TECHNICAL RESEARCH REPORT

Non-Standard Optimality Criteria for Stochastic Control Problems

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

T.R. 95-101

INSTITUTE FOR SYSTEMS RESEARCH

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

NON-STANDARD OPTIMALITY CRITERIA FOR STOCHASTIC CONTROL PROBLEMS

Emmanuel Fernández-Gaucherand
Systems & Industrial Eng. Dept.
The University of Arizona
Tucson, AZ 85721
emmanuel@sie.arizona.edu.

Steven I. Marcus
Institute for Systems Research and Electrical Engineering Department
The University of Maryland
College Park, MD 20742.
marcus@src.umd.edu

Abstract

In this paper, we survey several recent developments on non-standard optimality criteria for controlled Markov process models of stochastic control problems. Commonly, the criteria employed for optimal decision and control are either the discounted cost (DC) or the long-run average cost (AC). We present results on several other criteria that, as opposed to the AC or DC, take into account, e.g., a) the variance of costs; b) multiple objectives; c) robustness with respect to sample path realizations; d) sensitivity to long but finite horizon performance as well as long-run average performance.

I. Introduction and Motivation

Stochastic optimal control methods find significant applications in areas such as aerospace control problems, management and finance, manufacturing and production, communication/computer networks, military and service industry logistics, etc. Several of these areas fall in the general category of Discrete Event Stochastic Dynamic Systems (DESDS), which is a broad and interdisciplinary area of research at the intersection of systems and control, operations research, and knowledge-based systems. The focus of DESDS research are the complex, man-made systems that form the core of our technological society. The main distinguishing feature of DESDS is that their dynamic evolution is driven by the (random) occurrence of controlled and uncontrolled discrete events, e.g., system failures, service rate increase, etc. A broad view of models and methods for DESDS is given in [HO], and some recent books following this paradigm are, e.g., [CAS] and [GER]. The common objective for many of the problems mentioned before is to obtain rules for operating the system, i.e., *control policies* or *decision rules*, which optimize an appropriate performance measure, and which explicitly take into account the stochastic and discrete event-driven nature of the problems. Appropriate mathematical descriptions for the state evolution and control of these processes fall within the domain of **controlled Markov processes** (CMP). However, standard optimality criteria, e.g., expected discounted or average costs, frequently fail to capture important aspects associated with the stochastic control problems that arise in many applications. Thus the need arises for new paradigms, which lead to the formulation of some non-standard performance criteria that address some of the shortcomings of, and give alternatives to, the traditional expected long-run average and discounted sum-of-costs criteria [ABFGM], [FGM], [CFE1], [CFE2], [FEM]. In this paper, recent results in this area, as well as some open questions, are presented.

II. Methods and Procedures

A CMP is a discrete-time, discrete-event stochastic dynamic system specified by the five-tuple $\langle \mathbf{X}, \mathbf{U}, \mathcal{U}, P, c\rangle$, where $\mathbf{X}$ is the *state space;* $\mathbf{U}$ is the *action*, or *control* space; each pair (x,u) in $\mathbf{X}\times\mathbf{U}$ determines the *distribution law* $P(\cdot \mid x,u)$ of the next state X_{t+1}, when the current state and action are, respectively, $X_t = x$ and $U_t = u$; and $c : \mathbf{X}\times\mathbf{U} \to \mathbb{R}$ is the one-stage cost function. A control strategy, or policy, is a rule π for making decisions, based on the available information. At a given time t, the available information is the set h_t of observed states and actions taken up to that time, i.e., $h_t = (X_0, U_0, X_1, \ldots, U_{t-1}, X_t)$. Each policy π incurs a stream of costs $\{c(X_0,U_0), c(X_1,U_1),\ldots\}$. The two standard criteria used for optimal decision and control on an infinite planning horizon are the following.

