Chapter for
MARKOV DECISION PROCESSES
Models, Methods, Directions, and Open Problems

written by

Eugene A. Feinberg
Department of Applied Mathematics and Statistics
SUNY at Stony Brook
Stony Brook, 11794-3300, NY, USA

Adam Shwartz
Electrical Engineering
Technion—Israel Institute of Technology
Haifa 32000, Israel

May 28, 2000

Chapter 1

Mixed criteria

ch:WDC

Abstract

Mixed criteria are linear combinations of standard criteria which cannot be represented as standard criteria. Linear combinations of total discounted and average rewards as well as linear combinations of total discounted rewards are examples of mixed criteria. We discuss the structure of optimal policies and algorithms for their computation for problems with and without constraints.

1.1 Introduction

s:WDC-intro

The discounted cost criterion is widely and successfully used in various application areas. When modeling economic phenomena, the discount factor is determined by return rate (or interest rate) or, in a more general context, by the "opportunity costs" which presume that a dollar now is worth more that a dollar in a year. Being invested, current funds will bring an additional return in a year. In other areas, such as the control of communications networks, the discounted criterion may reflect the imprecise but fundamental principle that future occurrences are less important than immediate ones. In reliability, discounting models systems with geometric life time distributions.

Obviously, if some part of the cost decreases (in time) at an exponential rate, then a discounted cost arises. This is the case, for example, in production processes. When a new item is manufactured, we expect some production costs to decrease as production methods are improved. Obviously there is a learning curve for all involved, along which various costs decrease. If the effect of this learning diminishes geometrically (or exponentially), then the total cost (over the infinite time-horizon) is of the discounted type. This would also be the case if the cost of obtaining a component decreases at an exponential rate. Such an exponential decrease is evident in the computers industry (and one aspect goes by "Moore's law"): prices of various components decrease at an exponential rate, and since the processing speed increases at an exponential rate, the "unit cost" of processing power decreases exponentially fast.

However, the rates (or discount factors) for these different mechanisms are clearly unrelated. When combining several such costs (such as processing speed with economic considerations), we are naturally led to deal with several different discount factors, each applicable to a component of our optimization problem. Multiple discount factors also arise in control problems for systems with multiple parallel unreliable components. For details on applications see Feinberg and Shwartz FeS94, FeS95, FeS99 [14, 15, 17]. Similarly, in stochastic games it is natural to consider situations where each player has its own discount factor.

In contrast to the discounted criterion, the average cost measures long-term behavior, and de-emphasizes present conditions. Naturally, this criterion is appropriate for other applications, such as the long-term performance of systems. As before, if our criteria include system performance (measured through the average cost) as well as rewards (measured through a discounted cost), we are led to problems with mixed criteria.

In this paper we review results concerning such criteria, and point out some open questions. We shall mention specific, as well as general areas of potential applications. The main theoretical questions are:

- Existence of good (optimal, ε-optimal) policies for optimization, multi-objective optimization and in particular constrained optimization problems,
- Structure of "good policies," and
- Computational schemes for the calculation of the value and good policies.

We emphasize the mixed-discounted problem, where the criteria are all of the discounted type, but with several discount factors, since it possesses a rich structure and, in addition, much is already known. In Section 1.7 we shall review some other related results. We conclude this section with a brief survey of different mixed criterion problems.

There are several treatments of discounted models where the discounting is more general than the standard one. Hinderer [24] investigates a general model where the discounting is a function of the complete history of the model. In a number of papers, see e.g. Schäl [38] or Chitashvili [7], the discount rate depends on the current state and action. This type of discounting arises when a discounted semi-Markov Decision process is converted into an MDP; see Puterman [34] or Feinberg [13].

Mixed criteria and, in particular, mixed discounted criteria are linear combinations of standard criteria. The first two papers dealing with mixed criteria were published in 1982. Golabi, Kulkarni, and Way [21] considered a mixture of average reward and total discounted criteria to manage a statewide pavement system, see also [39]. Feinberg [10] proved for various standard criteria, by using a convex-analysis approach, that for any policy there exists a nonrandomized Markov policy with the same or better performance. In the same paper, Feinberg [10], proved that property for mixtures of various criteria.

The 1990's saw systematic interest in criteria that mix several standard costs. Krass, Filar and Sinha [26] studied a sum of a standard average cost and a standard discounted cost for models with finite state and action sets. They proved that for any $\varepsilon > 0$ there exists an ε-optimal randomized Markov policy which is stationary from some epoch N-onwards (so-called ultimately deterministic or (N, ∞)-stationary policies). They also provided an algorithm to compute such policies. As explained in Feinberg [11], the use of results from [10] simplifies the proofs in [26] and leads to the direct proof that for any $\varepsilon > 0$ there exists an ε-optimal (nonrandomized) Markov ultimately deterministic policy. The latter result can be derived from [26] but it was not formulated there.

Fernandez-Gaucherand, Ghosh and Marcus [18] considered several weighted as well as overtaking cost criteria. Ghosh and Marcus [20] considered a similar problem in the context of a continuous time diffusion model. Filar and Vrieze [19] considered a stochastic zero-sum game, and obtain the existence of ε-optimal policies.

Most of the papers on mixed criteria deal either with linear combinations of total discounted rewards with different discount factors or with a linear combination of total discounted rewards and average rewards per unit time. It appears that linear combinations of discounted rewards are more natural and easier to deal with than the weighted combinations of discounted and average rewards. The latter ones model the situation when there are two goals: a short-term goal modeled by total discounted rewards and a long-term goal modeled by average rewards per unit time. In the case of two different discount factors, the weighted discounted criterion models the same situation when one of the discount factors is close to 1. When the state and action sets are finite, optimal policies exist for mixed discounted criteria, they satisfy the Optimality Equation, and can be computed. Optimal policies may not exist for mixtures of total discounted rewards and average rewards per unit time [26]. Another advantage of dealing with mixed discounting is that for this criterion there is a well-developed theory for any finite number of discount factors [14, 15], while the papers that study linear combinations of discounted and average reward criteria usually deal with linear combinations of only two criteria: discounted rewards with a fixed discount factor and average rewards per unit time.

In this paper we concentrate on a mixed-discounted problem for Markov decision processes. The exposition is based on the detailed study of this problem, performed by Feinberg and Shwartz [14, 15, 17].

In Section 1.2 we describe the model more precisely, and show through example 1.4 that, although the weighted criterion seems like a small variation on a standard discounted problem, it induces quite different behavior on the resulting optimal policies. Then, in Section 1.3 we show that mixed-discounted problems can be reduced to standard discounted problems if we expand the state space $\mathbb{X}$ to $\mathbb{X} \times \mathbb{N}$, and that Markov policies are sufficient for one-criterion mixed-discounted problems. Section 1.4 obtains the characterization as well as an algorithm for the computation of optimal policies for the finite (state and action set) Weighted Discounts Optimization (**WDO**) problem 1.1. In Section 1.5 we treat multiple-criterion and in Section 1.6 we discuss finite constrained problems. In Section 1.7 we survey existing results for other relevant problems, related models of stochastic games, and discuss extensions and open problems.

