Finite State Approximations for Robust Markov Decision Processes

Kerem Uğurlu

Tuesday 10^{th} December, 2019

Abstract

We give a finite state approximation scheme to countable state infinite horizon robust Markov decision processes, whose transition density is not known with certainty. A convergence theorem to the optimal stationary policy along with the corresponding rate for this approximation is established. Our results show a fundamental difference between the finite state approximations for Markov decision processes whose transition densities are known with certainty and those with uncertain transition densities. In the latter case, depending on the level of uncertainty, the discount factor $0 < \beta < 1$ must be small enough to compensate for it, otherwise the convergence can not be guaranteed. This is not to be seen in finite state approximations for Markov decision processes whose transition kernel is given in advance.

Keywords: Robust optimization; Finite Approximations for Markov Decision Processes.

1 Introduction

Markov Decision Processes (MDP's) with expectation criteria are well-studied in the literature (see [5, 6] for an extensive treatment on the subject). However, using the expectation operator for the evaluation necessitates that the underlying transition density is known in advance. In particular, the decision maker assumes there is no ambiguity in conditional distributions of controlled Markov chain, i.e. the decision maker is *ambiguity-neutral*. On the other hand, we follow an *ambiguity-averse* approach, where we have a reference transition density $p(y|x,a)$ on the decision model (S, A, p, r, β) and an uncertainty ball with a fixed radius around $p(y|x,a)$. This implies that we have an estimate $p(y|x,a)$ but diffident about the underlying dynamics and want to evaluate the system in a *robust* way by considering alternative transition densities. Based on this robust framework, we give a finite state approximation scheme to a countable state ambiguity-averse controlled Markov chain in infinite horizon with discounted bounded rewards. We show the existence of the optimal stationary policy, and a finite scheme approximation to the optimal stationary policy is also provided. We observe a fundamental difference between approximations of ambiguity-neutral and ambiguity-averse Markov decision processes. In the former case, the discount factor $0 < \beta < 1$ is sufficient for the convergence of the finite approximations to the optimal value, whereas in the latter, this is not the case. In particular, in the latter case, the uncertainty in the underlying transition density should not be too large to cancel the discount factor β. Equivalently, for a given ambiguity-awareness level, the discount factor β should be small enough to compensate for it.

2 Decision Model

We first briefly summarize our decision model. We let (S, A, p, r, β) be the decision model, where $S = \{0, 1, 2, \ldots\}$ is the countable state space, and $A = \{0, 1, 2, \ldots, N\}$ is the finite action space. We denote the set of all probability measures on S and on A as $\mathcal{P}(S)$ and $\mathcal{P}(A)$, respectively. The initial state $X_0 = x_0 \in S$ is known and nonrandom. If action $a \in A$ is chosen in state $s \in S$, then the transition density is the conditional distribution on S denoted by $p(\cdot|s, a)$ for $(s, a) \in S \times A$. We condense these conditional probability distributions to a transition kernel $P_{sa} \in [\mathcal{P}(S)]^{S\times A}$, where $P_{sa} := p(\cdot|s, a)$ for $(s, a) \in S \times A$. The decision maker's reward is denoted by $r(x, a, y)$ for a chosen action $a \in A$ in state $x \in S$ and the subsequent state is $y \in S$. The MDP is controlled using $\pi = (\pi_n)_{n\geq 0}$, where $\pi_n : (S \times A)^n \times S \to \mathcal{P}(A)$ with $\pi_n(\cdot|s_0, a_0, s_1, a_1, \ldots, s_n)$ represents the conditional probability measure over A. The subsequent action is chosen according to $\pi_n(\cdot|s_0, a_0, \ldots, s_n)$, if the state-action history is given by $(s_0, a_0, \ldots, s_n)$. We denote the set of all policies by Π. A policy $\pi \in \Pi$ is called Markov, if $\pi_n(\cdot|s_0, a_0, s_1, a_1, \ldots, s_n) = \pi(\cdot|s_n)$ for all $n \geq 0$. A Markov policy is deterministic, if $\pi_n(s_n) = a_n$ for some mapping $\pi_n : S \to A$ for all $n \geq 0$. A deterministic Markov policy $\pi = (\pi_n)$ is called stationary, if there is a function $f : S \to A$ such that $\pi_n \equiv f$ for all $n \geq 0$ such that $\pi = (f, f, \ldots)$. In this paper, we take that only Markov deterministic policies are admissible and with some abuse of notation, we denote them as Π, as well. With the transition kernel P, $\pi \in \Pi$ induces stochastic process $(s_n, a_n)_{n\geq 0}$ on the space $(S \times A)^\infty$, whose existence is guaranteed by Ionescu-Tulcea theorem (see [5]). We use the notation $\mathbb{E}^{P,\pi}[\cdot|s_n, a_n]$ to denote conditional expectation with respect to probability measure P and policy π conditioned on (s_n, a_n). Finally, $0 < \beta < 1$ is the discount factor applied at each time in infinite time horizon.