Discounted Cost (DC): For $0 < \beta < 1$, the *discount factor,* and a policy π, the total discounted cost is given by

$$J_\beta(x,\pi) := E_x^\pi\left[\sum_{t=0}^{\infty}\beta^t c(X_t,U_t)\right].$$

Average Cost (AC): Given a policy π, we have that

$$J(x,\pi) := \limsup_{N\to\infty} E_x^\pi\left[\frac{1}{N}\sum_{t=0}^{N-1} c(X_t,U_t)\right].$$

III. Non-Standard Criteria

The AC and DC criteria suffer from several shortcomings. Among these shortcomings, we mention the following:

- They measure the *expected* sum of (discounted or averaged) cost, and thus are *insensitive* to, e.g., the variance of the sum of costs;
- They can be seen as two opposite extremes in the spectrum of possible criteria;
- In many situations, the optimal policy has to meet stringent robustness requirements, e.g., optimality *for almost all* sample paths, which are not obtained in general with AC or DC criteria;
- Stronger criteria than AC that exhibit selectivity with respect to long but finite horizons are desirable.

In this paper, we discuss several promising alternatives that address the above shortcomings.

III.a: Risk-Sensitive Criteria

In [FEM1], we studied the 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. Here $\mathbf{X} = \{1, 2, \ldots, N_{\mathbf{X}}\}$; $\mathbf{U} = \{1, 2, \ldots, N_{\mathbf{U}}\}$; $P(u) := \left[p_{i,j}(u)\right]$ is the $N_{\mathbf{X}} \times N_{\mathbf{X}}$ state transition matrix. In addition, $\mathbf{Y} = \{1, 2, \ldots, N_{\mathbf{Y}}\}$ is the set of observations (or messages), and $Q(u) := \left[q_{x,y}(u)\right]$ 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. Hence, based on $\mathcal{I}_t := (U_0, Y_1, U_1, Y_2, \ldots, U_t, Y_{t+1})$, a new decision U_{t+1} is selected at time $t+1$.

Let $\mathcal{C}_M := \sum_{t=0}^{M-1} c(X_t, U_t)$ be the sum of costs for the finite horizon M. 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:

$$J^\gamma(\pi, X_0) := sgn(\gamma)\mathbb{E}^\pi\left[exp(\gamma \cdot \mathcal{C}_M)\right], \qquad (1)$$

where $\gamma \neq 0$ is the *risk-factor*, and $sgn(\gamma)$ is the sign of γ; see [BER1], [FEM1], and [WHI] for details. If $\gamma > 0$, then the controller is *risk-averse* or *pessimistic*, whereas if $\gamma < 0$ then the controller is *risk-preferring* or *optimistic*.

The optimal control of HMM under standard, risk-neutral performance criteria, e.g., AC and DC, has received much attention in the past [ABFGM]. Controlled Markov chains with full state information and a risk-sensitive performance criterion have also received some attention [HOM], [CSO], [PUT]. 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. On the other hand, HMM under risk-sensitive criteria had not received attention until recently [BJ], [FEM1]. The reason for this was mainly due to the complexity of the problem, and the lack of *information states* that could serve a similar purpose as the conditional probability does for the case of HMM with risk-neutral criteria [ABFGM], [BE2], [KV]. Recently, James, Baras, and Elliott [BJ], [JBE] have derived information states for nonlinear, discrete-time partially observable stochastic systems, and in particular for HMM, under an "exponential of additive costs" criterion. They have also given dynamic programming equations from which optimal values and controls can be computed, for problems with a finite horizon. Nevertheless, the following remains mostly an open question:

How does risk-sensitivity manifest itself in a policy?

For the LEQG problem, Whittle [WHI] 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 [FEM1], a similar investigation was initiated for HMM, via an examination of a popular benchmark problem. The dynamic programming equations for both the risk-neutral and risk-sensitive cases were examined. The *threshold* structure of optimal controllers thus obtained was compared. Also, it was shown 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.

Recently, several general *structural* results have been obtained [FEM2]. In the HMM context, the information states are vectors $\sigma \in \mathbb{R}_+^{N_\mathbf{X}}$, where $\mathbb{R}_+^{N_\mathbf{X}} := \left\{\sigma \in \mathbb{R}^{N_\mathbf{X}} \mid \sigma_i \geq 0, \forall i\right\}$. Denote the value functions for the finite horizon M as $J^\gamma(\cdot, M-k) : \mathbb{R}_+^{N_\mathbf{X}} \to \mathbb{R}$, $k = 1, \ldots, M$. Among other results, the following is shown in [FEM1] and [FEM2].

Lemma 1: The value functions $J^\gamma(\cdot, M-k)$ are concave functions of $\sigma \in \mathbb{R}_+^{N_\mathbf{X}}$.