1.2 The mixed discounted problem

s:WDC-model

Throughout this chapter we deal with a discrete state model, as described in the Introduction, and follow notation introduced there. We fix L discount factors which, for convenience (and without loss of generality) we order as $1 > \beta_1 > \ldots > \beta_L > 0$, and $(K+1) \times L$ one-step reward functions $r_\ell^k,\ k = 0, \ldots, K,\ \ell = 1, \ldots, L$. We assume that all reward functions are bounded above. The index k will be used only for multi-objective problems, and is omitted otherwise. Let $v_\ell^k(x, \pi, \beta_\ell)$ denote the standard total expect discounted cost corresponding to discount factor β_ℓ and to the immediate reward r_ℓ^k. We formulate the optimization problems:

d:WDC-WDO **Definition 1.1** *The Weighted-Discount Optimization Problem* **WDO** *is to maximize the weighted-discount cost $v_{\mathcal{M}}$ over all policies $\pi \in \Pi^R$, where*

$$v_{\mathcal{M}}(x, \pi) \stackrel{def}{=} \sum_{\ell=1}^{L} v_\ell(x, \pi, \beta_\ell)\,. \tag{1.1}$$ e:WDC-defWDO

d:WDC-WDC **Definition 1.2** *The Constrained Weighted-Discount Optimization Problem* **WDC** *is the constrained optimization over all policies $\pi \in \Pi^R$*

$$\textit{maximize} \qquad v_{\mathcal{M}}^0(x, \pi) \stackrel{def}{=} \sum_{\ell=1}^{L} v_\ell^0(x, \pi, \beta_\ell) \tag{1.2}$$ e:WDC-defWDCm

$$\textit{subject to} \qquad v_{\mathcal{M}}^k(x, \pi) \stackrel{def}{=} \sum_{\ell=1}^{L} v_\ell^k(x, \pi, \beta_\ell) \ge C_k, \qquad k = 1, \ldots, K. \tag{1.3}$$ e:WDC-defWDCc

A policy π is feasible *if the constraints (1.3) are satisfied.*

Given any numbers $a_1, a_2, \ldots, a_q$ we use the notation $\bar{a} \stackrel{def}{=} (a_1, a_2, \ldots, a_q)$. In particular we define

d:WDC-Vec **Definition 1.3** *The* performance vectors *associated with problems* **WDO** *and* **WDC** *respectively are*

$$\bar{v}(x, \pi) \stackrel{def}{=} (v_1(x, \pi, \beta_1), \ldots, v_L(x, \pi, \beta_L))\,, \tag{1.4}$$

$$\bar{v}_{\mathcal{M}}(x, \pi) \stackrel{def}{=} \left(v_{\mathcal{M}}^0(x, \pi), \ldots, v_{\mathcal{M}}^K(x, \pi)\right)\,. \tag{1.5}$$

As was discussed in the Introduction to this book, for unconstrained problems in general and for problem **WDO** in particular, we consider optimality with respect to all initial states. For a constrained problem, including problem **WDC**, an initial state is fixed.

ss:WDC-x

1.2.1 An Example: job versus education dilemma

The use of different discount criteria induces a time-dependence on the model since the relative impact of different immediate costs changes over time. This implies, for example, that we cannot expect stationary policies to be optimal. This can be demonstrated with a very simple model. This negative result suggests that we search for different structural properties of optimal policies. It turns out that the structure suggested here is useful for other criteria as well, including single-discount constrained problems [16].

Example 1.4 illustrates the following dilemma. A person, say a high school or college graduate, has a choice: to accept a job offer or to continue his/her education and get a better job later. In a standard discounted model, stationary policies are optimal. Therefore, for standard models an optimal decision is either to accept a job or to continue education. For weighted discounted models, an optimal decision can suggest to accept a job for a limited period of time and then to continue education. This phenomenon cannot be modeled by standard discounted or average-reward criteria.

x:WDC-x

Example 1.4 ([14, Example 1.1]) *Consider the optimization problem* **WDO**. *Let $\mathbb{X} = \{x, y\}$ with deterministic transitions: under a we always go to state x, while under b we always go to state y. Set $r_\ell^k = r$, where*

$$r(x,a) = 1\,, \qquad r(y,a) = r(x,b) = 0\,, \quad \textit{and} \quad r(y,b) = 2\,. \tag{1.6}$$

It is then easy to calculate that for the standard discounted cost, if $\beta \le \frac{1}{2}$, then it is optimal to stay where you are, while for $\beta \ge \frac{1}{2}$ it is optimal to use only action b.
For the weighted problem with $\beta_1 = \frac{3}{5}$ and $\beta_2 = \frac{1}{5}$, an explicit calculation shows that the only optimal policy is to stay where you are at time 0, and use b at any later time.

Another illuminating conclusion from the same example is obtained by searching for the best stationary policy. This turns out to be a randomized one! Related examples [14, Examples 1.2–1.3] show that the best stationary (non-randomized) policy may depend on the initial state. In addition, this behavior can be observed in ergodic models.

Thus, it seems that much of the basic structure of MDPs is lost when mixed discounting is used. However, it turns out that a different structure arises. In fact, that the optimal policy in Example x:WDC-x 1.4 is Markov and becomes stationary after some initial period is a structure we shall discover in the following sections.

1.3 General properties

s:WDC-G

The first task in our search for optimal policies usually entails restricting attention to Markov policies. This can be justified under general circumstances by the following general result.

t:WDC-Markov **Theorem 1.5** (DS,Ho **[8, 23]**) *Let $\pi^1, \pi^2, \ldots$ be an arbitrary sequence of policies and $\lambda_1, \lambda_2, \ldots$ a sequence of positive numbers summing to 1. Fix an initial state x and define the randomized Markov policy π through (e:WDC-Markov 1.7) (if the denominator is 0 then choose π_t arbitrarily):*

$$\pi_t\left(C \mid y\right) \stackrel{def}{=} \frac{\sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i}\left(x_t = y, a_t \in C\right)}{\sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i}\left(x_t = y\right)}\,, \quad t \geq 0, \; y \in \mathbb{X}, \tag{1.7}$$

e:WDC-Markov

for all measurable subsets C of $\mathbb{A}(y)$. Then, for all $t \geq 0$, $y \in \mathbb{X}$ and measurable subsets C of $\mathbb{A}(y)$,

$$\mathbb{P}_x^{\pi}\left(x_t = y, a_t \in C\right) = \sum_{i=1}^{\infty} \lambda_i \, \mathbb{P}_x^{\pi^i}\left(x_t = y, a_t \in C\right). \tag{1.8}$$

e:WDC-MarkovP

In particular, setting $\lambda_1 = 1$ and $\lambda_i = 0, \; i > 1$, we find that for any given policy and initial state, we can find a Markov policy that produces the same one-dimensional distributions for the pair (x_t, a_t). Consequently, for any criterion depending only on such distributions (and in particular, for any linear combination of total and average costs), Markov policies suffice. Note that, in this generality, the Markov policy depends on the initial state.

The main difficulty with the mixed-discounted problem is that the immediate cost changes over time. Another technique of general applicability is the embedding of the problem into a larger one. Consider an auxiliary standard discounted model with the discount factor β_1 and with the state space $\mathbb{X} \times \mathbb{N}$. The one-step rewards in this model are equal to

$$r(x, n, a) \stackrel{def}{=} r_\ell(x, a) + \sum_{\ell=2}^{L} \left(\frac{\beta_\ell}{\beta_1}\right)^n r_\ell(x, a)\,. \tag{1.9}$$

e:WDC-AuxRew

If we then keep the same transition probabilities (but require a transition of one unit in the time component at each step), then we have a one-to-one correspondence between the original problem and the auxiliary problem started at $(x, 0)$. There is only one immediate reward and one discount factor in the larger model. The state space for the auxiliary problem remains countable. We therefore obtain immediately the following results [14, Section II].

t:WDC-embed

Theorem 1.6 ([14, Theorems 2.1–2.2]) *(i) For any $\varepsilon > 0$, the* **WDO** *problem possesses ε-optimal Markov policies. (ii) If the $\mathbb{A}(x)$ are compact subsets of a metric space, r_ℓ are upper semi-continuous and the $p(y|x,\cdot)$ are continuous in a then there are optimal Markov policies.*

Proof outline. This follows from the properties of the auxiliary problem. Note that a stationary policy for the auxiliary problem defines a Markov (but not necessarily stationary!) policy for the original problem. ∎

The above construction transforms a mixed-discounted problem with a finite or countable state spaces into a standard discounted problem with a countable state space. If the original problem has an uncountable Borel state space, so does the expanded problem.

