LP-Based Approaches to Stationary-Constrained Markov Decision Problems

Felisa Vázquez-Abad [*] **Pinhus Dashevsky** [**]
Matthew P. Johnson [***]

[*] *Hunter College, CUNY, New York, NY USA*
[**] *Columbia University, New York, NY USA*
[***] *Lehman College, CUNY, Bronx, NY USA*

Abstract: We study a class of Markov Decision Processes (MDPs) under stationary constraints (particularly but not exclusively, chance constraints). Our model focuses on problems where the state space is finite but the control at each state may take real values. It is straightforward to formulate the problem as a constrained nonlinear optimization problem, but solving it requires either projection, penalty or gradient-based methods that may show slow convergence, particularly because this type of problems may not be strictly convex.
In this paper we extend results that are known for the discrete control model and show that the solution of a binary randomized control problem is optimal under the assumption that the dependency on the control variable is linear. If the dependency is piecewise linear with several breakpoints, then the randomized problem is not equivalent to the original problem; however we show that a particular solution always exists that is also the solution to the original continuous problem. This special solution takes the form of a two-action control policy for each state. Solving each two-action randomized problem can be done in polynomial time by linear programming (LP). This leads to an efficient solution when the number of states is small, but for a large state space the complexity can be prohibitive. In a restricted yet interesting special case of piecewise linear, multiple actions setting where the actions are linearly ordered, we show that the problem can again be reduced to a single LP.
To overcome the computational complexity in the general case, we propose an alternative approach that involves solving a sequence of linear programs modeling the stationary measure and the optimal solution concurrently. We conclude the paper by discussing future extensions of this framework to application settings where the state space is continuous as well.

Keywords: Markov decision processes, convex optimization, linear programming

1. INTRODUCTION

We study Markov Decision Processes with continuous control. To motivate our model, consider the following illustrative example in healthcare. Suppose patients have one of N different "states" indicating the severity of a condition, such as diabetes (N will typically be three to five acuity levels). A drug has been found to have positive effects on humans and the problem is to find the optimal dosage for a life-long treatment. Different dosage amounts may affect the patient by changing her from one state to another, with different probabilities. The goal is to optimize for patient health. For safety reasons, the total amount of the drug a patient may take over time is limited to some maximum, which, as we will see, can be interpreted as a probability or stationary constraint. Specifically, we require that the amount of dose does not exceed a certain threshold with large enough probability. Note that while the actions are continuous (the possible dosages), the set of states is finite. Indeed, the problem can be formulated as a constrained Markov Decision Process (MDP) (Altman, 1999; Dolgov and Durfee, 2005; Puterman, 1994) with finite state (health level) and continuous actions (dosages); the cost (negative to health benefits) and the constraint implicitly refer to a "stationary" measure.

We begin by presenting a formal problem formation for this kind of problem (Section 2). After that, in succeeding sections we consider several approaches to attacking this problem. These approaches—linear programming when the relevant functions are linear (Section 3), reduction to multiple linear programs (Section 4), and approximation by piecewise linear functions or use of dynamic programming in the case of general nonlinear functions (Section 5)—offer different sorts of guarantees, depending on what assumptions we make about the nature of the functions involved. Next we propose two methods involving linear programming (LP): first, we formulate the piecewise linear setting as a single LP (Section 4), which is likely to be more tractable than the reduction to many LPs; second, we describe a method that involves a series of linear programs (Section 6) in order to more quickly narrow the search space. We believe such approaches will be a fruitful basis for future research. Finally, in conclusion we discuss a more general problem setting with continuous state space as well as continuous action space (Section 7), which will also be a topic for future research.

⋆ Partly supported by PSC-CUNY grant 65653-00 43.

2. FINITE-STATE MODEL

Consider a Markov Decision Process (MDP) $X_n \in S$, where the set of states S is finite. The possible actions u_n are continuous in $[0,1]$. Let $\mathfrak{F}_n = \sigma(X_n, u_n; \ldots; X_1, u_1)$ be the sigma algebra describing the history of the process (it defines the natural filtration of the process $\{(X_n, u_n)\}$) and assume that for $a \in [0,1]$

$$\mathbb{P}(X_{n+1} = j \,|\, X_n = i, u_n = a, \mathfrak{F}_{n-1}) = p_{ij}(a) \qquad (1)$$

is independent of the past and of the time n. We will assume that the functions $p_{ij}(\cdot)$ are continuous and piecewise continuously differentiable with uniformly bounded derivative. Let $c\colon S \times [0,1] \to \mathbb{R}$ and $g\colon S \times [0,1] \to \mathbb{R}$ be cost and constraint functions, respectively. We write the inputs i and/or a as either parameter or subscript, depending on whether we are treating these as functions (e.g., $c_i(\cdot)$) or as constants (e.g., c_{ia}).

