Risk-Sensitive Optimal Control of Hidden Markov Models: Structural Results

by E. Fernandez-Gaucherand, S.I. Marcus

T.R. 96-79

RISK-SENSITIVE OPTIMAL CONTROL OF HIDDEN MARKOV MODELS: STRUCTURAL RESULTS

Emmanuel Fernández-Gaucherand [†]
Systems & Industrial Engineering Department
The University of Arizona
Tucson, AZ 85721.

Steven I. Marcus [‡]
Institute for Systems Research and
Electrical Engineering Department
The University of Maryland
College Park, MD 20742.

[†] Research supported in part by a grant from the University of Arizona Foundation and the Office of the Vice President for Research; in part by the National Science Foundation under grant NSF-INT 9201430; and in part by The Engineering Foundation under grant RI-A-93-10. Email: emmanuel@sie.arizona.edu.

[‡] Research partially supported by the National Science Foundation under Grant EEC 9402384. Email: marcus@src.umd.edu.

1. INTRODUCTION

We consider a risk-sensitive optimal control problem for hidden Markov models (HMM), i.e., controlled Markov chains where state information is only available to the controller via an output (message) process. The optimal control of HMM under standard, risk-neutral performance criteria, e.g., discounted and average costs, has received much attention in the past. Many basic results and numerous applications have been reported in the literature in this subject; see [ABFGM], [BE2], [KV], and references therein. Controlled Markov chains with full state information and a risk-sensitive performance criterion have also received some attention, dating back at least to the work of Howard and Matheson [HOM]; see also [BSO], [CSO].

On the other hand, quite the opposite is the situation for HMM under risk-sensitive criteria, e.g., expected value of the exponential of additive costs. Whittle and others (see [WHI] and references therein) have extensively studied the risk-sensitive optimal control of partially-observable linear exponential quadratic Gaussian (LEQG) systems; see also [BVS]. More recently, James, Baras and Elliott [JBE], [BJ], have treated the risk-sensitive partially-observable optimal control problem of discrete-time non-linear systems.

The paucity of results in this subject area can be mostly attributed to the lack in the past of appropriate *sufficient statistics*, or *information states.* As is well known, if the cost criterion being considered is of the type "expected value of additive costs," then the posterior probability density, given all available information up to the present, constitutes a sufficient statistic for control (or information state); see [ABFGM], [BE2], [KV]. The latter result was originally proved by Shiryaev in the early sixties [SHI1]-[SHI2], who also proved that this was not the case for non-additive cost criteria [SHI3]. In particular, the posterior probability density is not a *sufficient* statistic for HMM under an "exponential of sum of costs" type of criterion, which is non-additive. This fact was overlooked in [GHE], thus invalidating the claims of optimality for the policies obtained in that paper. ¶

¶ Certainly, one can pose a well defined stochastic optimal control problem given *any* statistic. However, if the chosen statistic is not sufficient, then one cannot hope to obtain the "overall" optimal policy, except by serendipity.

Recently, James, Baras, and Elliott [BJ], [JBE] have derived information states for HMM under an "exponential of additive costs" criterion, and have also given dynamic programming equations from which optimal values and controls can be computed, for problems with a finite horizon. Building upon the results by Baras, James and Elliott, we report in this paper results of an investigation on the nature and structure of risk-sensitive controllers. We pose the following question:

How does risk-sensitivity manifest itself in the structure of a controller?

Whittle [WHI] has addressed a similar question for the LEQG problem, and he has shown that much insight can be gained from a comparison of the risk-neutral (i.e., the classical LQG) and risk-sensitive equations describing the optimal controller. In our context, one difficulty encountered is that optimal controllers are defined in terms of different information states for the risk-neutral and risk sensitive cases; see also [BJ], [JBE].

The paper is organized as follows. Section 2 summarizes some basic results about utility and risk theory. In section 3 we present our model, and recall the main results on information states from [BJ]-[JBE] that will be needed for our developments. Section 4 contains several general results, and in section 5 we present a particular case study of a popular benchmark problem. We obtain structural results for the optimal risk-sensitive controller, and compare it to that of the risk-neutral case. Furthermore, we show that indeed the risk-sensitive controller and its corresponding information state converge to the known solutions for the risk-neutral situation, as the risk factor goes to zero. We also study the infinite and general risk aversion cases.

2. RATIONAL PREFERENCES, UTILITY AND RISK

Consider the situation where a decision maker (DM) is faced with several choices, the outcomes of which are uncertain. That is, the DM can choose among several *gambles* or *lotteries* $\pi \in \Pi$, and the consequent reward $\mathcal{R}$ for the lottery is a random variable, with a known probability distribution $\mathcal{P}^\pi$. With no loss of generality for our purposes, we may associate $\mathcal{P}^\pi$ with the lottery itself. Utility theory, as developed by von Neuman and Morgenstern (vNM), see [BE1], [LR], tries to quantitatively describe the *preferences* of the DM, under the assumption that these are *rational*, in the following sense. Let $\mathcal{P}^{\pi_1}$ and $\mathcal{P}^{\pi_2}$ be two of the choices available to the DM, and suppose that he always prefers $\mathcal{P}^{\pi_1}$ over $\mathcal{P}^{\pi_2}$. Then, (vNM) assume that there exists a complete and transitive relation "$\preceq$" on the set of lotteries such that

$$\mathcal{P}^{\pi_1} \preceq \mathcal{P}^{\pi_2}. \tag{2.1}$$

Under some "continuity" condition for the relation "$\preceq$" (see [BE1], [LR, p. 27]) it can be shown that:

(*i*) there exists a real-valued *utility* function $\mathcal{U} : \mathbb{R} \to \mathbb{R}$ such that

$$\mathcal{P}^{\pi_1} \preceq \mathcal{P}^{\pi_2} \Longleftrightarrow \mathbb{E}^{\pi_1}\{\mathcal{U}(\mathcal{R})\} \leq \mathbb{E}^{\pi_2}\{\mathcal{U}(\mathcal{R})\} \qquad (2.2);$$

(*ii*) $\mathcal{U}(\cdot)$ is unique, up to an affine transformation of the form $\alpha\mathcal{U}(\cdot)+\beta$; with $\alpha > 0$ and $\beta \in \mathbb{R}$.

Let $\overline{\mathcal{R}}_\pi := \mathbb{E}^\pi\{\mathcal{R}\}$, which is sometimes called the "actuarial value" of the lottery π [PT]. Note that if the DM's utility function is affine, then his preferences only depend on the expected values of $\mathcal{R}$. Thus, he would then be indifferent as to whether to keep the lottery $\mathcal{P}^\pi$ (i.e., proceed with the gamble) or avoid it and instead be compensated with the quantity $\overline{\mathcal{R}}_\pi$, i.e., $\mathbb{E}^\pi\{\mathcal{U}(\mathcal{R})\} = \mathcal{U}(\overline{\mathcal{R}}_\pi)$. In this situation the DM is said to be **risk neutral**. On the other hand, the DM may prefer certainty over the uncertain lottery, i.e., he may be **risk averse or pessimistic**; in this case we have that:

$$\mathbb{E}^\pi\{\mathcal{U}(\mathcal{R})\} \leq \mathcal{U}(\overline{\mathcal{R}}_\pi). \tag{2.3}$$

Notice that (2.3) implies in general that $\mathcal{U}(\cdot)$ is a concave function. Conversely, if

$$\mathbb{E}^\pi\{\mathcal{U}(\mathcal{R})\} \geq \mathcal{U}(\overline{\mathcal{R}}_\pi), \tag{2.4}$$

which implies in general that $\mathcal{U}(\cdot)$ is convex, the DM is said to be **risk seeking or optimistic**.

