t, e_j > |\mathcal{G}_t]}{E[\Lambda_t|\mathcal{G}_t]} &\text{(2.43)} \\
&= \frac{\Lambda_{t-1} E[\lambda_t < Y_t, e_j > |\mathcal{G}_t]}{\Lambda_{t-1} E[\lambda_t|\mathcal{G}_t]} &\text{(2.44)} \\
&= E[\lambda_t < Y_t, e_j > |\mathcal{G}_t] &\text{(2.45)} \\
&= E[\prod_{i=1}^{N_\mathbf{Y}} (\frac{1}{N_\mathbf{Y} c_t^i})^{Y_t^i} < Y_t, e_j > |\mathcal{G}_t] &\text{(2.46)} \\
&= E[\sum_{i=1}^{N_\mathbf{Y}} (\frac{1}{N_\mathbf{Y} c_t^i}) Y_t^i Y_t^j |\mathcal{G}_t] &\text{(2.47)} \\
&= \frac{1}{N_\mathbf{Y} c_t^j} E[Y_t^j|\mathcal{G}_t] &\text{(2.48)} \\
&= \frac{1}{N_\mathbf{Y} c_t^j} c_t^j = 1/N_\mathbf{Y} = P^\dagger(Y_t^j = 1), &\text{(2.49)}
\end{aligned}
$$

which is a quantity independent of $\mathcal{G}_t$.

Observe that $E^\dagger[X_{t+1}|\mathcal{G}_t] = E[\Lambda_t X_{t+1}|\mathcal{G}_t]/E[\Lambda_t|\mathcal{G}_t] = E[\lambda_t X_{t+1}|\mathcal{G}_t] = \mathcal{A}^T X_t$ so that under the new measure $P^\dagger$, the process X_t remains a Markov chain with the transition matrix $\mathcal{A}$.

We now begin with a probability measure $P^\dagger$ on $(\Omega, \bigvee_{n=1}^\infty \mathcal{G}_n)$ such that

1) The process $\{X_t\}$ is a finite-state Markov chain with transition matrix $\mathcal{A}$ and

2) $\{Y_t\}$ is sequence of i.i.d. random variables and

$$
P^\dagger(Y_t^j = 1|\mathcal{G}_t) = P^\dagger(Y_t^j = 1) = 1/N_\mathbf{Y}.
$$

Suppose $\mathcal{C} = (c_{ij})$ is a matrix such that $c_{ij} > 0$ and $\sum_{j=1}^{N_\mathbf{Y}} c_{ij} = 1$.

We will construct a new measure P^* on $(\Omega, \bigvee_{n=1}^\infty \mathcal{G}_n)$ such that under P^*, the *discrete state space representation* still holds and $E[Y_t|\mathcal{G}_t] = \mathcal{C}^T X_t$. We again write

$$
c_t := \mathcal{C}^T X_t
$$

and $c_t^i := < c_t, e_i >$ which implies

$$\sum_{j=1}^{N_{\mathbf{Y}}} c_t^i = 1.$$

Write

$$\bar{\lambda}_l = \prod_{i=1}^{N_{\mathbf{Y}}} (N_{\mathbf{Y}} c_l^i)^{Y_l^i}, \tag{2.50}$$

and

$$\bar{\Lambda}_t = \prod_{l=0}^{t} \bar{\lambda}_l. \tag{2.51}$$

Lemma 2.2.8 With the above definitions

$$E^{\dagger}[\bar{\lambda}_t|\mathcal{G}_t] = 1. \tag{2.52}$$

Proof:

$$
\begin{aligned}
E^{\dagger}[\bar{\lambda}_t|\mathcal{G}_t] &= E^{\dagger}[\prod_{i=1}^{N_{\mathbf{Y}}} (N_{\mathbf{Y}} c_t^i)^{Y_t^i}|\mathcal{G}_t] \\
&= N_{\mathbf{Y}} \sum_{i=1}^{N_{\mathbf{Y}}} c_t^i P^{\dagger}(Y_t^i = 1|\mathcal{G}_t) \\
&= N_{\mathbf{Y}} \sum_{i=1}^{N_{\mathbf{Y}}} (c_t^i)/N_{\mathbf{Y}} = \sum_{i=1}^{N_{\mathbf{Y}}} c_t^i = 1.
\end{aligned}
$$

Next, define P^* by

$$\frac{dp^*}{dp^{\dagger}}|\ \mathcal{G}_{t+1} = \bar{\Lambda}_t. \tag{2.53}$$

Once again, P^* exists by Kolmogorov's extension theorem.

If two measures are absolutely continuous with respect to each other, the Radon-Nikodym derivative of one is the inverse of the other. Therefore,

$$\frac{dP^{\dagger}}{dP}|\ \mathcal{G}_{t+1} = \Lambda_t, \tag{2.54}$$

implies

$$\frac{dP}{dP^\dagger}\big|\ \mathcal{G}_{t+1} = 1/\Lambda_t = \bar{\Lambda}_t. \tag{2.55}$$

Thus,

$$\frac{dP^*}{dP^\dagger}\big|\ \mathcal{G}_{t+1} = \frac{dP}{dP^\dagger}\big|\ \mathcal{G}_{t+1} = \bar{\Lambda}_t, \tag{2.56}$$

which implies that for every set $B \in \mathcal{G}_t$

$$P^*(B) = \int_B \bar{\Lambda}_{t-1} dP^\dagger = P^\dagger(B).$$

Therefore, $P^* = P$.

Lemma 2.2.9 Under P,

$$E[Y_t|\mathcal{G}_t] = \mathcal{C}^T X_t, \tag{2.57}$$

and

$$E^\dagger[Z_{t+1}|\mathcal{G}_t] = 0. \tag{2.58}$$

Proof:

Using the conditional Bayes theorem,

$$
\begin{aligned}
P(Y_t^j|\mathcal{G}_t) &= E[< Y_t, e_j > |\mathcal{G}_t] \tag{2.59}\\
&= \frac{E^\dagger[\bar{\Lambda}_t < Y_t, e_j > |\mathcal{G}_t]}{E^\dagger[\bar{\Lambda}_t|\mathcal{G}_t]} \tag{2.60}\\
&= \frac{E^\dagger[\bar{\lambda}_t < Y_t, e_j > |\mathcal{G}_t]}{E^\dagger[\bar{\lambda}_t|\mathcal{G}_t]} \tag{2.61}\\
&= E^\dagger[\prod_{i=1}^{N_\mathbf{Y}}(N_\mathbf{Y} c_t^i)^{Y_t^i} < Y_t, e_j > |\mathcal{G}_t] \tag{2.62}\\
&= N_\mathbf{Y} E^\dagger[c_t^j < Y_t, e_j > |\mathcal{G}_t] = c_t^j, \tag{2.63}
\end{aligned}
$$

and writing in the matrix form gives the desired result.

Recall that under the new measure $P^\dagger$

$$E^\dagger[X_{t+1}|\mathcal{G}_t] = \mathcal{A}^T X_t.$$

Therefore,

$$
\begin{aligned}
E^\dagger[Z_{t+1}|\mathcal{G}_t] &= E^\dagger[X_{t+1} - \mathcal{A}^T X_t|\mathcal{G}_t] \\
&= E^\dagger[X_{t+1}|\mathcal{G}_t] - \mathcal{A}^T E^\dagger[X_t|\mathcal{G}_t] \\
&= \mathcal{A}^T X_t - \mathcal{A}^T X_t = 0.
\end{aligned}
$$

Lemma 2.2.10 With the above definitions

$$E^\dagger[Z_{t+1}|\mathcal{Y}_t] = 0 \tag{2.64}$$

Proof:

$$E^\dagger[Z_{t+1}|\mathcal{Y}_t] = E^\dagger[E^\dagger[Z_{t+1}|\mathcal{G}_t, \mathcal{Y}_t]|\mathcal{Y}_t]$$

We will show $E^\dagger[Z_{t+1}|\mathcal{G}_t, \mathcal{Y}_t] = 0$:

$$
\begin{aligned}
E^\dagger[Z_{t+1}|\mathcal{G}_t, \mathcal{Y}_t] &= \frac{E[\Lambda_t Z_{t+1}|\mathcal{G}_t, \mathcal{Y}_t]}{E[\Lambda_t|\mathcal{G}_t, \mathcal{Y}_t]} \tag{2.65} \\
&= \frac{E[\Lambda_t X_{t+1}|\mathcal{G}_t, \mathcal{Y}_t]}{E[\Lambda_t|\mathcal{G}_t, \mathcal{Y}_t]} - \frac{E[\Lambda_t \mathcal{A}^T X_t|\mathcal{G}_t, \mathcal{Y}_t]}{E[\Lambda_t|\mathcal{G}_t, \mathcal{Y}_t]} \tag{2.66} \\
&= \frac{\Lambda_t E[X_{t+1}|\mathcal{G}_t, \mathcal{Y}_t]}{\Lambda_t} - \frac{\Lambda_t \mathcal{A}^T X_t}{\Lambda_t} \tag{2.67} \\
&= E[X_{t+1}|\mathcal{G}_t, \mathcal{Y}_t] - \mathcal{A}^T X_t \tag{2.68} \\
&= E[X_{t+1}|\mathcal{G}_t] - \mathcal{A}^T X_t \tag{2.69} \\
&= E[Z_{t+1}|\mathcal{G}_t] = 0. \tag{2.70}
\end{aligned}
$$

Therefore, (2.64) follows.

Chapter 3

Product estimators and structured risk-sensitivity

It is often said that risk-sensitive filters take into account the "higher order" moments of the estimation error. Roughly speaking, this follows from the analytic property of the exponential $e^x = \sum_{k=0}^{\infty} x^k/k!$ so that if Ψ stands for the sum of the error functions over some interval of time then

$$E[exp(\gamma\Psi)] = E[1 + \gamma\Psi + (\gamma)^2(\Psi)^2/2 + ...].$$

Thus, at the expense of the mean error cost, the higher order moments are included in the minimization of the expected cost, reducing the "risk" of large deviations and increasing our "confidence" in the estimator. The parameter $\gamma > 0$ controls the extent to which the higher order moments are included. In particular, the first order approximation, $\gamma \to 0$, $E[exp(\gamma\Psi] \cong 1 + \gamma E\Psi$, indicates that the original minimization of the sum criterion or the risk-neutral problem is recovered as the small risk limit of the exponential criterion.

The exponential function, however, has the unique algebraic property of con-

verting a sum into a product. What if we consider a *product criterion* in the first place? In other words, what kind of estimators can we construct which are in some sense a generalization of the ones resulting from applying the sum criterion, if we begin with an estimation criterion based on the algebraic property of the exponential?

In this chapter, we will consider HMM's with discrete range observation and the Maximum A Posterior Probability (MAP) criterion. Later on, we will see that this is the "natural" criterion for the risk-sensitive HMM's because of its symmetric characteristics. The product estimators are introduced and studied in several examples. We show that product estimators allow for *dynamic* risk-sensitivity and attempt to understand the implications of this property. We study how risk-sensitivity parameter and the transition probabilities (which determine the underlying dynamics) are coupled in the behavior of the estimator. We consider the *multi-scale* representation of Markov chains and how this representation is dependent upon the risk-sensitivity parameter. This in turn suggests a coupling of the risk-sensitivity parameter, the underlying dynamics and the availability of information.

3.1   The MAP estimator and the sequential error accumulation

Consider a sequence of finite dimensional random variables X_t and observations Y_t defined on the probability space $(\mathbf{\Omega}, F, \mathbf{P})$. The Maximum A Posteriori Probability (MAP) estimator $\widehat{X}_t$ is a Borel measurable function of the observations up to Y_t denoted by $\mathcal{Y}_t$ which satisfies for $\omega \in \Omega$

$$\widehat{X}_t(\omega) = argmin_{\zeta \in \mathbf{S}_{N_{\mathbf{X}}}} \quad E[\rho(X_t - \zeta)| \, \mathcal{Y}_t = \mathcal{Y}_t(\omega)] \qquad t = 0, 1, ... \qquad (3.1)$$

where $\rho(\cdot, \cdot)$ is the discrete metric

$$\rho(X_t, \zeta) = \begin{cases} 0 & \text{if } X_t = \zeta; \\ 1 & \text{otherwise,} \end{cases} \qquad (3.2)$$

and $\mathbf{S}_{N_{\mathbf{X}}}$ of dimension $N_{\mathbf{X}}$ is the discrete range of X_t. In general, under mild conditions, one can show that almost surely such a measurable function exists. The usual definition of MAP as the argument with the highest probability given the observation follows from the above [8]. Because of the linearity of the expectation, MAP also results from the following additive cost minimization:

$$(\widehat{X}_0, ..., \widehat{X}_M)(\omega) = argmin_{\zeta_0 \cdots \zeta_N \in \mathbf{S}_{N_{\mathbf{X}}}^M} \quad E[\sum_{t=0}^{M} \rho(X_i - \zeta_t)| \, \mathcal{Y}_t = \mathcal{Y}_t(\omega)], \qquad (3.3)$$

where $\mathbf{S}_{N_{\mathbf{X}}}^M$ is the product space and each $\widehat{X}_t$ is $\mathcal{Y}_t$ measurable.

Thus, at each instant of time, our decision is not affected by our past or future decisions. This makes the MAP estimation insensitive to the accumulation of errors along sample paths. To see this, evaluate X_t and $\widehat{X}_t$ at some fixed value of $\omega \in \Omega$ to produce a sample path or a realization of the respective time series. The sequence of decisions $\widehat{X}_t \quad t = 0, 1, ...M$ partitions the sample space into 2^M subsets according to a binary tree with the branching criterion $X_t = \widehat{X}_t$ or $X_t \neq \widehat{X}_t$. Each ω belongs to exactly to one of these subsets. Some sample paths (See Figure 3.1) may end up on branches along which estimation errors accumulate for long stretches of time. The following simple yet instructive example illuminates the general case.

Consider a times series generated by a Markov chain depicted in Figure 3.2 with $a_{12} = 1/2 + \epsilon, a_{13} = 1/2 - \epsilon$ and $a_{23} = a_{32} = 0$. We assume that the initial state
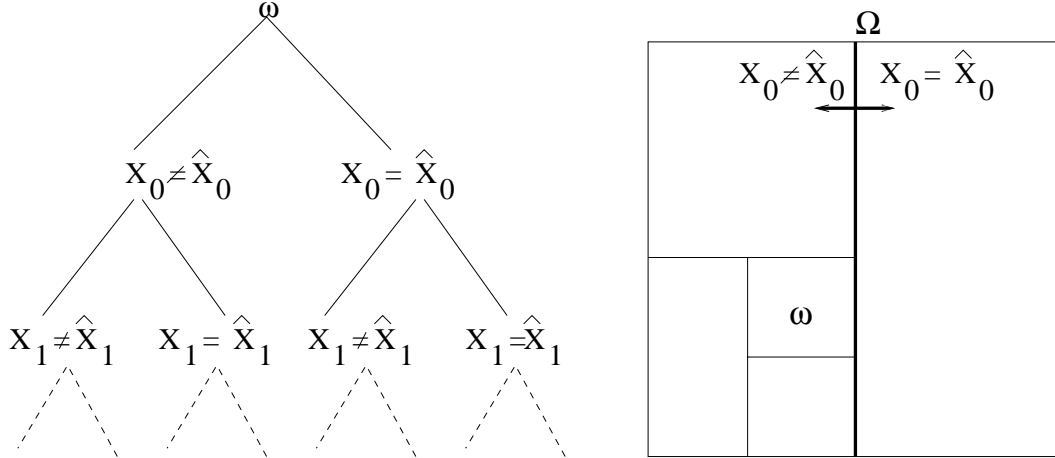
Figure 3.1: The MAP estimation along the sample paths of a time series.

is known to be 1 and that states 2 and 3 produce exactly the same observation; no useful observation is made. It is obvious that the MAP estimator will always choose state 2 of the chain after a transition from state 1 occurs. In this case sate 3 is never chosen. As it was shown in the previous chapter, we can associate the vertices of the unit simplex in R^3 to the states of this Markov chain so that e_1, e_2 and e_3 are associated with 1, 2 and 3 in that order. For $t > 1$ the following holds:

$$P(X_{t+1} = e_2 | X_t = e_2) = 1$$

$$P(X_{t+1} = e_3 | X_t = e_3) = 1.$$

Define $S_2^t := \{w | X_t(\omega) = e_2\}$ and $S_3^t := \{w | X_t(\omega) = e_3\}$. Clearly for $t > 1$, almost surely, $S_2^t = S_2^2 := S_2$, $S_3^t = S_3^2 := S_3$ and $S_2^t \bigcup S_3^t = \Omega$. Thus, sample paths associated with S_3 are always estimated incorrectly while sample paths associated with S_2 are always estimated correctly.

Our aim is to develop an estimator that balances the objective of finding the best estimate of the state in the sense of MAP at each instant of time, against the conflicting objective of preventing sequential error accumulation along sample
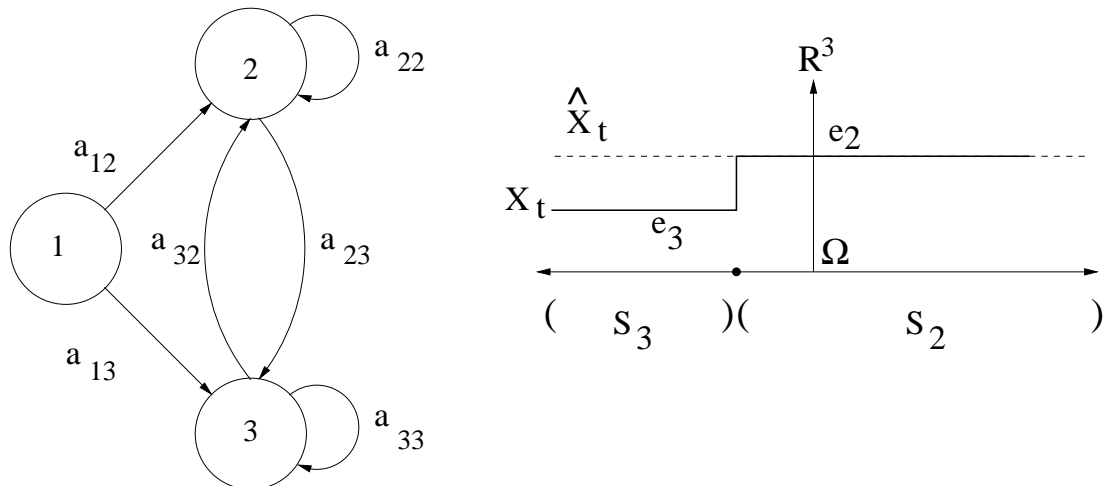
Figure 3.2: The MAP estimators sequential error accumulation.

paths. We shall develop a filter (generalized MAP or GMAP) that has the MAP estimator as its limit in some sense (when the accumulation of errors are ignored) and achieves this objective by a product cost structure that imposes higher costs when errors accumulate. We will show that risk-sensitive MAP estimation is a special case of GMAP.

3.2   Product estimators

Throughout this section and the subsequent ones, we will use the simplex representation of the HMM's developed in chapter 2.

Define $\widehat{X}_t$ as

$$\widehat{X}_t(\omega) = argmin_{\zeta \in \mathbf{s}_{N_\mathbf{X}}} \quad E[\Gamma_{0,t}(\zeta)| \, \mathcal{Y}_t = \mathcal{Y}_t(\omega)] \qquad t = 0, 1, ... \qquad (3.4)$$

The cost function $\Gamma_{0,t}(\zeta)$ is defined as

$$\Gamma_{0,t}(\zeta) = \widehat{\Gamma}_{0,t-1}.g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, X_t), \tag{3.5}$$

where

$$\widehat{\Gamma}_{0,t-1} = \prod_{i=0}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i). \tag{3.6}$$

Each $g_i(\omega)$ will determine an optimization or "cost" criterion at the time instant t scaled by the factor $\widehat{\Gamma}_{0,t-1}(\omega)$. In a sense, the product $\widehat{\Gamma}_{0,t-1}(\omega)$ is a "bias" which increases the estimation cost, at time t, for each sample path depending on the error accumulation in that path up to that time. This discourages the estimator to choose measurable functions of the observations that continue to estimate incorrectly sample paths with higher error accumulation.

In other words, by an appropriate choice of each random variable $g_i(\omega)$, the product $\widehat{\Gamma}_{0,t-1}$ contains a history of errors which have been made in the past along each sample path. The more errors we make, the higher the value of the product will be for that path. The product structure is not the only possible way to keep track of the past errors but it will make the computation of $\widehat{X}_t$ recursive. Morever, as we will show shortly, this product can be viewed as the dynamic generalization of the estimator achieved by the exponentiation of the MAP estimator. Therefore, we also solve the MAP risk-sensitive estimation problem as a special case of the product filters.

3.3 Generalized Maximum A Posteriori Probability estimator

Using the notation of chapter 2, define a new measure $P^\dagger$ by setting the Radon-Nikodym derivative equal to

$$\frac{dP}{dP^\dagger}\Big|\, \mathcal{G}_{t+1} = \bar{\Lambda}_t,$$

$$\bar{\Lambda}_t = \prod_{l=1}^{t} \bar{\lambda}_l \ , \qquad \bar{\lambda}_l = \prod_{i'=1}^{N_{\mathbf{Y}}} (N_{\mathbf{Y}}\, C_l^{i'})^{Y_l^{i'}},$$

$$C_t^{i'} = \sum_{j'=1}^{N_{\mathbf{X}}} c_{i'j'} <e_{j'}, X_t> .$$

We will need the following lemma [30]:

Lemma 3.3.1 Let $(\Omega, \mathcal{M}, P)$ be a probability space and let $\mathcal{F}_1, \mathcal{F}_2, \ \mathcal{F}_3$ be sub-σ-algebras of $\mathcal{M}$. Suppose also that $\mathcal{F}_3$ is independent of the σ-algebra $\mathcal{F}_1 \bigvee \mathcal{F}_2$, jointly generated by $\mathcal{F}_1$ and $\mathcal{F}_2$. Let f be measurable with respect to $\mathcal{F}_1$ then

$$E[f|\mathcal{F}_3, \mathcal{F}_2] = E[f|\mathcal{F}_2].$$

It can be shown, using Theorem 2.24, that under the new measure,

$$\widehat{X}_t = argmin_{\zeta \in \mathbf{S_X}}\ E^\dagger[\bar{\Lambda}_t \prod_{i=0}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i).g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, X_t)|\ \mathcal{Y}_t]. \quad (3.7)$$

To solve (3.7), we first define the following information state:

$$S_t := E^\dagger[\bar{\Lambda}_{t-1} \prod_{i=1}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i). \ X_t| \ \mathcal{Y}_{t-1}]. \tag{3.8}$$

Lemma 3.3.2 The information state can be computed according to

$$< S_{t+1} \ , \ e_h >= \sum_{j=1}^{N_\mathbf{X}} N_\mathbf{Y} \prod_{i'=1}^{N_\mathbf{Y}} (c_{i'j})^{Y_t^{i'}} g_t(\widehat{X}_0, ..., \widehat{X}_t, e_j) a_{jh} < S_t \ , \ e_j >. \tag{3.9}$$

Proof:

$$< S_{t+1} \ , \ e_h >= E^\dagger[\bar{\Lambda}_t \prod_{i=1}^{t} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i) < X_{t+1} \ , \ e_h >| \ \mathcal{Y}_t]. \tag{3.10}$$

But, $X_{t+1} = \mathcal{A}^T X_t + Z_{t+1}$ and thus $< X_{t+1}, e_h >=< \mathcal{A}^T X_t, e_h > + < Z_{t+1}, e_h >$.

Next, we will show that $< Z_{t+1}, e_h >$ drops out of (3.10).

