Operations Research Letters 51 (2023) 133–136

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Contractivity of Bellman operator in risk averse dynamic programming with infinite horizon

Miloš Kopa [a,*], Martin Šmíd [b]

[a] *Charles University, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovska 83, Prague, Czech Republic*
[b] *The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodarenskou vezi 4, Prague, Czech Republic*

ARTICLE INFO

Article history:
Received 11 May 2022
Received in revised form 19 December 2022
Accepted 6 January 2023
Available online 13 January 2023

Keywords:
Risk aversion
Dynamic programming
Infinite horizon

ABSTRACT

The paper deals with a risk averse dynamic programming problem with infinite horizon. First, the required assumptions are formulated to have the problem well defined. Then the Bellman equation is derived, which may be also seen as a standalone reinforcement learning problem. The fact that the Bellman operator is contraction is proved, guaranteeing convergence of various solution algorithms used for dynamic programming as well as reinforcement learning problems, which we demonstrate on the value iteration and the policy iteration algorithms.

1. Introduction

Risk averse variants of dynamic programming are widely studied. Our work is very close to [4] where, for a very similar setting, the convergence of the Value iteration algorithm is proved. The contribution of our work is two-fold. First, instead of complicated axiomatic definition, we define the conditional risk measures constructively by means of their risk envelopes; consequently, the whole exposition, including proofs, is much simpler. Second, instead of the convergence of a particular solution algorithm, we prove the contractive property of the Bellman operator, which can be then plugged into convergence proofs of many different algorithms; we demonstrate this with the Value iteration and the Policy iteration algorithms.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathbf{F} := (\mathcal{F}_0, \dots, \mathcal{F}_t, \dots)$ be a filtration, i.e. a sequence of increasing sigma algebras: $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots \subset \mathcal{F}$. Further, consider a real process $\{Z_t\}, t = 1, 2, \dots$, adapted to the filtration $(\mathcal{F}_1, \mathcal{F}_2, \dots)$, specifically $Z_t \in L_2(\mathcal{F}_t), t = 1, 2, \dots$. For most applications, it suffices to assume that $\Omega = (0,1) \times (0,1) \times \dots$, that $\mathbb{P} = U(0,1) \otimes U(0,1) \otimes \dots$ where U is the uniform distribution, and that $\mathcal{F}_t \subseteq \sigma(\Omega_1, \dots, \Omega_{2t+1})$ (see [3], Chp. 8).[1] In this case, we can put $Z_t = q_t(\Omega_{2t}|\mathcal{F}_{t-1})$ where, for any $t > 0$, q_t is the quantile function of the required conditional distribution of Z_t given $\mathcal{F}_{t-1}$.

For the process Z_t, $t > 0$, we use coherent conditional risk measures: $\sigma_{t|\mathcal{F}_{t-1}}(Z_t)$, $t > 0$. By saying that a conditional risk measure is coherent we mean it is measurable, monotonous ($\sigma_{t|\mathcal{F}_{t-1}}(X) \geq \sigma_{t|\mathcal{F}_{t-1}}(Y)$ for any random variables $X \geq Y$, $X, Y \in L_2$), sub-additive ($\sigma_{t|\mathcal{F}_{t-1}}(X+Y) \leq \sigma_{t|\mathcal{F}_{t-1}}(X) + \sigma_{t|\mathcal{F}_{t-1}}(Y)$ for any random variables $X, Y \in L_2$) translation invariant ($\sigma_{t|\mathcal{F}_{t-1}}(X + C)) = \sigma_{t|\mathcal{F}_{t-1}}(X) + C$ for any $C \in L_2(\mathcal{F}_{t-1})$, $X \in L_2$) and positively homogeneous ($\sigma_{t|\mathcal{F}_{t-1}}(\Lambda X)) = \Lambda \sigma_{t|\mathcal{F}_{t-1}}(X)$ for any $\Lambda \in L_2(\mathcal{F}_{t-1})$, $X \in L_2$). Next we construct nested risk measures:

$$\rho_t(Z_t) = \sigma_{1|\mathcal{F}_0}(\sigma_{2|\mathcal{F}_1}(\dots \sigma_{t|\mathcal{F}_{t-1}}(Z_t))), \qquad t = 1, 2, \dots.$$

It can be easily seen that, once $\mathcal{F}_0$ is trivial, $\rho_t(Z_t)$ is a deterministic coherent risk measure. We will be interested in its limit version:

$$\rho_\infty(Z) = \lim_{t \to \infty} \rho_t(Z_t) \qquad a.s. \tag{1}$$

Such a limit, however, may not exist in general. Next we state sufficient conditions for its existence, general enough for our purposes.

Definition 1. We say that process $\{Z_t\}, t = 1, 2, \dots$, has uniformly bounded support if there exist finite $a < b$ such that support$(Z_t) \subseteq \langle a, b \rangle$ for all t.

* Corresponding author.
E-mail address: kopa@karlin.mff.cuni.cz (M. Kopa).

