Robustness of policies in Constrained Markov Decision Processes

Alexander Zadorojniy and Adam Shwartz, *Senior Member, IEEE*

Abstract

We consider the optimization of finite-state, finite-action Markov Decision processes, under constraints. Cost and constraints are of the discounted type. We introduce a new method for investigating the continuity and robustness of the optimal cost and the optimal policy under changes in the constraints. This method is also applicable for other cost criteria such as finite horizon and infinite horizon average cost.

Index Terms

Markov Decision Processes, Constrained MDP, Discounted Cost, Sensitivity, Robustness.

I. Introduction

Consider the standard model of a Markov Decision Process (MDP) with finite state and action spaces. A natural generalization of the optimization problem is to include cost constraints. Such models arise in relation to resource-sharing systems. For example, in telecommunication networks which are designed to enable simultaneous transmission of different types of traffic: voice, file transfer, interactive messages, video, etc. Typical performance measures are transmission delay, power consumption, throughput, etc. [1]. A trade-off exists, for example, between minimizing delay and reducing power consumption: to minimize delay we should transmit with the highest possible power, since this increases the probability of successful transmission. Such problems are formulated as constrained MDP [2], where we wish to minimize the costs related to the delay subject to constraints on the average and peak power.

Work of Alexander Zadorojniy was performed while he was with the Faculty of Electrical Engineering, Technion, Israel Institute of Technology.

Adam Shwartz is with the Faculty of Electrical Engineering, Technion, Israel Institute of Technology. Work of this author was supported in part by the fund for promotion of research at the Technion.

While the continuity of the value with respect to the policy is a relatively simple matter, robustness is a more difficult issue. In general robustness means that under a small change in the parameters, the original policy still meets the requirements. This definition is not appropriate for optimization problems. Moreover, in constrained optimization the question arises as to whether under the new parameters the policy is required to meet the constrains, or perhaps is allowed to deviate by a small amount. Our definition of robustness is that a small change in some parameters requires a small change in the policy. Since the parameter we change is the constraints, we require that the new policy satisfies the new constraints, is optimal under the new constraints, and robustness means that the new policy is close to the original one.

Continuity of the optimizers does not hold in general. We develop a new technique to characterize and establish robustness with respect to changes in the values of the constraints. For a related study see [3], which deals with sensitivity of the cost and the policy to changes in the discount factor and in the transitions.

In Section II we introduce constrained MDP problems. Section III establishes the connection to linear programming and the formulation in terms of Karush-Kuhn-Tucker conditions. Section IV establishes the main results: we introduce a new method and give conditions so that an optimal policy is not sensitive to small enough changes in the constraint. Section V concludes our work.

Notation: For a vector v the notation $v \geq 0$ means that the inequality holds for each coordinate. Given a collection $\{v_k\}$ of constants or functions, we denote by $\overline{v}$ the vector with components v_k.

II. The Constrained Problem

A. The Model

A constrained Markov Decision process (CMDP) [2] is specified through a state space X and action space U, both assumed finite here, a set of transition probabilities $\{P_{yx}(u)\}$, where $P_{yx}(u)$ is the probability of moving from y to x when action u is taken, and immediate costs $c(x,u)$ and $d_k(x,u),\ k = 1,\dots,K$. It will be convenient to rename the states and actions so that $X = \{1,2,\dots,|X|\}$ and $U = \{1,2,\dots,|U|\}$.

We shall consider stationary policies π, which specify how actions are chosen. In a stationary policy, $\pi(u|y)$ is the probability for choosing action u if the process is in state y. The reason we restrict to stationary policies is given in Theorem 3 below.