Let $h = \bar{\Lambda}_t \prod_{i=1}^{t} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i)$. Observe that

$$E^\dagger[h \ Z_{t+1}|\mathcal{Y}_t] = E^\dagger[h E^\dagger[\ Z_{t+1}|\mathcal{Y}_t, \mathcal{G}_t]|\mathcal{Y}_t],$$

but by (2.70), $E^\dagger[\ Z_{t+1}|\mathcal{Y}_t, \mathcal{G}_t] = 0$. Thus,

$$E^\dagger[h \ Z_{t+1}|\mathcal{Y}_t] = 0.$$

We can write

$$\begin{aligned}
< \ &S_{t+1} \ , \ e_h > \\
= \ &E^\dagger[\prod_{i'=1}^{N_\mathbf{Y}} (N_\mathbf{Y} C_t^{i'})^{Y_t^{i'}} g_t(\widehat{X}_0, ..., \widehat{X}_t, X_t) \ \bar{\Lambda}_{t-1} \prod_{i=0}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i) < X_t, A e_h >|\mathcal{Y}_t] \\
= &\sum_{j=1}^{N_\mathbf{X}} E^\dagger[\prod_{i'=1}^{N_\mathbf{Y}} (N_\mathbf{Y} C_t^{i'})^{Y_t^{i'}} g_t(\widehat{X}_0, ..., \widehat{X}_t, X_t) \bar{\Lambda}_{t-1} a_{jh} \prod_{i=0}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i) < X_t, e_j >|\mathcal{Y}_t] \\
= &\sum_{j=1}^{N_\mathbf{X}} \prod_{i'=1}^{N_\mathbf{Y}} (N_\mathbf{Y} c_{i'j})^{Y_t^{i'}} g_t(\widehat{X}_0, ..., \widehat{X}_t, e_j) a_{jh} \\
&\qquad\qquad\qquad \cdot E^\dagger[\bar{\Lambda}_{t-1} \prod_{i=0}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i) < X_t, e_j >|Y_t, \mathcal{Y}_{t-1}]. \tag{3.11}
\end{aligned}$$

Applying Lemma 3.3.1 to (3.11) and by Theorem 2.2.7, (3.9) follows.

Lemma 3.3.3 $\widehat{X}_t$ is calculated according to

$$\widehat{X}_t = argmin_{\zeta \in \mathbf{S_X}} \sum_{j=1}^{N_\mathbf{X}} \prod_{i'=1}^{N_\mathbf{Y}} (N_\mathbf{Y} c_{i'j})^{Y_t^{i'}} g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, e_j) < S_t \ , \ e_j > \quad (3.12)$$

Proof:

$$E^\dagger[\bar{\Lambda}_t \prod_{i=0}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i).g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, X_t)| \ \mathcal{Y}_t]$$

$$= E^\dagger[\sum_{j=1}^{N_\mathbf{X}} \prod_{i'=1}^{N_\mathbf{Y}} (N_\mathbf{Y} C_t^{i'})^{Y_t^{i'}} g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, X_t)\bar{\Lambda}_{t-1}\prod_{i=0}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i) <X_t, e_j>|\mathcal{Y}_t]$$

$$= \sum_{j=1}^{N_\mathbf{X}} \prod_{i'=1}^{N_\mathbf{Y}} (N_\mathbf{Y} c_{i'j})^{Y_t^{i'}} g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, e_j)E^\dagger[\bar{\Lambda}_{t-1}\prod_{i=0}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i) <X_t, ej> |\mathcal{Y}_t]$$

$$= \sum_{j=1}^{N_\mathbf{X}} \prod_{i'=1}^{N_\mathbf{Y}} (N_\mathbf{Y} c_{i'j})^{Y_t^{i'}} g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, e_j) < S_t \ , \ e_j > . \quad (3.13)$$

The proof follows in steps similar to the the previous lemma and by the property

$$\sum_{j=1}^{N_\mathbf{X}} <X_t \ , \ e_j >= 1.$$

Notation: The operation $\odot$ in $a(P \odot Q)b$ produces "the statement P is true", if $a \leq b$, and "the statement Q is true", if $a \geq b$.

Theorem 3.3.4 (Generalized Maximum A Posteriori Probability Estimator) Assume that each g_i has the following form[1]:

$$g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i) = g_i(\gamma_i(\widehat{X}_{i-q}, ..., \widehat{X}_{i-1}), \rho(X_i, \widehat{X}_i)), \quad (3.14)$$

$$\rho(X_i, \widehat{X}_i) = \begin{cases} 0 & \text{if } X_i = \widehat{X}_i; \\ 1 & \text{otherwise}, \end{cases}$$

---

[1]with a slight abuse of notation

where each $g_i(.,.)$ is strictly positive and monotone increasing in the second variable. The functions $\gamma_i(\widehat{X}_{i-q}, ..., \widehat{X}_{i-1})$ are also strictly positive and q is a non-negative integer.

Then the estimate $\widehat{X}_t$ is calculated according to (GMAP)

$$\widehat{X}_t = argmax \ _{\zeta \in \mathbf{S}_{N_{\mathbf{X}}}} \prod_{i'=1}^{N_{\mathbf{Y}}} \left(c_{i' \ ind(\zeta)}\right)^{Y_t^{i'}} < S_t \ , \ \zeta >, \tag{3.15}$$

where $ind(e_i) = i$ and S_t evolves according to the following recursion:

$$S_{t+1} = \mathcal{A}^T G_t Q_t \ S_t \tag{3.16}$$

$\mathcal{A}$ is the transition matrix and

$$G_t = Diag\left\{...g_t(\gamma_t(\widehat{X}_{t-q}, ..., \widehat{X}_{t-1}), \rho(\widehat{X}_t, e_j))...\right\} \tag{3.17}$$

$$Q_t = Diag\left\{... \prod_{i'=1}^{N_{\mathbf{Y}}} (N_{\mathbf{Y}} \ c_{i'j})^{Y_t^{i'}} ...\right\} \tag{3.18}$$

$$j = 1, 2, ..., N_{\mathbf{X}}$$

Proof:

In the following, let "E_z" stand for the statement $\widehat{X}_t = e_z$.

The minimization developed in Lemma 3.3.2 is equivalent to a collection of binary comparisons

$$\sum_{j=1}^{N_{\mathbf{X}}} \prod_{i'=1}^{N_{\mathbf{Y}}} (N_{\mathbf{Y}} \ c_{i'j})^{Y_t^{i'}} g_t(\gamma_t(\widehat{X}_{t-q}, ..., \widehat{X}_{t-1}), \rho(e_s - e_j)) < S_t \ , \ e_j >$$

$(E_s \odot E_u)$

$$\sum_{j=1}^{N_{\mathbf{X}}} \prod_{i'=1}^{N_{\mathbf{Y}}} (N_{\mathbf{Y}} \ c_{i'j})^{Y_t^{i'}} g_t(\gamma_t(\widehat{X}_{t-q}, ..., \widehat{X}_{t-1}), \rho(e_u - e_j)) < S_t \ , \ e_j >$$

which simplifies to

$$\prod_{i'=1}^{N_{\mathbf{Y}}} (c_{i'u})^{Y_t^{i'}} < S_t \ , \ e_u > [g_t(\gamma_t(\widehat{X}_{t-q}, ..., \widehat{X}_{t-1}), 1) - g_t(\gamma_t(\widehat{X}_{t-q}, ..., \widehat{X}_{t-1}), 0)]$$

$(E_s \odot E_u)$

$$\prod_{i'=1}^{N_{\mathbf{Y}}} (c_{i's})^{Y_t^{i'}} < S_t \ , \ e_s > [g_t(\gamma_t(\widehat{X}_{t-q}, ..., \widehat{X}_{t-1}), 1) - g_t(\gamma_t(\widehat{X}_{t-q}, ..., \widehat{X}_{t-1}), 0)].$$

Since each g_i is positive and monotone increasing in the second variable, the above reduces to

$$\prod_{i'=1}^{N_{\mathbf{Y}}} (c_{i'u})^{Y_t^{i'}} < S_t, e_u > \qquad (E_s \odot E_u) \qquad \prod_{i'=1}^{N_{\mathbf{Y}}} (c_{i's})^{Y_t^{i'}} < S_t, e_s > . \qquad (3.19)$$

Therefore, the above *minimum* test reduces to the following *maximization* (termed generalized MAP because of a limiting result that will be stated in a corollary to this theorem):

$$\widehat{X}_t = argmax \ _{\zeta \in \mathbf{S}_{N_{\mathbf{X}}}} \prod_{i'=1}^{N_{\mathbf{Y}}} (c_{i' \ ind(\varsigma)})^{Y_t^{i'}} < S_t \ , \ \zeta > . \qquad (3.20)$$

Writing the information state in the matrix form completes the proof.

Corollary 3.3.5 Suppose $g_i(0, \cdot) = lim \ _{u \to 0} \ g_i(u, .) = 1$ uniformly over all i. Then MAP is recovered from GMAP when the functions $\gamma_i(\widehat{X}_{i-q}, ..., \widehat{X}_{i-1})$ converge almost surely to zero.

Proof:

The proof follows from the well known recursion formula for conditional expectation [26] and the following observation:

$$lim_{\gamma_t(\widehat{X}_{t-q},...,\widehat{X}_{t-1}) \to 0} \quad Diag\left\{...g_t(\gamma_t(\widehat{X}_{t-q},...,\widehat{X}_{t-1}),\rho(X_t,e_j))...\right\} = \mathbf{I},$$

where I is the identity matrix of the appropriate dimension. This implies that the information state converges to the unnormalized conditional expectation. Thus, the maximization in (3.20) becomes a maximization over the components of the conditional expectation which is precisely how MAP is defined [13].

Theorem 3.3.4 shows that the information state given by (3.8)

$$E^\dagger[\bar{\Lambda}_{t-1} \prod_{i=0}^{t-1} g_i(\widehat{X}_0,...,\widehat{X}_i,X_i). \ X_t| \ \mathcal{Y}_{t-1}]$$

is, in a sense, a generalized conditional expectation leading to a generalized notion of MAP for the filtering of HMM's. Motivated by (3.8), we could define a generalized conditional expectation for smoothing of HMM's by conditioning over an entire fixed time interval of size M, i.e., over the σ-algebra $\mathcal{Y}_M$ generated by all future and past observations

$$R_t := E^\dagger[\bar{\Lambda}_M \prod_{i=0}^{M} g_i(\widehat{X}_0,...,\widehat{X}_i,X_i). \ X_t| \ \mathcal{Y}_M]. \tag{3.21}$$

Using techniques similar to those used in the proof of Lemma 3.3.2, we can calculate R_t recursively according to

$$R_t = U_t \cdot S_t, \tag{3.22}$$

where S_t is the information state and U_t is a diagonal matrix whose components are calculated recursively backward in time according to

$$(U_t)_{jj} = \prod_{i'=1}^{N_{\mathbf{Y}}} (N_{\mathbf{Y}} \ c_{i'j})^{Y_t^{i'}} g_t(\gamma_t(\widehat{X}_{t-q},...,\widehat{X}_{t-1}),\rho(\widehat{X}_t - e_j)) \sum_{h=1}^{N_{\mathbf{X}}} a_{jh}(U_{t+1})_{hh} \tag{3.23}$$
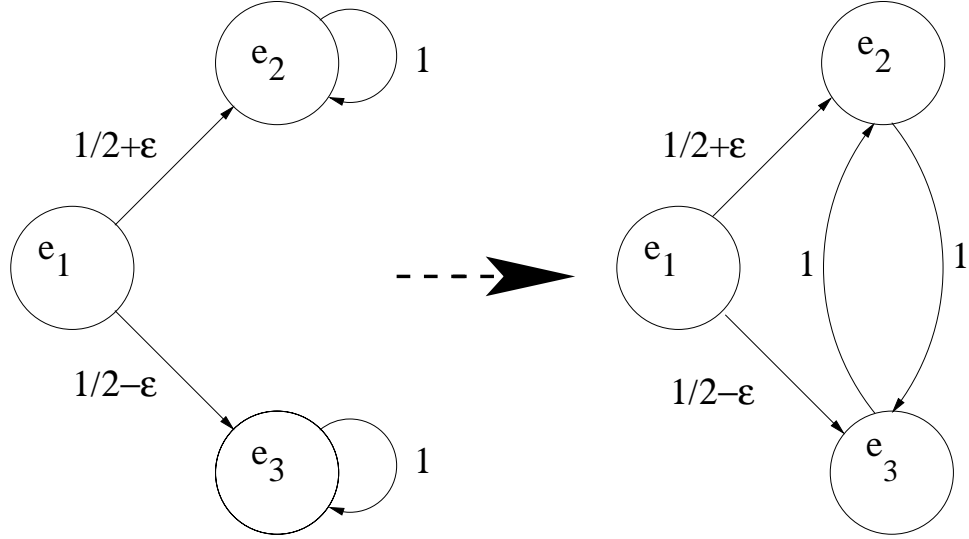
Figure 3.3: GMAP's limiting behavior .

with $(U_M)_{jj} = \prod_{i'=1}^{N_\mathbf{Y}} \ (N_\mathbf{Y} \ c_{i'j})^{Y_M^{i'}} g_M(\gamma_t(\widehat{X}_{M-q}, ..., \widehat{X}_{M-1}), \rho(\widehat{X}_M, e_j))$.

(3.23) is similar in form to the risk-neutral smoothing results for HMM's [26].

Next consider the Markov chain described earlier with $X_t \in \{e_1, e_2, e_3\}$. As before, assume no useful observation is made and that $X_0 = e_1$, $a_{11} = 0$, $a_{12} = a_{13} = 1/2$ and $a_{23} = a_{32} = 0$.

An acceptable solution to MAP for $t > 1$ is given by $\widehat{X}_t = e_2$. Let $g_t = 1 + \gamma * \rho(X_t, \widehat{X}_t)$. It turns out that for $\gamma > 0$ the only acceptable solutions to GMAP are oscillatory ones such as $\{e_2, e_1, e_2, e_1...\}$ and $\{e_1, e_2, e_1, e_2, ...\}$. The GMAP does not allow either of the states to be ignored. If we let $a_{12} = 1/2 + \epsilon$ and $a_{13} = 1/2 - \epsilon$ then the only acceptable solution to MAP is $\widehat{X}_t = e_2$ while it can be shown that for sufficiently large γ, GMAP still oscillates between e_2 and e_3 with a slight preference for e_2. If we set $a_{21} = 1/2 - \epsilon$ and $a_{31} = 1/2 + \epsilon$, then MAP changes dramatically to $X_t = e_3$ while GMAP still goes back and forth between e_2 and e_3 with a slight preference for e_3.

In the next section, we consider the Risk-sensitive Maximum A Posterior Probability (RMAP) and the Stochastic Risk-sensitive Maximum A Posterior Probability estimators (SRMAP) with the incremented-horizon, as special cases of GMAP and study their behavior.

3.4 RMAP, SRMAP and the behavior of their sample paths

3.4.1 RMAP, risk-sensitivity, information and the transition probabilities

Definition 3.4.1 SRMAP is defined as a special case of GMAP by setting

$$g_i = exp(\gamma(\widehat{X}_{i-q}, ..., \widehat{X}_{i-1})\rho(X_i, \widehat{X}_i)) \tag{3.24}$$

in (3.14) for every i.

Definition 3.4.2 RMAP is defined as a special case of SRMAP by setting

$$g_i = exp(\gamma\rho(X_i, \widehat{X}_i)) \tag{3.25}$$

in (3.14) for every i.

Obviously, every true statement for GMAP is also true for for SRMAP and RMAP including Theorem 3.3.4 and Corollary 3.3.5 proved in the previous section. In particular, we have shown that RMAP is obtained by a maximization criterion given by (3.20).

Let us first consider RMAP. The exponential function turns the sum of the error functions into a product whose value up to the last decision made for each

fixed ω is $e^{\gamma \cdot m}$, where m counts the number of times an error has been made in estimation of that particular sample path. This product is then multiplied by either 1 or $e^\gamma > 1$ depending on the next decision made based on minimizing the measure of the error set at the present time *and* taking note of the errors made in the past. Thus, the errors made in the past become more uniformly distributed over all sample paths. The information state condenses this error history and makes the computation recursive. Therefore, the MAP's additive cost structure is converted to RMAP's multiplicative one:

$$\sum : \begin{cases} 0 & \text{if } X_t = \zeta; \\ 1 & otherwise. \end{cases} \quad \rightarrow \quad \prod : \begin{cases} 1 & \text{if } X_t = \zeta; \\ e^\gamma & otherwise. \end{cases}$$

How does the "product count" of errors manifest itself in the behavior of the RMAP estimator of an HMM? Let us begin with a basic example.

Example 3.4.3 Let $N_\mathbf{Y} = N_\mathbf{X} = 2$, $c_{11} = c_{22} = q$. (See Figure 3.4.) Also, assume that the chain has an initial distribution $p = (p_1, p_2)$ with $p_1 + p_2 = 1$. q controls the amount of the available information, in particular $q = 1/2$ implies that no information is available, (the case of pure "prediction") and q=1 corresponds to the case of "perfect observation".

Let $q = 1/2$, $a_{12} = a_{21} = 0$ and allow γ to vary. Figure 3.5 and Table 3.1 show the behavior of the estimator. We have stacked up the sample paths starting at e_1 for a range of values of e^γ; on each path the lower value indicates e_1 and the higher value corresponds to e_2.

When $p_1 = p_2$, the oscillatory behavior appears only after one step. As we increase p_1, the onset of this behavior is delayed, but it is inevitable, unless $p_2 = 0$.

Now let $a_{12} = 0.1$ and $a_{21} = 0.15$ with everything else kept the same (Figure 3.6). The oscillatory behavior is delayed and relatively suppressed. This appears
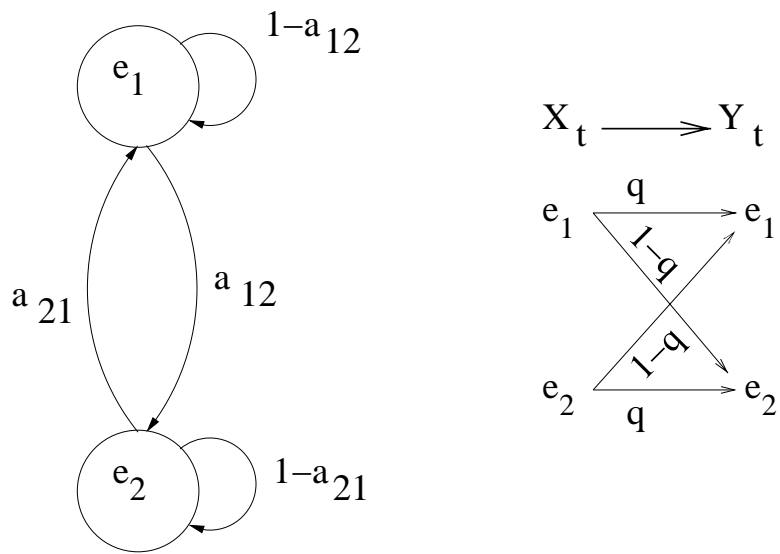
Figure 3.4: Example 3.4.3.
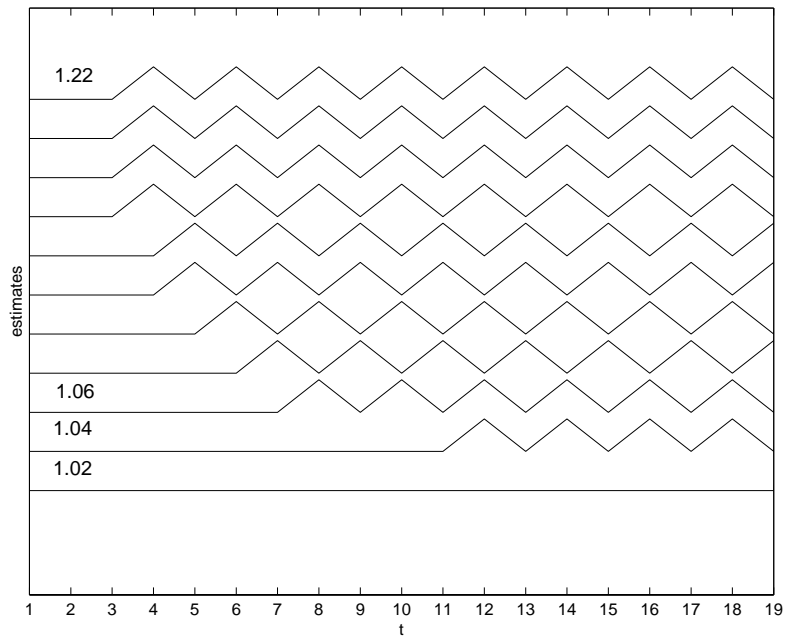
Figure 3.5: p_1=0.6, $a_{12} = a_{21} = 0$, $q = 1/2$, $1.02 \le e^\gamma \le 1.22$

$e^\gamma \mid t$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1.02	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1.04	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1	2	1	2	1
1.06	1	1	1	1	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1
1.08	1	1	1	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1	2
1.10	1	1	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1
1.12	1	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
1.14	1	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
1.16	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1
1.18	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1
1.20	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1
1.22	1	1	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1

Table 3.1: Example 3.4.3 with p_1=0.6, $a_{12} = a_{21} = 0$, $q = 1/2$, $1.02 \le e^\gamma \le 1.22$
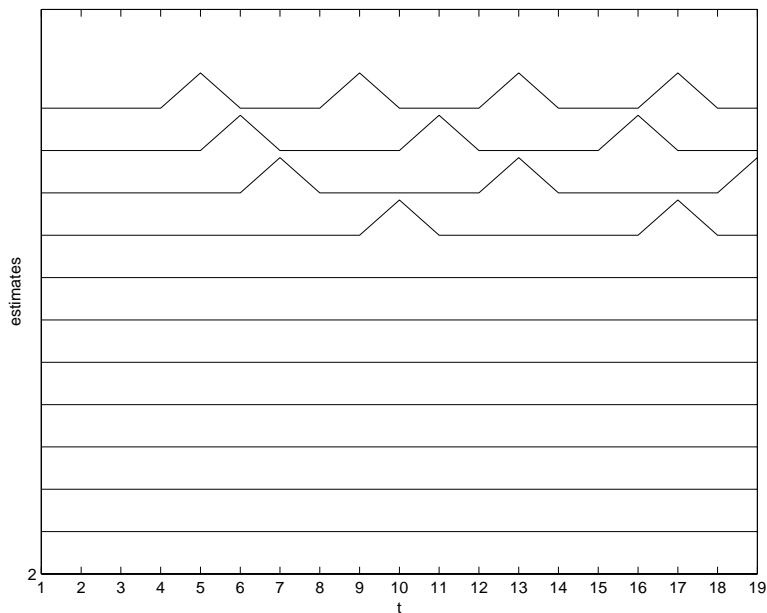
Figure 3.6: p_1=0.6, $a_{21} = 0.15, a_{12} = 0.1$, $q = 1/2$, $1.02 \le e^\gamma \le 1.22$

counter intutative: the initial setting $(a_{12} = a_{21} = 0)$ does not allow any transition between the states, but the RMAP estimator is oscillatory. The second set of parameters $(a_{12} = 0.1$ and $a_{21} = 0.15)$ allows transitions between the states, yet the estimator's behavior is less oscillatory and more similar to to that of MAP'. Why is this behavior altered by a modest change in the value of transition probabilities?

Now let p_1=0.6, $a_{12} = a_{21} = 0$, but $q = 0.6$. Figure 3.7 shows a series of sample paths with the state initialized at e_1 and $1.02 \le e^\gamma \le 1.22$ as before.

Once again, the oscillatations are supressed. It is evident that the transition probabilites representing the underlying dynamics, the risk-sensitivity parameter and the availability of information are coupled in the behavior of the sample paths. The difficulty, however, is that we cannot look at all sample paths at once. We must study these relations indirectly by looking at their influence on the properties of appropriately defined "averages" and "clusters" that represent the collective

41
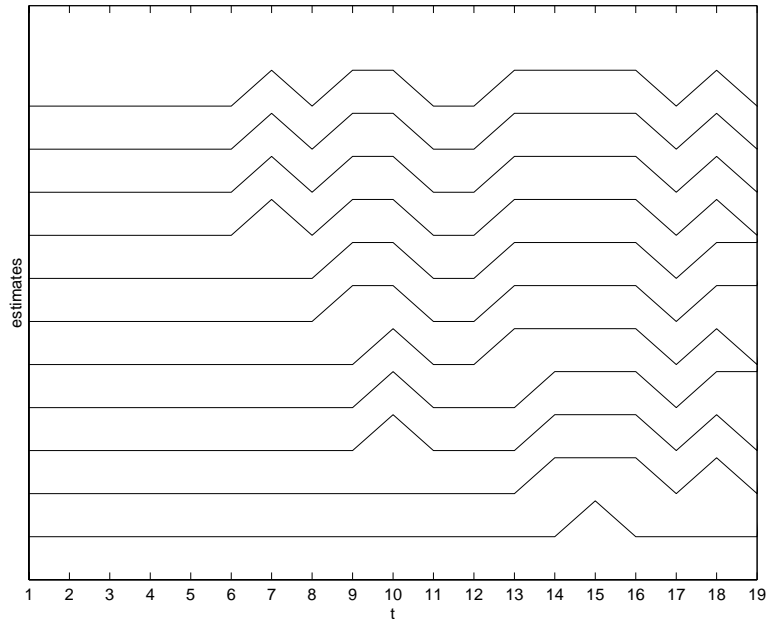
Figure 3.7: p_1=0.6, $a_{12} = a_{21} = 0$, $q = 0.6$, $1.02 \leq e^{\gamma} \leq 1.22$

properties of the sample paths. One view of conditional expectation comes from the notion of averaging or "smoothing" over the "atoms" of the conditioning σ-algebra [25]. Perhaps by computing the conditional expectation of appropriate random variables over various σ-algebras, these relations will come to light. In chapter 5, we will see how this idea follows naturally from certain dynamic programming equations for appropriately defined cost functions.