We end this section with some insight into why problems with mixed criteria are inherently more difficult. For simplicity, let the state and action sets be finite. Define the expected occupation vectors

$$f(\beta; x, \pi; y, a) \stackrel{def}{=} \sum_{t=0}^{\infty} \beta^t \, \mathbb{P}_x^\pi \left(x_t = y, a_t = a \big| x_0 = x \right) . \tag{1.10}$$

e:WDC-Freq

Then we can write the standard discounted cost as

$$v_\ell^k(x, \pi) = \sum_{y \in \mathbb{X}} \sum_{a \in \mathbb{A}(y)} f(\beta; x, \pi; y, a) r_\ell^k(y, a) \, . \tag{1.11}$$

That is, any discounted cost is a linear function of the occupation vectors $\{f(\beta; x, \pi; y, a) : y \in \mathbb{X}, \ a \in \mathbb{A}(y)\}$. Moreover, these occupation vectors obey a system of linear equalities, in terms of transition probabilities. It is therefor possible to transform the optimization problem, as well as the constrained optimization problem, into a linear program. This linear program is finite if the state and action spaces are finite. This approach is described in many papers and in the books by Kallenberg [25], Borkar [6], Piunovskiy [33], Altman [1], and Hernadez-Lerma and Lasserre [22]. However, the relation between $\{f(\beta_1; x, \pi; y, a) : y \in \mathbb{X}, \ a \in \mathbb{A}(y)\}$, the occupation vectors associated with discount β_1, and $\{f(\beta_2; x, \pi; y, a) : y \in \mathbb{X}, \ a \in \mathbb{A}(y)\}$, the occupation vectors associated with discount β_2, is non-linear (in fact, it is even non convex!). This makes the tools of mathematical programming much more difficult to apply; see Fe97 [12].

1.4 Single Criterion Models

s:WDC-WF

Motivated by the structure we found in Example x:WDC-x 1.4, we introduce some notions which will be fundamental beyond this section. The lost time-homogeneity of the model is partly recovered by the notion of a funnel.

d:WDC-SubM **Definition 1.7** *Given measurable subsets $\mathbb{A}_1(x) \subset \mathbb{A}(x),\ x \in \mathbb{X}$, the* submodel $\mathbb{A}_1$ *is the Markov decision process, where the actions at x are restricted to $\mathbb{A}_1(x)$.*

d:WDC-funnel **Definition 1.8** *Fix a positive integer N and subsets $\mathbb{A}_n(x) \subset \mathbb{A}(x),\ n \geq 0,\ x \in \mathbb{X}$, with the property that $\mathbb{A}_n(x) = \mathbb{A}_N(x),\ n \geq N,\ x \in \mathbb{X}$. The* funnel *associated with these data is the set of all randomized Markov policies π such that $\pi_n(\mathbb{A}_n(x)|x) = 1$ for all $n \geq 0$ and $x \in \mathbb{X}$.*

A funnel is thus defined by the number $N \in \mathbb{N}$ and sets $\mathbb{A}_n(x),\ n = 0, 1, \ldots, N,\ x \in \mathbb{X}$.

d:WDC-Ninfty **Definition 1.9** *Given a positive integer N, a Markov policy π is called (N, ∞)-stationary if there exists a stationary policy ϕ so that*

$$\pi_n(x) = \phi(x) \quad \textit{for all} \quad x \quad \textit{and} \quad n \geq N\ . \tag{1.12}$$

This generalizes the notion of stationarity, since obviously a $(0, \infty)$-stationary policy is stationary. Let each set $\mathbb{A}(x)$ be finite and all functions $r_\ell(x, a)$ be bounded on $\mathbb{X} \times \mathbb{A}$, $\ell = 1, \ldots, L$. Define recursively, for $\ell = 1, \ldots, L$,

$$\Gamma_0(x) = \mathbb{A}(x)\,, \quad N_0(x) = N_0 = 0, \tag{1.13}$$ First

$$d_\ell \stackrel{def}{=} \frac{\max_{x\in\mathbb{X}, a\in\Gamma_{\ell-1}(x)} r_\ell(x,a) - \min_{x\in\mathbb{X}, a\in\Gamma_{\ell-1}(x)} r_\ell(x,a)}{1-\beta_\ell}\,, \tag{1.14}$$ e:WDC-CostSpan

$$V_\ell(x) = \max\{v_\ell(x, \pi, \beta_\ell)\,|\,\pi \text{ in submodel } \Gamma_{\ell-1}\}\ , \tag{1.15}$$

$$\Gamma_\ell(x) = \{a \in \Gamma_{\ell-1}(x) : V_\ell(x) = r_\ell(x,a) + \beta_\ell P^a V_\ell(x)\}\ . \tag{1.16}$$ e:WDC-Dconserve

If $\Gamma_\ell(x) = \Gamma_{\ell-1}(x)$ set $N_\ell(x) = N_{\ell-1}(x)$. Otherwise define

$$\varepsilon_\ell(x) = V_\ell(x) - \max\left\{r_\ell(x,a) + \beta_\ell P^a V_\ell(x) : a \in \Gamma_{\ell-1}(x) \setminus \Gamma_\ell(x)\right\}\,, \tag{1.17}$$

e:WDC-OEgapK

$$N_\ell(x) = \min\left\{ t \geq N_{\ell-1}(x) | \sum_{j=\ell+1}^{L} \left(\frac{\beta_j}{\beta_\ell}\right)^t d_j < \varepsilon_\ell(x) \right\}, \quad \ell < L\,, \tag{1.18}$$

e:WDC-N1xK

$$N_\ell = \max_x N_\ell(x), \tag{1.19}$$

e:WDC-N1K

and set $N_L(x) \stackrel{def}{=} N_{L-1}(x)$, $N = N_L = N_{L-1}$. If the state space is finite, $N < \infty$.

The set $\Gamma_\ell(x)$ is the set of *conserving* actions for the discounted criterion $v_\ell(x, \pi, \beta_\ell)$ in submodel $\Gamma_{\ell-1}$. The basic structure of optimal policies derives from the following statement which follows from equalizing and thrifty properties; see Chapter ... "Total reward criteria."

l:WDC-DiscOE **Lemma 1.10** *A policy π is optimal for the criterion $v_\ell(x, \pi, \beta_\ell)$ in submodel Γ_ℓ, namely $v_\ell(x, \pi, \beta_\ell) = V_\ell(x)$ for all $x \in \mathbb{X}$, if and only if $a_t \in \Gamma_\ell(x_t)$ $\mathbb{P}_x^\pi$-a.s. for all $t = 0, 1, \ldots$ and for all $x \in \mathbb{X}$.*

In other words, all policies in submodel Γ_ℓ have the same v_ℓ-cost, and the value $V_\ell(x)$ is the optimal value in the $\Gamma_{\ell-1}$ model.

t:WDC-fun1 **Theorem 1.11** *Suppose that all action sets $\mathbb{A}(x)$ are finite and all reward functions r_ℓ are bounded. Consider problem* **WDO**. *If π is an optimal policy, then*

$$\pi_t(x_t) \in \Gamma_\ell(x_t) \quad \mathbb{P}_x^\pi - a.s.\ \textit{for any}\ t \geq N_\ell(x_t),\ \textit{for all}\ \ell,\ \textit{and for all}\ x \in \mathbb{X}\,. \tag{1.20}$$

Proof outline. Write

$$v_{\mathcal{M}}(x, \pi) = v_1(x, \pi, \beta_1) + \sum_{\ell=2}^{L} v_\ell(x, \pi, \beta_\ell)\,. \tag{1.21}$$

e:twoc

Suppose at time t we are at state y. By Lemma 1.10, if we choose an action outside $\Gamma_1(y)$, then our v_1 reward will be smaller by at least $\beta_1^t \varepsilon_1(y)$. On the other hand, in view of (1.9), the second summand in (1.21) can be made larger by at most

$$\sum_{j=2}^{L} \beta_j^t d_j\,. \tag{1.22}$$

The result for $\ell = 1$ follows from the definitions (1.17)–(1.18) since $\beta_j < \beta_1$ for $j > 1$. Therefore we know that, after time N_1, we must restrict to submodel Γ_1. By Lemma 1.10, the v_1 cost is the same for all policies in this model, so that this component of the cost may be ignored. Repeating the same argument establishes the result for $\ell = 2, \dots, L$. ∎