2.1 Uncertainty Ball

We assume that for any $\pi \in \Pi$, there exists a reference transition kernel $P \in [\mathcal{P}(S)]^{S\times A}$. In particular, if action $a \in A$ is chosen in state $x \in S$, conditional on (x, a), $p(\cdot|x, a)$ defines a probability measure on S. We further consider the alternative probability measures $Q \in [\mathcal{P}(S)]^{S\times A}$ that are absolutely continuous with respect to $p(\cdot|x, a)$, in particular $q(\cdot|x, a) = 0$, whenever $p(\cdot|x, a) = 0$ for $(x, a) \in S \times A$, and take that for any $(x, a) \in S \times A$, the corresponding density $q(\cdot|x, a)$ satisfies $\sup_{y\in S} \frac{q(y|x,a)}{p(y|x,a)} \leq C_{\mathcal{K}}$ for $C_{\mathcal{K}} > 1$ with $\sum_{y\in S} q(y|x, a) = 1$. We denote the set $\mathcal{K}_{xa}$ as the set of probability measures on S conditioned on $(x, a) \in S\times A$ with

$$\mathcal{K}_{xa} \triangleq \left\{ q(\cdot|x, a) \in \mathcal{P}(S) : \sup_{(x,a,y)} \frac{q(y|x, a)}{p(y|x, a)} \leq C_{\mathcal{K}} \right\} \tag{2.1}$$

Remark 2.1. *Our method does not assume there exists a true unknown transition density P^0. Instead, for any fixed π, we assume there exists a set of alternative transition densities of marginal sets $\mathcal{K}_{xa}$ in some ball of uncertainty with a pre-specified radius $C_{\mathcal{K}}$ with respect to a reference transition density $P \in [\mathcal{P}(S)]^{S\times A}$ as in* (2.1)*, and we evaluate the MDP with the worst case transition density among $\mathcal{K}_{xa}$ for each state-action pair $(x, a) \in S \times A$.*

2.2 Value Function

We define the value function as

$$\begin{aligned} v(\pi, x_0) := \lim_{T\to\infty} \rho_0 \Big(& r(x_0, a_0, X_1) + \beta\rho_1\big(r(x_2, a_2, X_3) \\ & + \ldots + \beta\rho_{T-1}(r(x_T, a_T, X_{T+1})\big) \ldots \Big), \end{aligned} \tag{2.2}$$

where $\rho_n(r(x_n, a_n, X_{n+1}))$ for $n \geq 0$ are defined as

$$\begin{aligned} \rho_n(r(x_n, a_n, X_{n+1})) &\triangleq \inf_{q\in\mathcal{K}_{xa}} \sum_{y\in S} r(x_n, a_n, y) \frac{q(y|x_n, a_n)}{p(y|x_n, a_n)} p(y|x_n, a_n) \\ &= \inf_{q\in\mathcal{K}_{xa}} \sum_{y\in S} r(x_n, a_n, y) q(y|x_n, a_n) \\ &= \inf_{q\in\mathcal{K}_{xa}} \mathbb{E}^{q,\pi}[r(x_n, a_n, X_{n+1})|(x_n, a_n)] \end{aligned} \tag{2.3}$$

Finally, the optimization problem reads as

$$v^*(x_0) = \sup_{\pi\in\Pi} v(\pi, x_0). \tag{2.4}$$

We note that $C_{\mathcal{K}} = 1$ in (2.1) corresponds at each $(x,a) \in S \times A$ to the single transition kernel $\mathcal{K}_{xa} = \{p(\cdot|x,a) \in \mathcal{P}(S)\}_{(x,a)\in S\times A}$. Hence, $\rho_n(\cdot)$ becomes the conditional expectation $\mathbb{E}^{p,\pi}[\cdot|x_n, s_n]$ in (2.2), where the expectation is taken with respect to the conditional probability measure $p(\cdot|x_n, a_n)$, and the problem turns into the classical performance evaluation using expectation operator (see e.g. [4, 7]).