The second idea of considering "clusters" is demonstrated in the *multi-scale representations* of the states in the next section.

3.4.2 Multi-scale represenation and risk-sensitivity

By a *Multi-scale representaion* of a Markov chain, we mean a partition of the states of that chain. Each element of the partition will be called a "scale" or a "cluster".
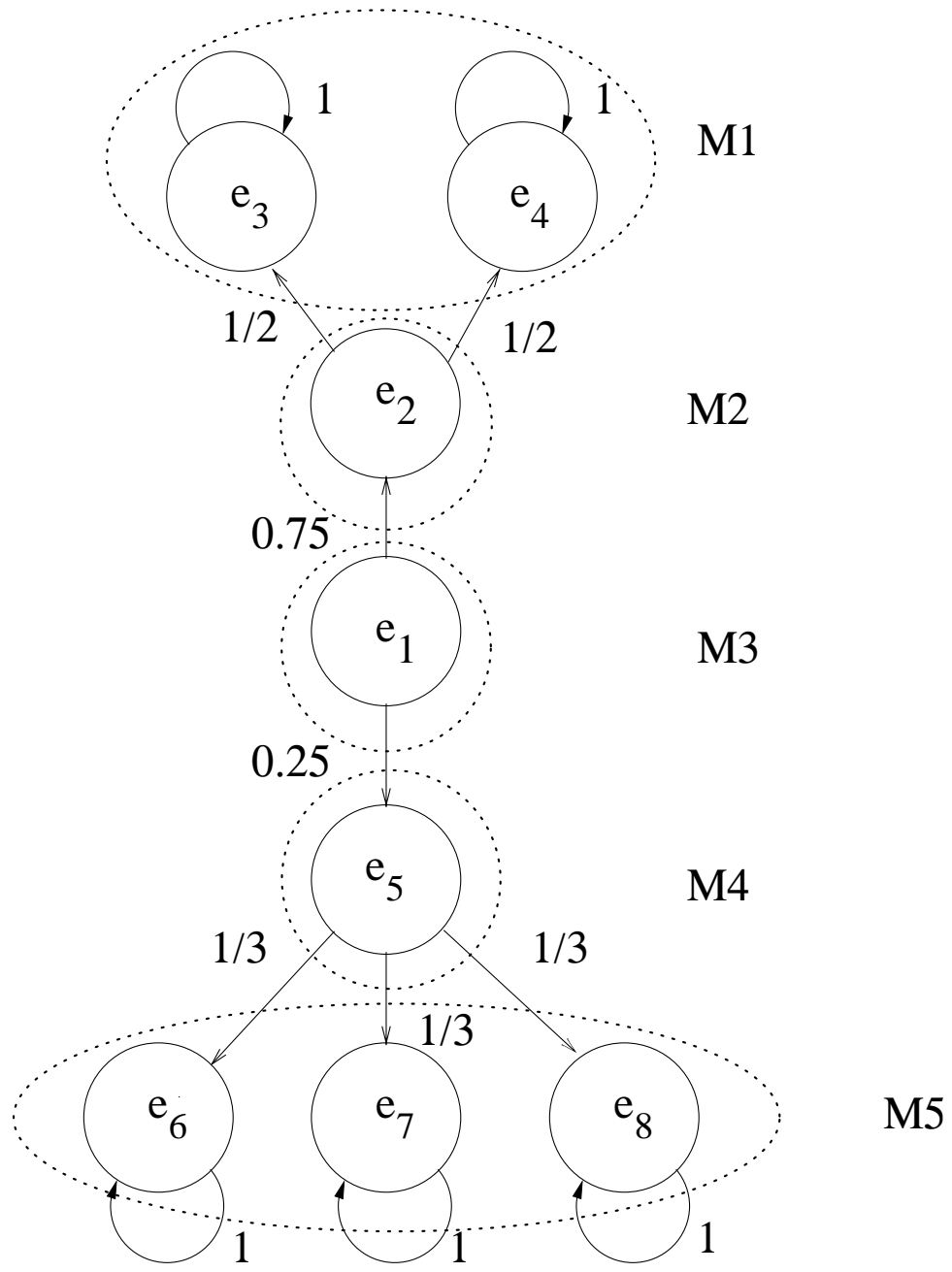
Figure 3.8: The multi-scale representation of example 3.3.4.

Example 3.4.4 Consider the Markov chain given by Figure 3.8. Consider the observation matrices $\mathcal{C}_u$ and $\mathcal{C}_l$:

$$
\mathcal{C}_u =
\begin{bmatrix}
991 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 991 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 701 & 291 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 291 & 701 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 991 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 496 & 496 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 496 & 496 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 496 & 496 & 1 & 1 & 1
\end{bmatrix} \times 10^{-3}
$$

$$
\mathcal{C}_l =
\begin{bmatrix}
991 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 991 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 496 & 496 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 496 & 496 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 991 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 991 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 991 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 991 & 1 & 1
\end{bmatrix} \times 10^{-3}
$$

We say that a "cluster" of states is resolved if it is possible from the observations to determine whether or not a sample path has assumed any of the states within that cluster. Examining $\mathcal{C}_u$ and $\mathcal{C}_l$ shows that both provide nearly perfect observation at the scales shown in the Figure 3.8 (M1 to M5). Within the clusters, $\mathcal{C}_u$ provides partial observation for the components of M1, e_3 and e_4 while keeping M5 "unresolved" so that it remains impossible to distinguish between the states e_6, e_7 and e_8. $\mathcal{C}_l$ leaves M1 unresolved while nearly resolving M5 into its components.
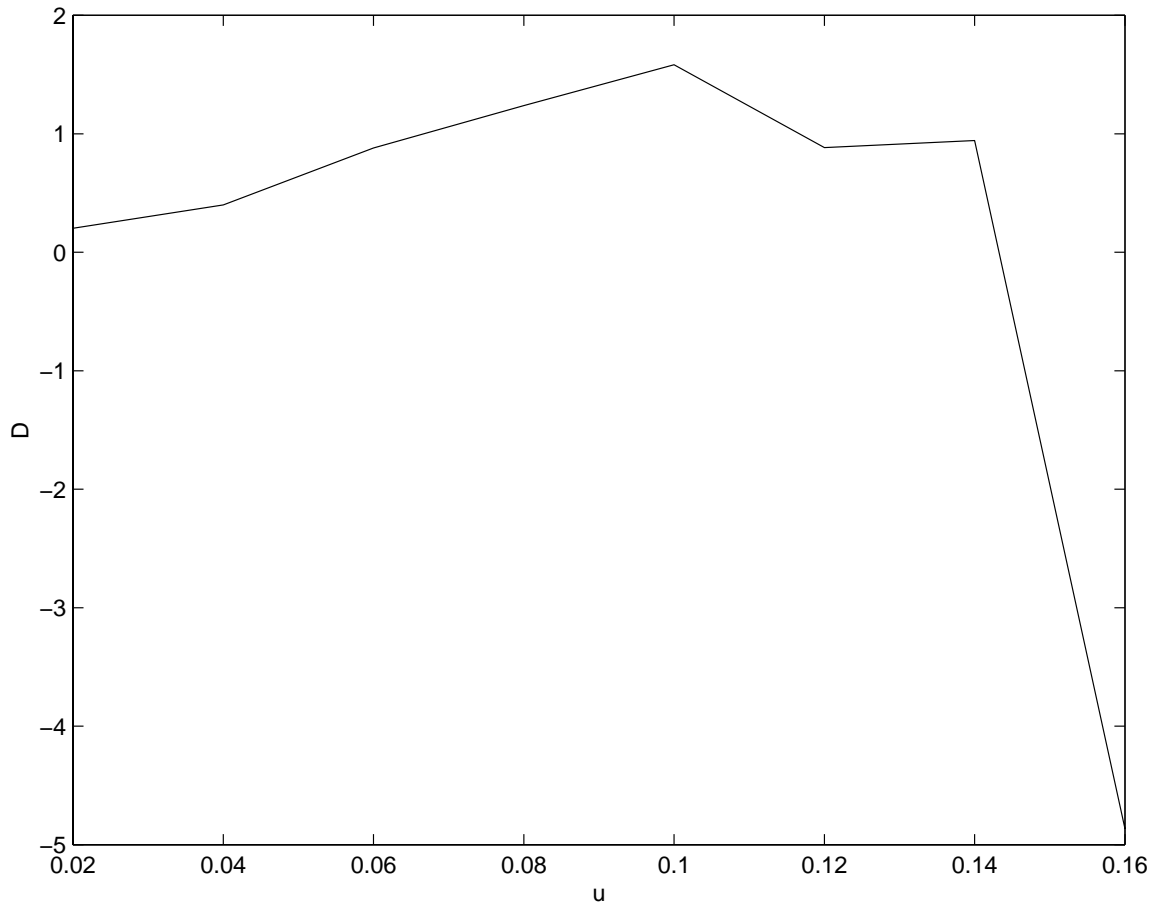
Figure 3.9: The cost difference vs. u, $t = 1, ..., 50$.

Which one of these matrices should we choose to "better" observe our Markov chain and is this choice invariant under risk-sensitivity?

$\mathcal{C}_l$ provides nearly perfect observation for the less likely path while $\mathcal{C}_u$ provides only partial observation but for the more likely path. Our simulations presented in Graph 3.9 (D is the cost under $\mathcal{C}_u$ subtracted from the cost under $\mathcal{C}_l$ via Monte Carlo simulation for the cost function (3.4) averaged over 10,000 sample paths) show that the choice depends on the amount of risk considered and is *not* invariant under risk-sensitivity. (On the x-axis, the parameter u is $e^{\gamma} - 1$.) Let us look at

the example and the simulations closely to understand why. For MAP and RMAP with small values of γ, $\mathcal{C}_u$, as the graph shows, is a better option but as γ increases choosing $\mathcal{C}_l$ quickly becomes the better option. The switch happens at a value of u^* between $u = 0.14$ and $u = 0.15$ (See Graph 3.9); changing the transition probabilities among the states of M1 and among those of M2 may change the value of u at which we switch from $\mathcal{C}_u$ to $\mathcal{C}_l$. For example, when we set $a_{34} = a_{43} = 0.1$ and $a_{33} = a_{44} = .9$, the value of u^* is less than half the previous value.

Do we want to see the "details" of M1 or M5? A more "conservative" estimator prefers to resolve M5 because M5 contains more branches (e_6 to e_8). The estimator can choose one branch at a time and will return to it only after it has visited the other branches. If we provide no information and leave M5 completely unresolved then, as γ increases, the estimator is forced to visit each state of M5 successively and thus each branch of M5 is ignored for two successive steps. This turns out to be costly as the value of γ is increased. If we increase the number of the branches in M5, the value of u^* at which the switch happens quickly decreases.

3.4.3 SRMAP and scale-dependent risk-sensitivity

Recall that SRMAP is defined as a special case of GMAP (3.14) by setting

$$g_i = exp(\gamma(\widehat{X}_{i-q}, ..., \widehat{X}_{i-1})\rho(X_i, \widehat{X}_i))$$

for every i.

We will study two examples of structured risk-sensitivity. First, consider the Markov chain shown in Figure 3.10 with the following transition and observation matrices:
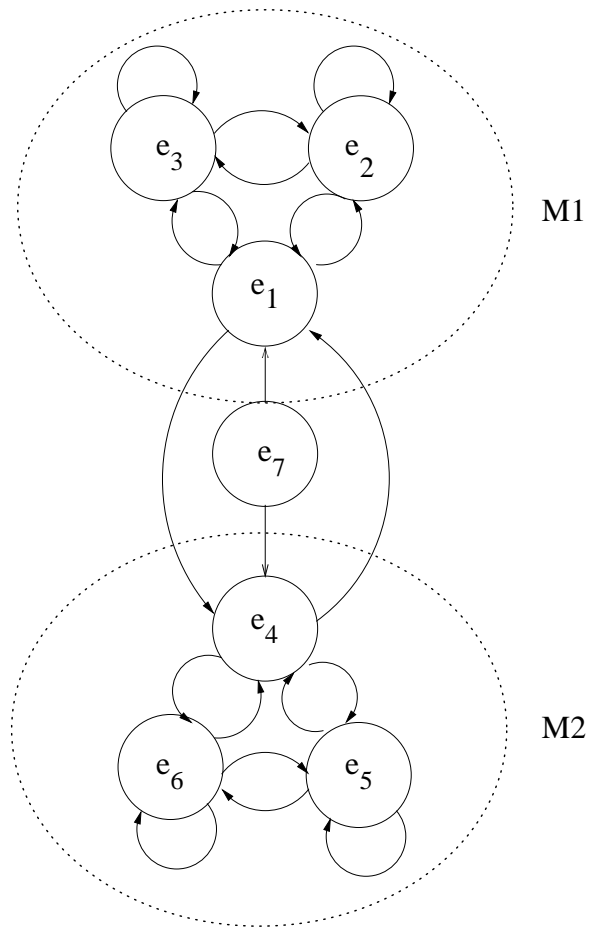
Figure 3.10: Scale dependent risk-sensitivity (parallel structure).

$$\mathcal{A} = \begin{bmatrix} 0.3 & 0.3 & 0.3 & 0.1 & 0.0 & 0.0 \\ 0.3 & 0.4 & 0.3 & 0.0 & 0.0 & 0.0 \\ 0.3 & 0.3 & 0.4 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.3 & 0.3 & 0.3 \\ 0.0 & 0.0 & 0.0 & 0.3 & 0.4 & 0.3 \\ 0.0 & 0.0 & 0.0 & 0.3 & 0.3 & 0.4 \\ 0.5 & 0.0 & 0.0 & 0.5 & 0.0 & 0.0 \end{bmatrix}$$

$$\mathcal{C} = \begin{bmatrix} 201 & 101 & 101 & 131 & 231 & 231 & 1 & 1 & 1 & 1 \\ 101 & 201 & 101 & 231 & 131 & 231 & 1 & 1 & 1 & 1 \\ 101 & 101 & 201 & 231 & 231 & 131 & 1 & 1 & 1 & 1 \\ 131 & 231 & 231 & 201 & 101 & 101 & 1 & 1 & 1 & 1 \\ 231 & 131 & 231 & 101 & 201 & 101 & 1 & 1 & 1 & 1 \\ 231 & 231 & 131 & 101 & 101 & 201 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 991 & 1 & 1 & 1 \end{bmatrix} \times 10^{-3}$$

For $t = 1$ to 100, let

$$\gamma(x) = \begin{cases} \gamma & \text{if } ind(x) \geq 5; \\ 10^{-18} & \text{otherwise.} \end{cases}$$

Graph 3.11 shows the average percentage of the time that the estimator resides in M2 vs. $u = e^{\gamma} - 1$. It is clear that applying higher risk within only one scale tends to "repel" the estimator toward the other cluster.
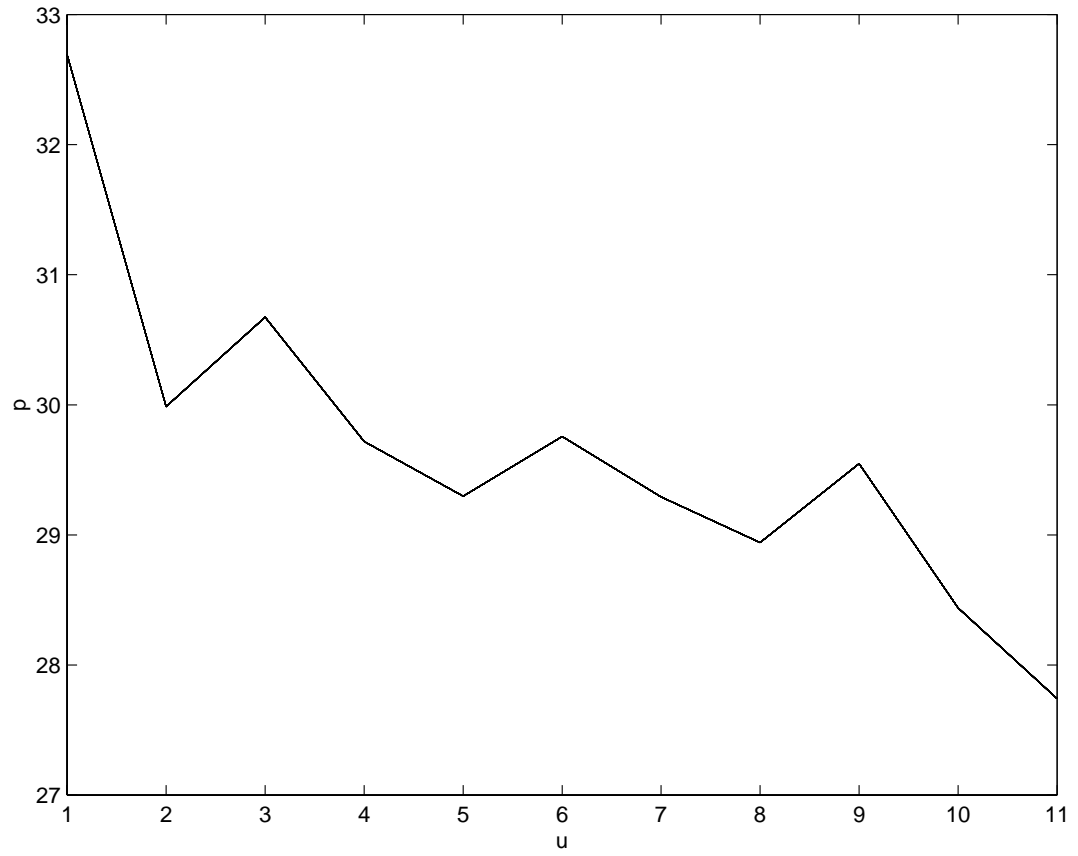
Figure 3.11: The percentage of M2

Now let the transition matrix be given by

$$
\mathcal{A} =
\begin{bmatrix}
0.4 & 0.3 & 0.3 & 0.0 & 0.0 & 0.0 \\
0.3 & 0.4 & 0.3 & 0.0 & 0.0 & 0.0 \\
0.3 & 0.3 & 0.4 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.4 & 0.3 & 0.3 \\
0.0 & 0.0 & 0.0 & 0.3 & 0.4 & 0.3 \\
0.0 & 0.0 & 0.0 & 0.3 & 0.3 & 0.4 \\
0.5 & 0.0 & 0.0 & 0.5 & 0.0 & 0.0
\end{bmatrix}
$$

and let the observations provide no information about the process. Then, for a constant risk of $e^\gamma = 2$ for both clusters, the sequence of predictions appears as follows: 4, 1, 3, 2, 1, 4, 3, 6, 2, 5, 1, 4, 3, 6, 2, 5, 1, 4, 3, 6, 2, 5, 1, ... But, if we rewrite this sequence at the higher scales M1 and M2, we get M2, M1, M1, M1, M1, M2, M1, M2, M1, M2, M1, M2, M1, M2, M1, M2, M1, M2, M1, M2, M1, M2, M1,.... Therefore, we see that after four steps an oscillatory behavior emerges at the higher scales.

In the previous example, the risk is made dependent on the scales in a "parallel" structure. This is because M1 and M2 communicate; i.e., it is possible to make transitions between them. Next, we will consider a "cascade" or "left-to-right" structure in which the risk depends on the scales which do not communicate (See

Figure 3.12) Consider the Markov chain with the following transition matrix:

$$
\mathcal{A} =
\begin{bmatrix}
0.4 & 0.3 & 0.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.4 & 0.2 & 0.0 & 0.4 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.2 & 0.0 & 0.2 & 0.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.4 & 0.0 & 0.5 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.4 & 0.0 & 0.5 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & .02 & 0.0 & 0.0 & .08 & 0.9 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.9 & 0.1 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0
\end{bmatrix}
$$

and observation matrix:

$$
\mathcal{C}^T =
\begin{bmatrix}
71 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
11 & 91 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
11 & 1 & 91 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 41 & 41 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 51 & 51 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 91 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 91 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 91 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 51 & 41 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 41 & 51
\end{bmatrix}
\times 10^{-2}
$$

Let the transition matrix used in the model be

$$\mathcal{M} = \begin{bmatrix}
0.4 & 0.3 & 0.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.4 & 0.2 & 0.0 & 0.4 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.2 & 0.0 & 0.2 & 0.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.4 & 0.0 & 0.5 & 0.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.4 & 0.0 & 0.5 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & .02 & 0.0 & 0.0 & .08 & 0.9 & 0.0 & 0.0 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.9 & 0.1 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.9 & 0.1 \\
0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.9
\end{bmatrix}$$

For large values of γ, we expect the limiting behavior of M2 (based on the model $\mathcal{M}$) to become oscillatory for reasons presented in the previous sections. Therefore, by applying a "high risk" to the entire chain we may be able to force our model to behave roughly like the actual process which is by construction oscillatory within M2. But, within M1, the model $\mathcal{M}$ and the transition matrix $\mathcal{A}$ are exactly the same. We consider the behavior of the whole chain under structured risk-sensitivity when the risk-sensitivity is stochastic and dependent upon the estimate of the state, and compare it to the constant risk-sensitive case. Which approach will create sample paths, on the average, closer to the actual process?