The following result is an extension of the corresponding result in Puterman (1994), noting that the proof does not in fact depend on the assumption of a discrete action space.

Lemma 1. Let $\{X_n, u_n)\}_\beta$ be a MDP on $(\Omega, \{\mathfrak{F}_n\}, \mathbb{P})$ with history-dependent randomized (HR) policy β, that is: $u_n = \phi_\beta(X_0, u_0, \ldots, X_n; \omega)$ for some function $\phi_\beta(\cdot)$. Then there is a Markov random (MR) policy β', for which

$$u'_n = u_n = \phi_\beta(X_n; \omega) \qquad (2)$$

such that the processes under β and β' have the same distribution, that is, $\{(X_n, u_n)\}_\beta \stackrel{d}{=} \{(X_n, u'_n)\}_{\beta'}$.

The above result means that we may restrict the optimization problem to policies of type MR. If (2) is satisfied, calling:

$$\beta_i(A) = \mathbb{P}(u_n \in A \,|\, X_n = i), \quad A \subset [0,1] \qquad (3)$$

then the embedded chain $\{X_n(\beta)\}$ is a Markov chain.

Lemma 2. For any MR policy satisfying (3) there is a Markov deterministic (MD) policy β' with degenerate probabilities for which the embedded chains have the same distribution.

Proof. For every $i, j \in S$, the embedded chain under policy (3) satisfies:

$$\mathbb{P}(X_{n+1} = i \,|\, X_n = i) = \int_0^1 p_{ij}(x) d\beta(x).$$

From the continuity of $p_{ij}(\cdot)$ and the mean value theorem, there exists a value $\theta_i \in [0,1]$ such that $p_{ij}(\theta_i) = \int_0^1 p_{ij}(x) d\beta(x)$. Defining the degenerate policy by $\beta_i(\theta_i) = 1$, we get a deterministic policy of the form $u_n = \theta_{X_n}$ that yields an embedded process with the same distribution as $\{X_n\}$, which proves the claim. □

Lemma 2 implies that we may solve the problem looking only at degenerate probabilities (deterministic policies) of the form $u_n = \theta_{X_n}$ that only depend on the state. The goal is to find the optimal values of the vector θ. We shall restrict our attention to MD policies that yield a *unichain* model, specifically those policies under which the embedded chain has a unique stationary measure. The optimization problem we study can now be stated as:

$$\begin{aligned} &\min \sum_{i \in S} c_i(\theta_i)\mu_i(\theta) \qquad (4)\\ &\text{s.t. } \sum_{i \in S} g_i(\theta_i)\mu_i(\theta) \le 0 \\ &\qquad \theta_i \in [0,1] \;\; \forall i \in S \end{aligned}$$

where $\mu_i(\theta)$ is the stationary distribution of the embedded chain $\{X_n(\theta)\}$ with control $u_n = \theta_{X_n}$. Introducing the notation $C(\theta) = \sum_{i \in S} c_i(\theta_i)\mu_i(\theta)$, $G(\theta) = \sum_{i \in S} g_i(\theta_i)\mu_i(\theta)$, the problem can be written more compactly as:

$$\min_{\theta_i \in [0,1], i \in S} C(\theta) \text{ s.t. } G(\theta) \le 0.$$

If the problem is convex, then a Lagrange method can be used:

$$\min_{\lambda \ge 0} \Big(\min_{\theta} (C(\theta) + \lambda\, G(\theta)) \Big)$$

This is a nonlinear optimization problem for which gradient-based methods converge to the optimum, when the problem is (strictly) convex. The main challenge in solving it is in evaluating the functions $C(\cdot), G(\cdot)$, which requires knowledge of $\mu(\cdot)$.