Lemma 2: The value functions $J^\gamma(\cdot, M-k)$ are piecewise linear functions in $\sigma \in \mathbb{R}_+^{N_\mathbf{X}}$.

Lemma 3: Optimal (separated) policies $\{\pi_0, \ldots, \pi_{M-1}\}$ are constant along rays through the origin, i.e., for $\sigma \in \mathbb{R}_+^{N_\mathbf{X}}$ then $\pi_t^*(\sigma') = \pi_t^*(\sigma)$, for all $\sigma' = \lambda\sigma$, $\lambda \geq 0$.

Definition: An action $\overline{u} \in \mathbf{U}$ is said to be a *resetting* action if, starting with $\sigma_t = \sigma \in \mathbb{R}_+^{N_\mathbf{X}}$ arbitrary, under action $\overline{u} \in \mathbf{U}$ the updated value of the information state is $\sigma_{t+1} = \lambda\sigma^{(i)}$, for some $\lambda > 0$, where $\sigma^{(1)} = (1,0,0,\ldots,0)$, $\sigma^{(2)} = (0,1,0,\ldots,0)$, ..., $\sigma^{(N_\mathbf{X})} = (0,0,0,\ldots,1)$.

Lemma 4: Let π^*_{M-k} be an optimal policy for stage $M-k$, and let $\overline{u} \in \mathbf{U}$ be a resetting action. Then the *control region* for action $\overline{u} \in \mathbf{U}$, $\mathcal{CR}_{M-k}(\overline{u}) := \{\sigma \in \mathbb{R}_+^{N_\mathbf{X}} \mid \pi^*_{M-k}(\sigma) = \overline{u}\}$, is a convex subset of $\mathbb{R}_+^{N_\mathbf{X}}$.

III.b: Weighted Cost Criteria

One possibility of obtaining a reasonable compromise of AC and DC criteria is by combining these in a weighted sum. CMP with a *generalized* WC criterion and with general state and control spaces and *several* one-stage cost functions being discounted at *different* rates, were studied in [FGM]. The consideration of several cost functions at the same time makes it necessary to explicitly include the cost function in the notation for both AC and DC, e.g., $J(x,\pi,c)$.

Generalized Weighted Cost (GWC): Let $K \in \mathbb{N}$ and $0 \leq \alpha \leq 1$ be given, and let $0 < \beta_K < \beta_{K-1} < \ldots < \beta_2 < \beta_1 < 1$ be given discount factors. In addition, for $k = 1,2,\ldots,K+1$, let $c_k(\cdot,\cdot)$ be given one-stage cost functions. The generalized weighted cost incurred by a policy π is given by

$$W_g(x,\pi) := \alpha(1-\beta_1)J_{\beta_1}(x,\pi,c_d) + (1-\alpha)J(x,\pi,c_{K+1}), \quad (2)$$

where $J_\beta(x,\pi,c)$ and $J(x,\pi,c)$ stand for the DC and AC cost, respectively, incurred by policy π, with initial state x and one-stage cost $c(\cdot,\cdot)$, and for $t \in \mathbb{N}$,

$$c_d(x,u,t) := \sum_{k=1}^{K} \left(\frac{1-\beta_k}{1-\beta_1}\right)\left(\frac{\beta_k}{\beta_1}\right)^t c_k(x,u). \quad (3)$$

Note that if each $c_k(\cdot,\cdot)$ is a bounded function, then $c_d(x,u,t)$ is (uniformly) bounded in $t \in \mathbb{N}$, and

$$J_{\beta_1}(x,\pi,c_d) = \sum_{k=1}^{K} \frac{(1-\beta_k)}{(1-\beta_1)} J_{\beta_k}(x,\pi,c_k). \quad (4)$$

Thus the first term on the right-hand side of (2) gives a weighted combination of discounted cost criteria. Note that $c_d(\cdot,\cdot,\cdot)$ is *nonstationary*, however by considering the augmented state space $\mathbf{X}\times\mathbb{N}$ the problem becomes stationary. One of the main results obtained in [FGM] is stated below, after some necessary assumptions.

Assumption 3.1: There exists an AC optimal policy π^*_a which is *stationary deterministic* (c.f. [FGM],[ABFGM]).

Assumption 3.2: There exists a constant $\rho^* \in \mathbb{R}$ such that $J^*(x,c_{K+1}) = \rho^*$, for all $x \in \mathbf{X}$.