Let

$$\gamma(x) = \begin{cases} \gamma & \text{if } ind(x) \geq 7; \\ 10^{-7} & \text{otherwise.} \end{cases}$$

The value 10^{-7} is sufficiently small so that the behaviors of SRMAP and MAP are identical within M1. As shown in the Graph 3.13 (the dashed curve), the prediction error decreases for sufficiently large values of γ to about 0.11. The solid curve shows what happens when γ is is applied uniformly to the whole model,
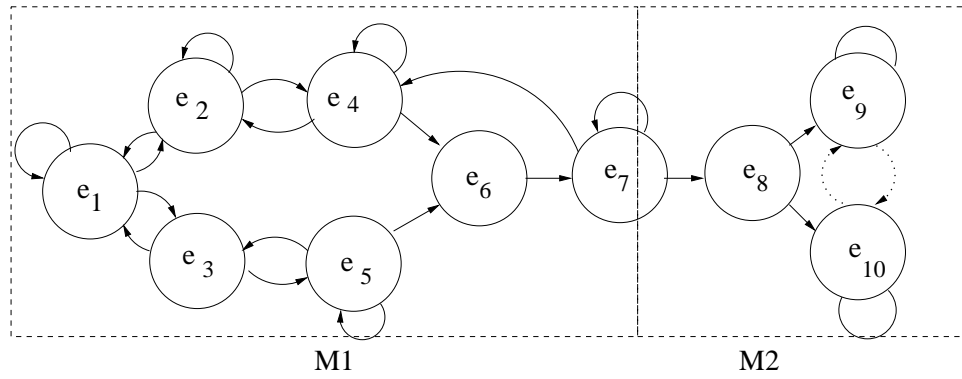
Figure 3.12: Scale dependent risk-sensitivity (cascade structure).
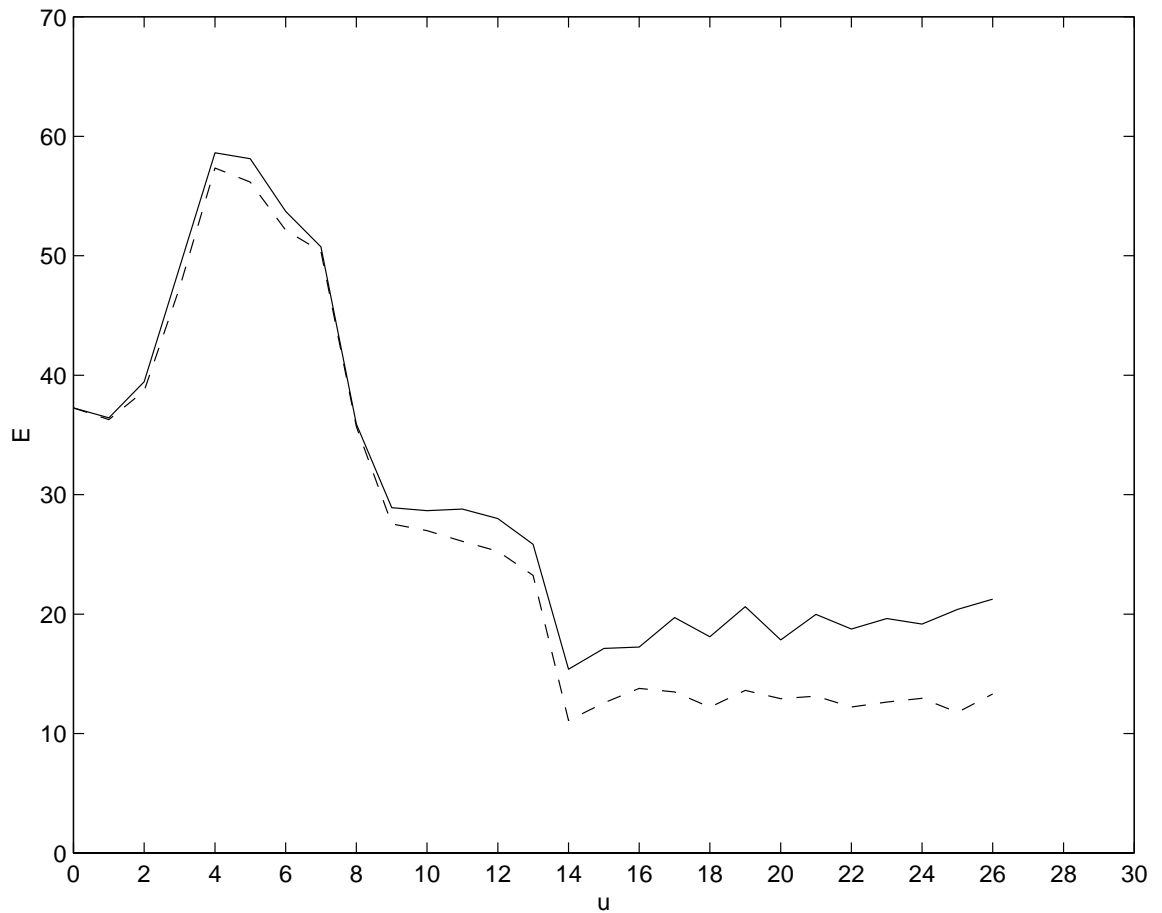
Figure 3.13: The error cost vs. $e^{\gamma} - 1$.

53

meaning γ is taken to be a constant function rather than a stochastic one. The error is considerably larger. By applying a constant γ to the whole model, we have perturbed the entire model (at both scales) while our model within M1 should not have been perturbed.

Chapter 4

Product estimators with continuous range observation

In this chapter, we consider product estimators of HMM's with observation under continuous range additive noise. As a special case, we will derive the Dey-Moore filter with the quadratic cost [10].

4.1 The state space model

The product estimators of the previous chapter can also be constructed for the estimation of HMM's with observation under continuous range additive noise. We let $\mathbf{S_X} = \{e_1, ..., e_{N_{\mathbf{X}}}\}$ be the range of the values of the process defined on the probability space $(\mathbf{\Omega}, F, \mathbf{P})$ and let $\mathcal{F}_t^0$ be defined as in chapter 2. The state space model is given by

$$\begin{cases} X_{t+1} = \mathcal{A}^T X_t + Z_{t+1} \\ Y_{t+1} = C(X_t) + W_{t+1} \end{cases},$$

where the process W_t is i.i.d. with a strictly positive density Φ taking values in

$R^{N_\mathbf{Y}}$. The transition matrix $\mathcal{A}$ is defined as before and $C(\cdot)$ is a real valued function in $R^{N_\mathbf{Y}}$. Although the results of this chapter hold for all strictly positive densities, for easing our presentation we assume that W_t is a zero mean unit variance Gaussian noise $N(0,1)$.

The Girsanov change of measure can also be carried out for the continuous range observation, yielding a new measure under which the observation sequence is i.i.d. and distributed according to $N(0,1)$ density. The development is similar to the one in chapter 2 for the discrete range observation; we will only state the results. As previously defined, $\mathcal{Y}_t$ is generated by the sequence of observations up to time t and $\mathcal{G}_t$ is the filtration generated by the process X_t up to time t and the observations up to time t-1.

It turns out that the right choice for the Radon-Nikodym derivative is

$$\frac{dP^\dagger}{dP}\big|\,\mathcal{G}_{t+1} = \Lambda_t$$

$$\Lambda_t = \prod_{l=0}^{t} \lambda_l \ , \qquad \lambda_t = \frac{\phi(Y_t)}{\phi(Y_t - C(X_t))}$$

and as shown in chapter 2; this implies that $\frac{dP}{dP^\dagger}|\mathcal{G}_{t+1} = \bar{\Lambda}_t$ where $\bar{\Lambda}_t = 1/\Lambda_t$.

4.2   The product estimator

Define $\widehat{X}_t$ as

$$\widehat{X}_t = argmin_{\zeta \in \mathbf{S_X}} \quad E[\Gamma_{0,t}(\zeta)|\,\mathcal{Y}_t] \qquad t = 0, 1, ...$$

The cost function $\Gamma_{0,t}(\zeta)$ is defined as

$$\Gamma_{0,t}(\zeta) \ = \widehat{\Gamma}_{0,t}(\zeta).g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, X_t)$$

where

$$\widehat{\Gamma}_{0,t} = \prod_{i=0}^{t-1} g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i)$$

As in the previous chapter,

$$E[\Gamma_{0,t}(\zeta)| \ \mathcal{Y}_t] = \frac{E^\dagger[\bar{\Lambda}_t \ \Gamma_{0,t}(\zeta)| \ \mathcal{Y}_t]}{E^\dagger[\bar{\Lambda}_t| \ \mathcal{Y}_t]}$$

Therefore, an equivalent problem is to find $\widehat{X}_t$ such that

$$\widehat{X}_t = argmin_{\zeta \in \mathbf{S_x}} \quad E^\dagger[\bar{\Lambda}_t \ \Gamma_{0,t}(\zeta)| \ \mathcal{Y}_t].$$

Definition 4.2.1 As in the discrete case, we define the information state

$$S_t = E^\dagger[\bar{\Lambda}_{t-1} \ \widehat{\Gamma}_{0,t-1} \ X_t| \ \mathcal{Y}_{t-1}]$$

Lemma 4.2.2 The information state S_t evolves according to

$$< S_{k+1} \ , \ e_i >= \sum_{j=1}^{N_\mathbf{x}} \frac{\phi(Y_t - C(e_j))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_t, e_j) a_{ji} < S_t \ , \ e_j > .$$

Proof:

$$
\begin{aligned}
< \ S_{k+1} \ , \ e_i >&= E^\dagger[\bar{\Lambda}_t \ \Gamma_{0,t} \ < X_{t+1} \ , \ e_i >| \ \mathcal{Y}_t] \\
&= \ E^\dagger[\frac{\phi(Y_t - C(X_t))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_t, X_t) \ \bar{\Lambda}_{t-1}\widehat{\Gamma}_{0,t-1} < X_t \ , Ae_i >| \ \mathcal{Y}_t] \\
&= \ \sum_{j=1}^{N_\mathbf{x}} E^\dagger[\frac{\phi(Y_t - C(X_t))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_t, X_t) \ \bar{\Lambda}_{t-1}\widehat{\Gamma}_{0,t-1} a_{ji} < X_t \ , e_j >| \ \mathcal{Y}_t] \\
&= \ \sum_{j=1}^{N_\mathbf{x}} \frac{\phi(Y_t - C(e_j))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_t, e_j) a_{ji} E^\dagger[\bar{\Lambda}_{t-1} \ \widehat{\Gamma}_{0,t-1} \ < X_t \ , \ e_j >|Y_t, \mathcal{Y}_{t-1}]
\end{aligned}
$$

$$= \sum_{j=1}^{N_{\mathbf{x}}} \frac{\phi(Y_t - C(e_j))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_t, e_j) a_{ji} E^{\dagger}[\bar{\Lambda}_{t-1} \; \widehat{\Gamma}_{0,t-1} \; < X_t \; , \; e_j > | \mathcal{Y}_{t-1}]$$

$$= \sum_{j=1}^{N_{\mathbf{x}}} \frac{\phi(Y_t - C(e_j))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_t, e_j) a_{ji} < S_t \; , \; e_j >$$

It can be shown that under the new measure

$$\widehat{X}_t = argmin_{\zeta \in \mathbf{S_x}} \quad E^{\dagger}[\bar{\Lambda}_t \; \widehat{\Gamma}_{0,t-1} \; .g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, X_t) | \; \mathcal{Y}_t].$$

Lemma 4.2.3 The estimate $\widehat{X}_t$ is calculated according to

$$\widehat{X}_t = argmin_{\zeta \in \mathbf{S_x}} \quad \sum_{j=1}^{N} \frac{\phi(Y_t - C(e_j))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, e_j) < S_t \; , \; e_j >$$

Proof:

$$E^{\dagger} \; [\bar{\Lambda}_t \; \widehat{\Gamma}_{0,t-1} \; .g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, X_t) | \; \mathcal{Y}_t]$$

$$= E^{\dagger}[\sum_{j=1}^{N} \frac{\phi(Y_t - C(X_t))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, X_t) \; \bar{\Lambda}_{t-1} \; \widehat{\Gamma}_{0,t-1} \; < X_t \; , e_j > | \; \mathcal{Y}_t]$$

$$= \sum_{j=1}^{N} \frac{\phi(Y_t - C(e_j))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, e_j) E^{\dagger}[\bar{\Lambda}_{t-1} \; \widehat{\Gamma}_{0,t-1} \; . < X_t \; , ej > | \; \mathcal{Y}_{t-1}]$$

$$= \sum_{j=1}^{N} \frac{\phi(Y_t - C(e_j))}{\phi(Y_t)} g_t(\widehat{X}_0, ..., \widehat{X}_{t-1}, \zeta, e_j) < S_t \; , \; e_j > .$$

4.3 Generalized Maximum A Posteriori Probability Estimator

Theorem 4.3.1 Assume that each g_i has the following form:

$$g_i(\widehat{X}_0, ..., \widehat{X}_i, X_i) = g_i(\gamma_i(\widehat{X}_{i-q}, ..., \widehat{X}_{i-1}), \rho(X_i, \widehat{X}_i))$$

$$\rho(X_i, \widehat{X}_i) = \begin{cases} 0 & \text{if } X_i = \widehat{X}_i; \\ 1 & \text{otherwise,} \end{cases}$$

where each $g_i(., .)$ is positive and monotone increasing in the second variable and where functions $\gamma_i(\widehat{X}_{i-q}, ..., \widehat{X}_{i-1})$ are strictly positive and q is a non-negative integer. Then $\widehat{X}_t$ is calculated according to (GMAP)

$$\widehat{X}_t = argmax_{\ \zeta \in \mathbf{S_X}} \ \frac{\phi(Y_t - C(\zeta))}{\phi(Y_t)} < S_t \ , \ \zeta >,$$

where S_t evolves according to the following recursion:

$$S_{t+1} = \mathcal{A}^T G_t Q_t \ S_t$$

A is the transition matrix and

$$G_t = Diag\Big\{ ... g_t(\gamma_t(\widehat{X}_{t-q}, ..., \widehat{X}_{t-1}), \rho(\widehat{X}_t - e_j))...\Big\}$$

$$Q_t = Diag\Big\{ ... \frac{\phi(Y_t - C(e_j))}{\phi(Y_t)}...\Big\}$$

$$j = 1, 2, ..., N_{\mathbf{X}}$$

Proof:

The proof is similar to that of Theorem 3.3.4 and will be skipped.

Corollary 4.3.2   Suppose $g_t(0, .) = lim_{\ u->0} \ g_t(u, .) = 1$ uniformly over all i. Then MAP is recovered from GMAP when the functions $\gamma_t(\widehat{X}_{t-q}, ..., \widehat{X}_{t-1})$ converge almost surely to zero.

Proof:

The proof is similar to that of Corollary 3.3.5 and will be skipped.

4.4 Quadratic cost risk-sensitive estimators

We will next derive the results in [10] as a special case of our product estimator.

The quadratic cost risk-sensitive problem is defined as

$$\widehat{X}_t = argmin_{\zeta \in \mathbf{S_X}} \quad E[exp(\ \gamma \Psi_{0,k}(\zeta)\ )|\ \mathcal{Y}_t] \qquad t = 0, 1, ...$$

$$\Psi_{0,t}(\zeta)\ =\ \widehat{\Psi}_{0,t}(\zeta)\ + (X_t - \zeta)^T U (X_t - \zeta)$$

where the matrix $U \geq 0$ is and

$$\widehat{\Psi}_{m,n}(\zeta)\ =\ \sum_{i=m}^{n} (X_i - \widehat{X}_i)^T U (X_i - \widehat{X}_i).$$

Theorem 4.4.1   The estimate $\widehat{X}_k$ is calculated according to

$$\widehat{X}_k = argmin_{\zeta \in \mathbf{S_X}}\ \sum_{j=1}^{N} \frac{\phi(Y_t - C(e_j))}{\phi(Y_t)} exp\ [\gamma(e_j - \zeta)^T U(e_j - \zeta)] < S_k\ ,\ e_j >,$$

where the information state S_t evolves according to the following recursion:

$$S_{t+1} = \mathcal{A}^T D_t B_t\ S_t, \tag{4.1}$$

$\mathcal{A}$ is the transition matrix and

$$D_t = diag\Big\{exp\ [\gamma(e_1 - \widehat{X}_t)^T U(e_1 - \widehat{X}_t)],\ ...,\ exp\ [\gamma(e_1 - \widehat{X}_t)^T U(e_n - \widehat{X}_t)]\Big\} \tag{4.2}$$

$$B_t = diag\Big\{ \frac{\phi(Y_t - C(e_1))}{\phi(Y_t)}\ ,\ ...\ ,\ \frac{\phi(Y_t - C(e_n))}{\phi(Y_t)} \Big\} \tag{4.3}$$

Proof:

Let $g_i(\cdot, \cdot) = g(\cdot, \cdot) = exp(\gamma(X_t - \zeta)^T U(X_t - \zeta)),$

and the proof follows from the Lemma 4.2.3.

4.5 Normalized information state and asymptotic periodicity

It is well known that under mild conditions [27] [19] the distribution of the states of a Markov chain is asymptotically stationary; i.e., any initial probability distribution of the states is transformed in the limit to a stationary probability distribution p^* which satisfies $\mathcal{A}^T p^* = p^*$. Note that if no information is available other than the initial probability distribution, the probability distribution of the chain is the information state for MAP, the so called "prediction" case. Therefore, the information state for the prediction case is asymptotically stationary.

As in chapter 3, RMAP is obtained by letting

$$g_i = exp(\gamma \rho(X_i, \widehat{X_i})) \tag{4.4}$$

for every i.

We would like to know if there is a similar asymptotic behavior for the information state of RMAP in the prediction case. We know that for small risk RMAP reduces to MAP. Inspired by this, we define the *generalized probability distribution* of the RMAP by simply normalizing the information state. Our calculations show that the generalized probability has the remarkable property of the *asymptotic periodicity* which can be considered as a generalization of the asymptotic stationarity.

But first, we consider the asymptotic stationarity of Markov chains.

Definition 4.5.1   A Markov chain is called *primitive* if there exists an integer N such that all the elements of $\mathcal{A}^n$ are strictly positive for every $n \geq N$.

We state the following without proof [27]:

Theorem 4.5.2 (*Perron*) Let $\mathcal{A}$ be a primitive $m \times m$ real matrix. Then, there exists an eigenvalue λ of $\mathcal{A}$ with the following properties: (i) $\lambda > 0$, (ii) λ is a simple eigenvalue, (iii) λ has a positive eigenvector, (iv) any other eigenvalue of $\mathcal{A}$ has modulus less than λ, (v) any other nonnegative eigenvector of $\mathcal{A}$ is a positive multiple of v.

Corollary 4.5.3 Let $\mathcal{A}$ be a primitive stochastic matrix. Then 1 is a simple eigenvalue of $\mathcal{A}$, both $\mathcal{A}$ and $\mathcal{A}^T$ have positive eigenvectors with eigenvalue 1 and any other eigenvalue of $\mathcal{A}$ has a modulus strictly less than 1.

It follows from the above that if our transition matrix $\mathcal{A}$ is primitive, then the dynamical system defined by

$$p_{n+1} = \mathcal{A}^T p_n, \tag{4.5}$$

for every choice of the initial probability distribution p_0, converges to p^* which satisfies $\mathcal{A}^T p^* = p^*$.

Next, note that when no observation is available Q_t becomes the identity matrix for both the discrete range and the continuous range RMAP (For the continuous range observation, let $C(\cdot) = 0$ which implies $\frac{\phi(Y_t - C(e_i))}{\phi(Y_t)} = 1$; a similar calculation holds for the prediction case with discrete range observation). Therefore, the information state S_t in Theorem 4.3.1 evolves according to the following recursion:

$$S_{t+1} = \mathcal{A}^T G_t \ S_t. \tag{4.6}$$

The above can be written explicitly:

$$S_{t+1} = \mathcal{A}^T \cdot diag \left(\ exp(\gamma < e_{argmax_i \ S_t^i} \ , \quad e_j >) \ \right) \cdot S_t \tag{4.7}$$

$$j = 1, 2, ..., N_{\mathbf{X}}, \quad i = 1, 2, ..., N_{\mathbf{X}}.$$

Unlike (4.5), the above is a non-linear dynamical system.

The generalized probability distribution U_t evolves according to

$$U_{t+1} = F(U_t) := H(\mathcal{A}^T G_t \ U_t), \tag{4.8}$$

where $H(X) = \frac{X}{\sum_i (X_i)}$ and $U_0 = p_0$. Finally, because $\mathcal{A}^T$ is stochastic, the above can be written as

$$U_{t+1} = \mathcal{A}^T \cdot H(diag \left(\ exp(\gamma < e_{argmax_i \ U_t^i} \ , \quad e_j >) \ \right) \cdot U_t). \tag{4.9}$$

Definition 4.5.4 A Cycle of Generalized Probabilities (CGP) is a finite set of probabilities $\{v^1, ..., v^m\}$ such that $F(v^i) = v^{i+1}$ with $F(v^m) = v^1$.

Conjecture 4.5.1 Let the stochastic matrix $\mathcal{A}$ be primitive; then, for every choice of the initial probability distribution p_0, the dynamical system

$$U_{t+1} = \mathcal{A}^T \cdot H(diag \left(\ exp(\gamma < e_{argmax_i \ U_t^i} \ , \quad e_j >) \ \right) \cdot U_t) \tag{4.10}$$

is asymptotically periodic, i.e., U_t approaches a CGP as $t \to \infty$.

Example 4.5.5 Let $\mathcal{A}$ be given by

$$\mathcal{A} = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \qquad e^\gamma = 1.078.$$
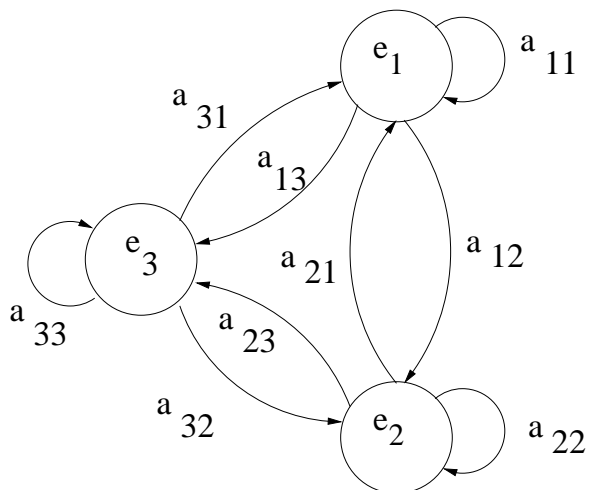
Figure 4.1: Example 4.5.7

We can write

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.4 \\ 0.2 & 0.6 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1.078 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

$$u_1 = \frac{0.8u_1 + 0.4312u_2}{u_1 + 1.078u_2} \qquad u_1 + u_2 = 1$$

$$0.078u_1{}^2 - 0.7092u_1 + 0.4312 = 0$$

$$u_1 = 0.65522 \quad u_2 = 0.34477$$

Our simulations show that (4.10), for every initial condition, approaches (0.65522, 0.34477) and so its CGP has only one element. The fixed point of (4.5) is

$$\mathcal{A}^T p^* = p^*$$

$$p_1 + p_2 = 1$$

$$p_1 = 0.66666 \quad p_2 = 0.33333.$$

Next consider

$$\mathcal{A} = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \qquad e^\gamma \to \infty$$

The CGP, for every initial condition, turns out to be

$$CGP : (v^1, v^2) \quad F(v^1) = v^2 \quad F(v^2) = v^1.$$

$$v^1 = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} \qquad v^2 = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}.$$

Example 4.5.6   Next consider

$$\mathcal{A} = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix} \qquad e^\gamma \to \infty.$$

The CGP's turn out to be made of two cycles

$$F(u) = u = \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix} \quad if \ u_1 \geq u_2$$

$$F(v) = v = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix} \quad if \ u_2 \geq u_1.$$

Thus, (4.10) unlike (4.5) could depend on the initial conditions.

Example 4.5.7   Consider the Markov chain given by Figure 4.1 with

$$\mathcal{A} = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.2 & 0.5 & 0.3 \\ 0.2 & 0.2 & 0.6 \end{bmatrix},$$

and $u_0 = (0.3\ 0.2\ 0.5)$. Let $e^\gamma = 2$. It turns out that (4.10) approaches the following CGP: $\{(0.2287\ 0.4243\ 0.3470),\ (0.2290\ 0.3679\ 0.4031)\}$ as $t \to \infty$. (4.5), on the other hand, approaches the stationary distribution $(0.2222\ \ 0.3810\ \ 0.3968)$. If we let $e^\gamma \to \infty$ the CGP becomes $\{(0.2484\ 0.3451\ 0.4065),\ (0.2418\ 0.5000\ 0.2582)\}$.

The period of CGP, in general, appears to increase with γ and approach a limiting CGP as $e^\gamma \to \infty$. If the CGP has only one element (m=1), the generalized probability is asymptotically stationary. This seems to be the case for small values of γ.

Asymptotic periodicity has been reported in the study of Markov operators of coupled map lattices [28]. Although the Markov operator in that context is linear and ours is not, the approach in [29] might provide us with insights as to how to prove Conjecture 4.5.1.