Note that the characterization given by (2.3)-(2.4) is not enough to measure the DM's sensitivity to risk. Pratt [PT] introduced a measure of risk that has been widely accepted, see also [BE1]. First, define the *risk premium* p_π as the quantity the DM is willing to pay in order to avoid the lottery, and instead receive its actuarial value, that is:

$$\mathcal{U}(\overline{\mathcal{R}}_\pi - p_\pi) = \mathbb{E}^\pi\{\mathcal{U}(\mathcal{R})\}. \tag{2.5}$$

Now, following [PT] (see also [BE1]), proceeding formally we have that

$$\mathcal{U}(\overline{\mathcal{R}}_\pi - p_\pi) = \mathcal{U}(\overline{\mathcal{R}}_\pi) - p_\pi \mathcal{U}'(\overline{\mathcal{R}}_\pi) + o[p_\pi], \tag{2.6}$$

where $o[\alpha]/\alpha \to 0$, as $\alpha \to 0$. Also

$$\begin{aligned}\mathbb{E}^\pi\{\mathcal{U}(\mathcal{R})\} &= \mathbb{E}^\pi\{\mathcal{U}(\overline{\mathcal{R}}_\pi) + (\mathcal{R} - \overline{\mathcal{R}}_\pi)\mathcal{U}'(\overline{\mathcal{R}}_\pi) + \frac{1}{2}(\mathcal{R} - \overline{\mathcal{R}}_\pi)^2\mathcal{U}''(\overline{\mathcal{R}}_\pi) + o[(\mathcal{R} - \overline{\mathcal{R}}_\pi)^2]\} \\ &= \mathcal{U}(\overline{\mathcal{R}}_\pi) + \frac{1}{2}\sigma_\pi^2\mathcal{U}''(\overline{\mathcal{R}}_\pi) + \mathbb{E}^\pi\{o[(\mathcal{R} - \overline{\mathcal{R}}_\pi)^2]\},\end{aligned} \tag{2.7}$$

where σ_π^2 denotes the variance of $\mathcal{R}$ with respect to $\mathcal{P}^\pi$. Thus, we obtain from (2.5)-(2.7) that

$$p_\pi \mathcal{U}'(\overline{\mathcal{R}}_\pi) = -\frac{1}{2}\sigma_\pi^2\mathcal{U}''(\overline{\mathcal{R}}_\pi) + o[p_\pi] + \mathbb{E}^\pi\{o[(\mathcal{R} - \overline{\mathcal{R}}_\pi)^2]\}. \tag{2.8}$$

Therefore, we see from (2.8) that the risk premium is proportional, up to first order (i.e. locally), to the variance of the reward, and the proportionality factor is one half the **risk aversion coefficient**:

$$r(z) := -\frac{\mathcal{U}''(z)}{\mathcal{U}'(z)} = -\frac{d}{dz}log(\mathcal{U}'(z)). \tag{2.9}$$

Notice that if $r(\cdot) > 0$ then the DM is risk averse, and risk seeking in the converse. Furthermore, if $r(z) = 0$, then the DM is **risk neutral**, i.e., his utility function is affine, and thus his preferences depend only on the actuarial (expected) values of the lotteries. The latter is the most common case (implicitly) studied in the literature, where expected values of rewards are maximized; see [ABFGM], [BE2], [KV].

2.1 CONSTANT RISK AVERSION

In many situations it is to be expected that the DM has either a decreasing or increasing risk aversion coefficient $r(\cdot)$, as a function of the DM's wealth. For example, it is to be expected that a portfolio manager may be more reticent to risk half his assets if the value of the portfolio is a billion dollars, than if it is $\$10,000$; see [BE2] for a very nice presentation of problems of this nature. On the other hand, if $r(\cdot) = constant$, then the DM's sensitivity to risk does not depend on the level of his current wealth. To gain more insight into this situation, consider the *certainty equivalent* for a lottery $\mathcal{P}^{\pi}$, denoted by $c_{\pi} \in \mathbb{R}$ and defined by:

$$\mathbb{E}^{\pi}\{\mathcal{U}(\mathcal{R})\} = \mathcal{U}(c_{\pi}). \tag{2.10}$$

Then, if the DM has constant risk aversion, his utility function must satisfy the so called "Δ property": if $\mathcal{R}$ is increased by a quantity $\Delta \in \mathbb{R}$, then one has that the certainty equivalent is increased by the same amount [HOM]. From (2.9), it is seen that if $r(\cdot) = \gamma \in \mathbb{R}$, then one has that the utility function for the DM must take one of the following forms (up to an affine transformation):

$$r(\cdot) = \gamma \Longrightarrow \begin{cases} \mathcal{U}(z) = z, & \gamma = 0; \\ \mathcal{U}(z) = -exp(-\gamma z), & \gamma > 0; \\ \mathcal{U}(z) = exp(-\gamma z), & \gamma < 0. \end{cases} \tag{2.11}$$

In the sequel, we will be concerned with negative net benefits, i.e., costs, instead of rewards, and thus we will seek to minimize *disutilities* instead of maximizing utilities. Hence, correspondingly the following disutility function $\mathcal{L}(\cdot)$ will be considered, for $\gamma \in \mathbb{R}$, $\gamma \neq 0$:

$$\mathcal{L}(c) := sgn(\gamma)exp(\gamma \cdot c), \tag{2.12}$$

where $sgn(\gamma)$ denotes the sign of γ.

3. THE CONTROLLED HIDDEN MARKOV MODEL

A *controlled hidden Markov model* is given by a five-tuple $\langle \mathbf{X}, \mathbf{Y}, \mathbf{U}, \{P(u) : u \in \mathbf{U}\}, \{Q(u) : u \in \mathbf{U}\}\rangle$; here $\mathbf{X} = \{1, 2, \ldots, N_{\mathbf{X}}\}$ is the finite set of (internal) states, $\mathbf{Y} = \{1, 2, \ldots, N_{\mathbf{Y}}\}$ is the set of observations (or messages), $\mathbf{U} = \{1, 2, \ldots, N_{\mathbf{U}}\}$ is the set of decisions (or controls). In addition, we have that $P(u) := \big[p_{i,j}(u)\big]$ is the $N_{\mathbf{X}} \times N_{\mathbf{X}}$ state transition matrix, and $Q(u) := \big[q_{x,y}(u)\big]$ is the $N_{\mathbf{X}} \times N_{\mathbf{Y}}$ state/message matrix, i.e., $q_{x,y}(u)$ is the probability of receiving message y when the state is x and action u has been selected. In the operations research literature similar models are called *partially observable Markov decision processes* [FAM1], [FAM2], and in the computer science literature *finite state stochastic automata* [DOB], [PAZ]. Two types of information patterns are of interest.

Information Pattern 1 (IP1):

At decision epoch t, the system is in the (unobservable) state $X_t = i$, a decision $U_t = u$ is taken, and the state evolves to $X_{t+1} = j$ with probability $p_{i,j}(u)$. Once the state has evolved to X_{t+1}, an observation Y_{t+1} is gathered, such that:

$$Prob\{Y_{t+1} = y \mid X_{t+1} = i, U_t = u\} = q_{x,y}(u). \tag{3.1.a}$$

Hence, based on $\mathcal{I}_t^{(1)} := (Y_0, U_0, Y_1, \ldots, U_t, Y_{t+1})$, a new decision U_{t+1} is selected.

Information Pattern 2 (IP2):

At decision epoch t, the system is in the (unobservable) state $X_t = i$, a decision $U_t = u$ is taken, and an observation Y_{t+1} is gathered, such that:

$$Prob\{Y_{t+1} = y \mid X_t = i, U_t = u\} = q_{i,y}(u). \tag{3.1.b}$$

The state then evolves to $X_{t+1} = j$ with probability $p_{i,j}(u)$. Hence, based on $\mathcal{I}_t^{(2)} := (U_0, Y_1, U_1, Y_2, \ldots, U_t, Y_{t+1})$, a new decision U_{t+1} is selected.

Hereafter we will simply write $\mathcal{I}_t$ and $\mathcal{Y}_t$ for a generic information pattern and the filtration generated by the available observations, respectively, up to decision epoch t.