Remark 3.1: If there is a *bounded solution* (ρ^*,h) to the average cost optimality equation (ACOE) with $\rho^* \in \mathbb{R}$ and $h : \mathbf{X} \to \mathbb{R}$, then Assumptions 3.1 and 3.2 are satisfied; see [ABFGM] and references therein.

Theorem 3.1: Let $K \in \mathbb{N}$ and $0 \leq \alpha \leq 1$ be given, and let $0 < \beta_K < \beta_{K-1} < \ldots < \beta_2 < \beta_1 < 1$ be given discount factors. In addition, for $k = 1,2,\ldots,K+1$, let $c_k(\cdot,\cdot)$ be given one-stage cost functions, each a bounded function. If Assumptions 3.1 and 3.2 hold, then
(i) For each $\varepsilon > 0$ there exists $N(\varepsilon) \in \mathbb{N}$, such that for all $x \in \mathbf{X}$ there is a GWC ε-x-optimal policy $\pi^{\varepsilon,x}$, with $\pi_t^{\varepsilon,x} = \pi^*_a$, for all $t \geq N(\varepsilon)$.
(ii) For each $\varepsilon > 0$ there exists $N(\varepsilon) \in \mathbb{N}$, such that if π^*_d is a Markov deterministic DC ε-optimal policy (see [ABFGM], [FGM]) for the one-stage cost function $c_d(\cdot,\cdot,\cdot)$ of (3), then for any $N \geq N(\varepsilon)$, the policy π^ε given by

$$\pi_t^\varepsilon := \begin{cases} \pi^*_{d,t}, & \text{if } 0 \leq t < N, \\ \pi^*_a, & \text{if } N \leq t, \end{cases}$$

is GWC ε-optimal.
(iii) The *utopian* lower bound is attained, i.e.,

$$W_g^*(x) = \alpha(1-\beta_1)J^*_{\beta_1}(x,c_d) + (1-\alpha)\rho^*, \qquad \forall x \in \mathbf{X}.$$

Remark 3.2: The structure of the ε-optimal policy obtained in (ii) above is intuitively appealing: use the best policy for the weighted DC criteria long enough at the beginning, and then switch to the policy which is best in the long run. Likewise, (i) above simply says to eventually use the policy which is best in the long run. These type of policies are sometimes called *ultimately (stationary) deterministic*. If $\alpha = 1$, then π^*_d above is GWC ε-optimal.

III.c: Overtaking Criteria

The use of the *overtaking criterion* (OC) is one way to incorporate sensitivity to finite time behavior, while preserving results obtained under an AC criterion. A policy π_1 is said to *overtake* another policy π_2 if

$$J_T(x,\pi_1) \leq J_T(x,\pi_2), \quad (5)$$

for all $x \in \mathbf{X}$, and for all T sufficiently large; $J_T(\cdot,\cdot)$ denotes the total expected cost up to time T. A policy is called *overtaking optimal* if it overtakes every other policy. Clearly, overtaking optimality implies average cost optimality, but the converse is not true in general. In [FGM], we studied models with an OC, countable state space and compact action space. Our approach was based on qualitative properties of the optimality equations for the AC criterion [ABFGM]. In [FGM] it was shown that under a *Lyapunov Function Condition* [ABFGM], OC optimal policies exists and can be characterized as the maximizers in a certain functional equation.

III.d: Strong Average Criteria

The *strong average cost* (SAC) criterion is also introduced to assess the performance of a policy over long but finite horizons, as well as in the long-run average sense. We say that policy π^* is *strong average cost* (SAC) optimal if

$$\frac{1}{n+1}\left[J_n(x,\pi^*) - J_n^*(x)\right] \underset{n\to\infty}{\longrightarrow} 0, \qquad \forall x \in \mathbf{X}; \quad (6)$$

$J_n^*(\cdot)$ denotes the optimal value function for horizon n. Note that (6) ensures that π^* induces good performance for long but finite horizons, and that every policy that is SAC optimal is also AC optimal. However the opposite is not necessarily true.

It was shown in [CFE2] that for bounded one-stage cost functions, conditions introduced by Sennott [SEN] (see also [ABFGM]) are sufficient to guarantee that *every* AC optimal policy is also SAC optimal. On the other hand, a detailed counterexample is given that shows that this result does not extend to the case of unbounded cost functions. The latter case was studied in [GM], under a different set of conditions.