It is possible to give a "flow" interpretation of the non-linearity in (4.10). In two dimensions consider

$$(u_1 > u_2), \qquad H(\begin{bmatrix} 1 & 0 \\ 0 & e^\gamma \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}) \Rightarrow$$

$$\begin{bmatrix} u_1 \\ e^\gamma u_2 \end{bmatrix} \to \begin{bmatrix} \frac{u_1}{e^\gamma u_2 + u_1} \\ \frac{e^\gamma u_2}{e^\gamma u_2 + u_1} \end{bmatrix}.$$

$$u_1 - \frac{u_1}{e^\gamma u_2 + u_1} = \frac{e^\gamma u_1 u_2 - u_1(1 - u_1)}{e^\gamma u_2 + u_1}$$

$$= \frac{(e^\gamma - 1)u_1 u_2}{e^\gamma u_2 + u_1} = \beta,$$

β can be considered as a "back-flow" from the first state with the higher probability to the state with the lower one. Note that, as expected, $\gamma \to 0 \Rightarrow \beta \to 0$ and $\gamma \to \infty \Rightarrow \beta \to u_1$. It turns out that the flow out of p_1 is maximized when

$$u_1 = \frac{e^{\gamma/2}}{e^{\gamma/2} + 1}.$$

Similar calculations can be carried in higher dimensions. Assume $u_1 > u_i$, $i \neq 1$. Then

$$\beta = \frac{(e^\gamma - 1)u_1 \sum_{i \neq 1} u_i}{e^\gamma \sum_{i \neq 1} u_i + u_1}.$$

The back-flow is then distributed among the states according to

$$\beta_i = \beta \cdot \frac{u_i}{\sum_{i \neq 1} u_i}$$

The non–linear flow

The linear flow

β_i

a_{ij}

$\underset{i}{\text{Max}}\ u_i = u_1$

β

e_i

e_1
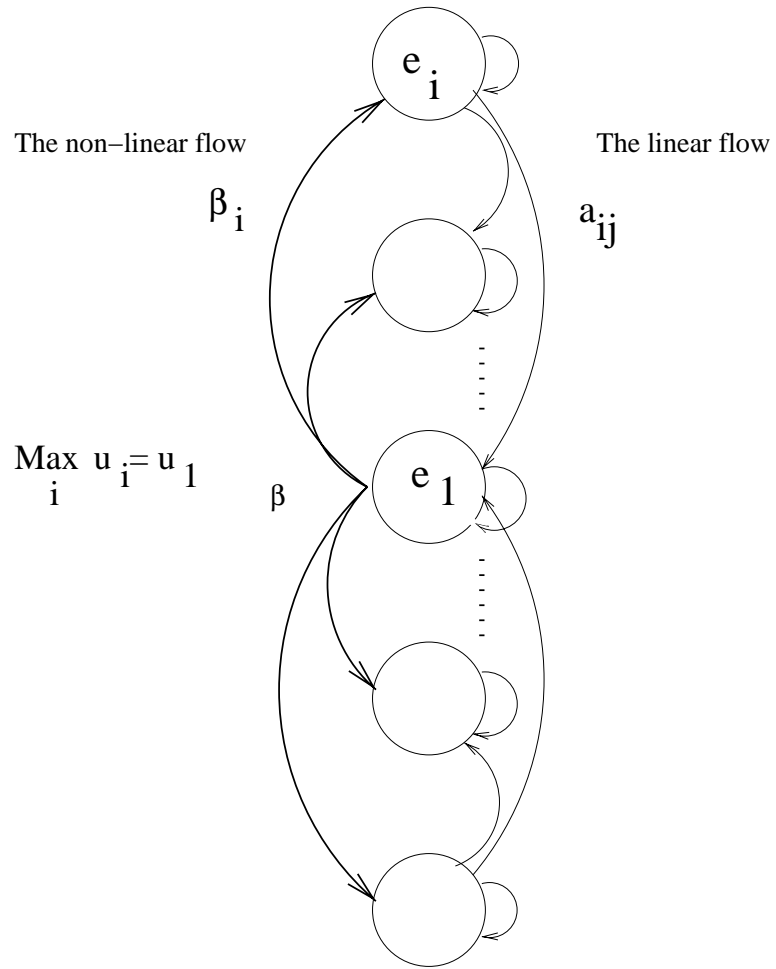
Figure 4.2: The back-flow.

ϵ	0.1	0.01	0.008	0.006	0.004	0.002	0.001	0.0002
Period	4	4	29	21	17	78	430	682

Table 4.1: Example 4.5.8

$$\beta = \sum_{i \neq 1} \beta_i.$$

See Figure 4.2.

Example 4.5.8 The period's dependence on the transition probabilities could be strong. Let

$$\mathcal{A} = \begin{bmatrix} 0.9 - \epsilon & 0.1 & \epsilon \\ 0.4 & 0.6 & 0.0 \\ 0.0 & \epsilon & 1.0 - \epsilon \end{bmatrix},$$

and $e^\gamma = 101$. The CGP's appear to be independent of the initial conditions but the period could depend strongly on ϵ as Table 4.1 shows. Graph 4.3 shows the values of the first component of the generalized probability vs. time for $\epsilon = 0.001$. (There are 10,000 data points and hence some apparent overlaps) Graph 4.4 shows the values of the second component of the generalized probability vs. time for $\epsilon = 0.0002$.
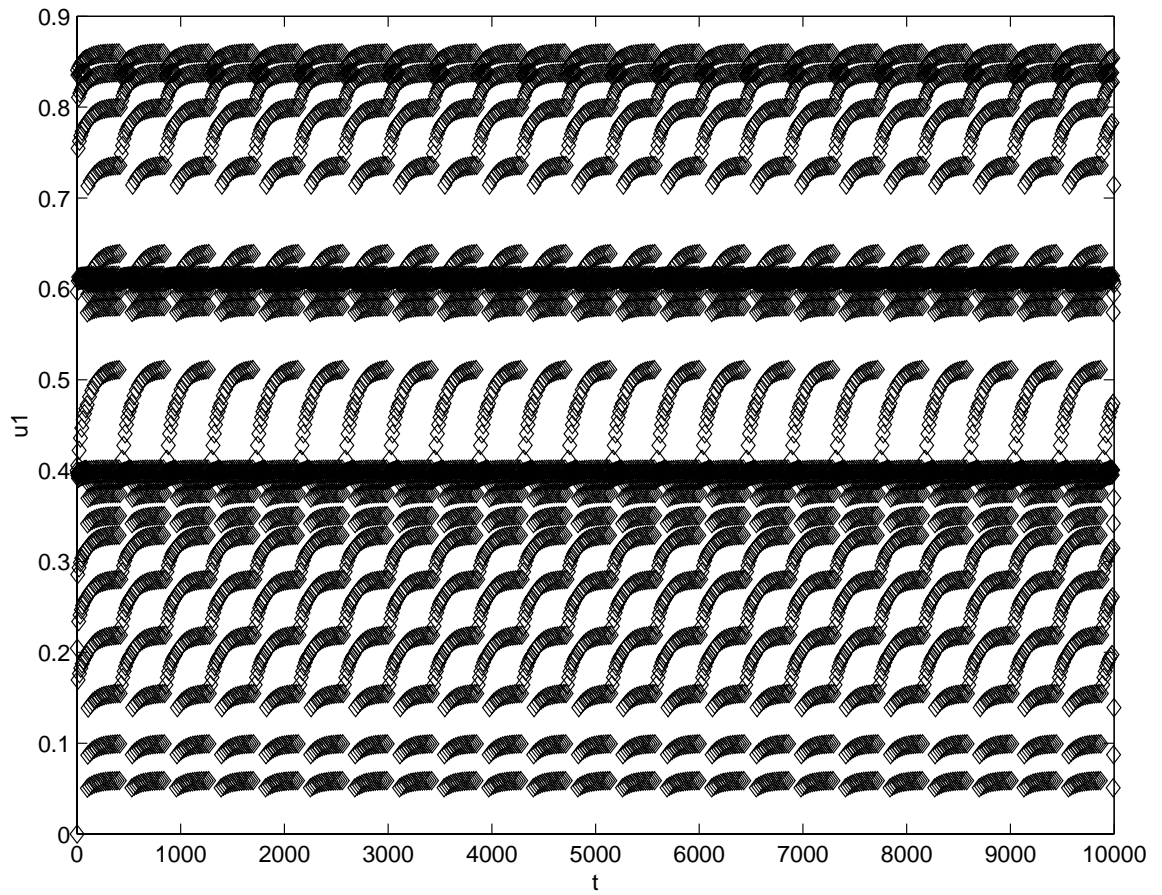
Figure 4.3: The first component of generalized probability vs. time for $\epsilon = 0.001$.
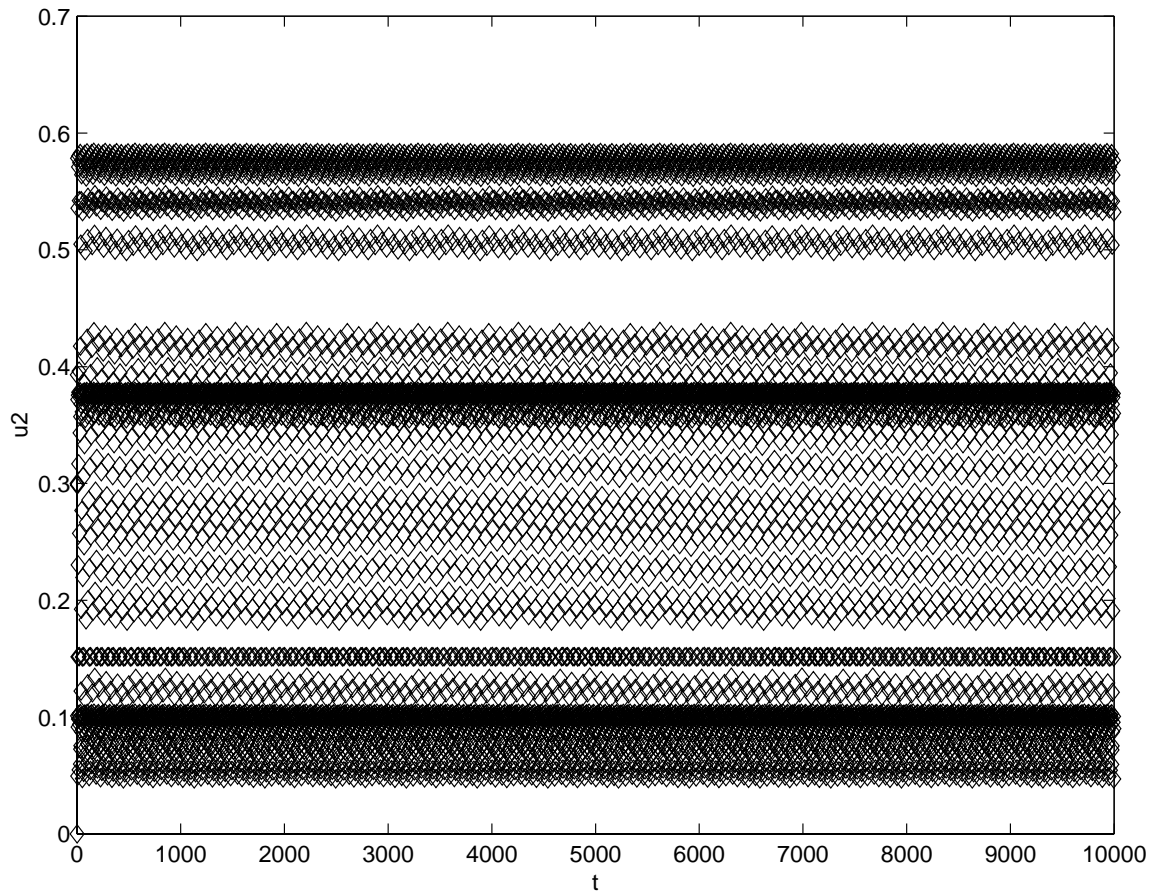
Figure 4.4: The second component of generalized probability vs. time for $\epsilon = 0.0002$.

Chapter 5

Risk sensitive filter banks and the qualitative analysis of their sample paths

In this chapter, a risk-sensitive generalization of the Maximum A Posterior Probability (MAP) estimation for partially observed Markov chains is presented. Using a change of measure technique, a sequential filtering scheme for the risk-sensitive state estimation is introduced. Structural results, the influence of the availability of information, mixing and non-mixing dynamics, and the connection with other risk-sensitive estimation methods are considered. A qualitative analysis of the sample paths clarifies the underlying mechanism. As in the previous chapters, we use the change of measure introduced in Chapter 2; however, the simplex is reserved for the *information* state and so to avoid confusion, the use of the simplex representation of the Markov chains will be implicit. We will use the simplex representation to carry out our proofs but sate the results without it.

5.1 Risk sensitive filter banks

5.1.1 The estimation of Hidden Markov Models

We define an estimated Hidden Markov Model as a five tuple $< \mathbf{X}, \mathbf{Y}, \mathbf{X}, \mathcal{A}, \mathcal{C} >$;
here $\mathcal{A}$ is the transition matrix, $\mathbf{Y} = \{1, 2, ..., N_{\mathbf{Y}}\}$ is the set of observations and
$\mathbf{X} = \{1, 2, ..., N_{\mathbf{X}}\}$ is the finite set of (internal) states as well as the set of estimates
or decisions. In addition, we have that $\mathcal{C} := [c_{i,j}]$ is the $N_{\mathbf{X}} \times N_{\mathbf{Y}}$ state/observation
matrix, i.e., $c_{i,j}$ is the probability of observing y=j when the state is x=i. We
consider the following information pattern. At decision epoch t, the system is
in the (unobservable) state $X_t = i$ and the corresponding observation $Y_t = j$ is
gathered, such that

$$P(Y_t = j | X_t = i) = c_{i,j} \tag{5.1}$$

The estimators V_t are functions of observations $(Y_0, ...Y_t)$ and are chosen according
to some specified criterion. Through out this paper, we use upper case letters to
denote estimators and script upper case letters to denote "estimation maps " from
observations to the set $\mathbf{X}$. If Y_t is an observation and V_t an estimator: $V_t = \mathcal{V}_t \circ Y_t$.
When it causes no confusion, we may use upper case letters for both.

5.1.2 Maximum A Posterior Probability Estimator

Consider a sequence of finite dimensional random variables X_t and the correspond-
ing observations Y_t defined on the common probability space $(\mathbf{\Omega}, F, \mathbf{P})$. The Max-
imum A Posteriori Probability (MAP) estimator $\widehat{X_t}$ is a Borel measurable function
of the filtration generated by observations up to Y_t denoted by $\mathcal{Y}_t$ which satisfies
for $\omega \in \Omega$:

$$\widehat{X}_t(\omega) = argmin_{\zeta \in \mathbf{X}} \quad E[\rho(X_t, \zeta) \mid \mathcal{Y}_t = \mathcal{Y}_t(\omega)] \qquad t = 0, 1, \ldots \tag{5.2}$$

where

$$\rho(X_t, \zeta) = \begin{cases} 0 & \text{if } X_t = \zeta; \\ 1 & otherwise. \end{cases}$$

The usual definition of MAP as the argument with the greatest probability given the observation follows from the above [13]. As in chapter 3, we will need the following Lemma:

Lemma 2.2.1: MAP also results from the additive cost minimization

$$(\widehat{X}_0, \ldots, \widehat{X}_N)(\omega) = argmin_{\zeta_0 \ldots \zeta_N \in \mathbf{X}^N} \quad E[\sum_{i=0}^{N} \rho(X_i, \zeta))\mid \mathcal{Y}_i = \mathcal{Y}_i(\omega)] \tag{5.3}$$

where $\mathbf{X}^N$ is the product space and each $\widehat{X}_i$ is $\mathcal{Y}_i$ measurable.

5.1.3   Change of measure

To carry out the computations, we will apply a change of measure technique introduced in chapter 2 and used in previous chapters. But this time, we let l=1 in (2.32) and use a slightly different notation for the Radon-Nikodym derivative. As before, let $(\mathbf{\Omega}, F, \mathbf{P})$ be the canonical probability space on which all of our time series are defined. Let $\mathcal{Y}_t$ be the filtration generated by the available observations up to decision epoch t, and let $\mathcal{G}_t$ be the filtration generated by the sequence of states and observations up to that time. Then the new probability measure $\mathcal{P}^\dagger$ is defined by the restriction of the Radon-Nikodym derivative on $\mathcal{G}_t$ to

$$\frac{d\mathcal{P}}{d\mathcal{P}^\dagger}|_{\mathcal{G}_t} = \bar{\Lambda}_t := N_{\mathbf{Y}}^t \cdot \Pi_{k=0}^t q_{X_k, Y_k} \tag{5.4}$$

where the function q_{X_k,Y_k} is a random variable defined by $q_{X_k=i,Y_k=j} = c_{ij}$.

A little thought should convince the reader that the right hand side of the above is equal to (2.51) and hence under the new measure $\{Y_t\}$ is independently and identically distributed (i.i.d.). Each distribution is uniform over the set $\mathbf{Y}$ and $\{Y_t\}$ is independent of $\{X_t\}$.

Before we can introduce our filter, we need an optimization result. Let V_t be measurable functions of observations up to t taking values in $\{X_t\}$ and $\rho(\cdot, \cdot)$ as above. Fix $V_0, ..., V_{k-1}$. We would like to find $\widehat{V}_k, ..., \widehat{V}_{H-1}$ such that the following criterion is minimized :

$$S^\gamma(V_0, ..., V_{H-1}) := E[exp(\gamma \cdot \mathcal{C}_H)], \tag{5.5}$$

where

$$\mathcal{C}_H := \sum_{t=0}^{H-1} \rho(X_t, V_t) \tag{5.6}$$

and γ is a strictly positive parameter. The minimum value will be denoted by $\bar{S}^\gamma(\widehat{V}_k, ..., \widehat{V}_{H-1})$.

This optimization problem can be solved via dynamic programming. A detailed analysis will be given shortly in Lemma 2.3.1. We need to define recursively an *information state*

$$\sigma_{t+1}^\gamma = N_\mathbf{Y} \cdot \overline{Q}(Y_{t+1})\mathcal{D}^T(V_t) \cdot \sigma_t^\gamma, \tag{5.7}$$

where $\overline{Q}(y) := diag(q_{i,y})$, $\mathcal{A}^T$ denotes the transpose of the matrix $\mathcal{A}$ and the matrix $\mathcal{D}$ is defined by

$$[\mathcal{D}(v)]_{i,j} := a_{i,j} \cdot exp(\gamma\rho(i,v)). \tag{5.8}$$

σ_0^γ is set equal to $N_{\mathbf{Y}} \cdot \overline{Q}(Y_0)p_0$, where p_0 is the initial distribution of the state and is assumed to be known. In the context of the risk-sensitive estimation of Markov chains, the meaning of σ_t^γ will become clear.

We define the matrix

$$L(v,y) := N_{\mathbf{Y}} \cdot \overline{Q}(y)\mathcal{D}^T(v). \tag{5.9}$$

We will show that

$$S^\gamma(V_0, ..., V_{H-1}) = E[exp(\gamma \cdot \mathcal{C}_H)] = E^\dagger[\sum_{i=1}^{N_X} \sigma_H^\gamma(i)], \tag{5.10}$$

where $E^\dagger$ is the expectation with respect to the new measure. Define the value functions $J^\gamma(\cdot, H-j) : R_+^{N_{\mathbf{X}}} \to R$, $j = H, ..., H-k$, as follows:

$$J^\gamma(\sigma, H-j) := \min_{V_{H-j}...V_{H-1}} \left\{E^\dagger\{\sum_{i=1}^{N_X} \sigma_H^\gamma(i) \mid \sigma_{H-j}^\gamma = \sigma\}\right\}. \tag{5.11}$$

Lemma 5.1.1   Let $V_0, ..., V_{k-1}$ be given. Then the value functions defined above are obtained from the following dynamic programming equations:

$$\begin{cases} \bar{J}^\gamma(\sigma, H) &= \sum_{i=1}^{N_{\mathbf{X}}} \sigma(i); \\ \\ \bar{J}^\gamma(\sigma, H-j) &= \min_{v \in \mathbf{X}}\{E^\dagger[\bar{J}^\gamma(L(v, Y_{H-j+1}) \cdot \sigma, H-j+1)]\}, \\ & \quad\quad j = 1, 2, ..., H-k \end{cases} \tag{5.12}$$

The estimation maps $\widehat{\mathcal{V}}_k, ..., \widehat{\mathcal{V}}_{H-1}$ obtained from (2.11) are risk optimal; i.e., $\widehat{\mathcal{V}}_k(\sigma_k^\gamma), ..., \widehat{\mathcal{V}}_{H-1}(\sigma_{H-1}^\gamma)$ achieve the minimum in (2.5).

Proof:

We have

$$S^\gamma(V_0, ..., V_{H-1}) = E^\dagger[\sum_{i=1}^{N_X} \sigma_H^\gamma(i)] = E^\dagger E^\dagger\{[\sum_{i=1}^{N_X} \sigma_H^\gamma(i)]|\sigma_k^\gamma\}.$$

Thus, to obtain $\bar{S}^\gamma(\widehat{V}_k, ..., \widehat{V}_{H-1})$, it is sufficient to minimize the expectation

$$E^\dagger\{[\sum_{i=1}^{N_X} \sigma_M^\gamma(i)]|\sigma_k^\gamma\}$$

which will show is given by the above dynamic programming equation.

We use the simplex representation of Markov chains developed in Chapter 2 by mapping the states of the chain onto the unit vectors of $R_{N_\mathbf{X}}$ denoted by $e_1....e_{N_\mathbf{X}}$. First, we define an information state and show that it evolves according to (2.7):

$$< \sigma_t, e_h > := E^\dagger[\bar{\Lambda}_t exp\{\gamma \sum_{i=0}^{t-1} \rho(X_i, V_i)\} <X_t, e_h> |\mathcal{Y}_t]. \tag{5.13}$$

Therefore,

$$< \sigma_{t+1}, e_h > \ = \ E^\dagger[\ \bar{\Lambda}_{t+1}\ exp\{\gamma \sum_{i=0}^{t} \rho(X_i, V_i)\} <X_{t+1}, e_h> |\mathcal{Y}_{t+1}]$$

$$= \ N_\mathbf{Y} E^\dagger[q_{X_{t+1}, Y_{t+1}}\ \bar{\Lambda}_t exp\{\gamma \sum_{i=0}^{t} \rho(X_i, V_i)\} <\mathcal{A}^T X_t + Z_{t+1}, e_h> |\mathcal{Y}_{t+1}],$$

where Z_{t+1} is a martingale increment defined by $Z_{t+1} := \mathcal{A}^T X_{t+1} - X_t$. Using the linearity of the inner-product and the expectation and the i.i.d. properties of the observations under the new measure, it can be shown that Z_{t+1} drops out of the above expression to give

$$< \ \sigma_{t+1}, e_h > = N_\mathbf{Y} E^\dagger[q_{e_h, Y_{t+1}}\ \bar{\Lambda}_t\ exp\{\gamma \sum_{i=0}^{t} \rho(X_i, V_i)\} <X_t, \mathcal{A}.e_h> |\mathcal{Y}_{t+1}]$$

$$
\begin{aligned}
&= N_{\mathbf{Y}} E^{\dagger}[\sum_{j=1}^{N_{\mathbf{X}}} q_{e_h, Y_{t+1}} \; \bar{\Lambda}_t \; exp\{\gamma \sum_{i=0}^{t} \rho(X_i, V_i)\} a_{jh} < X_t, e_j > |\mathcal{Y}_{t+1}] \\
&= N_{\mathbf{Y}} \sum_{j=1}^{N_{\mathbf{X}}} q_{e_h, Y_{t+1}} E^{\dagger}[\; \bar{\Lambda}_t \; exp\{\gamma \sum_{i=0}^{t} \rho(X_i, V_i)\} a_{jh} < X_t, e_j > |\mathcal{Y}_{t+1}] \\
&= N_{\mathbf{Y}} \sum_{j=1}^{N_{\mathbf{X}}} q_{e_h, Y_{t+1}} exp\{\rho(e_j, V_t)\} a_{jh} \\
&\qquad\qquad E^{\dagger}[\; \bar{\Lambda}_t \; exp\{\gamma \sum_{i=0}^{t-1} exp\{\rho(X_i, V_i)\} < X_t, e_j > |\mathcal{Y}_t, Y_{t+1}].
\end{aligned}
$$

But since Y_{t+1} is independent of the σ-algebra generated jointly by $\mathcal{Y}_t$ and all the random variables within the expectation, it too drops out to give

$$
N_{\mathbf{Y}} \sum_{j=1}^{N_{\mathbf{X}}} q_{e_h, Y_{t+1}} exp\{\gamma \rho(e_j, V_t)\} a_{jh} < \sigma_t, e_j >
$$

which results in (2.7) when put in the matrix form. The initial condition is given by

$$
\begin{aligned}
< \sigma_0, e_h > \; &:= \; E^{\dagger}[\bar{\Lambda}_0 < X_0, e_h > |\mathcal{Y}_0] \\
&= \; N_{\mathbf{Y}} E^{\dagger}[q_{X_0, Y_0} < X_0, e_h > |Y_0] \\
&= \; N_{\mathbf{Y}} q_{e_h, Y_0} E^{\dagger}[< X_0, e_h > |Y_0] \\
&= \; N_{\mathbf{Y}} q_{e_h, Y_0} E^{\dagger}[< X_0, e_h >].
\end{aligned}
$$

It is not hard to show that $E^{\dagger}[< X_0, e_h >] = E[< X_0, e_h >] =< p_0, e_h >$ where p_0 is the initial distribution of the state and so

$$
\sigma_0 = N_{\mathbf{Y}} \cdot \overline{Q}(Y_0) p_0.
$$

The information state is all we need to determine the optimal cost because of the following equality:

$$E[exp(\gamma \cdot \mathcal{C}_H)] = E^\dagger[\bar{\Lambda}_H exp(\gamma \cdot \mathcal{C}_H)] = E^\dagger[\sum_{i=1}^{N_X} \sigma_H^\gamma(i)].$$

The first equality is an application of discrete-time Girsanov Theorem ([see Theorem 3.2 in [6]]); the second follows directly from the definition of the information state as a conditional expectation and the observation that $\sum_{i=1}^{N_X} < X_t, e_i >= 1$ since in the simplex representation $X_t \in \{e_1, ...e_{N_X}\}$.