A choice of initial (state) distribution and a policy thus define the discrete time stochastic process $(X_t,\ U_t),\ t = 1,2,\dots$ and its distribution. We denote the probability and expectation that correspond to the initial distribution σ and policy π by P^π_σ and E^π_σ respectively. Throughout this paper we fix the initial distribution, and fix a discount factor $0 < \beta < 1$. We shall therefore omit both σ and β from the notation. The discounted cost and the value of each constraint under π are then defined as

$$C(\pi) \stackrel{\triangle}{=} (1-\beta)E^{\pi}\sum_{t=1}^{\infty}\beta^{t-1}c(X_t,U_t),$$
$$D_k(\pi) \stackrel{\triangle}{=} (1-\beta)E^{\pi}\sum_{t=1}^{\infty}\beta^{t-1}d_k(X_t,U_t).$$

B. The Constrained Problem

Given a set of constraints $V_1, \dots, V_K$ the Constrained Optimization problem COP is

COP: Find π that minimizes $C(\pi)$

Subject to $D_k(\pi) = V_k,\ 1 \le k \le K.$

Remark 1: In Section IV-A we comment on the constrained problem, where the constraints are of the form $D_k(\pi) \le V_k$.

Remark 2: Note that for constrained problems, optimal policies generally depend on initial conditions (there may be no feasible policy for some initial conditions). This is the reason we fix the initial condition throughout.

III. Constrained Optimization and Linear Programming

The approach we take relies on a Linear Programming formulation for COP.

A. Occupation measures

An occupation measure corresponding to a policy π is the total expected discounted time spent in different state-action pairs. It is thus a probability measure over the set of state-action pairs. More precisely, define for any policy π and any pair (x,u)

$$f(\pi;x,u) \stackrel{\triangle}{=} (1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}P^{\pi}(X_t = x, U_t = u).$$

$f(\pi)$ is then defined to be the set $\{f(\pi;x,u)\}_{x,u}$. It can be considered as a probability measure, called the occupation measure, that assigns probability $f(\pi;x,u)$ to the pair (x,u). The discounted cost can be expressed as the expectation of the immediate cost with respect to this measure [2]:

$$C(\pi) = \sum_{x\in X}\sum_{u\in U} f(\pi;x,u)c(x,u) = f\cdot c\,, \tag{1}$$

where in the last equality we consider f and c as vectors. Analogue expressions hold for D_k.

Given a set R of policies, denote $L_R \stackrel{\triangle}{=} \{f(\pi) : \pi \in R\}$. Let Π denote the set of all policies, S the set of stationary policies and D the set of deterministic policies (that is, the probability of using an action is either 1 or 0). Let $\overline{co}$ denote the closed convex hull, that is, all convex combinations and their limits. Then

Theorem 3 ([2, Theorem 3.2]): $L_\Pi = L_S = \overline{co} L_D$.

Since by (1) all costs are linear in $f(\pi)$, the first equality in the theorem shows that the restriction to stationary policies does not influence the optimal value, so that it is reasonable to impose this restriction, as we do here.

B. Linear Programming formulation

Define Q to be the following set of $\rho = \{\rho(y,u)\}$

$$Q = \left\{ \rho : \left(\begin{array}{l} \displaystyle\sum_{y \in X} \sum_{u \in U} \rho(y,u)(\delta_x(y) - \beta P_{yx}(u)) \\ \qquad\qquad\qquad = (1-\beta)\sigma(x), \ \forall x \in X \\ \\ \rho(y,u) \geqslant 0, \forall y,u. \end{array} \right) \right\} \tag{2}$$

By summing the first constraint in (2) over x we obtain that $\sum_{y,u} \rho(y,u) = 1$ for each $\rho \in Q$, so that any ρ satisfying (2) can be considered as a probability measure. We regard ρ as either a set of $\rho = \{\rho(y,u)\}$ as defined above, or as a vector of length $|X| \cdot |U|$. Below we represent COP in terms of elements of Q. To complete the picture we need to derive a stationary policy that corresponds to each ρ. So, given ρ define

$$\mu_y(u) = \rho(y,u) \left(\sum_{u \in U} \rho(y,u) \right)^{-1}, \quad y \in X, u \in U \tag{3}$$

provided $\sum_{u \in U} \rho(y,u) \neq 0$ (if the sum is 0 for some y then necessarily $\rho(y,u) = 0$ for each u. In this case choose $\mu_y(u) \geq 0$ arbitrarily but so that $\sum_u \mu_y(u) = 1$.)