The recursive method for solving (4) via gradient search follows the algorithm:

(1) Initialize $\lambda_0 > 0, \theta^0 \in [0,1]^d, n = 0, \epsilon_n$.
(2) Solve the system of linear equations:

$$x_j = \sum_{i \in S} x_i\, p_{ij}(\theta_i^n), \quad \sum_i x_i = 1$$

and notice that $x_j = \mu_j(\theta^n)$.
(3) Solve the system of linear equations:

$$\begin{aligned} y_{ij} &= \sum_{k \in S} p_{kj}(\theta_k^n) y_{ik} + x_i \frac{\partial}{\partial \theta_i} p_{ij}(\theta_i^n) \\ &\sum_{j \in S} y_{ij} = 0, \forall i \in S. \end{aligned}$$

Then $y_{ij} = \frac{\partial}{\partial \theta_i}\mu_j(\theta^n)$.
(4) Solve the system of linear equations:

$$z_i = \sum_{k \in S} \ell(k, \theta_k^n) y_{ik} + x_i \frac{\partial}{\partial \theta_i} \ell(i, \theta_i^n),$$

where $\ell(i, \theta) = c_i(\theta) + \lambda g_i(\theta)$ is the Lagrangian. Then $z_i = \frac{\partial}{\partial \theta_i}(C(\theta^n + \lambda G(\theta^n)$.
(5) Update using:

$$\begin{aligned} \theta_i^{n+1} &= \theta_i^n - \epsilon_n z_i \\ \lambda_{n+1} &= \lambda_n + \epsilon_n \sum_{i \in S} g_i(\theta^n) x_i \end{aligned}$$

(6) Set $n = n + 1$ and GO TO (2).

The above algorithm is guaranteed to converge to the solution when the problem (4) is strictly convex. Typically one chooses the step sizes to decrease as $\epsilon_n = \mathcal{O}(1/n)$ for better convergence (Bertsekas, 1999).

If the problem is not convex, however, then some technique may be found to convexify it. See Arrow and Hurwitz (1957), who proved convergence for this method, viewing it as a differential game, and suggested ways to modify the cost function to ensure convergence for non-convex problems.

3. TWO ACTIONS AND LINEAR FUNCTIONS

In this and succeeding sections, we will study other methods to solve this problem in special cases where the functions involved obey certain assumptions, in this section, linearity assumptions. In this case, we state a result that implies that the (continuous action) problem has the same solution as a binary-action model (discrete randomized actions). The randomized binary problem can be solved by linear programming and will provide a very "simple" solution.

Theorem 3. Consider the MDP defined by the transition probabilities (1). Assume that for each $i, j \in S$ the functions $p_{ij}(\theta_i), c_i(\theta_i), g_i(\theta_i)$ are linear in θ_i. Consider now a problem where we restrict the strategies to MR policies on the two-action model:

$$\theta_i = \beta_i(1) = \mathbb{P}(u_n = 1 \,|\, X_n = i) = 1 - \mathbb{P}(u_n = 0 \,|\, X_n = i),$$

so that the only actions are 0 or 1. Then the corresponding MDP problem is equivalent to (4).

Proof. By construction, the transition probabilities of the embedded chain for the two-action model satisfy:

$$\mathbb{P}(X_{n+1} = j \,|\, X_n = i) = p_{ij}(0)(1 - \theta_i) + p_{ij}(1)\theta_i,$$

and by linearity this is exactly the transition probability $p_{ij}(\theta_i)$ in (1). Thus for each value of the vector θ, the stationary measure μ is the same for both processes. Notice now that for this two-action problem, the expected cost is:

$$\sum_{i \in S} \mu_i(\theta) \mathbb{E}(c(X_n, u_n) \,|\, X_n = i)$$
$$= \sum_{i \in S} \mu_i(\theta)(c_i(0)\,(1 - \theta_i) + c_i(1)\,\theta_i) = C(\theta),$$

where the last equality follows from the assumption that C is linear. Similarly, the expected value of the constraint for the two-action randomized control is the same as $G(\theta)$. Therefore the two problems have the same solution. □

We remark that although the solution θ_i^* value may be the same, it is interpreted differently: it represents a deterministic continuous control in one model, and a randomized binary control in the other.

The randomized problem can be solved efficiently using linear programming as follows (see Ross (2006) for details). Let x_{ia} be LP variables solving for the stationary probabilities π_{ia}, for the multidimensional chain $Z_n = (X_n, u_n)$, and $i \in S, a \in \mathcal{A}$. In the two-action setting, we have $\mathcal{A} = \{0, 1\}$. Let $c_{ia} = c_i(a)$, $g_{ia} = g_i(a)$, $p_{ija} = p_{ij}(a)$ be constants set based on the corresponding functions, for all values i, a. Then consider the following linear program:

$$\min \sum_{i \in S, a \in \mathcal{A}_i} x_{ia} c_{ia} \tag{5}$$

$$\text{s.t.} \sum_{i \in S, a \in \mathcal{A}_i} x_{ia} = 1 \tag{6}$$

$$\sum_{a \in \mathcal{A}_j} x_{ja} = \sum_{i \in S, a \in \mathcal{A}_i} x_{ia} p_{ija} \quad \forall j \in S \tag{7}$$