Let $\mathcal{W}_{i,j}$ denote the set of estimates $V_i, ..., V_j$ where each V_i is $\mathcal{Y}_i$ measurable. With the notation and definitions of section 2.3 write

$$\widehat{S}^\gamma(t) := inf\,_{\mathcal{W}_{t,H-1}} E^\dagger\{[\sum_{i=1}^{N_X} \sigma_H^\gamma(i)]|\mathcal{Y}_t\}. \qquad (*)$$

Define recursively the functions

$$\begin{cases} \bar{J}^\gamma(\sigma, H) & = \sum_{i=1}^{N_\mathbf{X}} \sigma(i); \\ \\ \bar{J}^\gamma(\sigma, t) & = \min_{v \in \mathbf{X}}\{E^\dagger[\bar{J}^\gamma(L(V_t, Y_{t+1}) \cdot \sigma_t, t+1)|\sigma_t = \sigma, V_t = v]\}. \\ & t = k, k+1, \ldots, H-1 \end{cases} \qquad (**)$$

Now assume that for t+1,..., H-1, $(*)$ satisfies the above dynamic programming equation; i.e., assume that $inf\,_{\mathcal{W}_{t+1,H-1}} E^\dagger\{[\sum_{i=1}^{N_X} \sigma_H^\gamma(i)]|\mathcal{Y}_{t+1}\} = \bar{J}(\sigma_{t+1}^\gamma(\mathcal{Y}_{t+1}), t+1)$ with a choice of minimizing estimates $\widehat{\mathcal{V}}_{t+1}(\sigma_{t+1}^\gamma)....\widehat{\mathcal{V}}_{H-1}(\sigma_{H-1}^\gamma)$ obtained from $(**)$. We will show that the same statement holds true for t,...H-1.

$$\begin{aligned} \widehat{S}^\gamma(t) &= inf\,_{\mathcal{W}_{t,H-1}} E^\dagger\{[\sum_{i=1}^{N_X} \sigma_H^\gamma(i)]|\mathcal{Y}_t\} \\ &= inf\,_{\mathcal{W}_{t,H-1}} E^\dagger\{E^\dagger\{[\sum_{i=1}^{N_X} \sigma_H^\gamma(i)]|\mathcal{Y}_{t+1}\}|\mathcal{Y}_t\} \\ &= inf\,_{\mathcal{W}_{t,t}} E^\dagger inf\,_{\mathcal{W}_{t+1,H-1}} E^\dagger\{\{[\sum_{i=1}^{N_X} \sigma_H^\gamma(i)]|\mathcal{Y}_{t+1}\}|\mathcal{Y}_t\} \\ &= inf\,_{\mathcal{W}_{t,t}} E^\dagger\{\bar{J}(\sigma_{t+1}^\gamma(\mathcal{Y}_{t+1}), t+1)|\mathcal{Y}_t\} \end{aligned}$$

$$= \ \inf{}_{\mathcal{W}_{t,t}} E^{\dagger}\{\bar{J}(L(V_t, Y_{t+1})\sigma_t^{\gamma}(\mathcal{Y}_t),\ t+1)|\sigma_t^{\gamma}(\mathcal{Y}_t), V_t\}$$

$$= \ \bar{J}(\sigma_t^{\gamma}(\mathcal{Y}_t),\ t).$$

The interchange of the expectation and the infimum follows from the lattice property (Appendix C). The next two steps are true by the induction hypothesis and by the integral characterization of the conditional expectation from which one can show that in general $E[f(W, g(Z))|Z = z] = E[f(W, g(z))] = E[f(W, g(Z))|g(Z) = g(z)]$ if random vectors Z and W are independent (Appendix B). The last step is by $(**)$. The case t=H-1 is easily verified. Thus, by induction, the claim holds for k,...H-1. Showing that each minimizing V_t is σ_t^{γ} measurable can be done explicitly. The pair (σ_t^{γ}, V_t) is measurable with respect to $\mathcal{Y}_t$ which is independent of Y_{t+1} with a uniform distribution for Y_{t+1}. Thus $(**)$ reduces to the minimization of the following sum:

$$\tfrac{1}{N_{\mathbf{Y}}} \textstyle\sum_{y=1}^{N_{\mathbf{Y}}} [\bar{J}^{\gamma}(L(V_t, y) \cdot \sigma_t^{\gamma}, t + 1)].$$

It is now obvious that the minimizing V_t must be a function of σ_t.

The above shows that the minimizing sequence of the estimates depend only on the information state. We can write

$$\inf{}_{\mathcal{W}_{t,H-1}} E^{\dagger}\{[\sum_{i=1}^{N_X} \sigma_H^{\gamma}(i)]|\sigma_t\} \ = \ E^{\dagger}\{\inf{}_{\mathcal{W}_{t,H-1}} E^{\dagger}\{[\sum_{i=1}^{N_X} \sigma_H^{\gamma}(i)]|\mathcal{Y}_t\}|\sigma_t\}$$

$$= \ E^{\dagger}\{\bar{J}(\sigma_t^{\gamma}(\mathcal{Y}_t),\ t)|\sigma_t\} = \bar{J}(\sigma_t^{\gamma}(\mathcal{Y}_t))$$

which shows

$$\inf{}_{\mathcal{W}_{t,H-1}} E^{\dagger}\{[\sum_{i=1}^{N_X} \sigma_H^{\gamma}(i)]|\sigma_t\} = \inf{}_{\mathcal{W}_{t,H-1}} E^{\dagger}\{[\sum_{i=1}^{N_X} \sigma_H^{\gamma}(i)]|\mathcal{Y}_t\}.$$

It is straightforward to verify that the estimators, $\widehat{\mathcal{V}}_k(\sigma_k^{\gamma}), ..., \widehat{\mathcal{V}}_{H-1}(\sigma_{H-1}^{\gamma})$ are risk optimal, i.e., they achieve the minimum in (2.5).

Changing the index of the dynamic programming according to t= H-j (for easing our notation in the later sections) finishes the proof of the Lemma.

5.1.4 The T-step MAP risk-sensitive estimator (TMAP)

The T-step MAP risk-sensitive estimator (TMAP) is defined by the following criterion:

$$
\widehat{V}_{NT}, ..., \widehat{V}_{(N+1)T-1} = argmin_{V_t \in X, \ t=NT,...,(N+1)T-1}
$$
$$
E\{exp(\gamma \cdot [\sum_{t=0}^{NT-1} \rho(X_t, \widehat{V}_t) + \sum_{t=NT}^{(N+1)T-1} \rho(X_t, V_t)])\}, \quad (5.14)
$$

where V_t is $\mathcal{Y}_t$ measurable, T is the size of the filter and $N = 0, 1, ...,$ is the index of filtering segments. This exponential criterion is a generalization of the risk-sensitive filtering idea introduced in [10] for the quadratic cost with the filtering performed in single steps, i.e., for T=1; we will look at this special case for TMAP in sections 3 and 4 and show that it is essentially a "greedy algorithm".

Theorem 5.1.2   The TMAP can be computed recursively by the following procedure:

1) Set $\sigma_0 = N_{\mathbf{Y}} \cdot \overline{Q}(Y_0)p_0$.

2) Given that σ_{NT}, use the minimizing sequence of the value functions obtained from the following dynamic programming equations

$$
\begin{cases}
\bar{J}^\gamma(\sigma, T) & = \sum_{i=1}^{N_{\mathbf{X}}} \sigma(i); \\
\\
\bar{J}^\gamma(\sigma, T-j) & = \min_{v \in \mathbf{X}} \{E^\dagger[\bar{J}^\gamma(L(v, Y_{NT+T-j+1}) \cdot \sigma, T-j+1)]\} \\
& \quad\quad j = 1, 2, \ldots, T
\end{cases}
\quad (5.15)
$$

80
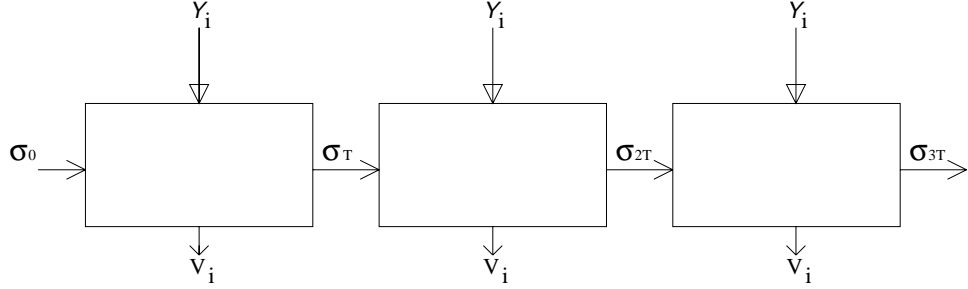
Figure 5.1: The T-step risk sensitive filter banks.

to determine the value of the optimum estimates $\widehat{V}_{NT}, ..., \widehat{V}_{(N+1)T-1}$ as a function of the information state $\sigma_{NT}, ..., \sigma_{(N+1)T-1}$ obtained by (2.7).

3) Apply (2.7) once more to obtain $\sigma_{(N+1)T}$ and repeat steps (2) and (3) starting at (N+1)T.

Furthermore, for any given N as $\gamma \to 0$, TMAP (i.e. the above algorithm) reduces to MAP.

Proof:

The proof follows from repeated applications of Lemma 2.3.1. We will skip the details. The limiting result follows from the first order approximation of the exponential function and the observation that as $\gamma \to 0$, the matrix $\mathcal{D}(v) \to \mathcal{A}$ element wise. This implies that in the limit the input to each filtering step is the unnormalized conditional distribution and thus by Lemma 2.2.1 the filtering process reduces to the well known MAP estimation of HMM's.

Note that although the size of the sum $\sum_{i=0}^{NT-1} \rho(X_t, \widehat{V}_t)$ increases with N, all we need to track is the information state, computed recursively. The optimal estimates $\widehat{\mathcal{V}}_{NT}(\sigma_{NT}), ..., \widehat{\mathcal{V}}_{(N+1)T-1}(\sigma_{(N+1)T-1})$ are measurable functions of the information state alone. Since our Markov chain is homogeneous and under the new measure the observations are i.i.d, (2.13) depends only on T and not on N. This justifies

the sequential filter banks representation of Figure 1.

We point out that theorem 5.1.2 has no control counterpart. In this case, the estimators $\{V_t\}$ have no influence on the dynamics and thus estimation can be broken down into separate segments with the information state reinitialized. The same cannot be said for a controlled Markov chain due to the influence of the controller on the dynamics; the separate segments cannot be joined to represent the entire process in the above limiting sense as the "decoupling" Lemma 2.2.1 no longer holds. Note also that controlled Markov chains are not homogeneous.

5.2 Structural results: the filter banks and the information state

It is clear from the above that to describe the behavior of TMAP we must understand the operation of each filtering segment and understand the meaning of the information state. The key in understanding the filter's operation is the analysis of the value functions which are obtained via dynamic programming.

Lemma 5.2.1 The value functions are continuous and concave functions of the information state $\sigma \in R_+^{N\mathbf{x}}$.

Proof:

Both statements are proved by induction. The continuity when j=0 is obvious. Since $\{Y_i\}$ is i.i.d., uniformly distributed and finite dimensional, then (5.15) is taking a minimum over the average of composition of functions, each of which is continuous by the continuity of the linear functions and the induction hypothesis, and is therefore continuous. For the second statement, once again the case j=0 is trivially verified. Assume concavity for j-1. Let $0 \leq \lambda \leq 1$ and $\sigma_1, \sigma_2 \in R_+^{N\mathbf{x}}$;

define $\tilde{\sigma} := \lambda\sigma_1 + (1-\lambda)\sigma_2$. Then by the induction hypothesis and by (5.15) we have

$$
\begin{aligned}
\bar{J}^\gamma(\tilde{\sigma}, T-j) &= \min_{v \in \mathbf{X}}\Big\{\frac{1}{N_\mathbf{Y}}\sum_{y=1}^{N_\mathbf{Y}}\bar{J}^\gamma(L(v,y)\cdot\tilde{\sigma}, T-j+1)\Big\} \\
&\geq \min_{v \in \mathbf{X}}\Big\{\frac{1}{N_\mathbf{Y}}\sum_{y=1}^{N_\mathbf{Y}}[\lambda\bar{J}^\gamma(L(v,y)\cdot\sigma_1, T-j+1) \\
&\quad + (1-\lambda)\bar{J}^\gamma(L(v,y)\cdot\sigma_2, T-j+1)]\Big\} \\
&\geq \lambda\bar{J}^\gamma(\sigma_1, T-j) + (1-\lambda)\bar{J}^\gamma(\sigma_2, T-j).
\end{aligned}
$$

Next, for P a finite set of vectors in $R_+^{N_\mathbf{X}}$, denote by $\mathcal{O}(P)$ the set

$$
\mathcal{O}(P) := \Big\{\frac{1}{N_\mathbf{Y}}\sum_{y=1}^{N_\mathbf{Y}}\alpha_y \cdot L(v,y) \mid \alpha_y \in P, v \in \mathbf{X}\Big\}. \tag{5.16}
$$

Note that if P is finite so is $\mathcal{O}(P)$, since $|\mathcal{O}(P)| \leq |P|^{N_\mathbf{Y}}\cdot N_\mathbf{X}$.

Lemma 5.2.2 The value functions given by (5.1.2) are piecewise linear functions (hyper-planes through the origin) of $\sigma \in R_+^{N_\mathbf{X}}$, such that if P_{j-1} indicates the vectors in $R_+^{N_\mathbf{X}}$ which specify the set of hyper planes for $\bar{J}^\gamma(\sigma, T-j+1)$ then

$$
\bar{J}^\gamma(\sigma, T-j+1) = \min_{\alpha \in P_{j-1}}\{\alpha \cdot \sigma\} \qquad \bar{J}^\gamma(\sigma, T-j) = \min_{\alpha \in \mathcal{O}(P_{j-1})}\{\alpha \cdot \sigma\}, \tag{5.17}
$$

where $P_0 = \bar{\mathbf{1}} := (\sum_{k=1}^{N_\mathbf{X}} e_k)^\mathbf{T}$ and $\{e_k\}$ are the unit vectors in $R^{N_\mathbf{X}}$.

Proof:

The statement of the Lemma is readily verified for $j = 0$. Assume the Lemma holds for j-1 then piecewise linearity implies that for each vector α_0 in P_{j-1} there is a point $\sigma^* \in R^{N_\mathbf{X}}$ and a disk $d(\sigma^*, r)$ such that on this disk $\bar{J}^\gamma(\sigma, T-j+1) = \alpha_0 \cdot \sigma$.

Consider a different point σ and $0 < t \leq 1$ small enough so that $t(\sigma - \sigma^*) + \sigma^* \in d(\sigma^*, r)$. Then by the concavity shown above

$$\bar{J}^\gamma(\sigma^* + t(\sigma - \sigma^*), T - j + 1) = \bar{J}^\gamma((1 - t)\sigma^* + t\sigma, T - j + 1)$$
$$\geq (1 - t)\bar{J}^\gamma(\sigma^*, T - j + 1) + t\bar{J}^\gamma(\sigma, T - j + 1). \tag{5.18}$$

Since $\bar{J}^\gamma(\sigma^* + t(\sigma - \sigma^*), T - j + 1) = \alpha_0 \cdot \sigma^* + t(\sigma - \sigma^*)$, after substitution and cancellations, we get

$$\bar{J}^\gamma(\sigma, T - j + 1) \leq \alpha_0 \cdot \sigma.$$

But α_0 was arbitrary and so we have the first equality. To prove the second equality, note that

$$
\begin{aligned}
\bar{J}^\gamma(\sigma, T - j) &= \min_{v \in \mathbf{X}}\Big\{\frac{1}{N_\mathbf{Y}} \sum_{y=1}^{N_\mathbf{Y}} \min_{\alpha \in P_{j-1}} \{\alpha \cdot L(v, y) \cdot \sigma\}\Big\} \\
&= \min_{v \in \mathbf{X}}\Big\{\Big[\frac{1}{N_\mathbf{Y}} \sum_{y=1}^{N_\mathbf{Y}} \tilde{\alpha}(v, y, \sigma) \cdot L(v, y)\Big] \cdot \sigma\Big\} \\
&= \min_{\alpha \in \mathcal{O}(P_{j-1})} \{\alpha \cdot \sigma\}, \tag{5.19}
\end{aligned}
$$

where $\tilde{\alpha}(v, y, \sigma) \in P_{j-1}$ minimizes $\alpha \cdot L(v, y) \cdot \sigma$ in the first equality, and the last equality follows since $\alpha \cdot L(u, y) \cdot \sigma > \tilde{\alpha}(v, y, \sigma) \cdot L(u, y) \cdot \sigma$, for all $\alpha \in P_{j-1}$, $v \in \mathbf{X}$, $y \in \mathbf{Y}$, $\sigma \in R_{N_\mathbf{X}}^+$.

Lemma 5.2.3   The optimal estimates $\{\widehat{V}_t\}$ are constant along rays through the origin, i.e., let $\sigma \in R_+^{N_\mathbf{X}}$ then $\widehat{V}_t(\sigma') = \widehat{V}_t(\sigma)$, for all $\sigma' = \lambda\sigma$, $\lambda > 0$.

Proof:

From Lemma 5.2.2, we see that $\bar{J}^\gamma(\sigma', T - j) = \lambda\bar{J}^\gamma(\sigma, T - j)$. The result follows from Theorem 5.1.2.

Definition 5.2.4   A cone in $R_+^{N_\mathbf{X}}$ is a set defined by $C_S := \{\sigma | \sigma = \lambda x, x \in S \subset R_+^{N_\mathbf{X}}, \lambda > 0\}$.

Definition 5.2.5   For j=1,2, ..., T and $v \in X$, let

$$\bar{J}_v^\gamma(\sigma, T - j) \ := \ E^\dagger[\bar{J}^\gamma(L(v, Y_{T-j+1}) \cdot \sigma, T - j + 1)]$$

$$= \ \frac{1}{N_\mathbf{Y}} \sum_{y=1}^{N_\mathbf{Y}} [\bar{J}^\gamma(L(v, y) \cdot \sigma, T - j + 1)]. \qquad (5.20)$$

Definition 5.2.6   The decision region $DR_v^j \subset R_+^{N_\mathbf{x}}$ for the estimate $v \in \mathbf{X}$, at the $T - j$ decision epoch, is defined as

$$DR_v^j := \{\sigma \mid \sigma \in R_+^{N_\mathbf{x}}, \bar{J}^\gamma(\sigma, T - j) = \bar{J}_v^\gamma(\sigma, T - j)\}. \qquad (5.21)$$

It follows from the definition of $\widehat{\mathcal{V}}_{NT+T-j}(\sigma)$ that

$$DR_v^j := \{\sigma \mid \sigma \in R_+^{N_\mathbf{x}}, \quad \widehat{\mathcal{V}}_{NT+T-j}(\sigma) = v\}. \qquad (5.22)$$

We say a decision is made "strictly", if it is the only possible decision.

Theorem 5.2.7   For each $v = i \in \{1, 2, ..., \mathbf{N_X}\}$ and for every $j = 1, 2, ..., T$, the decision region DR_i^j is always non-empty and includes a cone about the σ_i axis within which the decision made is (strictly) $\widehat{\mathcal{V}}_{NT+T-j}(\sigma) = i$.

Proof:

We state the proof for $N_X = 2$, from which the proof for the general case will become evident for the reader. On the σ_1 axis, we have by definition

$$\bar{J}_1^\gamma(\sigma, T - j) = \frac{1}{N_\mathbf{Y}} \sum_{y=1}^{N_\mathbf{Y}} [\bar{J}^\gamma(N_\mathbf{Y} \cdot \overline{Q}(y)\mathcal{A}^T \begin{bmatrix} 1 & 0 \\ 0 & e^\gamma \end{bmatrix} \begin{bmatrix} \sigma_1 \\ 0 \end{bmatrix}, T - j + 1)].$$

$$\bar{J}_2^\gamma(\sigma, T - j) = \frac{1}{N_\mathbf{Y}} \sum_{y=1}^{N_\mathbf{Y}} [\bar{J}^\gamma(N_\mathbf{Y} \cdot \overline{Q}(y)\mathcal{A}^T \begin{bmatrix} e^\gamma & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 \\ 0 \end{bmatrix}, T - j + 1)].$$

$$\bar{J}_1^\gamma(\sigma, T-j) = \frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}} [\bar{J}^\gamma(N_{\mathbf{Y}} \cdot \overline{Q}(y)\mathcal{A}^T \begin{bmatrix} \sigma_1 \\ 0 \end{bmatrix}, T-j+1)].$$

$$\bar{J}_2^\gamma(\sigma, T-j) = \frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}} [\bar{J}^\gamma(N_{\mathbf{Y}} \cdot \overline{Q}(y)\mathcal{A}^T \begin{bmatrix} e^\gamma \sigma_1 \\ 0 \end{bmatrix}, T-j+1)].$$

Applying Lemma 5.2.2 to each term of the summation on the right-hand side of $\bar{J}_2^\gamma(\sigma, T-j)$, we get

$$\bar{J}_2^\gamma(\sigma, T-j) = \frac{e^\gamma}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}} [\bar{J}^\gamma(N_{\mathbf{Y}} \cdot \overline{Q}(y)\mathcal{A}^T \begin{bmatrix} \sigma_1 \\ 0 \end{bmatrix}, T-j+1)].$$

Therefore, we can write

$$\bar{J}_2^\gamma(\sigma, T-j) - \bar{J}_1^\gamma(\sigma, T-j) = (e^\gamma - 1)\{\frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}} [\bar{J}^\gamma(N_{\mathbf{Y}} \cdot \overline{Q}(y)\mathcal{A}^T \begin{bmatrix} \sigma_1 \\ 0 \end{bmatrix}, T-j+1)]\}.$$

But $e^\gamma > 1$ since $\gamma > 0$ and for every j, the value functions are strictly positive being integrals of the exponential functions. Thus from the above on the σ_1 axis we have the strict inequality

$$\bar{J}_1^\gamma(\sigma, T-j) < \bar{J}_2^\gamma(\sigma, T-j)$$

which implies DR_1^j includes the σ_1 axis; fix σ on that axis, then by Lemma 3.1.1 and because sums and compositions of continuous functions are continuous, there exists a disk of positive radius in the R_+^2 metric, i.e., $d(r, \sigma) \cap R_+^2$, $r > 0$ such that $\bar{J}_1^\gamma(\sigma, T-j) < \bar{J}_2^\gamma(\sigma, T-j)$ for every $x \in d(r, \sigma) \cap R_+^2$. Therefore, Lemma 5.2.3 implies that on the cone $\mathbf{C}_{d(r,\sigma) \cap R_+^2} \subset DR_1^j$ the decision is strictly $v = 1$. The same proof works in higher dimensions by fixing an axis σ_l and making pairwise comparisons between $\bar{J}_l^\gamma(\sigma, T-j)$ and $\bar{J}_k^\gamma(\sigma, T-j)$, $k \neq l$ along the σ_l axis. The

"strict" cone around the σ_l axis will be the intersection of all cones obtained from pairwise comparisons.

In general, the boundaries among the decision regions are not of "threshold type" (unlike MAP). We give a two dimensional example. We consider the TMAP with $N_{\mathbf{X}} = 2$.

Remark: The transition cone $DR_1^j \cap DR_2^j$ is not, in general, of threshold type, i.e., the cone $DR_1^j \cap DR_2^j$ does not degenerate to a line; we give a simple counter example.

Let $a_{11} = a_{22} = 1$ and $q_{xy} = 1/2$; then it can be shown (by straightforward induction) that the transition cones are not degenerate.