$$C^L(\rho) \stackrel{\triangle}{=} \sum_{x,u} c(x,u)\rho(x,u), \quad D_k^L(\rho) \stackrel{\triangle}{=} \sum_{x,u} d_k(x,u)\rho(x,u). \tag{4}$$

LP: Find ρ that minimizes $C^L(\rho)$

Subject to $D_k^L(\rho) = V_k, \ 1 \leq k \leq K$ and $\rho \in Q$.

The last constraint is linear by definition (2). Now we can state the equivalence between *COP* and the *LP*.

Theorem 4 ([2, Theorem 3.3]): Consider a finite CMDP.

- For any $f(\pi)$ there is a $\rho \in Q$ such that $\rho = f(\pi)$, and conversely for any $\rho \in Q$ the policy μ defined in (3) satisfies $\rho = f(\mu)$.
- LP is feasible if and only if COP is. Assume that COP is feasible. Then there exists an optimal solution ρ^* for LP, and the stationary policy μ (3) is optimal for COP.

Let us rewrite LP as a generic linear program:

$$LP_G : \qquad Minimize \;\; z \cdot \rho \tag{5}$$

$$Subject\ to\ A \cdot \rho = b \tag{6}$$

$$and \qquad \rho \geq 0 \tag{7}$$

Remark 5: To cast LP in this form, we let z represent the cost c (written as a vector), so that $z \cdot \rho = \sum_{x,u} z(x,u)\rho(x,u)$ is the cost. Next, the matrix A has $|X| \cdot |U|$ columns and $|X| + K$ rows, where $|X|$ rows represent the left-hand-side of the equality constraints of (2), and K rows represent the value of the constraints; that is, row k represents d_k, with $1 \leq k \leq K$. The corresponding $|X|$ entries of b are then given by $(1-\beta)\sigma(x)$, and the remaining K entries take the values V_k. Note that only b depends on the value of V.

C. Karush-Kuhn-Tucker conditions

We need a standard tool in the theory of linear programming—the Karush-Kuhn-Tucker (KKT) conditions:

KKT: There exist w and v so that

$$A \cdot \rho = b, \quad \rho \geq 0 \tag{8}$$

$$w \cdot A + v = z, \quad v \geq 0 \tag{9}$$

$$v \cdot \rho = 0. \tag{10}$$

Theorem 6 ([4, KKT Conditions]): ρ satisfies conditions (8)–(10) for some v and w if and only if it is an optimal solution of the Linear Programming problem (5)–(7).

IV. Robustness Analysis

In this section we show that an optimal policy is not sensitive to small enough changes in the constraints, provided the changes retain feasibility. We start with a simple continuity result. Consider a policy π as a vector of dimension $|X| \cdot |U|$. Define the distance between two policies:

$$|\pi - \pi'| = \delta \quad \text{means} \quad \sum_{x,u} |\pi(u|x) - \pi'(u|x)| = \delta.$$

Lemma 7: C and D are continuous in π.

Proof: Fix π. Given $\varepsilon > 0$ we need to show that there is a δ so that $|\pi - \pi'| \leq \delta$ implies $C(\pi) - C(\pi') \leq \varepsilon$. The proof for D_k is identical. First, fix N so that

$$(1-\beta)\sum_{t=N}^{\infty}\beta^{t-1}\max_{x,u}|c(x,u)| < \frac{\varepsilon}{4}. \tag{11}$$