Given an expected cost per stage $(i,u) \mapsto c(i,u)$, the sum of costs for the finite horizon M is given by

$$\mathcal{C}_M := \sum_{t=0}^{M-1} c(X_t, U_t). \tag{3.2}$$

The *risk-sensitive optimal control* problem is that of finding a control policy $\pi = \{\pi_0, \pi_1, \ldots, \pi_{M-1}\}$, with $\mathcal{I}_t \mapsto \pi_t(\mathcal{I}_t) \in \mathbf{U}$, such that the following criterion is minimized:

$$J^\gamma(\pi) := sgn(\gamma)\mathbb{E}^\pi\big[exp\big(\gamma \cdot \mathcal{C}_M\big)\big], \tag{3.3}$$

where $\gamma \neq 0$ is the *risk-factor*, and $sgn(\gamma)$ is the sign of γ; here $\mathbb{E}^\pi$ denotes the expectation induced by policy π and, implicitly, the initial distribution of the state. By computing the Taylor series expansion of $J^\gamma(\pi)$, when γ is sufficiently small, the risk sensitivity of the above criterion becomes evident in that, in addition to the standard expected sum of costs, a second order term in the expansion measures the variance of $\mathcal{C}_M$; see [WHI] for details. If $\gamma > 0$, then the controller is *risk-averse* or *pessimistic*, whereas if $\gamma < 0$ then the controller is *risk-prefering* or *optimistic*.

3.1 INFORMATION STATES

As for the risk-neutral case [ABFGM], [BE], [KV], an equivalent stochastic optimal control problem can be formulated in terms of *information states* and *separated policies*. Here we follow the work of Baras, Elliott, and James [BJ]-[JBE], who derived information states both for problems with continuous [JBE] and discrete [BJ] state variables. First, we equivalently reformulate the stochastic control problem in terms of a canonical measure, as follows. 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 as given by (IP). Then the probability measure induced by a policy π is equivalent to a canonical distribution $\mathcal{P}^\dagger$, under which $\{Y_t\}$ is independently and identically distributed (i.i.d), uniformly distributed, independent of $\{X_t\}$, and $\{X_t\}$ is a controlled Markov chain with transition matrix as above. We have that

$$\frac{d\mathcal{P}^\pi}{d\mathcal{P}^\dagger}|_{\mathcal{G}_t} = \lambda_t^\pi, \tag{3.4}$$

where

$$\lambda_t^\pi = \begin{cases} N_{\mathbf{Y}}^t \cdot \Pi_{k=1}^t q_{X_k,Y_k}(U_{k-1}), & \mathcal{G}_t \text{ generated by (IP1)}; \\ N_{\mathbf{Y}}^t \cdot \Pi_{k=1}^t q_{X_{k-1},Y_k}(U_{k-1}), & \mathcal{G}_t \text{ generated by (IP2)}. \end{cases} \tag{3.5}$$

Then, the cost incurred by using the policy π is given by

$$J^\gamma(\pi) := sgn(\gamma)\mathbb{E}^\pi\big[exp\big(\gamma \cdot \mathcal{C}_M\big)\big] = sgn(\gamma)\mathbb{E}^\dagger\big[\lambda_M^\pi \cdot exp\big(\gamma \cdot \mathcal{C}_M\big)\big]. \tag{3.6}$$

Following [BJ], [EM] and [JBE], the information state for our problem is given by

$$\sigma_t^\gamma(i) := \mathbb{E}^\dagger\big[\mathbf{1}[X_t = i]exp\big(\gamma \cdot \mathcal{C}_t\big) \cdot \lambda_t^\pi \mid \mathcal{Y}_t\big], \tag{3.7}$$

where $\mathbf{1}[A]$ is the indicator function of the event A, and $\sigma_0^\gamma(i) = p_0$, where p_0 is the initial distribution of the state and is assumed to be known. Notice that $\sigma_t^\gamma \in \mathbb{R}_+^{N_{\mathbf{X}}} := \big\{\sigma \in \mathbb{R}^{N_{\mathbf{X}}} \mid \sigma(i) \geq 0, \forall i\big\}$. With this definition of information state, similar results as in the risk-neutral case can be obtained. In particular, one obtains a recursive updating formula for $\{\sigma_t^\gamma\}$, which is driven by the output (observation) path and evolves forward in time. Moreover, the value functions can be expressed in terms of the information state only, and dynamic programming equations give necessary and sufficient optimality conditions for *separated policies*, i.e., maps $\sigma_t^\gamma \mapsto \tilde{\pi}_t(\sigma_t^\gamma) \in \mathbf{U}$; see [BJ], [JBE]. In particular we have that:

$$J^\gamma(\pi) = sgn(\gamma)\mathbb{E}^\dagger\big[\sum_{i=1}^{N_X} \sigma_M^\gamma(i)\big], \tag{3.8}$$

where $\{\sigma_M^\gamma\}$ is obtained from (3.5) under the action of policy π. Hence, the original partially observed problem is equivalently expressed as one with complete state information, i.e., $\{\sigma_t^\gamma\}$. For ease of presentation, we consider hereafter the risk-averse case only ($\gamma > 0$); the risk-seeking case is treated similarly.

4. GENERAL RESULTS

As in the completely observed case [HOM], define the *disutility contribution matrix* as [§]:

$$[\mathcal{D}(u)]_{i,j} := p_{i,j}(u) \cdot exp(\gamma c(i,u)). \tag{4.1}$$

[§] Notice that we are using expected one-stage cost functions. If on the other hand a model using one-stage cost functions that depend explicitly on the current and next state, e.g, $c(i,j,u)$ is used, then (4.1) is modified accordingly.

The following lemma gives the recursions that govern the evolution of the information state.

Lemma 4.1: The information state process $\{\sigma_t^\gamma\}$ is recursively computable as:

$$\sigma_{t+1}^\gamma = \begin{cases} N_{\mathbf{Y}} \cdot \overline{Q}(Y_{t+1}, U_t)\mathcal{D}^T(U_t) \cdot \sigma_t^\gamma, & \mathcal{Y}_t \text{ generated by (IP1);} \\ N_{\mathbf{Y}} \cdot \mathcal{D}^T(U_t)\overline{Q}(Y_{t+1}, U_t) \cdot \sigma_t^\gamma, & \mathcal{Y}_t \text{ generated by (IP2);} \end{cases} \tag{4.2}$$

where $\overline{Q}(y,u) := diag(q_{i,y})(u)$, and A^T denotes the transpose of the matrix A.

Proof: Following [JBE], [BJ], we have that the operator governing the evolution of the information state is the matrix with (i,j) element given by $N_{\mathbf{Y}} \cdot p_{i,j}(u) \cdot exp(\gamma c(i,u))q_{i,y}(u)$, where u denotes the decision taken at decision epoch t, and y denotes the value of the most recent observation gathered by decision epoch $t+1$. Then (4.2) follows straightforwardly. □

Remark 4.1: The $N_{\mathbf{Y}}$ factor appears in (4.2) due to the use of the canonical measure $\mathcal{P}^\dagger$; see (3.5).

Remark 4.2: For risk-sensitive completely observed controlled Markov chains with finite state and control sets, it is well known that the disutility matrix $\mathcal{D}(u)$ governs the evolution of the disutility [HOM], [JAQ]. On the other hand, for risk-neutral HMM models, the information state used is the conditional probability distribution of the (unobservable) state, given the available observations [ABFGM], [BE2], [KV]. The unnormalized form of this conditional probability distribution is given by similar recursions as in (4.2), with $\mathcal{D}(u)$ replaced by $P(u)$. Moreover, observe that as $\gamma \to 0$, $\mathcal{D}(u) \to P(u)$ (elementwise). Therefore, we see that (4.2) is the "natural" extrapolation of the standard risk-neutral information state.