Theorem 1.11 is formulated for non-finite state spaces. However, in the finite case N is finite and it leads directly to existence as well as to an algorithm. Define the time-dependent immediate reward

$$r(t,x,a) \stackrel{def}{=} \sum_{\ell=1}^{L} \beta_\ell^t r_\ell(x,a)\,, \tag{1.23}$$

and the "tail reward"

$$v_\ell^{>}(x,N,y,\pi) \stackrel{def}{=} \mathbb{E}_x^\pi \left[\sum_{t=N}^{\infty} \beta_\ell^t r_\ell(x_t,a_t) \middle| x_N = y \right]. \tag{1.24}$$

If π is a Markov policy, then $v_\ell^{>}(x,N,y,\pi)$ does not depend on x.

t:WDC-FinWeight **Theorem 1.12** *Consider problem* **WDO** *where the state and action spaces are finite. Let Θ be the funnel defined by $\mathbb{A}_t(x) = \Gamma_\ell(x)$ for $N_{\ell-1}(x) \le t < N_\ell(x)$, and $\mathbb{A}_t(x) = \Gamma_L(x)$ for $t \ge N$. Let ϕ be any stationary policy in submodel Γ_L. Then (i) any optimal policy must satisfy $v_\ell^{>}(x,N,y,\pi) = v_\ell^{>}(x,N,y,\phi)$ for all x and all y such that $\mathbb{P}_x^\pi(x_N = y) > 0$; (ii) an optimal policy to* **WDO** *can be constructed as follows. Let $\pi^N = \{\pi_0, \pi_1, \dots, \pi_{N-1}\}$ solve the finite-horizon total cost problem with horizon N, immediate rewards $r(t,x,a)$, and terminal reward*

$$R(y) \stackrel{def}{=} \sum_{\ell=1}^{L} v_\ell^{>}(x,N,y,\phi)\,. \tag{1.25}$$

Then $\pi^* = (\pi_0, \pi_1, \dots, \pi_{N-1}, \phi, \phi, \dots)$ is optimal.

Proof outline. Part (i) follows from Theorem 1.11. This determines the "tail reward," and it remains to optimize over the finite horizon. ∎

The "tail" policy ϕ can be chosen stationary, and from the algorithm it follows that it does not depend on the initial state. Since these properties are also shared by solutions to finite-horizon problems, we obtain the following.

Corollary 1.13 *For the finite* **WDO** *problem there is an optimal (N,∞)-stationary policy (which does not depend on the initial state).*

Formulas (1.13)–(1.19) provide an algorithm that computes an integer N and sets $\Gamma_L(x)$, $x \in \mathbb{X}$, described in Theorem 1.12. In view of Theorem 1.12, the "tail" stationary policy ϕ can be selected as any stationary policy from submodel Γ_L. Theorem 1.12 also implies that, at steps $0, 1, \cdots, N-1$, an optimal policy can be constructed by a finite-horizon dynamic programming algorithm. Thus, formulas (1.13)–(1.19) and Theorem 1.12 provide an algorithm that computes an optimal (N, ∞)-stationary policy.

This algorithm requires the computation of at most L standard discounted problems, and then the solution of a finite horizon problem. The computational complexity is obviously influenced by the size of N. This in turn in determined by the data of our problem (and in particular the ratios $\beta_\ell/\beta_{\ell-1}$), as well as by the choice of bound (e.g. (1.14)). A more complex algorithm leading to a smaller value of N is in [14].

For any criterion $v(x,\pi)$ and set of policies Δ we denote

$$v(x,\Delta) = \{v(x,\pi) : \pi \in \Delta\} \ , \tag{1.26}$$

$$V(x,\Delta) = \sup \{v(x,\pi) : \pi \in \Delta\} \ , \tag{1.27}$$

$$\Delta_v^*(x) = \{\pi \in \Delta : v(x,\pi) = V(x,\Delta)\} \ . \tag{1.28}$$

Let Θ be a funnel. The embedding technique of Section 1.3 allows us to construct a finite model, where the time becomes part of the new state space, but only until time N. A funnel in this new model corresponds to a funnel in the original MDP. Therefore Theorem 1.12 implies the following result.

c:WDC-fun2fun **Corollary 1.14 ([15, Lemma 5.5])** *Consider an MDP with finite state and action sets. Fix x and let $\Theta \subset \Pi^R$ be a non-empty funnel. Then there exists a funnel Θ' so that $v_{\mathcal{M}}(x,\pi) = V_{\mathcal{M}}(x,\Theta)$ for all $\pi \in \Theta'$, and moreover $\bar{v}(x,\Theta') = \bar{v}(x,\Delta^*_{v_{\mathcal{M}}}(x))$.*

1.5 Multiple Criterion Optimization

s:WDC-MC

To describe some notions of optimality in the multiple criterion setting we need some definitions and notation. Recall the definition 1.3 of the *performance vectors* associated with problems **WDO** and **WDC** respectively.

Definition 1.15 *The* performance spaces *are, respectively,*

$$U_o(x) \stackrel{def}{=} \left\{\bar{v}(x,\pi) : \pi \in \Pi^R\right\} \ , \tag{1.29}$$

$$U_c(x) \stackrel{def}{=} \left\{\bar{v}_{\mathcal{M}}(x,\pi) : \pi \in \Pi^R\right\} \ . \tag{1.30}$$

For a vector $\bar{a}$ in $\mathbb{R}^q$ we write $\bar{a} \geq 0$ if and only if $a_i \geq 0$ for all i.

d:WDC-muC **Definition 1.16** *A point $\bar{u}$* dominates *a point $\bar{v}$ if $\bar{u} - \bar{v} \geq 0$. A point $\bar{u}$ is* Pareto optimal *in a set U if there is no other $\bar{v} \in U$ which dominates $\bar{u}$. We write $(u_1, u_2, \dots, u_q) = \bar{u} >_{\mathcal{L}} 0$ if $u_i = 0$ for $i = 1, \dots, j-1$ and $u_j > 0$ for some $1 \leq j \leq q$ ($j = 1$ implies $u_1 > 0$). Say $\bar{u}$ is* lexicographically larger *than $\bar{v}$ if $\bar{u} - \bar{v} >_{\mathcal{L}} 0$.*

Note that $\bar{u} - \bar{v} \geq 0$ and $\bar{u} \neq \bar{v}$ implies $\bar{u} - \bar{v} >_{\mathcal{L}} 0$, but the converse need not hold. We extend these notions from vectors to policies in the obvious way:

d:WDC-MulOptPol **Definition 1.17** *For a fixed initial state x, a policy π is called* Pareto optimal *if the corresponding performance vector is Pareto optimal. A policy π is called* lexicographically optimal *if the corresponding performance vector is lexicographically optimal.*

With these definitions, we have the following obvious statement.

c:WDC-Pareto **Lemma 1.18** *Any optimal policy for* **WDO** *is Pareto optimal for $\bar{v}(x, \pi)$.*

Theorem 1.12 immediately implies

c:WDC-lex **Corollary 1.19** *Any policy in submodel Γ_L, and in particular ϕ, is lexicographically optimal for $\bar{v}(x, \pi)$.*

The performance space has the following convenient structure.

t:WDC-PerSp **Theorem 1.20** *(i) The set $U_c(x)$ are convex, $x \in \mathbb{X}$. (ii) If the $\mathbb{A}(x)$ are compact, all reward and transition functions are continuous in a and the rewards are bounded, then $U_c(x)$ are compact. (iii) If Θ is a funnel then $\bar{v}_{\mathcal{M}}(x, \Theta)$ are convex and if, in addition, the conditions of (ii) hold then $\bar{v}_{\mathcal{M}}(x, \Theta)$ are compact.*

Proof outline. Convexity of $U_c(x)$ follows from theorem 1.5. Compactness follows from compactness of $\left\{\mathbb{P}_x^\pi : \pi \in \Pi^R\right\}$ and continuity; see [15, Lemma 3.5], which is based on [38]. The extension to $\bar{v}_{\mathcal{M}}(x, \Theta)$ follows by the argument preceding Corollary 1.14. ∎

1.5.1 Classes of policies

Multiple criterion problems often require randomization in order to achieve optimality. In order to quantify the amount of randomization, and to tie this with the notion of (N, ∞)-stationarity, we introduce the following classes of policies.