III.e: Sample Path Criteria

Departing from the use of *expected* values of costs, in [CFE1] we focused on a sample path analysis of the stream of costs. We have the following definition.

Sample Path Average Cost (SPAC): The long-run sample path average cost obtained by policy π, when the initial state of the system is $x \in \mathbf{X}$, is given by

$$J_S(x,\pi) := \liminf_{N\to\infty} \frac{1}{N+1}\sum_{t=0}^{N} c(X_t, A_t). \quad (7)$$

A policy $\overline{\pi}^*$ is said to be SPAC optimal if there exists a constant $\overline{\rho}$ such that for all $x \in \mathbf{X}$ we have that:

$$J_S(x,\overline{\pi}^*) = \overline{\rho}, \quad \mathcal{P}_x^{\overline{\pi}^*} - a.s., \quad (8)$$

while, for all policies π and for all $x \in \mathbf{X}$,

$$J_S(x,\pi) \geq \overline{\rho}, \quad \mathcal{P}_x^{\pi} - a.s.. \quad (9)$$

The constant $\overline{\rho}$ is the optimal sample path average cost.

Under a Lyapunov Function Condition (see, e.g., [ABFGM]), we showed in [CFE1] that stationary policies obtained from the average cost optimality equation are not only expected average cost optimal, but indeed sample path average cost optimal. For a summary of similar results, but under a different set of conditions as those used in [CFE1], see [ABFGM, Section 5.3].

Acknowledgements

The work of E. Fernández-Gaucherand was supported in part by a grant from the University of Arizona Foundation and the Office of the Vice President for Research; and in part by the National Science Foundation under grant NSF-INT 9201430. The work of S.I. Marcus was partially supported by the National Science Foundation under Grant EEC 9402384.

REFERENCES

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

[BER1] D.P. Bertsekas, *Dynamic Programming and Stochastic Control,* Academic Press, New York, 1976.

[BER2] 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, February 1993 and September 1994.

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

[CAS] C.G. Cassandras, *Discrete Event Systems,* Irwin and Aksen Assoc., 1993.

[CFE1] R. Cavazos-Cadena and E. Fernández-Gaucherand, "Denumerable Controlled Markov Chains with Average Reward Criterion: Sample Path Optimality," **ZOR**: *Methods and Models in Operations Research* **41**, (1995) 89-108.

[CFE2] R. Cavazos-Cadena and E. Fernández-Gaucherand, "Denumerable Controlled Markov Chains with Strong Average Optimality Criterion: Bounded & Unbounded Costs," in *Proc. 33rd IEEE Conference on Decision and Control,* Orlando, FL, (1994) 1456-1461. Also to appear in **ZOR**: *Methods and Models in Operations Research.*

[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.

[FEM1] E. Fernández-Gaucherand and S.I.Marcus, Risk-Sensitive Optimal Control of Hidden Markov Models: A Case Study, in *Proc. 33rd IEEE Conference on Decision and Control,* Orlando, FL, (1994) 1657-1662.

[FEM2] E. Fernández-Gaucherand and S.I.Marcus, Risk-Sensitive Optimal Control of Hidden Markov Models: Structural Results. University of Maryland at College Park, Institute for Systems Research Report, 1995.

[FGM] E. Fernández-Gaucherand, M.K. Ghosh and S.I. Marcus, Controlled Markov Processes on the Infinite Planning Horizon: Weighted and Overtaking Cost Criteria, **ZOR**: *Methods and Models in Operations Research* **39** (1994) 131-155.

[GER] S.B. Gershwin, *Manufacturing Systems Engineering*, Prentice-Hall, 1994.

[GM] M.K. Ghosh and S.I. Marcus, On Strong Average Optimality of Markov Decision Processes with Unbounded Costs, *Operat. Res. Lett.* **11** (1992) 99-104.

[HO] Y.C. Ho, Dynamics of Discrete Event Systems, *Proc. IEEE,* **77** (1989) 3-6.

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

[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.

[PUT] M.L. Puterman, *Markov Decision Processes,* Wiley, New York, 1994.

[SEN] L.I. Sennott (1989) Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. Oper. Res. 37:626–633.

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