Remark 2.2. *The framework can also be applied for cost minimization problems. In that case, the decision maker would look for a cost minimizing policy among the alternative maximizing transition densities. In particular, the results of this manuscript are valid via a sign change (see Example 3.1).*

Remark 2.3. *The construction in* (2.3) *corresponds to a general family of operators called dynamical coherent risk measures that is first introduced for one period case in [1], and later extended to multiperiod case via nested static operators conditioned on the history applied at each time epoch ([2, 3, 8]). Generally, they are defined as follows. Let Ω be an abstract space and $(\mathcal{F}_n)_{n\geq 0}$ be the filtration on Ω with $\mathcal{F}_0 = \{\Omega, \emptyset\}$. Let $(Z_n)_{n\geq 0}$ be an $\mathcal{F}_n$ adapted sequence of bounded random variables on some measurable space $(\Omega, \mathcal{F}_n, \mathbb{P})$. Then, dynamic coherent risk measures are operators that are of the form $\rho_n : \mathcal{F}_{n+1} \to \mathcal{F}_n$ such that for a given $\mathcal{F}_{n+1}$ measurable random variable Z_{n+1}, $\rho_n(Z_{n+1})$ is measurable with respect to $\mathcal{F}_n$. $(\rho_n)_{n\geq 0}$ are monotone, translation invariant, positive homogeneous and convex (See [10], Chapter 6). In particular, the dynamic operator in* (2.3) *corresponds to a variant of dynamic Average-Value-at-Risk. We refer the reader to [3, 9] for further exposure in this direction.*

3 Finite Dimensional Scheme

Assumption 3.1. *We put the following assumptions on the decision model (S, A, p, r, β).*

(A1) $\sup_{(x,a,y)\in S\times A\times S} |r(x,a,y)| \leq R < \infty$.

(A2) $\epsilon(n) := \sup_{(x,a,y)\in S\times A\times S} \sum_{y>n} p(y|x,a) \to 0,$ *as* $n \to \infty$.

(A3) The discount factor β and $C_{\mathcal{K}}$ in (2.1) *satisfy* $0 < \beta C_{\mathcal{K}} < 1$.

Next, we give the finite approximations denoted by v_n to v^* in (2.4) along with the optimal policies f_n on v_n, as follows.

$$
\begin{aligned}
&v_{-1} \equiv 0 \\
v_n(x) &:= \sup_{a\in A(x)} \left\{ \inf_{q\in\mathcal{K}_{xa}} \sum_{y<n} q(y|x,a)(r(x,a,y)+\beta v_{n-1}(y)) \right\}, \text{ if } x \le n. \qquad (3.5)\\
&:= 0 \text{ if } x > n, n \ge 0. \\
f_n(x) &:= \arg\max_{a\in A(x)} \left\{ \inf_{q\in\mathcal{K}_{xa}} \sum_{y<n} q(y|x,a)(r(x,a,y)+\beta v_{n-1}(y))(r(x,a,y)+\beta v_{n-1}(y)) \right\}, \text{ if } x \le n, \qquad (3.6)\\
&:= \text{ arbitrary point in } A(x) \text{ if } x > n, n \ge 0.
\end{aligned}
$$

We denote for any mapping u on S with $n \in \mathbb{N}$

$$
\begin{aligned}
\|u\|_n &:= \sup_{x\le n} |u(x)|, \\
\|u\| &:= \sup_{x\in S} |u(x)|.
\end{aligned}
$$

Theorem 3.1. *Under Assumption 3.1, the sequence* $\{v_n(\cdot)\}$ *in* (3.5) *converges to* $v^*(\cdot)$ *uniformly in* $x \le n$. *Namely, for some* C *independent of* n, *we have*

$$
\|v^* - v_n\|_n \le C \max\{C_{\mathcal{K}} \beta^{\lfloor n/2 \rfloor}, \epsilon(\lfloor n/2 \rfloor)\} \qquad (3.7)
$$