Let's look at the counter example more closely (Figure 5.2). The cone where the decision is strictly v=1, i.e., $R_+^2 \cap (DR_2^j)^c$ is given by $\sigma_1 > \sigma_2 \cdot exp(\gamma(j - 1))$ and by symmetry $R_+^2 \cap (DR_1^j)^c$ is given by $\sigma_2 > \sigma_1 \cdot exp(\gamma(j - 1))$. The transition cone (where either decision is acceptable) is given by the complement of the union of these two regions (the colored area). The value functions are given by $j + 1$ hyper-planes: $\sigma_1 + exp(\gamma(j))\sigma_2$, $\sigma_1 exp(\gamma(1)) + exp(\gamma(j - 1))\sigma_2$, $\sigma_1 exp(\gamma(2)) + exp(\gamma(j - 2))\sigma_2$ ..., $\sigma_2 + exp(\gamma(j))\sigma_1$ on the $j + 1$ cones beginning with $\sigma_1 > \sigma_2 \cdot exp(\gamma(j - 1))$ and ending with $\sigma_2 > \sigma_1 \cdot exp(\gamma(j - 1))$. The transition cone between them is the union of $j - 1$ cones whose boundaries are lines: $exp(-(j - 1)\gamma)\sigma_1 = \sigma_2$, $exp(-(j - 2)\gamma)\sigma_1 = \sigma_2$,..., $\sigma_1 = \sigma_2$,..., $exp(-(j - 2)\gamma)\sigma_2 = \sigma_1$, ..., $exp(-(j - 1)\gamma)\sigma_2 = \sigma_1$. When j is odd, the line $\sigma_1 = \sigma_2$ is a boundary (boundary in the sense of slope change in the cost and not decision); when j is even, it is not. (The solid cone which includes $\sigma_1 = \sigma_2$ for even values of j is meant to emphasize this). On the transition cone either decision is allowed. We can interpret this region as the zone of uncertainty. For MAP and TMAP with T=1 (we'll show
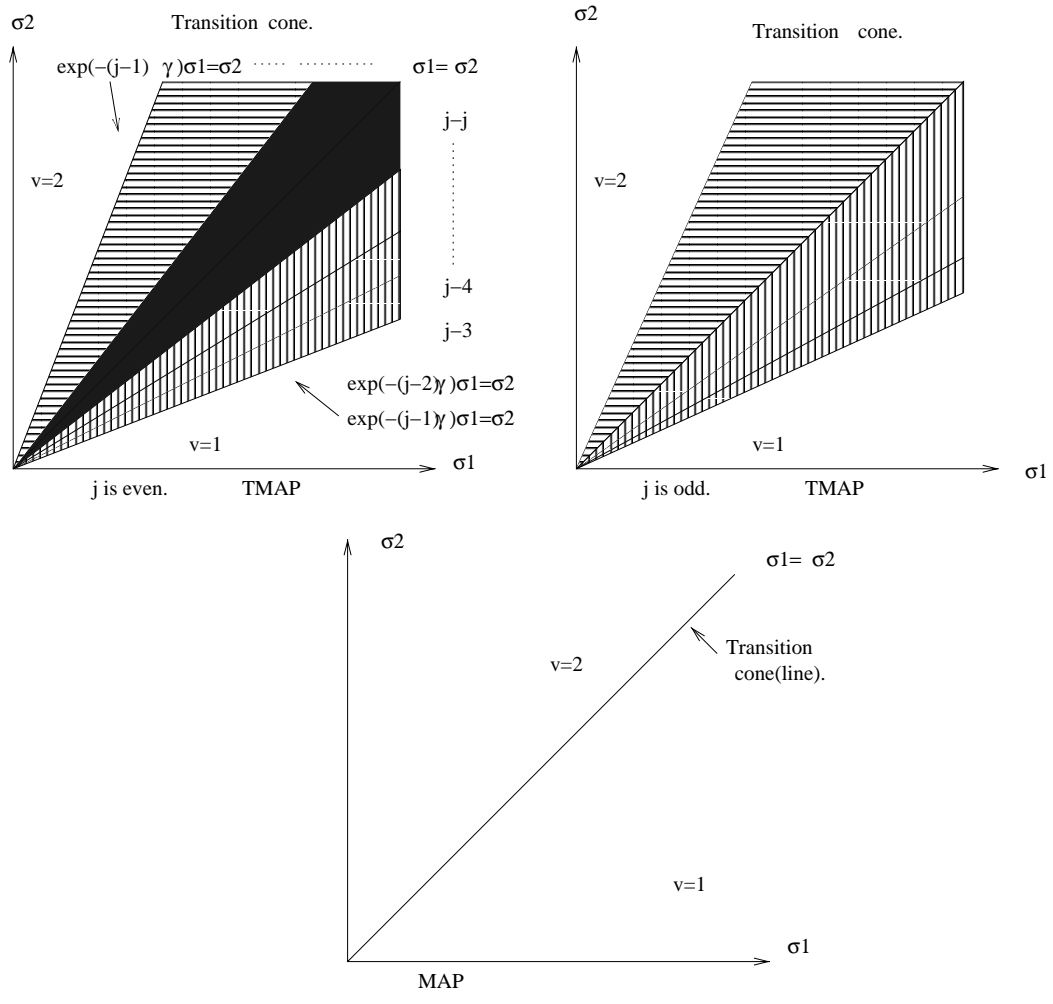
Figure 5.2: Decision cones for $a_{11} = a_{22} = 1, q_{xy} = 1/2$.

this later) this region is the threshold $\sigma_1 = \sigma_2$, but as the above example shows, for TMAP with $T > 1$, it may be a non-degenerate cone. We could interpret this as reflecting the "conservative" nature of the risk-sensitive estimation. We are expanding the zone of "uncertainty" at the expense of the region of "certainty". We will show in the subsequent sections that this is not always the manner in which risk-sensitivity manifests itself in the structure of decision regions. It is possible that the the transition cone remains degenerate and either of the two other cones expands at the expense of the other or the decision regions are not affected by risk sensitivity at all, i.e., they remain identical to that of MAP. In two dimensions for example, $DR_1^j = \{\sigma | \sigma_1 > \sigma_2\}$ and $DR_2^j = \{\sigma | \sigma_2 > \sigma_1\}$.

The above theorem only guarantees the existence of of non-degenerate cones around the σ_l axis but says nothing about their size. In fact, observe that in the above example the size of these cones becomes arbitrarily small as $\gamma \to \infty$ since the slopes of the lines $exp(-(j-1)\gamma)\sigma_1 = \sigma_2$ and $exp(-(j-1)\gamma)\sigma_2 = \sigma_1$, for every $j > 1$, converge to zero and infinity respectively.

Two special cases (N=0, T=M) and (T=1, N=0,..., M-1) are of interest. In both cases, the index t ranges from t=0 to t=M-1. In the first case, TMAP reduces to the exponential/sum criterion for HMM's which is the discrete and finite dimensional version of the risk-sensitive L^2 filter introduced by Speyer and others. The second case would be the MAP version of the quadratic cost risk-sensitive filtering introduced first to our best knowledge by Dey and Moore in [10]. Obviously, the structural results obtained so far apply to these special cases.

Theorem 5.2.8   Let $\mathcal{EX}(v)$ be the diagonal matrix $diag[exp(\gamma\rho(i,v))]$, $i = 1,..., N_{\mathbf{X}}$. Then the one step TMAP decision regions are given by $\widehat{\mathcal{V}}_t(\sigma) = i$ if $\sigma_i \geq \sigma_j$, $\forall j \neq i$.

Proof:

From the definition, we have: $\mathcal{D}^T(v) = \mathcal{A}^T \mathcal{E} \mathcal{X}(v)$ and that $\bar{J}^\gamma(\sigma, T) = <\sigma, \bar{\mathbf{1}}>$ and thus

$$
\begin{aligned}
\bar{J}_v^\gamma(\sigma, T-1) &= \frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}} [\bar{J}^\gamma(L(v,y) \cdot \sigma, \ T)] \\
&= \frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}} [\bar{J}^\gamma(N_{\mathbf{Y}} \cdot \overline{Q}(y) \mathcal{A}^T \mathcal{E} \mathcal{X}(v) \cdot \sigma, T)] \\
&= \sum_{y=1}^{N_{\mathbf{Y}}} < \overline{Q}(y)(\mathcal{A}^T \mathcal{E} \mathcal{X}(v) \cdot \sigma), \bar{\mathbf{1}} > \\
&= < (\sum_{y=1}^{N_{\mathbf{Y}}} \overline{Q}(y))(\mathcal{A}^T \mathcal{E} \mathcal{X}(v) \cdot \sigma), \bar{\mathbf{1}} > \\
&= < I \mathcal{A}^T \mathcal{E} \mathcal{X}(v) \cdot \sigma, \bar{\mathbf{1}} > \\
&= < \mathcal{E} \mathcal{X}(v) \cdot \sigma, \mathcal{A} \bar{\mathbf{1}} > \quad = \quad < \mathcal{E} \mathcal{X}(v) \cdot \sigma, \bar{\mathbf{1}} > . \quad (5.23)
\end{aligned}
$$

A little calculation shows that given σ the above is minimized, if we set v equal to the index of the largest component of σ, i.e., if σ_l is the largest component of σ then $v = l$. This is precisely how the decision regions for MAP are defined.

Note that TMAP for T=1 is not reduced to MAP; although the decision regions are the same, the information states are different. In the case of MAP, the information state is the conditional distribution, while in the case of TMAP the information state is given by (5.7). The conditional distribution has no memory of decisions made in the past while the TMAP's information state does depend on these decisions. On the other hand, when γ is very small, (5.7) becomes the unnormalized conditional distribution. We can think of TMAP's information state as the conditional distribution modified by the sequence of the decisions made in the past. This modified information state is then put through the next decision region which itself is calculated to minimize a certain cost structure based on the averaged behavior of the future sample paths: $(E^\dagger[\bar{J}^\gamma(L(v, Y_{NT+T-j+1}) \cdot \sigma, T-j+1)])$.

How far we look into the "averaged future" is determined by T. The information from the past is encapsulated in the information state. These ideas will become quantified in the next sections.

Theorem 5.2.9 The value function $\bar{J}^{\gamma}(\sigma, T - j)$ when $\forall (x, y)$, $q_{yx} = 1/N_{\mathbf{Y}}$ is given by

$$\bar{J}^{\gamma}(\sigma, T - j) =$$
$$\min_{v_{T-j},\ldots,v_{T-1}} <\sigma, \ \mathcal{EX}(v_{T-j}) \cdot \mathcal{A} \cdot \mathcal{EX}(v_{T-j+1}) \cdot \mathcal{A} \cdot \mathcal{EX}(v_{T-j+2})...\mathcal{A} \cdot \mathcal{EX}(v_{T-1}) \cdot \bar{1}> .$$
$$(5.24)$$

Proof:

The proof is based on the same technique used in the previous theorem and will be skipped.

In the above counter example when $\mathcal{A} = I_{2\times 2}$ and $q_{yx} = 1/2$, by the above theorem, we have

$$\bar{J}^{\gamma}(\sigma, T - j) = \min_{v_{T-j},\ldots,v_{T-1}} < \sigma, \mathcal{EX}(v_{T-j}) \cdot ...\mathcal{EX}(v_{T-j}) \cdot \bar{1} > .$$

If we let the number of times we choose $x = 2$ be n_2, and likewise for $x = 1$, $n_1 = T - n_2$, a little algebra will show that the total cost $\bar{J}^{\gamma}(\sigma, 0)$ is given by

$$\sigma_1 exp\{\gamma n_2\} + \sigma_2 exp\{\gamma(T - n_2)\}.$$

By differentiation with respect to n_2, a few rearrangements and taking logarithms, the minimum cost is obtained when (modulo the integer parts)

$$T/2 - \frac{1}{2\gamma}log(\sigma_1/\sigma_2) = n_2; \ \ 0 \le n_2 \le T. \tag{5.25}$$

91

This suggests that for large values of γ regardless of σ, we choose the two states an approximately equal number of times. To see why this limiting behavior occurs first let $\sigma_1 = \sigma_2 = 1/2$, then according to the decision regions, we could choose either 1 or 2. Suppose we choose 1, then the information state evolves to $(\sigma_1, \sigma_2 e^\gamma) = (1/2, e^\gamma/2)$. We could continue to choose v=1 and "move up" on the information state (successive) 2-dimensional planes. Note that the upper boundary of the transition cone is given by $\sigma_2 exp(\gamma) + exp(\gamma(j-1))\sigma_1$. Thus, going forward in time, the information state moves up while the upper boundary, $\sigma_2 exp(\gamma) + exp(\gamma(j-1))\sigma_1$, "moves down" since going forward, $j-1$ in the above changes to $(j-1)-1 = j-2$. Therefore, at about $j/2$ the information state moves out of the transition cone and enters the upper cone where we will have to choose 2 as our estimate. A similar argument shows that from that point on, the information state "moves forward" on the information state (successive) 2-dimensional planes and as the upper boundary $\sigma_2 exp(\gamma) + exp(\gamma(j-1))\sigma_1$ continues to decrease in slope, the information state σ remains on the decision region of the state 2. Hence, the decision sequence $\{1, 1, 1, ...2, 2, 2\}$ with about the same number of 1's and 2's (exactly if j is even) is a possible solution. When γ is large, every point not on the axes falls into the transition cone and a similar argument shows that it follows the same decision pattern.

Now consider the case T=1. In this case the transition cone is reduced for all values of j, to the line $\sigma_1 = \sigma_2$ and a similar argument shows that for large values of γ the decision is $\{1, 2, 1, 2, 1, 2...\}$. This simple example provides insight as to how the two special cases differ. The decision regions for $T = M > 1$ are more complex but it appears that this allows for a "smoother" solution. We will return to this subject in section 4.

Definition 5.2.10   A completely symmetric HMM is defined as an HMM whose transition probability matrix is symmetric in the sense that $a_{ii} = 1 - \epsilon$, $a_{ji} = \epsilon/(N_{\mathbf{X}} - 1)$ for all $j \neq i$, $N_{\mathbf{X}} = N_{\mathbf{Y}}$ (so that Q is a square matrix) and $q_{yx} = q$ for $x = y$ and otherwise $q_{yx} = (1 - q)/(N_{\mathbf{Y}} - 1)$.

Note that the discrete metric which induces MAP (and in turn TMAP) is symmetric in the sense that $d(x_1, y_1) = d(x_2, y_2)$ $\forall x_i \neq y_i$. Therefore, for a completely symmetric HMM, all criteria upon which the determination of the value functions at a particular σ depend are symmetrically defined and thus the value functions are unchanged under the permutation of components of σ.

In two dimensions, a completely symmetric HMM is given by

$$\mathcal{A} = \begin{bmatrix} 1 - \epsilon & \epsilon \\ \epsilon & 1 - \epsilon \end{bmatrix} \quad Q = \begin{bmatrix} q & 1 - q \\ 1 - q & q \end{bmatrix}. \tag{5.26}$$

In the above example we have set $\epsilon = 0$ and $q = 1/2$.

Theorem 5.2.11   For a completely symmetric HMM, the value functions $\bar{J}^\gamma(\sigma, T - j)\}$ restricted to the simplex (the set $\{\sigma | \sigma_1 + \sigma_2 .... + \sigma_{N_{\mathbf{X}}} = 1\}$) have their global maximum at the center of the simplex, i.e., at $\sigma_1 = \sigma_2 = ... = \sigma_{N_{\mathbf{X}}}$.

Proof:

The restriction of a concave function to a convex region is certainly concave. The simplex is a compact set, and so the value function restricted to the simplex is concave and has a maximum point. Let's begin with the 1-simplex. By the complete symmetry of the system, a value function will be symmetric with respect to the center of the simplex, namely around the point $\sigma_1 = \sigma_2 = 1/2$. To see this let $\sigma = (\sigma_1, \sigma_2)$, $\sigma_1 + \sigma_2 = 1$ be an arbitrary point on the simplex. Consider this point and its mirror image $\sigma' = (\sigma_2, \sigma_1)$ obtained from σ by a permutation

of its components ($\sigma_1 \rightarrow \sigma_2, \sigma_2 \rightarrow \sigma_1$) which leaves the value function unchanged. Because the value functions are concave we can write

$$\bar{J}^\gamma(1/2, T-j) = J((\sigma+\sigma')/2, T-j) \geq 1/2\{\bar{J}^\gamma(\sigma, T-j) + \bar{J}^\gamma(\sigma', T-j)\} = \bar{J}^\gamma(\sigma, T-j).$$
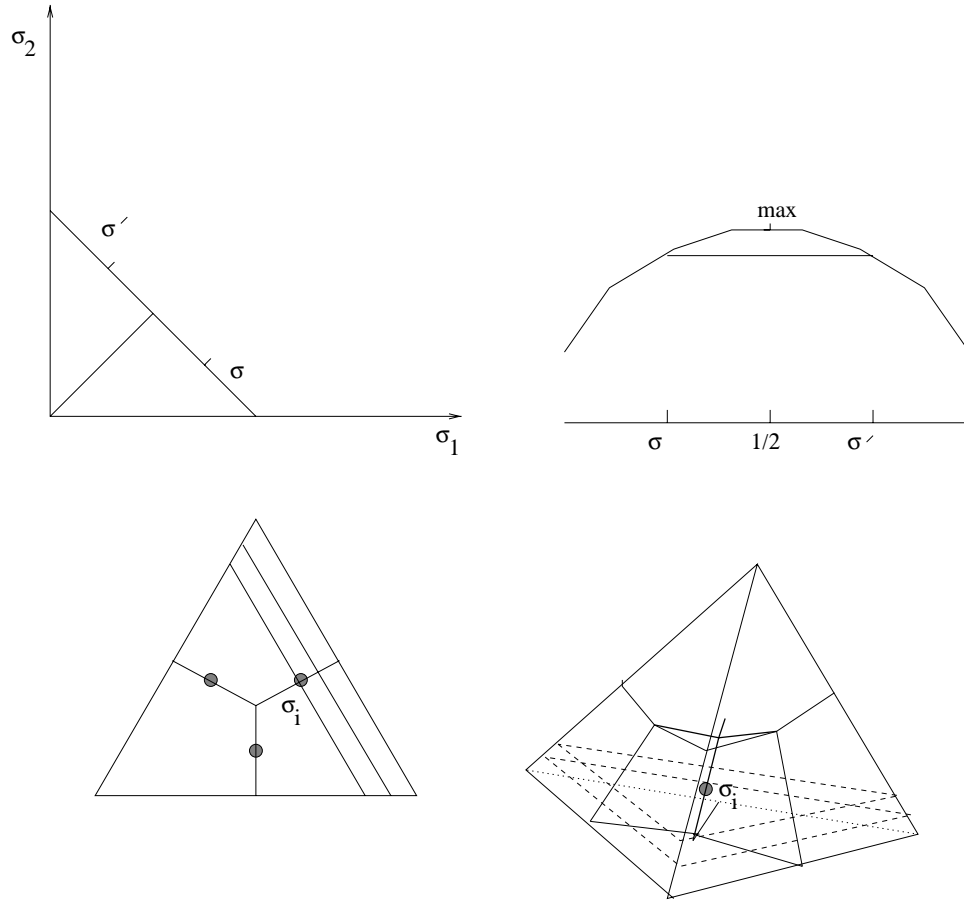

Figure 5.3: The value functions' global maximum under complete symmetry.

Therefore, the center of the simplex is the global maximum.

In general, for an $(N_\mathbf{X} - 1)$-simplex, consider $N_\mathbf{X}$ permutations of σ

$$\sigma_1 \rightarrow \sigma_2, \sigma_2 \rightarrow \sigma_3..., \sigma_i \rightarrow \sigma_{i+1}, \quad \sigma_{N_\mathbf{X}} \rightarrow \sigma_1$$

94

repeated $N_{\mathbf{X}}$ times and generating $N_{\mathbf{X}}$ points σ^k $k = 1, ..., N_{\mathbf{X}}$. Then by symmetry and by Jensen's inequality:

$$\bar{J}^\gamma([\sigma^1 + ... + \sigma^{N_{\mathbf{X}}}]/N_{\mathbf{X}}, T - j) \geq$$
$$1/N_{\mathbf{X}} \cdot \bar{J}^\gamma(\sigma^1, T - j) + ... + 1/N_{\mathbf{X}} \cdot \bar{J}^\gamma(\sigma^{N_{\mathbf{X}}}, T - j) = \bar{J}^\gamma(\sigma, T - j).$$

Let $\sigma^* = [\sigma^1 + ... + \sigma^{N_{\mathbf{X}}}]/N_{\mathbf{X}}$. Then $\sigma_i^* = 1/N_{\mathbf{X}} \sum_{l=1}^{l=N_{\mathbf{X}}} \sigma_l = 1/N_{\mathbf{X}}$ since σ is a point of the simplex. Therefore, σ^* is the center of the simplex and by the above the maximal point. This completes the proof of the theorem.

In fact a stronger version of the theorem holds. We will only outline the proof. Consider the 2-simplex which is bounded by three 1-simplices. Complete symmetry of system implies that along these simplices and all the line segments parallel to them, the value function has its maximum at the middle of such lines. Thus the maximum on the whole simplex can be found by considering only the values on the set of the mid-points of all these line segments joined at the center of the simplex. Similarly, a $(N_{\mathbf{X}} - 1)$-simplex is bounded by $N_{\mathbf{X}}$, $(N_{\mathbf{X}} - 2)$-simplices. Along these sub-simplices and hyper-planes parallel to them and restricted to the simplex, the value functions have their maximum at the central points, reaching their global maximum at the center of the $(N_{\mathbf{X}} - 1)$-simplex. Note that the maximum point need not be unique. In fact, in the above example for even values of j the maximum point is not unique.

If the assumption of complete symmetry is removed, Theorem 5.2.10 no longer holds. We will show this shortly.

5.3 Qualitative analysis of the sample paths and the information state

5.3.1 Sample path error smoothing

It is often said that risk sensitive estimators take into account "the higher order moments" of the cost. How do higher order moments manifest themselves in the behavior of the estimator? To understand what risk sensitivity take into account, we first explain what MAP does not.

Given a single observation and a single random variable, the MAP estimation is a reasonable thing to do: minimize the measure of the set where the estimate and the random variable disagree. But a problem arises when we consider a stochastic process (in our case a time series). The obvious thing to do then is to minimize the expected summation of the error functions for the whole time series. As shown before, this reduces back to finding the MAP estimate at each instant of time. Thus, at each instant of time our decision is not affected by our past or future decisions. This makes the MAP estimation insensitive to the accumulation of errors along sample paths. To see this evaluate X_k and the estimate $\widehat{X}_k$ at some fixed value of $\omega \in \Omega$ to produce a sample path or a realization of the respective time series and its estimate. The sequence of decisions $\widehat{X}_k \quad k = 0, 1, ...N$ partitions the sample space into 2^N subsets according to a binary tree with the branching criterion $X_k = \widehat{X}_k$ or $X_k \neq \widehat{X}_k$. Each ω belongs to exactly to one of these subsets. Some sample paths may end up on branches along which estimation errors accumulate for long stretches of time. Now consider TMAP with T=1 (which is the Dey-Moore filter equivalent for MAP). The exponential function turns the sum of the error functions into a product, the value of this product up to the last decision made for

each fixed ω is $e^{\gamma \cdot m}$, where m counts the number of times an error has been made in estimation of that particular sample path. This product is then multiplied by either 1 or $e^{\gamma} > 1$ depending on the next decision made based on minimizing the measure of the error set at the present time *and* taking note of the errors made in the past. Thus the errors made in the past become more uniformly distributed over all sample paths. The information state condenses this error history and makes the computation recursive. For this reason perhaps *sample path error smoothing estimators* could be an alternative name for the risk-sensitive estimators. Dey-Moore filter is (in a sense) a greedy algorithm which does not consider (on the average) the accumulation of errors in the future but has the important benefit of computational simplicity. The exponential/sum filter does so for the entire path and in general TMAP looks into the future for T steps. In Figure 5.4, observe that a particular sample path of our familiar example becomes more oscillatory as T is made smaller. Our simulations show that this is a general behavior no matter how complicated the chain: the smaller the filter size T, the bigger is the burden of error smoothing on the T next estimates. But making T large comes at the cost of increased computational complexity.

The underlying mechanism of all these estimators is the coupling of the estimation errors in the product form.