$$\sum_{i \in S, a \in \mathcal{A}_i} x_{ia} g_{ia} \leq 0 \tag{8}$$

$$x_{ia} \geq 0 \quad \forall i \in S, a \in \mathcal{A}$$

This formulation minimizes the cost of the stationary probabilities x_{ia}, weighted by c_{ia}, subject to the constraints that (a) the probabilities x_{ij} sum to one; (b) the total stationary probability for each state j is consistent with other states' stationary probabilities and the transition probabilities; (c) the constraint based on g_{ia} is obeyed; and finally (d) the x_{ia} are nonnegative.

Given an optimal solution to (5-8), we compute the probability, when in state i, of taking action a as:

$$\beta_i(a) = \frac{x_{ia}}{\mu_i}, \text{ where } \mu_i = \sum_{b \in \mathcal{A}} x_{ib} \tag{9}$$

Then an optimal solution to (4) is given by the vector of values $\theta_i^* = \beta_i(1)$.

It is known (see Puterman (1994), Corollary 8.9.7) that a basic optimal solution to (5-7) (omitting constraint (8)) will be deterministic, i.e., will satisfy $\theta_i \in \{0, 1\}$ for all $i \in S$, and that a basic solution to formulation (5-8) will be nondeterministic in at most one state i, in which case it will randomize between only two actions. [1] We call such solutions *semi-deterministic*. This follows because a basic solution to (5-7) (respectively, (5-8)) will have at most $|S|$ (respectively, $|S| + 1$) nonzero variables x_{ia}, but for each state i, $\sum_a x_{ia}$ must be nonzero without loss of generality. (Otherwise state i could simply be deleted from the formulation.)

This result provides insight into the original problem (4): it follows from the assumption of linearity that absent constraint (8), all decisions would necessarily be at the local (state by state) best value, which is attained at either $\theta_i = 0$ or $\theta_i = 1$. The inclusion of the constraint (8) has the effect (in general) of changing the optimal value and rendering the optimal policies non-unique. Such non-uniqueness was exploited in Vázquez-Abad and Mason (1999) to achieve accelerated convergence of a gradient-based nonlinear stochastic optimization approach.

4. MANY ACTIONS AND PIECEWISE LINEAR FUNCTIONS

We now extend this analysis, generalizing from the setting with two discrete actions and linear functions characterizing the randomized combinations of them to a setting with an arbitrary finite number of *linearly ordered* discrete actions $\mathcal{A}$, and where the functions characterizing them are piecewise linear, with breakpoints corresponding to the members of $\mathcal{A}$. First we show how to reduce the problem in this setting to a collection of problems in the linear two-action setting, i.e., to solving a series of linear programs. Second, we show that in the narrower generalization in which $c_i(\cdot), g_i(\cdot)$ are piecewise linear and convex, and $p_{ij}(\cdot)$ is linear, the problem can be solved by reducing to solving a *single* linear program.

4.1 Non-convex functions

We begin by recalling some facts about the randomized control problem with a finite number of actions and specifying the assumptions of the current setting.

[1] Although this latter statement is vacuous when $\mathcal{A}$ is binary, we will use it next to confine our study to two-action models.

Let $\mathcal{A} = \{h_m : m = 0, \ldots, \kappa\} \subset [0,1]$, numbered in (strictly) increasing order, with $h_0 = 0$ and $h_\kappa = 1$, be the set of discrete actions. We permit solutions randomizing between each pair of actions h_m, h_{m+1}, corresponding to points in the subinterval $[h_m, h_{m+1})$. We call such solutions *adjacent*. We do not permit solutions to randomize between nonadjacent members of $\mathcal{A}$, or *nonadjacent*. This restriction is implemented as follows.