Proof. For any f and g real-valued and bounded functions on S, we have

$$
|\sup_{x\in S} f(x) - \sup_{y\in S} g(y)| \le \sup_{x\in S} |f(x)-g(x)|.
$$

Let $u = -f$ and $v = -g$ to conclude that

$$
|\inf_{x\in S} u(x) - \inf_{y\in S} v(y)| \le \sup_{x\in S} |u(x)-v(x)|.
$$

By Assumption 3.1, v^* and v_n are uniformly bounded for all n. Then

$$
\begin{aligned}
C_1 &:= \frac{R}{1-\beta} \qquad (3.8)\\
\|v^*\| \le C_1, &\text{ and } \|R\sum_{k=0}^{n} \beta^k\| \le C_1.
\end{aligned}
$$

Hence, for $x \leq n+1$, we have

$$
\begin{aligned}
&|v_{n+1}(x) - v^*(x)| \\
&\quad = |\sup_{a\in A(x)} \big\{ \inf_{q\in\mathcal{K}_{xa}} \sum_{y\leq n} q(y|x,a)(r(x,a,y) + \beta v_n(y))\big\} \\
&\qquad\quad - \sup_{a\in A(x)} \big\{ \inf_{q\in\mathcal{K}_{xa}} \sum_{y\in S} q(y|x,a)(r(x,a,y) + \beta v^*(y))\big\}| \\
&\leq \sup_{a\in A(x)} | \inf_{q\in\mathcal{K}_{xa}} \sum_{y\leq n} q(y|x,a)(r(x,a,y) + \beta v_n(y)) \\
&\quad - \inf_{q\in\mathcal{K}_{xa}} \sum_{y} q(y|x,a)(r(x,a,y) + \beta v^*(y))| \\
&\leq \beta \sup_{a\in A(x)} \sup_{q\in\mathcal{K}_{xa}} \sum_{y\leq n} q(y|x,a)|v_n(y) - v^*(y)| \\
&\quad + \beta \sup_{a\in A(x)} \sup_{q\in\mathcal{K}_{xa}} \sum_{y>n} q(y|x,a)|v^*(y)| \\
&\leq \beta \sup_{a\in A(x)} C_{\mathcal{K}} \sum_{y\leq n} p(y|x,a)|v_n(y) - v^*(y)| \\
&\quad + \beta \sup_{a\in A(x)} C_{\mathcal{K}} \sum_{y>n} p(y|x,a)|v^*(y)|
\end{aligned}
$$

Hence, we get

$$\|v_{n+1} - v^*\|_{n+1} \leq C_{\mathcal{K}}\beta\|v_n - v^*\|_n + C_{\mathcal{K}}\beta\|v^*\|\epsilon(n)$$

By iterating, we get for $m \geq 1$

$$
\begin{aligned}
\|v_{n+m} - v^*\|_{n+m} &\leq (C_{\mathcal{K}}\beta)^m\|v_n - v^*\|_n + \|v^*\| \sum_{k=1}^{m} (C_{\mathcal{K}}\beta)^k \epsilon(n+m-k) \\
&\leq (C_{\mathcal{K}}\beta)^m\|v^n - v^*\| + \epsilon(n)\|v^*\| \sum_{k=1}^{m} (C_{\mathcal{K}}\beta)^k \\
&\leq 2C_1(C_{\mathcal{K}}\beta)^m + \epsilon(n)C_1\frac{C_{\mathcal{K}}\beta}{1 - C_{\mathcal{K}}\beta}.
\end{aligned}
$$

where C_1 is as in (3.8). Thus,

$$\|v_{n+m} - v^*\|_{n+m} \leq 2\max\{2C_1, C_1\frac{C_{\mathcal{K}}}{1 - C_{\mathcal{K}}\beta}\}\max\{(C_{\mathcal{K}}\beta)^m, \epsilon(n)\}$$

We plug $C = 2\max\big\{2C_1, C_1\frac{C_{\mathcal{K}}}{1-C_{\mathcal{K}}\beta}\big\}$, $k = n+m$, $n = \lfloor k/2 \rfloor$ and $m = k - \lfloor k/2 \rfloor \geq \lfloor k/2 \rfloor$ in (3.7). Hence, we conclude the proof. $\square$

