Dynamic Programs as Convex Optimization Problems

Tunç Şimşek, Laurent El-Ghaoui

November 30, 1999

Abstract

It is well known that the Bellman Equation may be cast as a convex optimization problem. Specifically, when there are finite number of states and control inputs the resulting problem is a linear program. We survey other cases that result in tractable convex optimization problems and propose approximations when a tractable formulation does not exist. Special attention is given to approximations by semi-definite programs.

1 Introduction

We consider dynamic systems modeled as discrete-time controlled Markov Decision Processes (MDP). An MDP is the tuple

$$M = (X, U, p, c) \tag{1}$$

where: X is the state space, U the set of control inputs, $p : X \times X \times U \to [0,1]$ the conditional transition probabilities and $c : X \times U \to \mathbb{R}$ the cost function.[1]

Given an MDP the goal is to minimize the expected infinite horizon discounted cost:

$$J = \mathrm{E}\sum_{k=0}^{\infty} \beta^k c\big(x(k), u(k)\big), \tag{2}$$

where $x(k)$ is the state of the system at time k and $u(k)$ is the control input applied at time k. We denote the optimal (i.e. minimal) average cost achievable with J^*. The Bellman equation

$$V(x) = \min_{u\in U}\big\{c(x,u) + \beta(x,u)\sum_{x'\in X} p(x' \mid x, u)V(x')\big\} \tag{3}$$

relates a *value*[2] $V(x)$ to each state $x \in X$ so that:

$$V\big(x(0)\big) = \mathrm{E}\Big(\sum_{k=0}^{\infty} c\big(x(k), u(k)\big)\Big| x(0)\Big) \tag{4}$$

and

$$\mathrm{E}\, V\big(x(0)\big) = J^* \text{ w.p. } 1. \tag{5}$$

Computation of the Bellman equation 3 is referred to as *dynamic programming*.

[1] If we are interested in a reward then we replace C with R to emphasize this fact.

[2] Also called the *cost-to-go* function.

2 Semi-Markov Decision Processes

A discrete-time controlled Semi-Markov Decision Process (SMDP) is the tuple

$$S = (X, U, p, c, F) \tag{6}$$

where X, U, p, c are defined exactly as in 1 and $F : \mathbb{R} \times X \times U \to [0, 1]$ is a probability disrtibution. $F(t \mid x, u)$ is the probabilty that a transition from state x will be taken within t seconds when control u is applied. When a transition is taken, the SMDP behaves exactly like an MDP.

Example. *Let $F(t \mid x, u) \equiv F(t) = 1(t > 0)$. Then, the SMDP reduces to an MDP.*

In [1, 2] SMDP's were shown to arise naturally in hierarchical control of MDP's. Essentially an SMDP is seen as a temporal abstraction of an MDP with a collapsed state space and scaled time axis. The temporal abstraction is reflected by modifying the discount factor, β, in the discounted cost 2. Namely, w.l.o.g., the Bellman equation for an SMDP becomes:

$$V(x) = \min_{u \in U} \{c(x, u) + \beta(x, u) \sum_{x' \in X} p(x' \mid x, u)V(x')\} \tag{7}$$

where

$$\beta(x, u) = \int_{t=0}^{\infty} \beta^t dF(t \mid x, u). \tag{8}$$

Again, 4 applies. Thus, as far as optimization is concerned, we need only be concerned with the solution of the Bellman equation 7. The rest of this paper is devoted to the solution of 7 by convex optimization techniques.

3 Basic formulation: Convex optimization

A great deal of research is devoted to efficient dynamic programming; i.e. efficient computation of the value function V in the Bellman equation 7. This computation is compactly represented as a linear program (LP) in the decision variables $x \in X; V(x)$ when the state space X and the control inputs U are finite:

$$\begin{aligned} &\text{maximize} \sum_{x \in X} V(x) \\ &\text{subject to} \\ &\qquad \forall x \in X,\, u \in U: \\ &\qquad\qquad V(x) \le c(x, u) + \beta(x, u) \sum_{x' \in X} p(x' \mid x, u)V(x'). \end{aligned} \tag{9}$$

Here we will be concerned with 9 when U is finite but large, non-finite U, uncertain p and uncertain F. Note that in all cases, the optimization problem 9 remains convex however intractable.

4 Survey of basic approximations

Consider the convex optimization problem 9 where $U \subset \{0,1\}^m$ with large m. The problem will be an LP with $O(2^m)$ constraints. This problem quickly becomes intractable as m increases. Approximations of large LP's by semi-definite programs (SDP's) are well known in the literature. **Need example here.**

Now let U be non-finite. Consider the feasability set of the LP 9:

$$\begin{aligned}&\forall x \in X,\, u \in U: \\ &\qquad V(x) \leq c(x,u) + \beta(x,u) \sum_{x' \in X} p(x' \mid x,u) V(x').\end{aligned}$$

Fix x. Observe that $V(x)$ is feasible if and only if the inequality holds for all $u \in U$. Specifically, if U is a polytope (e.g. a k-cell) and $c(x,\cdot)$ and $\forall x' \in X$; $\beta(x,\cdot)p(x' \mid x,\cdot)$ are concave in u then $V(x)$ is feasible if and only if the inequality holds on the vertices of the polytope. Thus, in this case, the problem reduces to an LP whose complexity is linear in the number of vertices of U[3].