The functions $p_{ij}(\cdot), c(i,\cdot), g(i,\cdot)$ are piecewise linear with breakpoints $\mathcal{A}$, so that $p_{ij}(\theta)$ in (1) behaves as follows. For $\theta \in [0,1] - \mathcal{A}$, let h_m be such that $\theta \in [h_m, h_{m+1})$, with $\Delta_m = h_{m+1} - h_m$. Then the weights used to interpolate values at θ are:

$$w_m = (1 - w_{m+1}) = \frac{\theta - h_m}{\Delta_m}, \quad w_{m+1} = \frac{h_{m+1} - \theta}{\Delta_m} \tag{10}$$

and in particular the value of $p_{ij}(\theta)$ interpolated between $p_{ij}(h_m)$ and $p_{ij}(h_{m+1})$ is:

$$p_{ij}(\theta) = p_{ij}(h_m) \cdot (1 - w_{m+1}) + p_{ij}(h_{m+1}) \cdot w_{m+1}$$

Consider now a binary randomized MDP formulation where for each vector θ, the MR policy is of the form:

$$\beta_i(h_m) = \mathbb{P}(u_n = h_m \,|\, X_n = i) = w_m$$
$$\beta_i(h_{m+1}) = \mathbb{P}(u_n = h_{m+1} \,|\, X_n = i) = w_{m+1}$$

for the interval such that $\theta \in [h_m, h_{m+1})$.

By construction, the transition probabilities of the embedded chain of the randomized problem are:

$$\mathbb{P}(X_{n+1} = j \,|\, X_n = i) = \sum_{h_m \in \mathcal{A}} \beta_i(h_m) p_{ij}(h_m). \tag{11}$$

which are the same as $p_{ij}(\theta)$.

Thus we have defined a problem setting (with a finite, linearly ordered set of actions, adjacent pairs of which can be randomized between) in such a way that the solution component corresponding to each state i is not a sequence of probabilities π_{ia} for all $a \in \mathcal{A}$ but rather a pair of values (θ_i, μ_i), where $\theta_i \in [0,1]$ is interpreted as (depending on its value) a single discrete action of $\mathcal{A}$ or a randomization of two succeeding ones, performed when in state i, and $\mu_i \geq 0$ is the stationary probability of state i.

In any solution, in particular an optimal solution, each θ_i will lie within some such half-open subinterval. (We make may the last subinterval closed to avoid omitting value 1.) Observe that by rescaling these subintervals, and restricting $c_i(\cdot), g_i(\cdot), p_{ij}(\cdot)$ to them, we can reinterpret this problem as an instance of problem (5-8): for each $\theta_i \in [h_m, h_{m+1})$, scale $[h_m, h_{m+1})$ to $[0,1]$ (or equivalently set $\mathcal{A}_i = \{h_m, h_{m+1}\}$).

This immediately suggests a method of solving the current problem setting by reduction to a series of linear programs: for each of the $\kappa^{|S|}$ ways of choosing subinterval the states' θ_i values to lie in, solve the corresponding problem (5-8), and keep the best of all solutions. For small S this may be practical, but because this is exponential in $|S|$, the complexity will be prohibitive in general.

4.2 Convex Functions

In a narrower problem setting, a more efficient solution is possible, where we solve only a single linear program. Now assume that $c_i(\cdot), g_i(\cdot)$ are both piecewise linear and convex, and that $p_{ij}(\cdot)$ is linear, all still defined on $[0,1] = [h_0, h_\kappa]$. For purposes of clarity, for the remainder of this section call this problem **P**:

$$\min \sum_{i \in S} \mu_i c_i(\theta_i) \tag{12}$$

$$\text{s.t.} \sum_{i \in S} \mu_i = 1 \tag{13}$$

$$\sum_{i \in S} \mu_i g_i(\theta_i) \leq 0 \tag{14}$$

$$\theta_i \in [0,1], \ \mu_i \geq 0 \ \ \forall i \in S$$

In order to solve **P**, we construct a new instance of problem (5-8), setting the constants c_{ia}, p_{ija}, g_{ia} based again on the corresponding functions, and using $\mathcal{A} = \{h_0, ..., h_\kappa\}$. Call this problem $\mathbf{P_R}$. We claim that an optimal solution to $\mathbf{P_R}$ can be interpreted as an optimal for **P**, that is, that it yields a minimum-cost feasible solution of **P**. Note that in order to be feasible for **P**, a solution must satisfy the problem constraints and must be semi-deterministic.