We define the optimality equation as

$$\phi(x,a) \triangleq \inf_{q\in\mathcal{K}_{xa}} \sum_{y\in S} q(y|x,a)(r(x,a,y)+\beta v^*(y)) - v^*(x), \tag{3.9}$$

which gives the difference between the optimal action in state $x \in S$ and an arbitrary action $a \in A(x)$. The optimality equation (3.9) can be equally stated as

$$\max_{a\in A(x)} \phi(x,a) = 0,\ x \in S. \tag{3.10}$$

(3.10) gives also the existence of optimal Markovian stationary policy. Indeed, since the available controls in each state $x \in S$ with $A(x) \subset A$ is finite, there exists an optimal control $a^*(x)$ for (3.10) for each $x \in S$. By S being countable, defining $g^*(x) = a(x)$ for each x, we reach the optimal stationary Markov policy $g^* : S \to A$. Next, we show that the finite scheme in (3.6) is convergent to the optimal stationary policy g^*.

Theorem 3.2. *Let $\pi^* = \{f_n(\cdot)\}$ be the Markov policy that is defined as in* (3.6)*. Then, under Assumption 3.1 and $0 < C_{\mathcal{K}}\beta < 1$, we have*

$$\begin{aligned}\|\phi\|_n &\triangleq \sup_{x\le n}\Big| \inf_{q\in\mathcal{K}_{xa}} \sum_{y\in S} q(y|x,f_n(x))(r(x,f_n(x),y)+\beta v^*(y)) - v^*(x)\Big| \\ &\to 0, \quad \text{as } n\to\infty.\end{aligned}$$

Proof. For $x \le n$ and for the maximizing f_n as in (3.6), we have

$$\begin{aligned}\phi(x,f_n(x)) &= \phi(x,f_n(x)) - v_n(x) + v_n(x) \\ &= [\inf_{q\in\mathcal{K}_{xa}} \sum_{y\in S} q(y|x,f_n(x))(r(x,f_n(x),y)+\beta v^*(y))] - v^*(x) - v_n(x) + v_n(x).\end{aligned}$$

Then, we have

$$\begin{aligned}\phi(x,f_n(x)) - v_n(x) &= [\inf_{q\in\mathcal{K}_{xa}} \sum_{y\in S} q(y|x,f_n(x))(r(x,f_n(x),y)+\beta v^*(y))] \\ &\quad - v^*(x) - v_n(x) \\ &= [\inf_{q\in\mathcal{K}_{xa}} \sum_{y\in S} q(y|x,f_n(x))(r(x,f_n(x),y)+\beta v^*(y)) \\ &\quad - \inf_{q\in\mathcal{K}_{xa}} \sum_{y\in S} q(y|x,f_n(x))(r(x,f_n(x),y)+\beta v_{n-1}(y))] - v^*(x)\end{aligned}$$

Hence,

$$\phi(x,f_n)-v_n(x)+v_n(x)\leq\beta\sup_{q\in\mathcal{K}_{xa}}\sum_{y\in S}q(y|x,f_n(x))(f_n(x))|v^*(y)-v_{n-1}(y)|+v_n(x)-v^*(x)$$

$$\begin{aligned}&\leq C_{\mathcal{K}}\beta\sum_{y<n}p(y|x,f_n(x))|v^*(y)-v_{n-1}(y)|\\&\quad+C_{\mathcal{K}}\beta\sum_{y\geq n}p(y|x,f_n(x))|v^*(y)|+[v_n(x)-v^*(x)].\end{aligned}$$

Thus, by Theorem 3.1,

$$\begin{aligned}\|\phi\|_n&\leq C_{\mathcal{K}}\beta\|v^*_{n-1}-v^*\|_{n-1}+C_{\mathcal{K}}\beta\|v^*\|\epsilon(n-1)+\|v^*-v_n\|_n\\&\to 0,\ \text{as } n\to\infty.\end{aligned}$$

Hence, we conclude the proof. □

Finally, we give an example to show that the approximation scheme is not necessarily convergent, if the discount factor together with the uncertainty radius is not sufficiently small, even though the same scheme is convergent, if it is applied on the expected value criteria.