As in [BJ], [JBE], define value functions $J^\gamma(\cdot, M-k) : \mathbb{R}_+^{N_{\mathbf{X}}} \to \mathbb{R}$, $k = 1, \ldots, M$, as follows:

$$J^\gamma(\sigma, M-k) := \min_{\pi_{M-k} \cdots \pi_{M-1}} \Big\{ \mathbb{E}^\dagger \{ \sum_{i=1}^{N_X} \sigma_M^\gamma(i) \mid \sigma_{M-k}^\gamma = \sigma \} \Big\}. \tag{4.3}$$

For ease of presentation, we will hereafter consider exclusively (IP1). Simple modifications to the results in the sequel give the corresponding results for (IP2). Furthermore, we will denote by $T(u,y)$ the matrix

$$T(u,y) := N_{\mathbf{Y}} \cdot \overline{Q}(y,u)\mathcal{D}^T(u). \tag{4.4}$$

The next result follows directly from [BJ], [JBE].

Lemma 4.2: The dynamic programming equations for the value functions in this problem are given as:

$$\begin{cases} J^\gamma(\sigma, M) & = \sum_{i=1}^{N_{\mathbf{X}}} \sigma(i); \\ J^\gamma(\sigma, M-k) & = min_{u\in\mathbf{U}}\big\{\mathbb{E}^\dagger\big[J^\gamma(T(u, Y_{M-k+1})\cdot\sigma, M-k+1)\big]\big\} \quad k=1,2,\ldots,M. \end{cases} \tag{4.5}$$

Furthermore, a separated policy $\pi^* = \{\pi_0^*, \ldots, \pi_{M-1}^*\}$ that attains the minimum in (4.5) is risk-sensitive optimal.

Recall that $\mathbb{E}^\dagger[\cdot]$ is the expectation with respect to the canonical measure $\mathcal{P}^\dagger$, and thus for a given function $f : \mathbf{Y} \to \mathbb{R}$,

$$\mathbb{E}^\dagger\big[f(Y_t)\big] = \frac{1}{N_{\mathbf{Y}}}\sum_{y=1}^{N_{\mathbf{Y}}}\big[f(y)\big]. \tag{4.6}$$

Next, we present several general results for the risk-sensitive case that have similar counterparts in the standard risk-neutral case [ABFGM], [BE2], [FAM1], [KV], [SSO].

Lemma 4.3: The value functions given by (4.5) are concave functions of $\sigma \in \mathbb{R}_+^{N_{\mathbf{X}}}$.

Proof:

We proceed by induction in k, with the case $k = 0$ being trivially verified from (4.5). Assume that the claim holds true for $0 \le \overline{k} = k-1 < M$. Let $0 \le \lambda \le 1$ and $\sigma_1, \sigma_2 \in \mathbb{R}_+^{N_{\mathbf{X}}}$, and define $\tilde{\sigma} := \lambda\sigma_1 + (1-\lambda)\sigma_2$. Then we have that:

$$\begin{aligned} J^\gamma(\tilde{\sigma}, M-k) &= \min_{u\in\mathbf{U}}\Big\{\frac{1}{N_{\mathbf{Y}}}\sum_{y=1}^{N_{\mathbf{Y}}} J^\gamma(T(u,y)\cdot\tilde{\sigma}, M-k+1)\Big\} \\ &\ge \min_{u\in\mathbf{U}}\Big\{\frac{1}{N_{\mathbf{Y}}}\sum_{y=1}^{N_{\mathbf{Y}}}\big[\lambda J^\gamma(T(u,y)\cdot\sigma_1, M-k+1) \\ &\qquad + (1-\lambda)J^\gamma(T(u,y)\cdot\sigma_2, M-k+1)\big]\Big\} \\ &\ge \lambda J^\gamma(\sigma_1, M-k) + (1-\lambda)J^\gamma(\sigma_2, M-k), \end{aligned} \tag{4.7}$$

where the first inequality follows due to the induction hypothesis, and the second inequality due to (4.5). □

Remark 4.3: For the risk-neutral case, a similar result was initially pointed out by Shiryaev [SHI1]-[SHI2], and shown in detail by Åström [AST], for normalized information states. This result has been fundamental in showing optimality of structured policies in the risk-neutral case; see [FAM1], [LOV], [WCC], and references therein.

Next, define recursively sets of vectors in $\mathbb{R}_+^{N_{\mathbf{X}}}$ as follows:

$$\begin{aligned} A_0 &:= \{\mathbf{1} = (1, 1, \dots, 1)\}, \\ A_k &:= \Big\{ \frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}} \alpha_y \cdot T(u, y) \mid \alpha_y \in A_{k-1}, u \in \mathbf{U} \Big\}. \end{aligned} \tag{4.8}$$

Note that the cardinality of the sets defined in (4.8) obeys the recursion $|A_k| \leq |A_{k-1}|^{N_{\mathbf{Y}}} \cdot N_{\mathbf{U}}$. In the risk-neutral case, the counterpart of the following result has been shown to have important computational implications [ABFGM], [FAM1], [SSO]. It will play a key role in our subsequent developments.

Lemma 4.4: The value functions given by (4.5) are piecewise linear functions in $\sigma \in \mathbb{R}_+^{N_{\mathbf{X}}}$, such that:

$$J^\gamma(\sigma, M-k) = \underset{\alpha \in A_k}{min} \big\{ \alpha \cdot \sigma \big\}. \tag{4.9}$$

Proof:

We proceed by induction in k, with the case $k = 0$ being trivially verified from (4.5). Assume that the claim holds true for $0 \leq overlinek = k - 1 < M$, then from (4.5) above we have:

$$\begin{aligned} J^\gamma(\sigma, M-k) &= \underset{u \in \mathbf{U}}{min} \Big\{ \frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}} \underset{\alpha \in A_{k-1}}{min} \big\{ \alpha \cdot T(u, y) \cdot \sigma \big\} \Big\} \\ &= \underset{u \in \mathbf{U}}{min} \Big\{ \Big[\frac{1}{N_{\mathbf{Y}}} \sum_{y=1}^{N_{\mathbf{Y}}} \tilde{\alpha}(u, y, \sigma) \cdot T(u, y) \Big] \cdot \sigma \Big\}, \\ &= \underset{\alpha \in A_k}{min} \Big\{ \alpha \cdot \sigma \Big\} \end{aligned} \tag{4.10}$$

where $\tilde{\alpha}(u,y,\sigma) \in A_{k-1}$ denotes a minimizer in the expression on the right of the first equality above. The last equality follows since $\alpha \cdot T(u,y) \cdot \sigma > \tilde{\alpha}(u,y,\sigma) \cdot T(u,y) \cdot \sigma$, for all $\alpha \in A_{k-1}$, $u \in \mathbf{U}$, $y \in \mathbf{Y}$, $\sigma \in \mathbb{R}^+_{N_\mathbf{X}}$. □

Lemma 4.5: Optimal separated policies $\{\pi^*_t\}$ are constant along rays through the origin, i.e., let $\sigma \in \mathbb{R}^{N_\mathbf{X}}_+$ then $\pi^*_t(\sigma') = \pi^*_t(\sigma)$, for all $\sigma' = \lambda\sigma$, $\lambda \geq 0$.