The problem $\mathbf{P_R}$ is a relaxation of problem **P** in the sense that any feasible solution to **P** is also a solution to $\mathbf{P_R}$ of the same cost, or rather is translatable into such a solution as follows (where $\theta_i \in [h_m, h_{m+1})$; see (10)):

$$x_{ih_\ell} = \mu_i \cdot \begin{cases} 1 - w_{m+1}, & \text{if } \ell = m \\ w_{m+1}, & \text{if } \ell = m+1 \\ 0, & \text{otherwise} \end{cases}$$

Therefore it immediately follows that an optimal solution to $\mathbf{P_R}$ will be lower-bound the optimal solution cost of **P**. Note also that any adjacent solution to $\mathbf{P_R}$ can be translated back into into a feasible solution of **P** of the same cost. The stationary probability μ_i is computed as in (9), and

$$\theta_i = \frac{x_{ih_m}}{\mu_i} \cdot h_m + \frac{x_{ih_{m+1}}}{\mu_i} \cdot h_{m+1}$$

where m is the smallest index such that x_{ih_ℓ} is nonzero (and $x_{ih_{m+1}}$ may or may not be nonzero). (Feasibility of **P** simply consists of satisfying constraints (13,14), which follow from the constraints (6,8) being satisfied by feasible solutions of $\mathbf{P_R}$; restricting to solutions randomizing between adjacent actions only is built-in.) Therefore it suffices to show that an optimal solution to $\mathbf{P_R}$ can be found that is translatable back into problem **P**.

Lemma 4. Any nonadjacent solution to problem **P** can be transformed into a solution of **P** that is adjacent and of only lower cost.

Proof. Let θ be a basic optimal solution to (5-8) that for state i randomizes, with probabilities p and $1-p$, between states a_m and a_n, where $m < n-1$. Let $\bar{a} = p \cdot a_m + (1-p) \cdot a_n$, with $\bar{a} \in [a_{m'}, a_{m'+1})$ for some $m' \in \{m, ..., n-1\}$, and let q be the corresponding weighting probability with which $\bar{a}$ can be expressed as a convex combination of the breakpoints defining this subinterval, i.e., $\bar{a} = q \cdot a_{m'} + (1-q) \cdot a_{m'+1}$. Let θ' be the same as θ except that it randomizes thus between $a_{m'}$ and $a_{m'+1}$.

Transforming θ into θ' can only lower the cost, because the $c_i(\cdot)$ is convex.

θ' is shown to be feasible by the following lemma. $\square$

Lemma 5. The feasibility of θ implies the feasibility of θ'.

Proof. Constraint (6) clearly remains satisfied, by construction.

Now, consider the constraints (7). The value of the LHS in one such constraint is not affected if j is not the state whose actions are modified; if j is, then the value still does not change because

$$\mu_j \cdot (p + (1-p)) = \mu_j \cdot (q + (1-q))$$

The RHS also does not change. Due to the linearity of $p_{ij}(\cdot)$ we have:

$$\begin{aligned}\sum_{a\in\{a_m,a_n\}} x_{ia}p_{ij}(a) &= \mu_i \cdot (p \cdot p_{ij}(a_m) + (1-p)\cdot p_{ij}(a_n)) \\ &= \mu_i \cdot (p_{ij}(p \cdot a_m + (1-p)\cdot a_n)) \\ &= \mu_i \cdot (p_{ij}(\bar{a})) \\ &= \mu_i \cdot (p_{ij}(q \cdot a_{m'} + (1-q)\cdot a_{m'+1})) \\ &= \sum_{a\in\{a_{m'},a_{m'+1}\}} x_{ia}p_{ij}(a)\end{aligned}$$

Finally, consider constraint (8). We have:

$$\begin{aligned}\sum_{a\in\{a_{m'},a_{m'+1}\}} x_{ia}g_{ia} &= \mu_i \cdot (q \cdot g_{ia_{m'}} + (1-q)\cdot g_{ia_{m'+1}}) \\ &\le \mu_i \cdot (p \cdot g_{ia_m} + (1-p)\cdot g_{ia_n}) \\ &= \sum_{a\in\{a_m,a_n\}} x_{ia}g_{ia}\end{aligned}$$

The inequality follows from the convexity of $g_i(\cdot)$. □

By repeated application of Lemma 4, we thus obtain:

Theorem 6. Problem **P** (12-14) can be solved in polynomial time.

The insight that we gain from this result is that two-action randomization is in a sense optimal. Vázquez-Abad and Mason (1996) introduced the concept of the N-action versus the two-action automata for a network flow and routing control problem. Although that work does not consider MDP formulations, we recover the same type of behavior for the solutions of randomized actions.

5. NONLINEAR OBJECTIVES

One way of dealing with the more general setting of nonlinear functions is to approximate them by piecewise linear functions, and then apply one of the methods discussed above. Assuming the functions $p_{ij}(\cdot), c(i,\cdot), g(i,\cdot)$ are continuous and piecewise continuously differentiable with (a.e.) uniformly bounded derivative, and also that they are monotonic (nondecreasing or nonincreasing) in θ, it is possible to approximate such functions to any degree of accuracy (in the sup norm) by piecewise linear functions.