We say that a randomized Markov policy π is *discrete* if all probabilities $\pi_t(\cdot|x)$ are discrete, $t \in \mathbb{N}$, $x \in \mathbb{X}$. We recall that any randomized stationary policy is randomized Markov.

d:WDC-policies **Definition 1.21** *A randomized stationary policy ϕ is called* M-randomized stationary *if it is discrete and*

$$\sum_{x \in \mathbb{X}} \left[\sum_{a \in \mathbb{A}(x)} \mathbf{1}[\phi(a|x) > 0] - 1 \right] = M\,. \tag{1.31}$$

e:WDC-Mrand

A randomized Markov policy π is called randomized Markov of order M *if it is discrete and*

$$\sum_{t=0}^{\infty} \sum_{x \in \mathbb{X}} \left[\sum_{a \in \mathbb{A}(x)} \mathbf{1}[\pi_t(a|x) > 0] - 1 \right] = M\,. \tag{1.32}$$

e:WDC-randOrdM

Note that the terms in square brackets are always non-negative. An M-randomized stationary policy randomizes every time when the process reaches a state where the support of $\phi(x)$ contains more than one point. By contrast, a randomized Markov policy of order M makes at most M randomizations over the entire time-horizon.

Definition 1.22 *A Markov policy π is called an* (m, N)-policy *if it is randomized Markov of order m and, in addition, it is (N, ∞)-stationary, that is, it agrees with some stationary policy ϕ after time N. An (m, N)-policy π is called* a strong (m, N)*-policy if, in addition, there is an m-randomized stationary policy ψ such that*

$$\pi_t(a|x) > 0 \quad \textit{implies} \quad \psi(a|x) > 0\,. \tag{1.33}$$

Thus (m, N)-policies have the simple structure of only m randomizations over the entire time horizon, as well as stationarity beyond time N. For a strong (m, N)-policy, the total number of actions is further restricted in that the total number of actions beyond those of a stationary (non-randomized!) policy does not exceed m.

1.6 Finite Models: Constrained Optimization

s:WDC-CWF

1.6.1 Finite horizon models

As we saw, construction of optimal policy in the weighted-discount problem goes through the computation of finite-horizon problems. We note in passing that Theorem t:WDC-PerSp 1.20 applies to finite-horizon problems with arbitrary time-dependence of the reward functions.

To define the finite horizon constrained optimization problem, we use Definition d:WDC-WDC 1.2. However, we let the immediate rewards depend on time by setting

$$r_\ell^k(t,x,a) = \begin{cases} r_\ell^k(x,a) & \text{if } t < N, \\ f_\ell^k(x) & \text{if } t = N, \\ 0 & \text{if } t > N, \end{cases} \tag{1.34}$$ e:WDC-FinHorCost

where $f_\ell^k(x)$ is a terminal reward for the reward function r_ℓ^k and discount factor β_ℓ. We usually set $f_\ell^k(x) = v_\ell^k(x, \phi, \beta_\ell)$, where ϕ is a stationary policy. For the finite horizon problem, it is possible to use a Linear Programming approach. Here, the mixed criteria is not a hindrance: if fact, this approach applies for a time-dependent cost structure, by an embedding technique as in Section s:WDC-G 1.3.

t:WDC-finHorLP **Theorem 1.23 (FeS95 [15, Theorem 4.1])** *The finite horizon, finite state and action constrained optimization problem is feasible if an only if the associated LP FeS95 [15, (4.1)–(4.5)] is feasible. If it is feasible then it has an optimal randomized Markov policy of order K.*

1.6.2 Infinite horizon models

The proof of the existence of optimal (K, N)-policies requires some convex analysis. We shall relate special subsets of performance spaces to funnels. We need the following definition.

Definition 1.24 *Let W be a convex subset of a convex set E. Call W* extreme *if the relation $u_3 = \lambda u_1 + (1-\lambda)u_2$, where $0 < \lambda < 1$, $u_1, u_2 \in E$, and $u_3 \in W$ implies that necessarily $u_1, u_2 \in W$. Call W* exposed *if there is a supporting hyperplane H of E so that $W = H \cap E$.*

An exposed set is extreme, but the converse may not hold: take

$$E = \{(x,y) : -1 \le x \le 0, |y| \le 1\} \cup \{(x,y) : x \ge 0, x^2 + y^2 \le 1\}$$

and $W = \{(0,1)\}$. Then W is obviously extreme, but the only supporting hyperplane containing W satisfies $H \cap E = \{(x,y) : -1 \le x \le 0, y = 1\} \ne W$, so that W is not exposed. Note that $(0,1)$ is a Pareto-optimal point, and is a solution of the constrained optimization problem of maximizing y subject to $x \ge 0$. However, is not isolated by any exposed subset of E.

Our plan is to show that Pareto optimal points of $U_c(x)$ are achieved by (K,N)-policies. We first show that boundary points of $U_c(x)$ are achieved by points of the set $\bar{v}_{\mathcal{M}}(x,\Theta)$ for some funnel Θ. We then show that any boundary point is achieved by a convex combinations of performances of (N,∞)-stationary policies that utilize the same stationary policy from epoch N onwards. The properties of the finite-horizon problems of Theorem t:WDC-finHorLP 1.23 are used to conclude optimality of (K,N)-policies.

t:FUN **Theorem 1.25** *Let Θ be a funnel and W an exposed subset of $\bar{v}_{\mathcal{M}}(x,\Theta)$. Then there exists a funnel Θ' such that $W = \bar{v}_{\mathcal{M}}(x,\Theta')$. If $E \ne \bar{v}_{\mathcal{M}}(x,\Theta)$ is an extreme subset, then then there exists a funnel Θ' such that $E = \bar{v}_{\mathcal{M}}(x,\Theta')$. In particular, these statements hold for $\Theta = \Pi^R$ and $U_c(x) = \bar{v}_{\mathcal{M}}(x,\Theta)$.*

Proof outline. A supporting hyperplane H and exposed subset W are defined by some $b, b_0, \dots, b_K$ so that

$$H = \left\{ \bar{u} : \sum_{k=0}^{K} b_i u_i = b \right\},$$
$$W = \left\{ \bar{u} \in \bar{v}_{\mathcal{M}}(x,\Theta) : \sum_{k=0}^{K} b_i u_i = b \right\},$$

and

$$\sum_{k=0}^{K} b_i u_i \le b \quad \text{for all} \quad \bar{u} \in \bar{v}_{\mathcal{M}}(x,\Theta).$$

Apply Corollary c:WDC-fun2fun 1.14 to conclude the first result. For the extreme subset we use the fact that for a proper extreme subset E of a compact convex set $\tilde{W}$ in an Euclidean space there is a finite sequence of sets $W_0, \dots, W_j$ such that $W_0 = \tilde{W}$, $W_j = E$, and W_{i+1} is an exposed subset of W_i; see the proof of Lemma 6.3 in FeS95 [15]. The first result, applied repeatedly to the sets $\tilde{W} = \bar{v}_{\mathcal{M}}(x,\Theta), W_1, \dots, W_{j-1}$, leads to the second statement of the theorem. ∎

When $\tilde{W} = U_c(x)$ or, in a more general situation, $\tilde{W} = \bar{v}_{\mathcal{M}}(x, \Theta)$, where Θ is a funnel, then Theorem 1.25 implies that for any proper extreme subset E of $\tilde{W}$ there is a funnel Θ' such that $E = \bar{v}_{\mathcal{M}}(x, \Theta')$. If $\bar{u}$ is an extreme point of some funnel $\bar{v}_{\mathcal{M}}(x, \Theta)$ (that is, the singleton $\{\bar{u}\}$ is an extreme subset), then its performance is achieved by a funnel, and in particular we can choose any policy π from this funnel and have $\bar{v}_{\mathcal{M}}(x, \pi) = \bar{u}$. For N large enough, we select π being (N, ∞)-stationary.