$P_M(\pi)$ denotes the Markov transition matrix induced by π,

$$\{P_M(\pi)\}_{yx} = \sum_u P_{yx}(u)\pi(u|y).$$

Note that P_M is linear (hence continuous) in π. Now

$$E^{\pi}c(X_t, U_t) = \sum_{x,u} P^{\pi}(X_t = x, U_t = u)c(x,u).$$

But $P^{\pi}(X_t = x, U_t = u) = P^{\pi}(X_t = x)\pi(u|x)$ and

$$P^{\pi}(X_t = x) = \sum_y \sigma(y)P_M^t(\pi)_{yx}.$$

Thus, since $P_M(\pi)$ is linear in π, we have that $E^{\pi}c(X_t, U_t)$ is a polynomial of degree t in π and so

$$(1-\beta)E^{\pi}\sum_{t=1}^{N-1}\beta^{t-1}c(X_t, U_t)$$

is a polynomial in π, of degree at most $N-1$. This together with the approximation in (11) proves the continuity. ■

This continuity means that a small change in π entails a small change in C and D_k. However, suppose π is optimal for COP, and suppose π' is close to π. Define

$$V_k' = D_k(\pi'), \quad 1 \leq k \leq K.$$

Then by continuity, V_k' is close to V_k. But it is not difficult to construct examples in which π' is not optimal for problem COP with constraints V_k', regardless of how close V_k and V_k' are. There may be a better policy, and it may be quite far from π. This is a particular case of a general phenomenon: the minimizing point is in general not continuous in the parameters of the problem.

The following key Theorem gives conditions under which π' is in fact optimal.

Theorem 8: Let π_V be an optimal policy for COP. Define

$$U_V(x) \stackrel{\triangle}{=} \{u : \pi_V(u|x) = 0\}. \tag{12}$$

Let π' be any stationary policy such that $\pi'(u|x) = 0$ for all $u \in U_V(x)$. Denote $V'_k = D_k(\pi')$. Then π' is optimal for COP with constraints $V'_k,\ 1 \le k \le K$.

Remark 9: The condition on π' means that if π_V never uses action u in state x, then the same holds for π'. Thus π' differs from π only in the value of the randomization, at those states where π uses randomization.

The optimality means that $C(\pi') \le C(\pi)$ for any π that satisfies $D_k(\pi) = V'_k,\ 1 \le k \le K$.

Proof: Recall from Remark 5 that b is the only coefficient that depends on the value of $\overline{V}$: let us make this explicit using the notation b_V. A change in $\overline{V}$ does not change the matrix A or the vector z. To help the exposition, let us consider each $\rho \in Q$ as a a vector of dimension $|X| \cdot |U|$. Since π_V is optimal for COP, by Theorem 4 $\rho_V = f(\pi_V)$ is optimal for LP, and by Theorem 6 ρ_V satisfies the KKT conditions (8)–(10) for some v and w. Consider $\rho' = f(\pi')$. We claim that ρ' satisfies the KKT conditions (8)–(10) with constants b', and with the same v and w. b' is obtained by replacing the constraints V_k with V'_k.

Note first that since the elements of ρ_V are non-negative, Condition (10) holds if and only if $v(x, u)$ satisfies $v(x, u) = 0$ whenever $\rho_V(x, u) \neq 0$. Now ρ' satisfies Condition (8) since $\rho' \in Q$ and by definition $D^L_k(\rho') = V'_k$. Condition (9) is unchanged—it does not depend on ρ. As for the last condition, it suffices to show that $\rho'(x, u) = 0$ whenever $\rho_V(x, u) = 0$; since for other (x, u), we have $v(x, u) = 0$ by the optimality of ρ_V.

Using $\rho_V = f(\pi_V)$, it follows that if

$$\rho_V(x, u) = f(\pi : x, u) = P^{\pi_V}(X_t = x) \cdot \pi_V(u|x) = 0$$

then one of the following holds.

(i) $\pi_V(u|x) = 0$, that is, action u is never used in state x, or

(ii) $P^{\pi_V}(X_t = x) = 0$ for all t, that is, state x is never visited under π_V.