If all the functions are analytically known then it is possible to find for each function the set of breakpoints for a piecewise linear interpolation that makes the L_2 error small to within a pre-specified tolerance. Specifically, let $\epsilon > 0$ and define, for each function $f \in \{p_{ij}, c, g\}$: $h_0 = 0$, and recursively:

$$h_{m+1} = \inf\{x \in \chi :\ |f(h_m) + f'(h_m)x - f(h_m + x)| \le \epsilon\} \tag{15}$$

until $h_{m+1} = 1$, where:

$$\chi = \{x \ge 0 \colon h_m + x \le 1\},$$

and then define the piecewise linear approximation $\tilde{f}(\cdot)$ with breakpoints $\{h_m\}$.

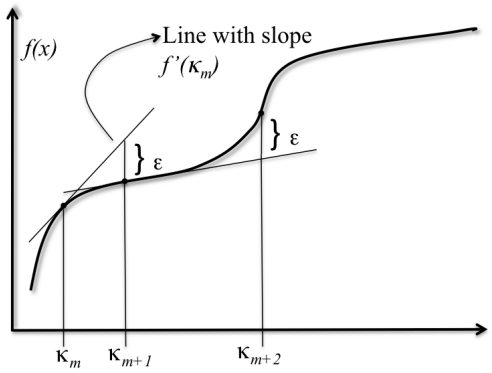

Fig. 1. Breakpoints for piecewise linear approximation.

Fig. 1 illustrates the iterative choice of breakpoints for two iterations of the procedure. Using this criterion, both the sup norm and the L_2 norm of the difference between $f(\cdot)$ and $\tilde{f}(\cdot)$ will be absolutely bounded by ϵ, because $\|f - \tilde{f}\|_2 \le \sqrt{\sum_m \epsilon^2 (h_{m+1} - h_m)^2} \le \epsilon$ (using the fact that $\sum_m (h_{m+1} - h_m) = 1$).

6. A NESTED LP APPROACH

We now turn to the method which is one focus of our ongoing research. The method that we propose here starts narrowing down consecutive multi-action problems and uses information on the curvature of the functions in order to better place a smaller number of candidate breakpoints while computing the optimum. Our emphasis is on speed of convergence.

Given an error tolerance $\epsilon > 0$, our first step for $n = 0$ is to find the solution θ^n to the randomized problem with actions $\{0, 1\}$ that corresponds to the linear interpolation on the whole interval $[0, 1]$.

For each $i \in S$, let $h_i = \theta_i^n$ and $\alpha = f'(h_i)$. Find the two adjacent breakpoints $h_i^{\pm}$ as follows. Let:

$$\chi^+ = \{x \ge 0 \colon h_i + x \le 1\}, \quad \chi^- = \{x \ge 0 \colon h_i - x \ge 0\}$$

and find the new breakpoints:

$$h_i^{\pm}(h_i) = h_i \pm \inf\{x \in \chi^{\pm} \colon\ |f(h_i) \pm \alpha x - f(h_i \pm x)| \ge \epsilon\} \tag{16}$$

For each state i, we find the candidate breakpoints for all functions $p_{ij}(\cdot), c_i(\cdot), g_i(\cdot) : j \in S$ and use the maximum h_i^- and the minimum h_i^+ as new interval. Next, solve the LP problem with randomized actions on the sets $\mathcal{A}(i) = \{0, h_i^-, h_i, h_i^+, 1\}$ (for each state i) imposing a two-action policy. Call the solution θ^{n+1} and repeat the procedure until the new interval for the breakpoints is smaller than the previous one.

In ongoing research, we are working to prove the following conjecture.

Conjecture 7. Assume that all functions are monotone functions in each argument θ_i for every $i \in S$. Furthermore, assume that all functions have uniformly bounded

second derivative except possibly on a finite number of breakpoints. Then the nested LP procedure terminates after a finite number of iterations and the error in the approximation is $\mathcal{O}(\epsilon)$.

The optimization subproblems (16) may have to be solved by approximations, or by local searches. For many problems, constraints or cost functions are linear, so (16) only has to be evaluated for few functions.

7. DISCUSSION

In this paper, motivated by several probability-constrained application settings, we gave a problem formulation subsuming a patient health optimization problem. We then explored several approaches to solving such problems, culminating in a method we propose involving a series of linear programs. In ongoing work, we are working to prove Conjecture 7, and we are performing experiments to evaluate the different methods in simulation.