If $\bar{u}$ is a Pareto optimal point of $U_c(x)$ then it belongs to the boundary of the closed convex set $U_c(x)$. Therefore, it belongs to an exposed subset E of $U_c(x)$ and, according to Theorem 1.25, E can be represented as a performance set of a funnel; $E = V(x, \Theta')$. If $\bar{u}$ is an extreme point of E, $\bar{u} = \bar{v}_{\mathcal{M}}(x, \pi)$ for some (N, ∞)-stationary policy π. If $\bar{u}$ is a relatively inner point of E, Caratheodory theorem implies that it can be represented as a convex combination of at most K extreme points of E.

Such representation holds if each extreme point is approximated by an element of E close to it. By selecting N large enough, we can approximate (N, ∞)-stationary policies, whose performance vectors are extreme points of E, with (N, ∞)-stationary policies coinciding with the same stationary policy ϕ from epoch N onwards. Thus, if $\bar{u}$ is a Pareto optimal element of $U_c(x)$, it can be represented as a convex combination of performance vectors of (N, ∞)-stationary policies with the same "tail," i.e. these policies act as the same stationary policy from some epoch N onwards; see Figure 1.1, where a, b, and c are extreme points of $E = V(x, \Theta')$ and a', b', and c' are their approximations which are performance vectors of (N, ∞)-stationary policies with the same "tail."

t:WC-WDCkN **Theorem 1.26 ([15, Theorems 6.6–6.8])** *(i) If $\bar{u}$ is a Pareto optimal point of $U_c(x)$ then for some $N < \infty$ there exists a (K, N)-policy π such that $\bar{v}_{\mathcal{M}}(x, \pi) = \bar{u}$. (ii) If the* **WDC** *is feasible then for some $N < \infty$ there exists an optimal (K, N)-policy.*

Proof outline. (i) Fix N and the "tail" stationary policy ϕ described in the paragraph preceding the theorem. Set $\pi_t(y) = \phi(y)$ for all $y \in \mathbb{X}$ and for all $t \geq N$. In order to determine the policy π at steps $t = 0, \ldots, N$, one has to solve a constrained finite-horizon problem with $C_k = u^k$, $k = 1, \ldots, K$, where $\bar{u} = (u^0, \ldots, u^K)$. By Theorem 1.23, π is a (K, N)-policy.

(ii) Any solution of the **WDC** defines either a Pareto optimal point of $U_c(x)$ or it is dominated by a solution with this property. Therefore, (i) implies (ii). ∎

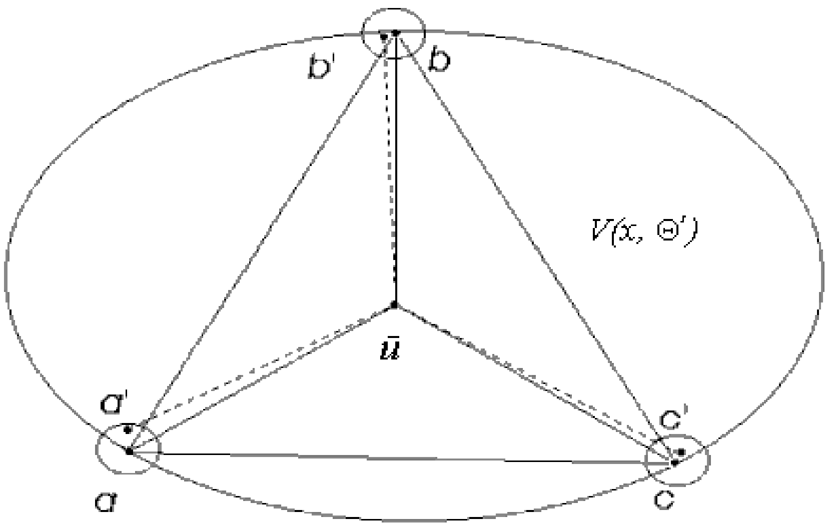

Figure 1.1: Representation of a Pareto optimal point $\bar{u}$ as a convex combinations of performance vectors a', b', and c' of (N, ∞)-stationary policies with the same "tail." Fig1

Theorem 1.27 *Fix x and consider performance vectors $\bar{v}_{\mathcal{M}}(x, \pi)$ and performance space $U_c(x)$. If $\bar{u}$ belongs to the boundary of $U_c(x)$ then there is a (K, N)-policy π with $\bar{v}_{\mathcal{M}}(x, \pi) = \bar{u}$. If $\bar{v}$ is any point in $U_c(x)$ then there is a $(K+1, N)$-policy σ with $\bar{v}_{\mathcal{M}}(x, \sigma) = \bar{v}$.*

Proof outline. Since U is convex and compact, any point on the boundary of $U_c(x)$ can be represented as the unique solution to a constrained optimization problem, with K constraints, so Theorem 1.26 implies the result. Any point in the interior of $U_c(x)$ can be represented by a similar constrained problem, but with $K+1$ constraints, and similar arguments apply. ∎

In general, it is not possible to achieve a given performance with (K, N)-policies, so that the result above is sharp.

1.6.3 Calculation of optimal policies

The computation of optimal policies for the constrained problem is, in general, an open problem. It is easy to compute approximate policies, provided that by "approximate" we mean that we allow the constraints to be "slightly violated.” To do this, given ε we fix a large N so that

$$\frac{K\beta_1^N \max_{a,x} |r_\ell^k(x,a)|}{1-\beta_1} < \varepsilon \tag{1.35}$$

and solve the finite horizon problem, ignoring all costs after N. This would put costs and constraints within ε of the desired values; see [15] for details.

A relaxation technique can be used to decrease the error, either in the constraints, or the value, or both. However, such algorithms are iterative, and it is difficult to obtain information about their accuracy.

Consider the case $K = 1$, and where

$$v^0_{\mathcal{M}}(x,\pi) = v_1(x,\pi,\beta_1)\,, \tag{1.36}$$

$$v^1_{\mathcal{M}}(x,\pi) = v_2(x,\pi,\beta_2). \tag{1.37}$$

That is, each criterion is a simple discounted one, but the discounts are different, $\beta_1 \neq \beta_2$. For the next result we do not assume that $\beta_1 > \beta_2$. Let the problem be feasible, that is

$$\max_\pi v^1_{\mathcal{M}}(x,\pi) \geq C\,. \tag{1.38}$$

We say a policy π is (a,b)-lexicographic optimal if it is lexicographic optimal for the vector $(v^a_{\mathcal{M}}(x,\pi), v^b_{\mathcal{M}}(x,\pi))$.

Theorem 1.28 ([17]) *(i) If $\max_\pi v^1_{\mathcal{M}}(x,\pi) = C$ then the $(1,0)$-lexicographic optimal policy solves the constrained optimization problem. (ii) Let σ be the $(0,1)$-lexicographic optimal policy, and suppose $v^1_{\mathcal{M}}(x,\sigma) \geq C$. Then σ is an optimal solution. (iii) If neither conditions hold and $\beta_1 > \beta_2$, then there is a finite algorithm for the computation of the optimal policy. The complexity of the algorithm is similar to the solution of the* **WDO** *problem. (iv) If on the other hand $\beta_1 < \beta_2$ than there is an iterative algorithm, that terminates in a finite number of steps, for the computation of optimal policies.*

We note that even for this simple problem, in case (iv) we have no prior estimate of the complexity of this calculation.