Example 3.1. *We consider a Markov chain with two states $S=\{0,1\}$ and take that only one control is possible in each state with $A(x)=\{1\}$ for $x\in S$ and let $\beta=\frac{1}{2}$. We further suppose that the Markov chain has the following transition probabilities: $p(j|i,1)=1/2$ for $i,j=0,1$. The decision maker wants to minimize the discounted cost in infinite horizon. The cost of transition from state 0 equals 2, i.e. $c(0,1)=2$, whereas $c(1,1)=0$. Let the uncertainty set $\mathcal{K}_{xa}$ be defined as all those transition probabilities $q(\cdot|i,1)$ with $0\leq\frac{q(j|i,1)}{p(j|i,1)}\leq 2$ for $i,j=0,1$, i.e. $C_{\mathcal{K}}=2$, and $\sum_{j=0}^{j=1}q(j|i,1)=1$ for $i=0,1$. Let $x_0=0$. Thus,*

$$\begin{aligned}\beta\rho_n(c(X_{n+1}))&=\beta\max_{q\in\mathcal{K}_{xa}}\mathbb{E}^q[c(X_{n+1},1)|x_n,1],\\&=\beta\max_{j=0,1}c(j,1)\\&=1,\end{aligned}\tag{3.11}$$

Note that (A1) and (A2) satisfied, whereas (A3) is violated with $C_{\mathcal{K}}\beta=1$ in Assumption 3.1. Our value function reads as

$$v(x_0)=c(0,1)+\lim_{T\to\infty}\rho_0\bigg(\beta\big(c(X_1,1)+\beta\rho_1(c(X_2,1)+\ldots+\rho_{T-1}(\beta c(X_T,1)\big)\ldots\bigg).$$

Hence, by (3.11), $\rho_n(\frac{1}{2}c(X_{n+1},1))=1$ for all $n\geq 1$ and $v(0)\to\infty$. Also, the scheme in (3.5) does not converge neither with $v_0(0)=2$ and $v_k(0)=2+k$ for $k=1,2,\ldots$.

On the other hand, we note that, when there is no uncertainty in transition probabilities, namely, when $C_{\mathcal{K}} = 1$, we have

$$
\begin{aligned}
v(0) &= c(0,1) + \lim_{T\to\infty} \mathbb{E}_0^p\Big(\beta\big(c(X_1,1) + \mathbb{E}_1^p(\beta c(X_2,1) + \ldots + \mathbb{E}_{T-1}^p(\beta c(X_T,1)\big)\ldots\Big) \\
&= 2 + \sum_{i=1}^{\infty}(\frac{1}{2})^i = \frac{7}{2}
\end{aligned}
$$

and the scheme is convergent, as well. This is indeed so, since $\frac{1}{2} = \beta C_{\mathcal{K}} < 1$, when we use the expectation for performance evaluation.

References

[1] Artzner, P., Delbaen, F., Eber, J.M., Heath, D. (1999). *Coherent measures of risk*. Mathematical Finance 9, 203-228.

[2] Ruszczynski, A., Shapiro, A. (2006) *Conditional risk mappings*. Mathematics of Operations Research, 31, 544-561.

[3] Ruszczynski, A. (2010). *Risk-averse dynamic programming for Markov decision processes*. Mathematical Programming Series B, 125:235-261.

[4] Hernandez-Lerma, O.(1986) *Finite-state approximations for denumerable multidimensional state discounted Markov decision processes*. Journal of Mathematical Analysis and Applications, 382-389.

[5] Hernandez-Lerma O, Lasserre, JB. (1996) *Discrete-time markov control processes. Basic optimality criteria* Springer, New York.

[6] Hernandez-Lerma, O. and Lasserre, J. B., (1999) Further topics on discrete-time markov control processes, Springer, New York.

[7] White, D.J.(1980) *Finite state approximations for denumerable state infinite horizon discounted Markov decision processes: The method of successive approximations*, Recent Developments in Markov Decision Processes Academic Press, New York.

[8] Shapiro, A. (2011) *A dynamic programming approach to adjustable robust optimization*, Operations Research Letters, 39, pp. 83-87.

[9] Linwei, X. and Shapiro, A. (2012) *Bounds for nested law invariant coherent risk measures*, Operations Research Letters, vol. 40, pp. 431-435.

[10] Shapiro, A., Dentcheva, D.,Ruszczyski, A., (2014) *Lectures on stochastic Programming: modeling and theory*, SIAM, Philadelphia.