Proof:

¿From Lemma 4.4 we see that $J^\gamma(\sigma', M-k) = \lambda J^\gamma(\sigma, M-k)$. Hence, the result follows from Lemma 4.2. □

Definition 4.1: From (4.5) and (4.6), for $u \in \mathbf{U}$ and $k = 1, 2, \ldots, M$, let

$$\begin{aligned} J^\gamma_u(\sigma, M-k) &:= \mathbb{E}^\dagger\big[J^\gamma(T(u, Y_{M-k+1}) \cdot \sigma, M-k+1)\big] \\ &= \frac{1}{N_\mathbf{Y}} \sum_{y=1}^{N_\mathbf{Y}} \big[J^\gamma(T(u, Y_{M-k+1}) \cdot \sigma, M-k+1)\big]. \end{aligned} \tag{4.11}$$

The *control region* $CR^k_u \subseteq \mathbb{R}^{N_\mathbf{X}}_+$ for action $u \in \mathbf{U}$, at the $M-k$ decision epoch, is defined as:

$$CR^k_u := \big\{\sigma \mid \sigma \in \mathbb{R}^{N_\mathbf{X}}_+, J^\gamma(\sigma, M-k) = J^\gamma_u(\sigma, M-k)\big\}. \tag{4.12.a}$$

Furthermore by Lemma 4.2 if π^*_{M-k} is an optimal separated policy for stage $M-k$ then, for $u \in \mathbf{U}$,

$$CR^k_u := \{\sigma \in \mathbb{R}^{N_\mathbf{X}}_+ \mid \pi^*_{M-k}(\sigma) = u\}. \tag{4.12.b}$$

Definition 4.2: An action $\overline{u} \in \mathbf{U}$ is said to be a *resetting* action if there exists $j^* \in \mathbf{X}$ such that $p_{i,j^*}(\overline{u}) = 1$, for all $i \in \mathbf{X}$. Therefore from (4.1)-(4.2) and (4.4) we note that, for any $\sigma \in \mathbb{R}^{N_\mathbf{X}}_+$ and $y \in \mathbf{Y}$,

$$T(\overline{u}, y) \cdot \sigma = q_{j^*,y} \sum_{\ell=1}^{N_\mathbf{X}} \big[exp(c(\ell, \overline{u})) \cdot \sigma(\ell)\big] \cdot \nu_{j^*}, \tag{4.13}$$

where $\nu_1 = (1, 0, 0, \ldots, 0)^T$, $\nu_2 = (0, 1, 0, \ldots, 0)^T$, ..., $\nu_{N_\mathbf{X}} = (0, 0, 0, \ldots, 1)^T$. We further denote

$$\Lambda(\sigma, \overline{u}) := q_{j^*,y} \sum_{\ell=1}^{N_\mathbf{X}} \big[exp(c(\ell, \overline{u})) \cdot \sigma(\ell)\big]. \tag{4.14}$$

Theorem 4.1: Let $\overline{u} \in \mathbf{U}$ be a resetting action. Then $CR^k_{\overline{u}}$, is a convex subset of $\mathrm{I\!R}^{N_\mathbf{X}}_+$.

Proof:

Recall from Lemma 4.3 that the optimal cost-to-go functions $J^\gamma(\cdot, M-k)$ are concave. Since the maps $T(u,y)$ in (4.4) are linear then $J^\gamma_u(\cdot, M-k)$ are also concave, for all $u \in \mathbf{U}$. Furthermore, $\mathrm{I\!R}^{N_\mathbf{X}}_+$ is a convex domain. Then, by Lemma 1 in [LOV] we have that: if $J^\gamma_{\overline{u}}(\cdot, M-k)$ is a linear function in $\mathrm{I\!R}^{N_\mathbf{X}}_+$, then $CR^k_{\overline{u}}$ is convex. Thus all that remains to be proven is the linearity of $J^\gamma_{\overline{u}}(\cdot, M-k)$.

Let $\sigma \in \mathrm{I\!R}^{N_\mathbf{X}}_+$, we have by (4.2), (4.4), Definitions 4.1-4.2, and Lemma 4.4 that:

$$\begin{aligned} J^\gamma_{\overline{u}}(\sigma, M-k) &= \frac{1}{N_\mathbf{X}} \sum_{y=0}^{1} \underset{\alpha \in A_{k-1}}{min} \left\{ \alpha \cdot T(\overline{u}, y)\sigma \right\} \\ &= \Lambda(\sigma, \overline{u}) \cdot \underset{\alpha \in A_{k-1}}{min} \left\{ \alpha \cdot \nu_{j^*} \right\} \\ &= \Lambda(\sigma, \overline{u}) \cdot \alpha^*(j^*), \end{aligned} \tag{4.15}$$

where

$$\alpha^*(j^*) := min\left\{ \alpha(j^*) \mid \left(\alpha(1), \alpha(2), \ldots \alpha(j^*), \ldots \alpha(N_\mathbf{X}) \right) \in A_{k-1} \right\}.$$

Hence, since by (4.14) $\Lambda(\cdot, \overline{u})$ is linear in $\mathrm{I\!R}^{N_\mathbf{X}}_+$, so is $J^\gamma_{\overline{u}}(\cdot, M-k)$. □

5. A CASE STUDY

We consider a popular benchmark problem for which much is known in the risk-neutral case. This is a two-state replacement problem which models failure-prone units in production/manufacturing systems, communication systems, etc. The underlying state of the unit can either be *working* ($X_t = 0$) or *failed* ($X_t = 1$), and the available actions are to *keep* ($U_t = 0$) the current unit or *replace* ($U_t = 1$) the unit by a new one. The cost function $(x,u) \mapsto c(x,u)$ is as follows: let $R > C > 0$, then $c(0,0) = 0$, $c(1,0) = C$, $c(x,1) = R$. The messages received have probability $1/2 < q < 1$ of coinciding with the true state of the unit. The state transition matrices are given as:

$$P(0) = \begin{bmatrix} 1-\theta & \theta \\ 0 & 1 \end{bmatrix}; \quad P(1) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \tag{5.1}$$

with $0 < \theta < 1$; see [WCC], [FAM1], [FAM2] for more details. With the above definitions, the matrices used to update the information state vector are given by:

$$\begin{aligned} T(0,y) &= 2\begin{bmatrix} q_y(1-\theta) & 0 \\ (1-q_y)\theta & (1-q_y)e^{\gamma C} \end{bmatrix}; \\ T(1,y) &= 2\begin{bmatrix} q_y e^{\gamma R} & q_y e^{\gamma R} \\ 0 & 0 \end{bmatrix}, \end{aligned} \tag{5.2}$$

where $q_y := q(1-y) + (1-q)y$, $y = 0, 1$.

For this case $\sigma = (\sigma(1), \sigma(2))^T \in \mathbb{R}^2_+$, and the dynamic programming recursions (4.5) take the form:

$$\begin{cases} J^\gamma(\sigma, M) & = \sigma(1) + \sigma(2); \\ J^\gamma(\sigma, M-k) & = min\big\{J_0^\gamma(\sigma, M-k); J_1^\gamma(\sigma, M-k)\big\}. \end{cases} \tag{5.3}$$

Define the *replace* control region $CR^k_{replace}$ and the *keep* control region CR^k_{keep} in the obvious manner, c.f. Definition 4.1.

Lemma 5.1. For all decision epochs the *replace* control region is a (possibly empty) conic segment in $\mathbb{R}^2_+$.

Proof:

¿From (5.1) the *replace* action resets the system to the *working* state. Then the result follows immediately from Lemma 4.5 and Theorem 4.1. □

The next result establishes an important *threshold* structural property of the optimal control policy. This is similar to well known results for the risk neutral case [FAM1], [FAM2], [LOV], [WCC].

Theorem 5.1. If $CR^k_{replace}$ is nonempty, then it includes the $\sigma(2)$-axis, i.e., $\mathbb{R}^2_+$ is partitioned by a line through the origin such that for values of $\sigma \in \mathbb{R}^2_+$ above the line it is optimal to *replace* the unit, and it is optimal to *keep* the unit otherwise.

Proof:

We proceed to show that if it is optimal to *keep* the unit in the $\sigma(2)$-axis (see Lemma 4.5), then the optimal policy is to *keep* the unit for all values of $\sigma \in \mathbb{R}^2_+$.

Hence, by contradiction, we can then conclude from Lemma 5.1 that if $CR^k_{replace}$ is nonempty, then it must include the $\sigma(2)$-axis, and the statement of the theorem then follows.