If (i) holds then, by (12), π', $\pi'(u|x) = 0$. Therefore $\rho'(x, u) = f(\pi : x, u) = 0$.

If (ii) holds then note that π' does not introduce any new transitions to the process: it merely changes the probability of transitions. But transitions that have probability 0 under π_V will also have the same probability under π'. Thus if $f(\pi_V : x, u) = \rho_V(x, u) = 0$ then $f(\pi' : x, u) = \rho'(x, u) = 0$ and the proof is complete. ∎

Clearly, the larger $U_V(x)$ is, the simpler it is to implement the policy. While not much can be said at each x, there is a general result on the combined size over all states. Let $\mathbf{1}[A]$ be the indicator of the event A: that is, it is equal 1 if A holds, and zero otherwise.

Theorem 10 ([2], Theorem 3.8): There exists an optimal policy π^* for COP so that the number of randomizations is at most K. That is,

$$\sum_x \left(\sum_u \mathbf{1} \left[\pi^*(u|x) > 0 \right] - 1 \right) \leq K.$$

In particular, if $K = 1$ then there is an optimal policy π^* which chooses one action in every state, except in one state, say x_0. This allows to say more about the case with one constraint.

Corollary 11: Consider the case $K = 1$. Let π^* be an optimal policy for COP, and suppose $\pi^*(u|x)$ is either 0 or 1 except at x_0 and that

$$\pi^*(u|x_0) = \begin{cases} q_V & \text{if } u = u' \\ 1 - q_V & \text{if } u = u''. \end{cases} \tag{13}$$

Let π^q denote the policy which agrees with π^*, except that at x_0 it chooses between u' and u'' with probability q and $1 - q$ respectively. Let

$$V_{min} \stackrel{\triangle}{=} \inf_{0 \leq q \leq 1} D_1(\pi^q), \quad V_{max} \stackrel{\triangle}{=} \sup_{0 \leq q \leq 1} D_1(\pi^q).$$

Then, for each $V_{min} \leq \alpha \leq V_{max}$ there is a q_α so that π^{q_α} is optimal for COP with constraint α.

Proof: By Lemma 7, $D_1(\pi^q)$ is continuous in q. The proof now follows from Theorem 8. ■

A. Inequality constraints

With MDP, constrained optimization with inequality constraints are more common. We now extend our results to this case. Define

COPi: Find π that minimizes $C(\pi)$

Subject to $D_k(\pi) \leq V_k,\ 1 \leq k \leq K$.

Let π_V be optimal for $COPi$ and suppose that

$$D_k(\pi_V) = V_k, \quad k \leq K_1 \quad \text{and} \quad D_k(\pi_V) < V_k, \quad k > K_1.$$

Lemma 12: π_V is optimal for problem $COPi$ with constraints $k \leq K_1$, and with the constraints for $k > K_1$ omitted.

The point is that the constraints that are not biding may be omitted, and optimality still holds. The proof is immediate and is omitted.

Recall now the definition (12) of U_V and define

$$\Pi_V = \{\pi \text{ stationary, } \pi(u|x) = 0 \text{ for all } u \in U_V(x), \\ D_k(\pi) \leq V_k,\ k > K_1\}. \tag{14}$$

By continuity, this set of policies is not empty, and contains all policies which are close enough to π_V and do not introduce new actions.

Recall the one-to-one correspondence between policies and occupation measures—Theorem 4. Let $\{\pi_i\}$ be the (finite) collection of all deterministic policies without any actions in U_V. By Theorem 3 for any $\pi \in \Pi_V$ we can write

$$f(\pi) = \sum_i \alpha_i f(\pi_i) \tag{15}$$

for some $\alpha_i \geq 0$ with $\sum_i \alpha_i = 1$. That is, $f(\pi)$ is a convex combination of occupation measures corresponding to deterministic policies.