5.3.2   Risk-sensitivity, information and mixing

Through the sample path perspective, we can explain the behavior of the risk-sensitive estimators. In HMM's all sample paths pass through a finite number of states. We can think of the transition probabilities as determining a flow in the system. So far the only example we have considered was a non-mixing dynamical
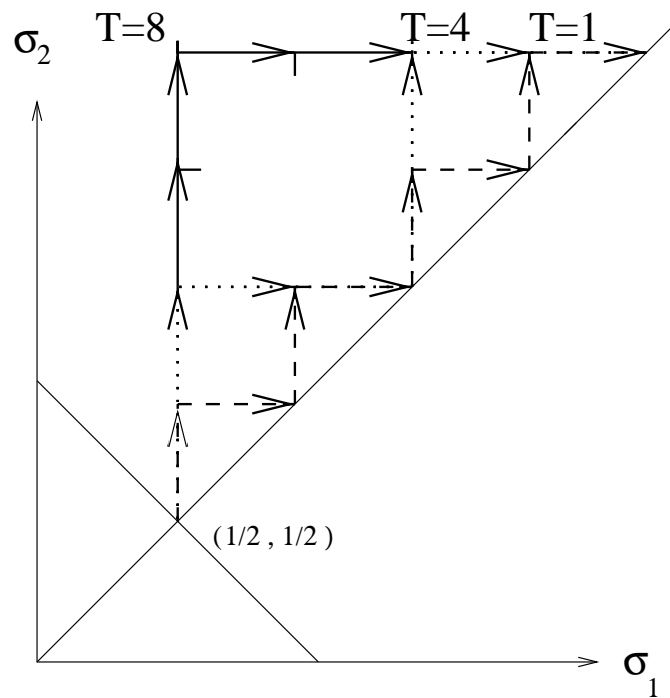
Figure 5.4: Sample paths for $a_{11} = a_{22} = 1, q_{xy} = 1/2$ starting from $(1/2, 1/2)$.

system. The transition probabilities were set to be zero and there was no flow between the states. This is important from the error smoothing point of view, for as the flow passes through the states, so does the history of the errors. If sample paths that have accumulated estimation error remain in a particular state, the estimator will be "attracted" to the state in order to relieve the error accumulated there. This explains the oscillatory behavior of our two state example in which no flow between the states was allowed. On the other hand, if the transition probabilities are non-zero this "attraction" is somewhat relieved; we have verified through simulations that mixing indeed inhibits the oscillatory behavior. This will also have an effect on the decision regions as we will see shortly. But if we go through a state "too quickly", we cannot use that state to smoothen the error accumulated in the path effectively. Both these cases lead to certain type of singularities in the decision regions.

The second issue is the role of information. If we expect to receive good information about the system in the future, which will in turn reduce the error accumulation, we are likely to be less conservative at the present about our decisions. This means that we expect TMAP's decision regions to become less conservative and look more like MAP's under increased availability of information. This too will be shown in the following example.

We will study the decision regions for T=2 TMAP for an HMM with $N_{\mathbf{X}} = N_{\mathbf{Y}} = 2$ and

$$\mathcal{A} = \begin{bmatrix} 1/2 & 1/2 \\ \delta & 1-\delta \end{bmatrix}; \quad Q = \begin{bmatrix} 1/2 + I & 1/2 - I \\ 1/2 - I & 1/2 + I \end{bmatrix}. \tag{5.27}$$

The parameter I controls the availability of information. When $I = 0$, no information is available (the case of pure prediction) and as $I \to 1/2$, the HMM becomes

perfectly observable. The parameter δ determines the transition probabilities of the second state and in particular $\delta = 0$ will make the Markov chain non-mixing. As shown before for T=1, the decision regions are identical to those of MAP. First let $I = 0$, then it can be shown that for T=2, the decision regions of j=2 (the first stage of the two step filter) are of the threshold type determined by a transition line L with the slope $m(\delta)$ followed by the equi-partition decision regions (identical to the decisions regions of MAP). The decision regions of the first stage are given by $\sigma_2 < m(\delta)\sigma_1$ choose 1 and if $\sigma_2 > m(\delta)\sigma_1$ choose 2. The slope $m(\delta)$ is given by

$$
m(\delta) = \begin{cases} \frac{e^\gamma + 1}{2} \cdot \frac{1}{\delta e^\gamma + 1 - \delta} & \delta < 1/2; \\ \\ \frac{e^\gamma + 1}{2} \cdot \frac{1}{(1-\delta)e^\gamma + \delta} & \delta > 1/2. \end{cases}
$$

Simple calculations show that the slope is always greater than or equal to 1 (only when $\delta = 1/2$), so that the decision region of the first state is enlarged at the expense of the second. As expected when $\gamma \to 0$, the decision regions equalize. When $\gamma \to \infty$, the slope is given by

$$
m(\delta) = \begin{cases} \frac{1}{2\delta} & \delta < 1/2; \\ \\ \frac{1}{2(1-\delta)} & \delta > 1/2. \end{cases}
$$

When either $\delta = 0$ or $\delta = 1$, the slope becomes infinite. These are the singularities that we mentioned earlier. The equalization of the two regions at $\delta = 1/2$ is a general property which holds true even when no constraint is put on the available information as the following theorem demonstrates:

Theorem 5.3.1 Consider the HMM described by (5.1). The risk-sensitive decision regions are equalized under uniform flow: $a_{ij} = 1/N_\mathbf{X} \quad \forall(i, j)$. Furthermore, TMAP reduces to MAP for every choice of T and γ.

Proof:

Fix T and γ. Consider the calculations of the decision regions according to (5.15). For j=1, we already saw that the decision regions are equalized. For $j > 1$ we can write

$$
\begin{aligned}
\bar{J}^{\gamma}(\sigma, T - j) &= \min_{v \in \mathbf{X}}\{E[\bar{J}^{\gamma}(L(v, Y_{T-j+1}) \cdot \sigma, T - j + 1)]\} \\
&= \min_{v \in \mathbf{X}}\{\frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}}[\bar{J}^{\gamma}(N_{\mathbf{Y}} \cdot \overline{Q}(y)\mathcal{A}^{T}\mathcal{E}\mathcal{X}(v) \cdot \sigma, T - j + 1)]\} \\
&= \min_{v \in \mathbf{X}}\{\frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}}[\bar{J}^{\gamma}(N_{\mathbf{Y}} \cdot \overline{Q}(y)\{< \mathcal{E}\mathcal{X}(v) \cdot \sigma, \bar{\mathbf{1}} >\}\mathcal{A}^{T} \cdot \bar{\mathbf{1}}, T - j + 1)]\} \\
&= \min_{v \in \mathbf{X}}\{< \mathcal{E}\mathcal{X}(v) \cdot \sigma, \bar{\mathbf{1}} > \frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}}[\bar{J}^{\gamma}(N_{\mathbf{Y}} \cdot \overline{Q}(y)\mathcal{A}^{T} \cdot \bar{\mathbf{1}}, T - j + 1)]\} \\
&= \min_{v \in \mathbf{X}}\{< \mathcal{E}\mathcal{X}(v) \cdot \sigma, \bar{\mathbf{1}} >\}\{\frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}}[\bar{J}^{\gamma}(N_{\mathbf{Y}} \cdot \overline{Q}(y)\mathcal{A}^{T} \cdot \bar{\mathbf{1}}, T - j + 1)]\}.
\end{aligned}
$$

The second term does not depend on v and so the minimization reduces to

$$
\min_{v \in \mathbf{X}}\{< \mathcal{E}\mathcal{X}(v) \cdot \sigma, \bar{\mathbf{1}} >\}.
$$

This shows that the decision regions are equalized. Similar techniques applied to the evolution of the information state show that TMAP's and MAP's information states, under uniform flow, are multiples of each other. This fact together with Lemma 3.1.3 and the above result regarding the decision regions completes the proof.

In the above result, the observation matrix $\overline{Q}(y)$ plays no role under the assumption of uniform flow in the determination of the decision regions. But this is the exception. In general, the availability of information appears to have an elegant relation to the structure of the decision regions as the following shows.

Proposition 5.3.2   In 4.1, let $\delta = 0$ and $I \geq \frac{1}{2(1+e^{-\gamma})}$; then the decision regions for TMAP, T=2 are equalized.

Proof:

The proof follows from the solution of a system of simultaneous inequalities defined by (5.15) with the constraints $\delta = 0$ and $I \geq \frac{1}{2(1+e^{-\gamma})}$. We skip the tedious algebra.

As we mentioned earlier, this does not imply that TMAP for T=2 reduces to MAP because the recursions governing the evolution of the information states for the two cases are different. But if for some T the decision regions are equalized then TMAP with filter size T does reduce to the TMAP with filter size T=1. This would be significant both conceptually and computationally if we could determine conditions under which the decision regions are equalized. Note in the above for computational reasons, we had constrained the observation matrix to be symmetric. This produces only a sufficient condition as stated in Proposition 5.3.2. The right measure of the minimum quantity of information needed must be free of such constraints (for example, Shannon's mutual information among the states and the observations). Observe that in the above example, the amount of needed information grows with increasing γ. Clearly $\frac{1}{2(1+e^{-\gamma})} \to 1/2$ as $\gamma \to \infty$ which implies under infinite risk, we need perfect observation to equalize the decision regions.

We proved earlier that the maximum expected error "cost to go" for a completely symmetric HMM occurs at the center of the simplex. This is not true in general. It can be shown that the maximum cost in the first case of the above example (I=0) occurs along the transition line L, with slope $m(\delta)$, which does not cross the center of the simplex $\sigma = (1/2, 1/2)$ unless $\delta = 1/2$, a special case of complete symmetry with $q = \epsilon = 1/2$. When $\delta = 0$ as $\gamma \to \infty$, this maximum

point is moved arbitrarily closer to $\sigma = (0,1)$. This is another example of what we termed a singularity under infinite risk.

We saw that under the condition of uniform flow TMAP reduces to MAP; i.e., the estimated process based on the discrete metric with the assumption of uniform flow is invariant under risk-sensitivity. We may be led to believe that perhaps for large values of γ, risk-sensitivity tends to move the estimator toward this stable invariance, making the estimator look more and more "uniform". One has to be careful about what this means. In fact, risk-sensitivity tends to increase oscillatory behavior and not relieve it. A more conservative estimator tends to move around the state space more rapidly from sample path to sample path and not for too long "trust" the correctness of the sample path it may be following. It is in this sense that the estimates are made more "uniform".

Finally, we point out that many of the results of this chapter depend on the properties of the discrete metric (used to define MAP) which is not the natural metric for R^n. Therefore, our structural results do not directly illuminate the linear-Gaussian risk-sensitive estimation case. However, the intuition gained in the discrete finite dimensional setting about the behavior of the sample paths may lead to a better understanding of the linear-Gaussian risk-sensitive estimator as well.

Chapter 6

Conclusions and future research problems

We set out to study the risk-sensitive estimation of HMM's from a dynamical systems point of view. We did so by first generalizing the risk-sensitive estimators to the broader class of product estimators in which risk-sensitivity parameter was replaced by state dependent scaling functions. From this point of view, risk-sensitive estimators are a special case of dynamic *sample path error smoothing* estimators. Our theorems and examples demonstrated that this error smoothing action manifests itself in relations among transition probabilities, risk-sensitivity and the behavior of risk-sensitive sample paths. In chapter 3 and 4, we looked at these relations from the perspective of the "forward dynamics"; in chapter 5, we studied them by looking at the structure of the decision regions that in some average sense capture the dynamics in the "backward" direction.

In chapter 5, we unified the ideas of Speyer and Dey-Moore risk-sensitive filtering in the context of risk-sensitive filter banks and showed how and why the two may coincide. The *sample path error smoothing* perspective enabled us to

understand and explain the meaning of the information state in both cases and how the "burden" of error smoothing depends on the size of the filter banks. It is likely that interesting problems remain.

The first set of problems are the "classification" ones. For example, we conjecture that if a Markov-chain is ergodic with non-zero self-transitions, the decision regions are of the threshold type. Furthermore, the characterization of the decision regions in terms of the coupling of information, risk-sensitivity and the transition probabilities is incomplete and Conjecture 4.5.1 requires a proof. Going further and deeper in these directions, in our opinion, will prove fruitful to future researchers.

A less tangible idea was demonstrated in example 3.4.4 (and there are other similar examples we have not included for the sake of brevity): information and "how much" of it we have depends on how "risk-conscious" and "conservative" we are. We have not been able to formulate this problem. Roughly speaking, a generalized notion of information is needed here which in the limit (for small risk) must correspond to the notion of information in the sense of Shanon's. Perhaps solving the classification problems can show us the way.

Appendix A

Kolmogrov's Extension Theorem and the change of measure

Kolmogrov's Extension Theorem together with the Radon-Nikodym Derivative were used to define the Girsanov's measure in chapter 2.

Theorem A.1: (Kolmogrov's Extension Theorem) For all $\tau_1, ..., \tau_k,\ k \in \mathbf{N}$ and τ the time index, let $P_{\tau_1,...,\tau_k}$ be probability measures on $\mathbf{R}^{nk}$ such that

$$P_{\tau_\sigma(1),...,\tau_\sigma(k)}(F_1 \times ... \times F_k) = P_{\tau_1,...,\tau_k}(F_{\sigma^{-1}(1)} \times ... \times F_{\sigma^{-1}(k)})$$

for all permutations σ on $\{1, 2, ..., k\}$ and

$$P_{\tau_1,...,\tau_k}(F_1 \times ... \times F_k) = P_{\tau_1,...,\tau_k,...,\tau_{k+m}}(F_1 \times ... \times F_k \times \mathbf{R}^n \times ... \times \mathbf{R}^n)$$

for all $m \in \mathbf{N}$, and the set on the right-hand side has a total of $k + m$ factors. Then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a stochastic process $\{X_\tau\}$ on Ω into $\mathbf{R}^n$ such that

$$P_{\tau_1,\dots,\tau_k}(F_1 \times \dots \times F_k) = P[X_{\tau_1} \in F_1, \dots, X_{\tau_k} \in F_k],$$

for all τ_i in the time set $k \in \mathbf{N}$ and all Borel sets F_i.

Theorem A.2: (Radon-Nikodym) If P and $P^\dagger$ are two probability measures on a measurable space $(\Omega, \mathcal{M})$ such that for each $B \in \mathcal{M}, P(B) = 0$ implies $P^\dagger(B) = 0$, then there exists a nonnegative random variable Λ, such that $P^\dagger(C) = \int_C \Lambda dP$ for all $C \in \mathcal{M}$. We write

$$\frac{dp^\dagger}{dP}|\mathcal{M} = \Lambda.$$

Λ is the Radon-Nikodym derivative of $P^\dagger$ with respect to P.

Furthermore, if for each $B \in \mathcal{M}, P^\dagger(B) = 0$ implies $P(B) = 0$ then

$$\frac{dP}{dp^\dagger}|\mathcal{M} = (\Lambda)^{-1}.$$

In Chapter 2, we let $\frac{dP^\dagger}{dP}| \mathcal{G}_{t+1} = \Lambda_t$. A little thought should convince the reader that the conditions of the Kolmogrov's Extension Theorem applied to this construction require that

$$\int_B \Lambda_t dP = \int_B \Lambda_{t-1} dP \quad B \in \mathcal{G}_t.$$

Also since we had assumed $c_{ij} > 0$ for every i and j in the observation matrix C, the measures $P^\dagger$ and P are absolutely continuous with respect to each other; therefore,

$$\frac{dP}{dp^\dagger} = (\frac{dp^\dagger}{dP})^{-1}$$

by the above.

Appendix B

The integral characterization of the conditional expectation and the independence property.

The integral characterization of the conditional expectation [19] together with Fubini's theorem [24] can be used to show the following Theorem:

Theorem: If random variables Z and W are independent

$$E[f(X, g(Y))|Y] = E[f(X, g(Y))] = E[f(X, g(Y))|g(Y)].$$

Proof:

We will give a formal proof using Dirac's delta function. The proof can be made rigorous and is equivalent to the use of Fubini's theorem.

By the integral characterization of the conditional expectation

$$E[f(X, g(Y))|Y] = \int f(x, g(y))dP(x, y|Y),$$

where $P(x, y|Y)$ is the measure defined by the indicator functions.

$$P(x, y|Y) = E[I_{X<x,Y<y}|Y] = E[I_{X<x} \cdot I_{Y<y}|Y]$$
$$= I_{Y<y}E[I_{X<x}|Y]$$

Note that $I_{Y<y}$ as a function of y is a step function at Y, and $E[I_{X<x}|Y]$ is nothing but $P(x|Y)$. Therefore,

$$\int f(x, g(y))dP(x, y|Y) = \int \int f(x, g(y))\delta(Y - y)dP(x|Y)$$
$$= \int f(x, g(Y)))dP(x|Y).$$

But, if X and Y are independent

$$\int f(x, g(Y))dP(x|Y) = \int f(x, g(Y))dP(x).$$

This proves the first equality. The second equality follows from the fact that functions of independent random variables are themselves independent.

Appendix C

The Lattice Property

The lattice property is used to show that conditional expectation and the essential infimum commute. The result is due to Striebel. We will only quote the lemma and the results leading to it; the proofs can be found in [18].

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\phi_\gamma : \gamma \in \Gamma\}$ a subset of $L^1(\Omega, F, P)$ such that $\phi_\gamma(\omega) \geq 0$ a.s. for each $\gamma \in \Gamma$. Write

$$\phi = \bigwedge_\gamma \phi_\gamma$$

for the lattice infimum and θ_γ for the measure defined by

$$\frac{d\theta_\gamma}{dP} = \phi_\gamma.$$

Lemma: For $A \in \mathcal{F}$ define

$$\theta(A) = inf \sum_{i=1}^n \theta_{\gamma_i}(A_i),$$

where the infimum is taken over n, $\gamma_1, ..., \gamma_n \in \Gamma$, and finite disjoint partitions $A_1, ..., A_n$ of A. Then θ is a finite nonnegative measure on $\mathcal{F}$.

Lemma: Suppose, as above, that $\{\phi_\gamma : \gamma \in \Gamma\}$ is a subset of $L^1(\Omega, F, P)$ and that $\phi_\gamma(\omega) \geq 0$ a.s. for each $\gamma \in \Gamma$. Then

$$\frac{d\theta}{dP} = \phi = \bigwedge_\gamma \phi_\gamma.$$

Definition: The set $\{\phi_\gamma\} \subset L^1(\Omega, F, P)$ has the $\epsilon-$lattice property for $\epsilon > 0$ if, given $\gamma_1, \gamma_2 \in \Gamma$ there is a $\gamma_3 \in \Gamma$ such that

$$\phi_{\gamma_3} \leq \phi_{\gamma_1} \wedge \phi_{\gamma_2} + \epsilon \ a.s.$$

Lemma: Suppose the set $\{\phi_\gamma : \gamma \in \Gamma\} \subset L^1(\Omega, F, P)$ of nonnegative functions has the $\epsilon-$lattice property for every $\epsilon > 0$. If θ is a measured defined above then, for every $A \in \mathcal{F}$,

$$\theta(A) = inf_\gamma \ \theta_\gamma(A).$$

We can now state the lattice property:

Lemma: Suppose that $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$ and the set $\{\phi_\gamma : \gamma \in \Gamma\} \subset L^1(\Omega, F, P)$ of nonnegative functions has the $\epsilon-$lattice property for every $\epsilon > 0$. Then

$$E[\bigwedge_\gamma \phi_\gamma | \mathcal{G}] = \bigwedge_\gamma E[\phi_\gamma | \mathcal{G}] \ a.s.$$

BIBLIOGRAPHY

[1] R. K. Fish, M. Ostendorf, G. D. Bernard, David Castanon and H. Shivakumar. Multilevel classification of milling tool wear with confidence estimation (pre-print).

[2] Private communication with M. Ostendorf.

[3] I. B. Collings, M. R. James and J. B. Moore. An information-state approach to risk-sensitive tracking problems. *Journal of Mathematical Systems, Estimation, and Control,* Vol. 6, no. 1, pp. 1-13, 1996.

[4] S. I. Marcus. Nonlinear estimation *Systems and control encyclopedia: theory, technology, applications* Madan G. Singh (Editor). *Elsevier Science* pp. 3293-3303, September 1987.

[5] D. H. Jacobson Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic games. *IEEE Trans. Aut. Control,* vol. AC-18, no. 2, April 1973

[6] R. J. Elliott, L. Aggoun, and J. B. Moore. *Hidden Markov Models: Estimation and control,* Springer, Berlin, Heidelberg, New York, 1994.

[7] M. R. James, J. S. Baras, and R. J. Elliot. Risk sensitive Control and dynamic games for partially Observed Discrete-Time Nonlinear Systems. *IEEE Trans. Aut. Control,* vol. 39 (4), pp. 780-792, 1994.

[8] J. S. Baras and M. R. James. Robust and Risk-sensitive Output Feedback Control for Finite State Machines and Hidden Markov Models. *Institute for Systems Research, University of Maryland at college Park, Technical Research Reports* T. R. 94-63.

[9] J. L. Speyer, C. Fan and R. N. Banava. Optimal Stochastic Estimation with Exponential Cost Criteria. *Proceedings of 31th conference on decision and control,* pp. 2293-2298, 1992.

[10] S. Dey and J. Moore. Risk sensitive filtering and smoothing for hidden Markov models. *Systems and control letters,* vol. 25, pp. 361-366, 1995.

[11] S. Dey and J. Moore. Risk sensitive filtering and smoothing via reference probability methods. *IEEE Trans. Aut. Control,* vol. 42, no.11 pp. 1587-91, 1997.

[12] R. K. Boel, M. R. James, and I. R. Petersen. Robustness and risk-sensitive filtering. *Proceedings of the 36th IEEE Conference on Decision and Control,* Cat. No. 97CH36124, 1997.

[13] H. Vincent Poor. *An introduction to signal detection and estimation.* Springer Verlag, 1994.

[14] P. Whittle. *Risk-sensitive optimal control.* John Woley & sons, 1990.

[15] P. Whittle. Risk-sensitive linear/quadratic/Gaussian control. *Adv. Applied probability.* 13, 746-777, 1981, applied probability trust.

[16] E. Fernandez-Gaucherand and S. I. Marcus. Risk-sensitive optimal control of Hidden Markov Models: a case study. IEEE Transactions on Automatic Control, vol. 42, no. 10, pp. 1418-1422, 1997.

[17] L. Aggoun, A. Bensoussan, R. J. Elliot and J. B. Moore. Finite-dimensional quasi-linear risk-sensitive control *Systems and control letters* vol. 25, no. 2, pp. 151-157, May 1995

[18] R. J. Elliot. *Stochastic Calculus and Applications.* New York: Springer-Verlag, 1982.

[19] A. N. Shiryayev. *Probability.* New York: Springer-Verlag, 1984.

[20] T. Basar and P. bernhard. H^∞-Optimal control and related minimax design problems: a dynamic game approach. Birkhauser, Boston.

[21] R. K. Boel, M. R. James and I.R. Petersen. Robustness and risk-sensitive filtering. *Proceedings of the 36th IEEE Conference on Decision and Control.* Cat. No.97CH36124, 1997.

[22] J. C. Doyle, Glover, P.P. Khargonekar and B.A. Francis. State-Space Solutions to standard H_2 and H_∞ control problems. *IEEE Trans. Aut. Control.* AC-34 (8) (1989) 831-847.

[23] J. W. Helton and M. R. James. A framework for generalizing H_∞ control to Non-linear systems.

[24] H. R. Royden. *Real Analysis.* Prentice Hall PTR 1988.

[25] E. Wong and B. Hajek. *Stochastic processes in engineering systems.* Springer-Verlag, 1985.

[26] L. R. Rabiner. A Tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE,* vol. 77, pp. 257-285, 1989.

[27] M. Brin and G. Stuck. Dynamical systems. *University of Maryland, lecture notes.*

[28] J. Losson and M. Mackey. Thermodynamic properties of coupled map lattices. *Stochastic and Spatial Structures of Dynamical Systems.* (eds.: S.J. van Strien, S.M. Verduyn Lunel)North Holland, Amsterdam (1996).

[29] J. Komornik. Asymptotic periodicity of the iterates of weakly constrictive Markov operators. *Tohoku Math. J., 38:15-17, 1986.*

[30] K. Chung. *A course in probability theory.* Academic press, 1974.