Let $\sigma' = (0, \sigma(2))^T$, $\sigma(2) > 0$. Then, for $0 < k \leq M$, we have from (5.2), (5.3), and Lemma 4.4 that:

$$J_0^\gamma(\sigma', M-k) = e^{\gamma C}\sigma(2)\alpha^*(2),$$

$$J_1^\gamma(\sigma', M-k) = e^{\gamma R}\sigma(2)\alpha^*(1),$$

where $\alpha^*(i)$ denotes the componentwise minimum over A_{k-1}. Suppose that

$$\begin{aligned} &J_0^\gamma(\sigma', M-k) < J_1^\gamma(\sigma', M-k) \\ &\Longleftrightarrow e^{\gamma C}\alpha^*(2) < e^{\gamma R}\alpha^*(1). \end{aligned} \tag{5.4}$$

Now, for any other $\sigma \in \mathbb{R}^2_+$,

$$J_1^\gamma(\sigma, M-k) = e^{\gamma R}(\sigma(1) + \sigma(2))\alpha^*(1),$$

and

$$J_0^\gamma(\sigma, M-k) = \sum_{y=0}^{1} \min_{\alpha \in A_{k-1}} \big\{q_y(1-\theta)\sigma(1)\alpha(1) + (1-q_y)\theta\sigma(1)\alpha(2) + (1-q_y)\sigma(2)e^{\gamma C}\alpha(2)\big\}. \tag{5.5}$$

Since $\gamma > 0$ and $R > C > 0$, then

$$\begin{aligned} J_0^\gamma(\sigma, M-k) < &\sum_{y=0}^{1} \min_{\alpha \in A_{k-1}} \big\{q_y(1-\theta)\sigma(1)e^{\gamma R}\alpha(1) + (1-q_y)\theta\sigma(1)e^{\gamma C}\alpha(2) \\ &+ (1-q_y)\sigma(2)e^{\gamma C}\alpha(2)\big\}. \end{aligned} \tag{5.6}$$

Now, define $\tilde{A}_{k-1} \subseteq A_{k-1}$ as:

$$\tilde{A}_{k-1} := \Big\{\alpha \in A_{k-1} \mid e^{\gamma C}\alpha(2) < e^{\gamma R}\alpha^*(1)\Big\}, \tag{5.7}$$

which is nonempty by (5.4). Then by minimizing over $\tilde{A}_{k-1}$ the terms on the right hand side in (5.6) we obtain an upper-bound for this expression, and we finally get that:

$$J_0^\gamma(\sigma, M-k) < e^{\gamma R}\sigma(1)\alpha^*(1) + e^{\gamma R}\sigma(2)\alpha^*(1) = J_1^\gamma(\sigma, M-k),$$

and therefore it is optimal to *keep* the unit at all $\sigma \in \mathbb{R}_+^2$. □

Using the dynamic programming recursions (5.3), the structure of optimal policies can be further elucidated. First we need a simple technical result, the proof of which is presented in the Appendix. Define:

$$\begin{aligned} \alpha_0 &:= 1; \\ \alpha_1 &:= 1 = (1-\theta)\alpha_0 + \theta e^{0\cdot\gamma C}; \\ &\vdots \\ \alpha_{k+1} &:= (1-\theta)\alpha_k + \theta e^{k\gamma C}; \qquad k = 0, 1, \ldots, M. \end{aligned} \tag{5.8}$$

Lemma 5.2: $\alpha_{k+1} > \alpha_k$, and $e^{\gamma R}\alpha_k > \alpha_{k+1}; k = 1, 2, \ldots, M$.

The following theorem gives more precise results on the structure of optimal policies, and its proof is presented in the Appendix.

Theorem 5.2: Let $0 < \overline{K} \le M$ be given.

(i) The necessary and sufficient condition for the policy with $\pi^*_{M-1}(\cdot) = \ldots = \pi^*_{M-\overline{K}}(\cdot) = 0$ (i.e., always *keep* the unit in the last $\overline{K}$ stages) to be optimal is that:

$$\frac{e^{\overline{K}\gamma C}}{\alpha_{\overline{K}-1}} \le e^{\gamma R}, \tag{5.9.a}$$

or equivalently,

$$R \ge \overline{K}C - \frac{ln(\alpha_{\overline{K}-1})}{\gamma}. \tag{5.9.b}$$

(ii) If (5.9) holds, then:

$$\begin{aligned} J^\gamma(\sigma, M-\overline{K}) &= J_0^\gamma(\sigma, M-\overline{K}) = \alpha_{\overline{K}}\sigma(1) + e^{\overline{K}\gamma C}\sigma(2); \\ J_1^\gamma(\sigma, M-\overline{K}) &= \alpha_{\overline{K}-1}e^{\gamma R}\Big(\sigma(1) + \sigma(2)\Big). \end{aligned} \tag{5.10}$$

(iii) If $1 \leq \overline{K} \leq M$ is the smallest integer for which (5.9) fails to hold, then $\pi^*_{M-\overline{K}}(\cdot)$ is of threshold type, with $\mathbb{R}^2_+$ being partitioned by the line:

$$\frac{e^{\gamma R}\alpha_{\overline{K}-1} - \alpha_{\overline{K}}}{e^{\overline{K}\gamma C} - e^{\gamma R}\alpha_{\overline{K}-1}}\sigma(1) = \sigma(2), \tag{5.11}$$

such that the region to the left (above) the line is the *replace* control region.

Remark 5.1. Note that the simplest nontrivial decision process corresponds to the case $M = 2$, since (5.9) is always satisfied for $\overline{K} = 1$.

5.1. SMALL AND LARGE RISK LIMITS

In order to build a better understanding as to how risk sensitivity manifests itself in the structure of the optimal control strategies, we analyze both the small risk sensitivity limiting case ($\gamma \to 0$) and the infinite risk aversion case ($\gamma \to \infty$). First, we give some insight on conditions (5.9).

Remark 5.2. Let $\overline{K}$ be a positive integer. Then if $R \geq \overline{K}C$, we have that

$$sgn(\gamma)e^{\gamma R} \geq sgn(\gamma)e^{\gamma \overline{K}C},$$

and hence the DM would never prefer to *replace* the unit, at any point within the last $\overline{K}$ stages. Therefore, $R \geq \overline{K}C$ is a *sufficient* condition for the optimal policy to be $\pi^*_{M-k}(\sigma) = 0$, $\forall \sigma \in \mathbb{R}^*_+$, $k = 1, 2, \ldots, M - \overline{K}$, i.e., it is **optimal to *keep* the unit** for the last $\overline{K}$ stages; c.f. (5.9).

Infinite Risk Aversion Case ($\gamma \to \infty$).

Consider the situation $\gamma \to \infty$ in (5.8). Note that for γ large enough:

$$\alpha_{\overline{K}-1} \approx \theta \cdot e^{\left(\overline{K}-2\right)\gamma C}.$$

Therefore, as $\gamma \to \infty$, we have from (5.9) that the necessary and sufficient condition for it to be always optimal to *keep* the unit in the last $\overline{K}$ stages approaches $R \geq 2C$, which is the same condition for it to be always optimal to *keep* the unit in the last two stages. Furthermore, as is readily verified from (4.10), if $R < 2C$ and $\gamma \to \infty$ then it is always optimal to *replace* the unit at stage $M - k$, for all $2 \leq k \leq M$, i.e., the threshold line tends to the $\sigma(1)$-axis. Hence the DM becomes *myopic* in the sense that, except for the last one, all decision epochs appear to be the same. The DM appears to always face a *two*-stage decision process, the simplest one possible; c.f. Remark 5.1. In the jargon of Whittle [WHI], it could be then said that an infinitely risk averse DM exhibits "neurotic" behavior, his optimal strategy being of the "*bang bang*" type with respect to the parameter R: if $R \geq 2C$, then $\pi^*_{M-k}(\cdot) = 0$, and otherwise $\pi^*_{M-k}(\cdot) = 1$, for all $2 \leq k \leq M$. This behavior can be partly explained by noting that at most one change will then occur in the stream of costs, thus achieving least variability in the cumulative cost.

Small Risk Aversion Case ($\gamma \to 0$).