Now take p uncertain. If the uncertainity is polytopic then this problem may be cast as that of uncertain U with $\beta(x,\cdot)p(x' \mid x,\cdot)$ linear in u.

Finally, consider uncertainities in F. The nature of the resulting problem is not immediate. We must make simplifying assumptions on the quality of the uncertainity and the next section proceeds to do so.

5 Uncertain waiting distributions

Consider the discount factor $\beta(x,u)$. We notice that this is nothing but $\mathrm{E}(\beta^t \mid x,u)$. Supposing that $dF(t \mid x,u)$ is concentrated around the mean, t_0, of the distribution we have a good approximation of $\beta(x,u)$:

$$\begin{aligned}\beta(x,u) &= \int_{t=0}^{\infty} \beta^t dF(t \mid x,u) \\ &\approx \beta^{t_0} + \left(\beta^t\right)'' \frac{\sigma(x,u)^2}{2} + \cdots + \left(\beta^t\right)^{(n)} \frac{\mu_n(x,u)}{n!} \\ &= \mu(x,u)^T b.\end{aligned} \tag{10}$$

Thus we obtain $\beta(x,u)$ as a linear function of the central moments of F:

$$\begin{gathered}\mu(x,u) = \left(\mu_0(x,u)\mu_1(x,u)\ldots\mu_n(x,u)\right)^T \mu_n(x,u) \equiv \mathrm{E}\Big\{(t-t_0)^n \;\Big|\; x,u\Big\} \\ \mu_0(x,u) = 1, \quad \mu_1(x,u) = 0, \quad \mu_2(x,u) = \sigma(x,u)^2.\end{gathered}$$

[3]The number of vertices is combinatorial in the number of facets of U.

Now, we may formulate LP 9 as a robust optimization problem:

$$\begin{aligned} &\text{maximize } 1^T V \\ &\text{subject to} \\ &\qquad \forall x \in X,\, u \in U: \\ &\qquad \max_{\substack{H(x,u,\mu)\succeq 0 \\ |\mu_i(x,u)|\le m_i(x,u)}} \left\{ \Big(\varepsilon(x,u)^T - \big(\mu(x,u)^T bP(x,u)^T\big)\Big) V - c(x,u) \right\} \le 0, \end{aligned} \tag{11}$$

where

$$\begin{aligned} \varepsilon(x,u) &= (0,\ \ldots,\ 0,\ 1,\ 0,\ \ldots,\ 0)^T \\ P(x,u) &= \big(\ldots,\ p(x' \mid x,u),\ \ldots\big)^T \end{aligned}$$

and the Hankel matrix

$$H(x,u,\mu) = \begin{pmatrix} \mu_0(x,u) & \mu_1(x,u) & \ldots & \mu_n(x,u) \\ \mu_1(x,u) & \mu_2(x,u) & & \\ \vdots & & \ddots & \vdots \\ \mu_n(x,u) & & \ldots & \mu_{2n}(x,u) \end{pmatrix}.$$

To avoid notational confusion, consider a single inequality in the robust optimization 11 and drop the arguments x,u; i.e. fix x,u:

$$\max_{\substack{H(\mu)\succeq 0 \\ |\mu_i|\le m_i}} \left\{ \big(\varepsilon - (\mu^T bP^T)\big) V - c \right\} \le 0. \tag{12}$$

First of all, note that this constraint is convex; it is the point-wise maximum of a set of convex constraints. Thus problem 11 is a convex optimization problem. However, it is not an LP, nor is it tractable. As an approximation of the problem consider Lagrangian relaxations:

$$\begin{aligned} &\exists \lambda \succeq 0,\ S \succeq 0 \text{ so that } \forall \mu \\ &\qquad \varepsilon^T V - c - \mu^T bP^T V \le \sum_{i=1,\ldots,n} \lambda_i ({\mu_i}^2 - {m_i}^2) - \mathrm{Trace}\big(SH(\mu)\big). \end{aligned} \tag{13}$$

To see this, suppose there exists such λ and S. In particular, when $H(\mu) \succeq 0$ and $i = 1, \ldots, n$; $|\mu_i| \le m_i$ the r.h.s. of 13 is negative so that 13 implies 12. Thus, maximizing $1^T V$ w.r.t. 13 in the optimization variables λ, S and V will provide a lower bound for 11. On the other hand, 13 may be written a linear matrix inequality (LMI). Taking $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$

and $m = (m_1, \dots, m_n)^T$:

$$\forall \mu\colon$$
$$\varepsilon^T V - c - \mu^T b P^T V \leq \sum_{i=1,\dots,n} \lambda_i({\mu_i}^2 - {m_i}^2) - \mathrm{Trace}\big(SH(\mu)\big)$$
$$\Longleftrightarrow$$
$$\forall \mu\colon$$
$$\begin{pmatrix}\mu\\1\end{pmatrix}^T \begin{pmatrix} -\Lambda & \frac{1}{2}\big(H^*(S) - bP^T V\big) \\ \frac{1}{2}\big(H^* - bP^T V\big)^T & m^T \Lambda m + \varepsilon^T - c \end{pmatrix} \begin{pmatrix}\mu\\1\end{pmatrix} \leq 0,$$

and the last condition is satisfied if and only if:

$$\begin{pmatrix} -\Lambda & \frac{1}{2}\big(H^*(S) - bP^T V\big) \\ \frac{1}{2}\big(H^*(S) - bP^T V\big)^T & m^T \Lambda m + \varepsilon^T V - c \end{pmatrix} \preceq 0. \tag{14}$$

Here, H^* is the dual map of H:

$$H^* = (\mathrm{Trace}\, SH_0,\ \mathrm{Trace}\, SH_1,\ \dots,\ \mathrm{Trace}\, SH_n)^T$$

where

$$H_0 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & 0 \end{pmatrix},\ H_1 = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & & & \\ 0 & & & & \\ \vdots & & & \ddots & \vdots \\ 0 & & & \dots & 0 \end{pmatrix},\ \dots,\ H_n = \begin{pmatrix} 0 & \dots & 0 & 1 \\ \vdots & & 1 & 0 \\ 0 & 1 & & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix}.$$

Notice that LMI 14 makes no reference to μ. Thus, we may re-write the robust optimization problem 11 as an SDP in Λ, V and S:

maximize $1^T V$

subject to

$$\Lambda \succeq 0$$
$$S \succeq 0$$
$$\forall x \in X,\ u \in U\colon$$
$$\begin{pmatrix} -\Lambda & \frac{1}{2}\big(H^*(S) - bP^T(x,u) V\big) \\ \frac{1}{2}\big(H^*(S) - bP^T(x,u) V\big)^T & m(x,u)^T \Lambda m(x,u) + \varepsilon^T(x,u) V - c(x,u) \end{pmatrix} \preceq 0. \tag{15}$$

The interpretation of the robust SDP 15 is that we are computing an upper bound on the minimal cost for an SMDP whose waiting distribution, F, is uncertain. The uncertainity of F is given by bounds on its first n central moments μ. The LMI $H(\mu(x,u)) \succeq 0$ ensures that there exists a probability distribution $F(t \mid x, u)$ that generates the moments $\mu(x,u)$.

6 Conclusions

We have summarized results on basic approximations to uncertainities in the Bellman equation and derived an SDP for the case of uncertain waiting distributions. The impact of the latter is the yield of a computationally plausible abstraction framework for hierarchically controlled Markov decision processes. Specifically, the uncertainity of the waiting distribution, i.e. the temporal abstraction, is given in terms of a meaningful quantity: the moments of an unknown distribution.

Intuition suggests that casting dynamic programs as convex optimization problems will be useful also in the game theoretic framework. As a simple example consider the *minimax* matrix problem on the probability simplex:

$$\min_{\substack{1^T x = 1 \\ x \succeq 0}} \max_{\substack{1^T y = 1 \\ y \succeq 0}} x^T A y. \tag{16}$$

We will illustrate that this is equivalent to an LP. Consider the dual of the *maximization* problem:

$$\begin{aligned} &\text{minimize } \mu \\ &\text{subject to} \\ &\qquad \mu 1^T - x^T A \succeq 0. \end{aligned}$$

The dual achieves the optimal value for the primal problem, $\max x^T A y$, if and only if either the primal or the dual is feasible. In this case the minimax problem 16 may be rewritten as the LP:

$$\begin{aligned} &\text{minimize } \mu \\ &\text{subject to} \\ &\qquad \mu 1^T - x^T A \succeq 0 \\ &\qquad x \succeq 0 \\ &\qquad 1^T x = 1. \end{aligned}$$

The simplicity of this derivation suggests further research to the view of dynamic programs as convex optimization problems. Specifically, for the derivation of tractable problems for non-trivial abstractions, perturbations and approximations of dynamic programs and game-theoretic problems. As a concluding remark we refer the reader to [3, 4] for a comprehensive treatment of SDP's in the analysis of perturbed dynamic systems.

References

[1] J. Forestier and P. Varaiya. Multilayer Control of Large Markov Chains. *IEEE Trans. Automatic Control*, AC-23(2):298–304, April 1978.

[2] R. E. Parr. *Hierarchical Control and Learning for Markov Decision Processes*. PhD thesis, University of California at Berkeley, Berkeley, California, 1998.

[3] S. P. Boyd and L. Vandenberghe. Introduction to convex optimization with engineering applications. Lecture notes for EE364, Stanford University, 1996. Available via anonymous ftp into `isl.stanford.edu`, under `/pub/boyd`.

[4] L. El Ghaoui. Convex optimization and applications. Lecture notes for EE290N, University of California at Berkeley, 1999. Available at `www.eecs.berkeley.edu/~elghaoui`.