Theorem 13: Let π' be any stationary policy in Π_V. Denote $V'_k = D_k(\pi')$ for $k \leq K_1$ and set $V'_k = V_k$ for $k > K_1$. Then π' is optimal for COPi with constraints $V'_k,\ 1 \leq k \leq K$.

Note that $D_k(\pi') \leq V_k$ for $k > K_1$ by definition.

Proof: Let us represent π_V using (15) with the coefficients $\{\alpha_i\}$ and π' with the coefficients $\{\alpha'_i\}$. Define $\gamma = \min_i\{\alpha_i/\alpha'_i\}$ and note that $\gamma < 1$ and so $\gamma\alpha'_i \leq \alpha_i$ for all i. Recall that each occupation measure corresponds to a ρ in Q (Equation (2)), which is convex.

If π' is not optimal, then there exists some $\tilde{\pi}$ so that $D_k(\tilde{\pi}) \leq V'_k$ for all k, and $C(\tilde{\pi}) < C(\pi')$. Note that

$$\rho \triangleq \gamma \left(f(\tilde{\pi}) - \sum_i \alpha'_i f(\pi_i) \right) + f(\pi_V) \tag{16}$$

$$= \gamma f(\tilde{\pi}) + \sum_i \left(\alpha_i - \gamma\alpha'_i\right) f(\pi_i) \tag{17}$$

is in Q. This is the case since $\alpha_i - \gamma\alpha'_i \geq 0$ and $\gamma + \sum_i(\alpha_i - \gamma\alpha'_i) = 1$, so that ρ is a convex combination of $f(\tilde{\pi})$ and the $f(\pi_i)$. From ρ, define μ through (3). Now by (16) and Theorem 4, for $k \leq K_1$

$$D_k(\mu) = \gamma \left(D_k(\tilde{\pi}) - D_k(\pi')\right) + D_k(\pi_V) \tag{18}$$

$$\leq D_k(\pi_V) \tag{19}$$

since for such k we have $D_k(\mu) \leq V'_k = D_k(\pi')$. For $k > K_1$ we have that $D_k(\pi_V) < V_k$ and so, by making γ smaller if necessary, we obtain $D_k(\mu) \leq V_k$ is this case as well. Thus we conclude that μ is feasible for the constraints V. Now

$$C(\mu) = \gamma \left(C(\tilde{\pi}) - C(\pi')\right) + C(\pi_V) \tag{20}$$

$$< C(\pi_V), \tag{21}$$

by assumption, a contradiction to the optimality of π_V. ■

V. Conclusions

We introduced a new method to establish robustness of policies in constrained MDPs. The method is clearly applicable to finite-horizon problems, and is also applicable to the average cost problem under some recurrence conditions. With a small change in the values of the constrains, only a small number of parameters need to be adjusted in order to retain optimality. This method was applied to telecommunication networks in [7].

References

[1] E. Altman, Applications of Markov Decision Processes in Communication Networks, in *Handbook of Markov Decision Processes: Methods and Applications*, E. Feinberg and A. Shwartz Eds., pp. 489–536, Kluwer, Boston, 2002.

[2] E. Altman, Constrained Markov Decision Processes, Chapman&Hall/CRC,1998.

[3] E. Altman and A. Shwartz, Sensitivity of Constrained Markov Decision Processes, Ann. Operations Research 32pp 1-22, 1994.

[4] M.S. Bazaraa, J.J. Jarvis, H.D. Sherali, ”Linear Programming and Network Flows”, John Wiley and Sons, 1990.

[5] D.P. Bertsekas, Dynamic Programming and Optimal Control, Athena Scientific, Belmont, Massachusetts, 1995.

[6] Martin L. Puterman, Markov Decision Processes, John Wiley&Sons, 1994.

[7] A. Zadorojniy, Constrained Markov Decision Processes with Application to Wireless Communications, M.Sc. Thesis, Electrical Engineering, Technion, 2004. Available at http:\\www.ee.technion.ac.il/˜adam/Students .