Next, we examine the question: **How do the results in Theorems 4.1, 5.1-5.2 compare to known results for the risk-neutral case?** The answer is that the risk-sensitive controller obtained here has as its small risk limit the known risk-neutral controller, and both controllers have in general a similar structure. Similarly as in [WCC], the dynamic programming equations for the risk-neutral case can be written, with the conditional probability distribution of the state as the information state. Then, it can be shown that the optimal risk-neutral controller has a structure similar to the risk-sensitive controller given in Theorem 5.1. Furthermore, it can be shown that the necessary and sufficient condition in the risk-neutral case for the separated policy $\pi^*_{M-1}(\cdot) = \ldots = \pi^*_{M-\overline{K}}(\cdot) = 0$ to be optimal is:

$$R > \overline{K}C - \alpha'_{\overline{K}-1}, \tag{5.12}$$

where $\alpha'_{\overline{K}-1}$ is obtained as the derivative with respect to γ, evaluated as $\gamma \to 0$, of (5.8). As can be easily verified, the above is nothing but the small risk limit (i.e., as $\gamma \to 0$) of (5.9).

General Risk Aversion Case ($\gamma > 0$).

The following result helps bring to light a manifestation of aversion to risk in the DM; its proof is given in the Appendix.

Lemma 5.3. Let $\gamma > 0$, then for all $k > 1$:

$$ln(\alpha_k) > \gamma \cdot \alpha'_k. \tag{5.13}$$

□

Notice that the decision to *replace* a unit involves an uncertain, and therefore a risky, investment in that the unit being replaced may actually be in *working* condition, or it may subsequently fail. This is reflected in (5.9), (5.12) and (5.13) in that a risk neutral DM or controller may decide to *replace* a unit for values of R higher than a risk averse DM or controller would not.

REFERENCES

[ABFGM] A. Arapostathis, V. Borkar, E. Fernández-Gaucherand, M.K. Ghosh and S.I. Marcus, Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey, *SIAM Journal on Control & Optimization* **31** (1993) 282-344.

[AST] K.J. Åström, Optimal Control of Markov Processes with Incomplete State Information, II. The Convexity of the Loss Function, *J. Math. Anal. Appl.*, **26** (1969) 403-406.

[BE1] D.P. Bertsekas, *Dynamic Programming and Stochastic Control,* Academic Press, New York, 1976.

[BE2] D.P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models,* Prentice-Hall, Englewood Cliffs, 1987.

[BJ] J.S. Baras and M.R. James, Robust and Risk-sensitive Output Feedback Control for Finite State Machines and Hidden Markov Models, preprint, August (1994).

[BSO] M. Bouakiz and M.J. Sobel, Inventory Control with an Exponential Utility Criterion, *Operations Research* **40** (1992) 603-608.

[BVS] A. Bensoussan and J.H. Van Schuppen, Optimal Control of Partially Observable Stochastic Systems with an Exponential-of-Integral Performance Index, *SIAM Journal on Control & Optimization* **23** (1985) 599-613.

[CSO] K-J. Chung and M.J. Sobel, Discounted MDP's: Distribution Functions and Exponential Utility Maximization, *SIAM Journal on Control & Optimization* **25** (1987) 49-62.

[DOB] E.-E. Doberkat, *Stochastic Automata: Stability, Nondeterminism, and Prediction*, Springer-Verlag, Berlin, 1981.

[EM] R.J. Elliot and J.B. Moore, Discrete Time Partially Observed Control, in *Proceedings 12th IFAC World Congress*, Sidney, Australia (1993).

[FAM1] E. Fernández-Gaucherand, A. Arapostathis, and S.I. Marcus, On the Average Cost Optimality Equation and the Structure of Optimal Policies for Partially Observable Markov Decision Processes, *Annals of Operations Research* **29** (1991) 439–470.

[FAM2] E. Fernández-Gaucherand, A. Arapostathis and S.I. Marcus, Analysis of an Adaptive Control Scheme for a Partially Observed Controlled Markov Chain, *IEEE Transactions on Automatic Control* **38** (1993) 987-993.

[FEM] E. Fernández-Gaucherand and S.I. Marcus, Risk-sensitive Optimal Control of Hidden Markov Models: a Case Study, in *Proceedings 33rd IEEE Conf. Decision and Control*, Orlando, FL, (1994) 1657-1662.

[GHE] A. Gheorghe, Partially Observable Markov Processes with a Risk-sensitive Decision Maker, *Rev. Roum. Math. Pures et Appl.* **22** (1977) 461-482.

[HOM] R.A. Howard and J.E. Matheson, Risk-sensitive Markov Decision Processes, *Management Science* **18** (1972) 356-370.

[JAQ] S.C. Jaquette, A Utility Criterion for Markov decision Processes, *Management Science* **23** (1976) 43-49.

[JBE] M.R. James, J.S. Baras and R.J. Elliott, Risk-sensitive control and dynamic games for Partially Observed Discrete-time Nonlinear Systems, *IEEE Transactions on Automatic Control* **39** (1994) 780-792.

[KV] P.R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, 1986.

[LOV] W.S. Lovejoy, On the Convexity of Policy Regions in Partially Observed Systems, *Operations Research*, **35** (1987) 619-621.

[LR] R.D. Luce and H. Raiffa, *Games and Decisions: Introduction and Critical Survey,* Dover Publications, New York, 1989 (reprint from 1957 edition).

[PAZ] A. Paz, *Introduction to Probabilistic Automata*, Academic Press, New York, 1971.

[PT] J.W. Pratt, Risk Aversion in the Small and in the Large, *Econometrica* **32** (1964) 122-136.

[SHI1] A. N. Shiryaev, On the Theory of Decision Functions and Control by an Observation Process with Incomplete Data, in *Trans. 3rd Prague Conf. on Inform. Theory, Statistical Decision Functions and Random Processes,* (1964) 657–681, Publ. House Czech. Acad. Sci., Prague (in Russian). English transl., *Selected Translations in Mathematical Statistics and Probability* **6** (1966) 162–188, American Mathematical Society, Providence, RI.

[SHI2] A. N. Shiryaev, Some New Results in the Theory of Controlled Random Sequences, *Trans. 4th Prague Conf. on Inform. Theory, Statistical Decision*

Functions and Random Processes, (1965) 131–203, Publ. House Czech. Acad. Sci., Prague (in Russian). English transl.,*Selected Translations in Mathematical Statistics and Probability* **8** (1970) 49–130, American Mathematical Society, Providence, RI.

[SHI3] A. N. Shiryaev, On Markov Sufficient Statistics in Non-additive Bayes Problems of Sequential Analysis, *Theory Probab. Appl.* **9** (1964) 604–618.

[SSO] R.D Smallwood and E.J. Sondik, The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon, *Operations Research* **21** (1973) 1071-1088.

[WCC] C.C. White, A Markov Quality Control Process Subject to Partial Observation, *Management Science* **23** (1977) 843-852.

[WHI] P. Whittle, *Risk-sensitive Optimal Control*, Wiley, England, 1990.

Appendix

Proof of Lemma 5.2:

Note that $e^{\gamma C} > 1 = \alpha_1$. Proceeding by induction, suppose that $e^{k\gamma C} > \alpha_k$. Then, from (5.8), we have:

$$\begin{aligned} e^{(k+1)\gamma C} &= (1-\theta)e^{(k+1)\gamma C} + \theta e^{(k+1)\gamma C} \\ &> (1-\theta)e^{k\gamma C} + \theta e^{k\gamma C} \\ &> (1-\theta)\alpha_k + \theta e^{k\gamma C} \\ &= \alpha_{k+1}. \end{aligned} \tag{A.1}$$

Then the first result follows since:

$$\alpha_{k+1} - \alpha_k = \theta\big[e^{k\gamma C} - \alpha_k\big] > 0. \tag{A.2}$$

On the other hand, we have that

$$\begin{aligned} e^{\gamma R}\alpha_k &= (1-\theta)e^{\gamma R}\alpha_k + \theta e^{\gamma R}\alpha_k \\ &> (1-\theta)\alpha_k + \theta e^{\gamma C} = \alpha_{k+1}. \end{aligned} \tag{A.3}$$

□

Proof of Theorem 5.2. It is instructive to first perform a few steps of the dynamic programming recursions of (5.3), which we proceed to do next. Recall (4.11), (5.2), and (5.8) which will be used repeatedly.