In the health problem, let $p_{ij}(a)$ be the probability of changing from state i to j when an amount $a \in [0, 1]$ of drug is administered. The patient state is monitored each day. Each state and dose have an associated economic value of health benefit $W(i, a)$; dosage is priced following an increasing function $P(a)$. The cost is $c(i, a) = W(i, a) + P(a)$. The total cumulative dosage is bounded by $A \in (0, 1)$. We remark that in the discretized version of the problem where $a \in \{0, 1\}$, the corresponding constraint is a probability constraint: $\mathbb{P}(a = 1) \leq A$.

This work assumed a continuous action set but a finite state set. Two infrastructure problems we plan to study in the future, however, involve both continuous action and state. Preliminary experiments suggest it may be possible to extend the nested LP approach to problems of this kind. We now briefly discuss these problems, which concern energy management and resale in two environmental infrastructure settings: 1) wind turbines and 2) hydroelectric dams. The optimizer's goal is to maximize profits made by selling excess energy to the electrical grid, consistent with obeying constraints on the legal operation of the infrastructure, as specified by the municipality. An interesting characteristic shared by these problem formulations is a probabilistic constraint of the following form: in valid solutions, a certain value must be above a specified threshold with some probability $1 - p$, or equivalently it must be so a $1 - p$ fraction of the time.

First consider wind turbines, i.e., windmills that convert (kinetic) wind energy into electrical power. Such windmills are linked to the electrical grid to provide clean energy for customer usage, ideally a relatively steady amount of every over time. Wind energy is by nature bursty, however, so some mechanism must be used to avoid large differentials in the amounts of energy released over short periods of time, without letting the energy obtained during bursts go to waste. Suppose this energy is buffered through a finite-capacity battery before being passed through the grid. One way of expressing a stationarity requirement is to ensure that the battery level will never go above or below certain levels (and that all arriving energy passes through the battery). (A similar requirement appears, for somewhat different reasons, in peak-shaving problems (Johnson et al., 2011).) Wind power is very significant on a year-to-year basis, but can have high variation over short time periods. Because it may therefore be impossible to ensure that such overflow/underflow situations never occur, some municipalities require of utilities that these unfavorable events happen only rarely, with at most some specified probability. Consistent with this requirement, the windmill operator must decide, over time, how much energy to sell to the grid when, with the goal of maximizing total profit.

For a second application area, consider dams on rivers. These also can be used to generate electrical power, which is sold to the grid. The lower bound to be satisfied now is put in terms of water level: in order to open a dam for recreational activities during the summer months, the water level must exceed a certain minimum. Again there is nondeterminism, due now to the amount of rainfall, and so the requirement is that the water level be violated only with a certain low probability. Consistent with that, the dam operator manages the dam, i.e., decides when and how much energy to produce and sell to the grid, which affects the water level, with the goal of maximizing profit.

ACKNOWLEDGEMENTS

This work has been motivated by joint work with Professors Guy Cohen (ENPC (Ecole nationale des ponts et chausées)), Pierre Carpentier (ENSTA (Ecole nationale supérieure des techiques avancées)), and Owen Jones (University of Melbourne), as well as with Dr. Laetitia Andrieu (OSIRIS, EDF (Electricité de France)).

REFERENCES

Altman, E. (1999). *Constrained Markov Decision Processes*. Chapman and Hall/CRC.

Arrow, K. and Hurwitz, L. (1957). Gradient methods for constrained maxima. *Operations Research*, 5(2), 258–265.

Bertsekas, D. (1999). *Nonlinear programming*. Athena Scientific, Belmont, MA.

Dolgov, D.A. and Durfee, E.H. (2005). Stationary deterministic policies for constrained mdps with multiple rewards, costs, and discount factors. In *IJCAI-05*, 1326–1332. Edinburgh, Scotland.

Johnson, M.P., Bar-Noy, A., Liu, O., and Feng, Y. (2011). Energy peak shaving with local storage. *Sustainable Computing*, 1(3), 165–256.

Puterman, M.L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

Ross, S.M. (2006). *Introduction to Probability Models, Ninth Edition*. Academic Press, Inc., Orlando, FL, USA.

Vázquez-Abad, F.J. and Mason, L.G. (1996). Adaptive decentralized control under non-uniqueness of the optimal control. *Discrete Event Dynamic Systems*, 6(4), 323–359.

Vázquez-Abad, F.J. and Mason, L.G. (1999). Decentralized adaptive flow control of high-speed connectionless data networks. *Journal of Operations Research*, 47, 928–942.