1.6.4 Single-discount constrained optimization

The problem of constrained optimization with the standard discounted criterion has been extensively studied; see the books by Kallenberg [25], Borkar [6], Piunovskiy [33], and Altman [1]. However, the non-stationary policies introduced in this chapter give this problem a different perspective. Indeed, if $\mathbb{X}$ and $\mathbb{A}$ are finite then Theorem 1.26 states the existence of optimal (K,N)-policies. This is a new result for problems with a single discount factor! However, for problems with a single discount factor, this result can be strengthened. If $\mathbb{X}$ and $\mathbb{A}$ are finite and a problem is feasible, then there exist an optimal randomized stationary policy [25]. Standard linear programming arguments [36] imply that, if the problem is feasible, then there exists a K-randomized stationary optimal policy. Combined with Theorem 1.26, this result implies the existence of strong (K,N)-policies for some $N<\infty$ for models with finite state and action sets.

If the state space is infinite, optimal (N,∞)-stationary policies may not exist for unconstrained mixed discounted problems [14]. Therefore, optimal (K,N)-policies may not exist for constrained mixed discounted models with infinite state spaces. However, as was proved in [16], optimal strong (K,N)-policies exist for constrained discounted problems with countable state spaces if these models satisfy standard continuity assumptions. We give here a brief survey of the results of [16].

We treat the countable state model of this chapter, and make the continuity assumptions of Theorem 1.6(ii). We consider a constrained problem with K constraints. We consider constrained problem (1.2,1.3) when, instead of a vector $\bar{v}_{\mathcal{M}}(x,\pi)$, the performance of a policy π is evaluated by a vector $\bar{v}(x,\pi)=(v^0(x,\pi),\ldots,v^K(x,\pi))$, where $v^k(x,\pi)$ are expected discounted total rewards for reward functions $r^k(x,a)$ and the common discount factor $\beta\in[0,1[, k=0,\ldots,K$. Note that our definitions of (N,∞)-stationary, K-randomized, randomized Markov of order K, and (K,N)-policies are all well-posed. We note that, in this generality, the set $\bar{v}\left(x,\Pi^R\right)$ may not be compact because it may not bounded. However, out assumptions suffice for the following.

Lemma 1.29 *If $\bar{u}$ belongs to the closure of $\bar{v}\left(x,\Pi^R\right)$ then there exists $\bar{u}'\in\bar{v}\left(x,\Pi^R\right)$ that dominates $\bar{u}$. Consequently, there exists a policy whose performance dominates $\bar{u}$.*

Theorem 1.30 *If π is Pareto optimal then (i) there exists an K-randomized stationary policy with the same performance, and (ii) there exists a strong (K,N)-policy with the same performance.*

t:WDConed **Theorem 1.31** *If problem* **WDC** *is feasible, then (i) there exists an optimal K-randomized stationary policy, and (ii) there exists an optimal strong (K,N)-policy.*

The strengthening of the conclusions from (K,N)-policies is worth a comment. Using conclusion (i) of Theorem 1.31, we obtain an K-randomized optimal policy σ. Consider now the submodel $\mathbb{A}'$ where

$$\mathbb{A}'(x) = \{a \in \mathbb{A}(x) : \sigma(a|x) > 0\} . \tag{1.39}$$

In this submodel, all but at most K of the $\mathbb{A}'(x)$ are singletons. We now obtain an optimal (K, N)-policy in the new model. By definition, this policy is a strong (K, N)-policy.

1.7 Related problems and criteria

s:WDC-EX

In this section we survey some related MDP problems and some extensions to stochastic games. Other related models are mentioned in the introduction.

1.7.1 MDP models

Average cost criteria are usually of the expected type. However, it is well known that for ergodic models that the value can be achieved with probability one. This is also the case for constrained MDPs; see Borkar [6] and Altman and Shwartz [4]. Ross and Varadarjan [37] consider a finite MDP and maximize the expected average cost subject to a constraint that another average cost does not exceed a given bound with probability one. For a general multichain MDP they establish that if the problem is feasible, then there is an ε-optimal stationary policy. An algorithm for its computation is provided.

Reiman and Shwartz [35] consider a mixed-criteria problem that arises in telecommunications. Arriving users may be rejected. If accepted, they generate communication packets according to an independent random process until they leave. There are two average per unit time optimization criteria determining the Quality of Service. The percentage of lost packets by a user that is accepted at a given state (given number of users) should be below or equal to a given bound. The probability of blocking (rejecting an arriving user) should be minimized. Due to the nature of the model, only stationary policies are relevant, and the fact that users leave after a geometric session time implies that the first criterion is actually of the discounted type. Since the bound must hold for every initial state, we have a mixed criterion problem with average cost optimization and a countable number of discounted constraints. The authors provide an algorithm for the computation of optimal policies and derive a relation to a mathematical program.

1.7.2 Stochastic games with mixed criteria

Filar and Vrieze [19], Altman, Feinberg, and Shwartz [3], and Altman, Feinberg, Filar, and Gaitsgory [2] investigated stochastic games with weighted criteria. Filar and Vrieze [19] considered zero-sum games with finite state and action sets. They considered two criteria: a mixture of two discounted criteria and a mixture of a discounted and average reward criteria. In both cases, they proved the existence of the value and the existence of randomized (N, ∞)-stationary ε-optimal policies for $\varepsilon > 0$. Altman, Feinberg, and Shwartz [3] provided an example when optimal (N, ∞)-stationary policies do not exist in a mixed-discounted problem and proved the existence of such policies in models with perfect information.

Altman, Feinberg, and Shwartz [3] also introduced lexicographic games when the players play the game with the payoff function r_1 and discount factor β_1 first. We denote this game by Γ_1. Then the players play the game with the discount factor β_2 and reward functions r_2 on the set of optimal policies of the game Γ_1. We denote this game by $\Gamma_{1,2}$. This construction can be repeated L times and, as the result, the players play the game $\Gamma_{1,\dots,L}$. In games with perfect information, when players play optimal (N, ∞)-stationary policies for mixed-discounted games, from epoch N onwards they play any optimal policy for the game $\Gamma_{1,\dots,L}$.

We recall that a mixed-discounted game with finite or countable state space can be reduced to a standard discounted countable state game with essentially the same action spaces ([14]). This construction is briefly described in Section 1.3. Therefore, if the action sets are finite, the set of optimal actions at each step for each player exists at each state. This set is a polytope which is a subset of the set of all probability distributions on the sets of all actions $A(x)$ for player one and $B(x)$ for player two. We denote theses sets of optimal actions by $\mathbf{A}_n(x)$ and $\mathbf{B}_n(x)$ respectively. Altman, Feinberg, Filar, and Gaitsgory [2] proved for repeated mixed-discounted games that the sequence of sets $\mathbf{A}_n$ (there is only one state in repeated games and therefore $\mathbf{A}_n(x)$ do not depend on x) converges to a subset of the set of optimal policies of the game $\Gamma_{1,2}$. Whether this result holds for stochastic games with finite state and action sets is an open question. We also remark that the examples in [2] show that the limit may not be equal to the set of optimal policies in game $\Gamma_{1,2}$ and that this limit may not be a subset of the sets of optimal policies for the game $\Gamma_{1,2,3}$.

In general, the existence of values for zero-sum games and equilibrium values for non-zero sum games are nontrivial questions. For mixed criteria, we are aware of two general methods to prove the existence of such values.