STAGE k = 0.

$$J^\gamma(\sigma, M) = \sigma(1) + \sigma(2). \tag{A.4}$$

STAGE k = 1.

$$J^\gamma(\sigma, M-1) = min\Big\{J_0^\gamma(\sigma, M-1), J_1^\gamma(\sigma, M-1)\Big\}; \tag{A.5}$$

where

$$\begin{aligned} J_0^\gamma(\sigma, M-1) &= (1-\theta)\sigma(1) + \theta\sigma(1) + e^{\gamma C}\sigma(2) = \alpha_1\sigma(1) + e^{\gamma C}\sigma(2); \\ J_1^\gamma(\sigma, M-1) &= \alpha_0 e^{\gamma R}\big(\sigma(1) + \sigma(1)\big). \end{aligned} \tag{$A.6$}$$

Therefore, the region in $\mathbb{R}_+^2$ where it is optimal to *keep* the unit is defined by the condition:

$$\begin{aligned} & J_0^\gamma(\sigma, M-1) \leq J_1^\gamma(\sigma, M-1) \\ \iff & \alpha_1\sigma(1) + e^{\gamma C}\sigma(2) \leq \alpha_0 e^{\gamma R}\big(\sigma(1) + \sigma(2)\big) \\ \iff & \big[\alpha_1 - e^{\gamma R}\alpha_0\big)\sigma(1) \leq \big(e^{\gamma R} - e^{\gamma C}\big]\sigma(2) \\ \iff & \frac{e^{\gamma R}\alpha_0 - \alpha_1}{e^{\gamma C} - e^{\gamma R}\alpha_0}\sigma(1) \leq \sigma(2). \end{aligned} \tag{$A.7$}$$

Since $e^{\gamma C} \leq e^{\gamma R}$ and $\sigma(i) \geq 0, i = 1, 2$, then (A.7) implies that it is **always optimal to keep the unit**, i.e., $\pi_{M-1}^*(\sigma) = 0, \forall \sigma \in \mathbb{R}_+^2$.

STAGE k = 2.

We have that $J^\gamma(\cdot, M-1) = J_0^\gamma(\cdot, M-1)$. Therefore:

$$\begin{aligned} J_0^\gamma(\sigma, M-2) &= (1-\theta)\sigma(1) + e^{\gamma C}\big(\theta\sigma(1) + e^{\gamma C}\sigma(2)\big) \\ &= \alpha_2\sigma(1) + e^{2\gamma C}\sigma(2); \end{aligned} \tag{$A.8.a$}$$

$$J_1^\gamma(\sigma, M-2) = \alpha_1 e^{\gamma R}\big(\sigma(1) + \sigma(2)\big). \tag{$A.8.b$}$$

Then the condition for it to be optimal to *keep* the unit is:

$$\begin{aligned} & J_0^\gamma(\sigma, M-2) \leq J_1^\gamma(\sigma, M-2) \\ \iff & \alpha_2\sigma(1) + e^{2\gamma C}\sigma(2) \leq \alpha_1 e^{\gamma R}\big(\sigma(1) + \sigma(2)\big) \\ \iff & \big[\alpha_2 - e^{\gamma R}\alpha_1\big]\sigma(1) \leq \big[e^{\gamma R}\alpha_1 - e^{2\gamma C}\big]\sigma(2). \end{aligned} \tag{$A.9$}$$

Therefore, using the results in Lemma 5.2, we conclude that it is **always optimal to keep the unit**, i.e., $\pi^*_{M-2}(\sigma) = 0, \forall \sigma \in \mathbb{R}^2_+$, if and only if $R \geq 2C$. Otherwise, $\mathbb{R}^2_+$ is partitioned into two control regions by a the line, as described by:

$$\frac{\left[e^{\gamma R}\alpha_1 - \alpha_2\right]}{e^{2\gamma C} - e^{\gamma R}\alpha_1}\sigma(1) = \sigma(2). \tag{A.10}$$

Thus the optimal policy is of the *threshold type*, such that it is optimal to *keep* the unit for σ values below and on the line determined by (A.10), and to *replace* the unit otherwise.

STAGE $\overline{\mathbf{K}}$, $\mathbf{2} < \overline{\mathbf{K}} \leq \mathbf{M}$.

Assume that

$$R \geq \left(\overline{K} - 1\right)C - \frac{ln(\alpha_{\overline{K}-2})}{\gamma},$$

and therefore: a) $\pi^*_{M-1}(\cdot) = \ldots = \pi^*_{M-\overline{K}+1}(\cdot) = 0$, and b) $J^\gamma(\sigma, M - \overline{K} + 1) = \alpha_{\overline{K}-1}\sigma(1) + e^{(\overline{K}-1)\gamma C}\sigma(2)$. Thus, we have:

$$\begin{aligned} J^\gamma_0(\sigma, M - \overline{K}) &= \alpha_{\overline{K}-1}(1-\theta)\sigma(1) + e^{(\overline{K}-1)\gamma C}\left(\theta\sigma(1) + e^{\gamma C}\sigma(2)\right) \\ &= \alpha_{\overline{K}}\sigma(1) + e^{\overline{K}\gamma C}\sigma(2); \end{aligned} \tag{A.11.a}$$

$$J^\gamma_1(\sigma, M - \overline{K}) = e^{\gamma R}\alpha_{\overline{K}-1}\left(\sigma(1) + \sigma(2)\right) \tag{A.11.b}$$

Then, the condition for it to be optimal to *keep* the unit is:

$$\begin{aligned} & J^\gamma_0(\sigma, M - \overline{K}) \leq J^\gamma_1(\sigma, M - \overline{K}) \\ \iff & \alpha_{\overline{K}}\sigma(1) + e^{\overline{K}\gamma C}\sigma(2) \leq e^{\gamma R}\alpha_{\overline{K}-1}\left(\sigma(1) + \sigma(2)\right) \\ \iff & \left[\alpha_{\overline{K}} - e^{\gamma R}\alpha_{\overline{K}-1}\right]\sigma(1) \leq \left[e^{\gamma R}\alpha_{\overline{K}-1} - e^{\overline{K}\gamma C}\right]\sigma(2). \end{aligned} \tag{A.12}$$

Thus, the necessary and sufficient condition for it to be optimal to *keep* the unit for all $\sigma \in \mathbb{R}^2_+$ is:

$$\frac{e^{\overline{K}\gamma C}}{\alpha_{\overline{K}-1}} \leq e^{\gamma R}. \tag{A.13}$$

If (A.13) is not satisfied, then the optimal policy will be of the threshold type, with $\mathbb{R}_+^2$ being partitioned into *keep* and *replace* regions by a line, as defined by (5.11). □

Proof of Lemma 5.3. From the recursion (5.8) it follows that:

$$\alpha_k = \sum_{l=0}^{k-2}(1-\theta)^l\theta e^{(k-1-l)\gamma C} + (1-\theta)^{k-1}, \tag{$A.14.a$}$$

$$\alpha'_k = \sum_{l=0}^{k-2}(1-\theta)^l\theta(k-1-l)C. \tag{$A.14.b$}$$

Now, it also follows from (5.8) that:

$$\alpha_k|_{\gamma=0} = 1, \tag{$A.14.c$}$$

and therefore (A.14.a) is a convex combination of exponentials (the last term corresponding to $e^{0\cdot\gamma C}$. Therefore, since $ln(\cdot)$ is a strictly convex function, we have that:

$$\begin{aligned} ln(\alpha_k) &= ln\big(\sum_{l=0}^{k-2}(1-\theta)^l\theta e^{(k-1-l)\gamma C} + (1-\theta)^{k-1}\big) \\ &> \sum_{l=0}^{k-2}(1-\theta)^l\theta ln\big(e^{(k-1-l)\gamma C}\big) \\ &= \gamma\alpha'_k. \end{aligned} \tag{$A.15$}$$

□