The first method is to represent a mixed criterion as a limsup criterion and use the results by Maitra and Sudderth [28, 29, 30]. The second method, which can be applied directly to mixed-discounted criteria, is to consider an expanded model described in Section 1.3 and then to apply the results for standard discounted criteria [27, 32, 9]. The existence of Nash equilibria is often established using fixed-point methods. Altman and Shwartz [5] consider the following stochastic game. We have L players. Player ℓ has a discount factor β_ℓ (where $\beta = 1$ means that the average cost is used) and immediate costs r_ℓ^k, $0 \le k \le B_\ell$. A policy π is called *feasible* if

$$v_\ell^k(x, \pi, \beta_\ell) \le V_\ell^k \quad \text{for } 1 \le \ell \le L,\ 1 \le k \le B_\ell, \tag{1.40}$$

where V_ℓ^k are given numbers. It is established in [5] that, if this problem is feasible then (under some regularity conditions) there exists a Nash equilibrium. An ergodicity condition is required if the average cost is used by some player. The proof uses fixed point methods.

1.7.3 Extensions and open problems

s:WDC-EX2

The authors are currently considering the extension of the mixed-discounted problem to the semi-Markov setting.

Except for the unconstrained problems, the algorithmic aspects of mixed-discounted criteria are still open: we do not have an algorithm for the computation of optimal, or even feasible ε-optimal policies.

Convergence of solutions for zero-sum games to the subsets of solutions of lexicographic games, established in [2] for repeated games, is an open question for stochastic games with finite state and action sets.

Acknowledgement. Research of the first author was partially supported by NSF Grant DMI-9908258. Research of the second author was supported in part by the fund for promotion of research at the Technion, in part by the fund for promotion of sponsored research at the Technion.

Bibliography

Al [1] E. Altman, *Constrained Markov Decision Processes,* Chapman & Hall/CRC, London, 1999.

AFFG [2] E. Altman, E. Feinberg, J.A. Filar, and V.A. Gaitsgory, "Perturbed Zero-sum Games with Applications to Dynamic Games," in *Proc. 8-th International Symposium on Dynamic Games and Aplications*, pp. 45–51, Maastricht, The Netherlands, 1998.

AFS [3] E. Altman, E.A. Feinberg, and A. Shwartz, "Weighted Discounted Stochastic Games with Perfect Information," in *Proc. 7-th International Symposium on Dynamic Games and Applications* **1**, pp. 18–31, Kanagawa, Japan, 1995; to appear in *Annals of the International Society of Dynamic Games.*

AS2 [4] E. Altman and A. Shwartz, "Sensitivity of constrained Markov decision processes," *Ann. Operations Research* **32** pp. 1–22, 1994.

AS [5] E. Altman and A. Shwartz, "Constrained Markov Games: Nash Equilibria," CC Pub. #143, Electrical Engineering, Technion, April 1996; to appear, Annals of the International Society for Dynamic Games,

Bo [6] V.S. Borkar, *Topics in Controlled Markov Chains*, Longman Scientific & Technical, Harlow, 1991.

Ch [7] R. Ya. Chitashvili, "A Finite Controlled Markov Chain with Small Break Probability," *SIAM Theory Probability Appl.* **21**, pp. 157–163, 1976.

DS [8] C. Derman and R.E. Strauch, "A note on memoryless rules for controlling sequential processes," *Ann. Math. Stat.* **37** pp. 272–278, 1966.

Fed [9] A. Federgruen, "On N-person stochastic games with denumerable state spaces," *Ad. Appl. Prob* **10**, pp. 452–471, 1978.

Fe [10] E.A. Feinberg, " Controlled Markov Processes with Arbitrary Numerical Criteria," *SIAM Theory Probability Appl.* **27**, pp. 486–503, 1982.

Fe96 [11] E.A. Feinberg, "Letter to the Editor," *Oper. Res.* **44**, p. 526, 1996.

Fe97 [12] E.A. Feinberg, "Constrained Discounted Markov Decision Processes and Hamiltonian Cycles," *Math. of Operations Research,* to appear.

Fe98 [13] E.A. Feinberg, "Continuous Time Discounted Jump Markov Decision Processes: Discrete-Event Approach," State University of New York at Stony Brook, Preprint, 1998.

FeS94 [14] E.A. Feinberg and A. Shwartz, "Markov decision models with weighted discounted criteria," *Math. of Operations Research* **19** pp. 152–168, 1994.

FeS95 [15] E.A. Feinberg and A. Shwartz, "Constrained Markov decision models with weighted discounted rewards," *Math. of Operations Research* **20** pp. 302–320, 1995.

FeS96 [16] E.A. Feinberg and A. Shwartz, "Constrained discounted dynamic programming," *Math. of Operations Research* **21** pp. 922–945, 1996.

FeS99 [17] E.A. Feinberg and A. Shwartz, "Constrained dynamic programming with two discount factors: applications and an algorithm," *IEEE Transactions on Automatic Control* **TAC-44** pp. 628–630, 1999.

FGM [18] E. Fernandez-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 of Operations Research* **39** pp. 131–155, 1994.

FiV [19] J. Filar and O. Vrieze, "Weighted reward criteria in competitive Markov decision programming problems," *ZOR—Methods and Models of Operations Research* **36** pp. 343–358, 1992.

GhM [20] M.K. Ghosh and S.I. Marcus, "Infinite horizon controlled diffusion problems with some nonstandard criteria," J. Math. Systems, Estimation and Control **1** pp. 45–69, 1991.

GKW [21] K. Golabi, Ram B. Kulkarni and G.B. Way, "A statewide pavement management system," *Interfaces*, January 1982.

HLLa [22] O. Hernandez-Lerma and J. Lasserre, *Future Topics in Markov Decision Processes,* Springer, New York, 1999.

Ho [23] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Math. Centre Tracts **51**, Math. Centrum, Amsterdam, 1974.

Hi [24] K. Hinderer, *Foundations of Non Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Operations Research **33**, Springer-Verlag, NY, 1970.

Ka [25] L.C.M. Kallenberg, *Linear Programming and Finite Markovian Problem*, Math. Centre Tracts **148**, Math. Centrum, Amsterdam, 1983.

KrFS [26] D. Krass, J. Filar and S.S. Sinha, "A weighted Markov decision process," *Oper. Res.* **40**, pp. 1180–1187 1992.

MaP [27] A.P. Maitra and T. Parthasarathy, "On stochastic games," *Journal of Optimization Theory and Applications* **5**, pp. 289–300, 1970.

MaS92 [28] A.P. Maitra and W.D. Sudderth, "An operator solution of stochastic games," *Israel Journal of Mathematics* **78**, pp. 33–49, 1992.

MaS93 [29] A.P. Maitra and W.D. Sudderth, "Borel stochastic games with limsup payoff," *Annals of Probability* **21**, pp. 861–885, 1993

MaS96 [30] A.P. Maitra and W.D. Sudderth, *Discrete Gambling and Stochastic Games,* Springer, New York, 1996.

No [31] A.S. Nowak, "Universally measurable strategies in zero-sum stochastic games," *Annals of Probability* **13**, pp. 269–287.

PaR [32] T. Parthasarathy and E.S. Raghavan, *Some Topics in Two-Person Games,* Elsevier, New York, 1967.

Pi [33] A.B. Piunovskiy, *Optimal Control of Random Sequences in Problems with Constraints*, Kluwer, Boston, 1997.

Pu [34] M.L. Puterman, *Markov Decision Processes*, Wiley, New York, 1994.

RS [35] M.I. Reiman and A. Shwartz, "Call Admission: A new Approach to Quality of Service," Center for Communication and Information Technologies, Technion, Israel, CC Pub # 216, 1999.

Ro [36] K.W. Ross, "Randomized and past dependent policies for Markov decision processes with finite action sets," *Oper. Res.* **37** pp. 474–477, 1989.

RV [37] K.W. Ross and R. Varadarajan, "Multichain Markov decision processes with a sample path constraint: a decomposition approach," *Math. Operations Research* **16** pp. 195–207, 1991.

Sc [38] M. Schäl, "Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal," *Z. Wahr. verw. Gebiete* **32** pp. 179–196, 1975.

WZ [39] K.C.P. Want and J.P. Zaniewski, "20/30 hindsight: the new pavement optimization," *Interfaces*, 1996.