[1] In this setting, the 2nd, 4th, …, underlying variables "drive" the process Z while the 1st, 3rd, …, ones "stand behind" the additional stochasticity generating the corresponding sigma algebras. This setting could not be used only if $\mathcal{F}_t$ would be generated by a non-Borel random element while e.g. real random sequences, processes or measures are Borel random elements.

https://doi.org/10.1016/j.orl.2023.01.008
0167-6377/© 2023 Elsevier B.V. All rights reserved.

Definition 2. We say that conditional risk measure $\sigma_{t|\mathcal{F}_{t-1}}$ is *support-bounded* if, for every $X \in L_2$, we have $\sigma_{t|\mathcal{F}_{t-1}}(X) \in \langle \mathrm{essinf}(X), \mathrm{esssup}(X) \rangle$.

Remark 1. All $\mathbb{E}(\bullet|\mathcal{F}_{t-1})$, $\mathrm{CVaR}_\alpha(\bullet|\mathcal{F}_{t-1})$, $\alpha \in (0,1)$, and $\mathrm{esssup}(\bullet)$ are support-bounded conditional risk measures.

Theorem 1. *Let process $\{Z_t\}$, $t = 1, 2, \dots$, be a.s. non-increasing, and let it have uniformly bounded support. Assume that a conditional risk measure $\sigma_{t|\mathcal{F}_{t-1}}(Z_t)$ is coherent and support-bounded for all t. Then $\rho_\infty(Z)$ exists.*

Proof. First, thanks to the nested form of ρ_t and the fact that conditional risk measures $\sigma_{t|\mathcal{F}_{t-1}}$ are support-bounded, we have bounded sequence $\rho_t(Z_t)$, $t = 1, 2, \dots$. Second,

$$\begin{aligned}\rho_t(Z_t) - \rho_{t+1}(Z_{t+1}) &= \rho_t(Z_t) - \rho_t(\sigma_{t+1|\mathcal{F}_t}(Z_{t+1})) \\ &\geq \rho_t(Z_t) - \rho_t(\sigma_{t+1|\mathcal{F}_t}(Z_t)) \\ &= \rho_t(Z_t) - \rho_t(Z_t) = 0,\end{aligned}$$

where the inequality follows from (i) coherency of $\sigma_{t+1|\mathcal{F}_t}$ and (ii) the fact that $\{Z_t\}$ is a.s. non-increasing, $t = 1, 2, \dots$. Hence, the sequence $\rho_t(Z_t)$, $t = 1, 2, \dots$ is non-increasing. Since every bounded non-increasing sequence has a limit, $\rho_\infty(Z)$ exists, which completes the proof. □

It is well known (see [1]) that every coherent risk measure σ can be expressed in a dual form: $\sigma(X) = \sup_{Q \in \mathcal{M}} \int X(\omega) Q(\omega) P(d\omega)$, where P is an underlying probability measure, and $\mathcal{M} = \{Q \in L_2 : Q \geq 0, \mathbb{E}_P(Q) = 1, \mathbb{E}_P(XQ) \leq \sigma(X), X \in L_2\}$ is a set of probability distributions known as risk envelope. Especially, if $P = U(0,1)$, then $\sigma(X) = \sup_{Q \in \mathcal{M}'} \int_0^1 q_X(\omega) Q(\omega) d\omega$ where $\mathcal{M}'$ is a (possibly different) risk envelope and q_X is a quantile function of X. Clearly, as conditional risk measures become coherent risk measures once the conditioning random element is fixed, they can be expressed by means of a dual representation, too.

Generally, the risk envelope can depend on the conditioning random element; however, we shall further study a special case when this is not considered and, moreover, the risk envelope does not depend on t:

$$\sigma_{t|\mathcal{F}_{t-1}}(X) = \Sigma(\mathcal{L}(X|\mathcal{F}_{t-1})), \quad t \geq 1,$$

$$\Sigma(\mathcal{L}) = \sup_{Q \in \mathcal{M}} \int_0^1 q_{\mathcal{L}}(x) Q(x) dx, \tag{2}$$

where $\mathcal{M}$ is a (deterministic) risk envelope and $q_{\mathcal{L}}$ is a quantile function corresponding to $\mathcal{L}$. Assumption (2) is a clear restriction: without this assumption, for instance, it could be $\sigma_{2|\mathcal{F}_1} = \mathbf{1}[Z_1 > 0]\mathrm{CVaR}_\alpha(Z_2|Z_1) + \mathbf{1}[Z_1 \leq 0]\mathrm{esssup}(Z_2)$; however, having (2), all the conditional measures have to be "of the same type", e.g., CVaR with level α, which is, however, usually the case in practice. The following Proposition states a recursive property of ρ_∞.

Proposition 1. *Assume (2). Let $\mathcal{F}_1$ be trivial (implying that Z_1 is deterministic) and let*

$$0 \leq Z_t - Z_{t+1} \leq \epsilon_t, \qquad t > 0, \tag{3}$$

where ϵ_t is deterministic with $\sum_t \epsilon_t$ finite. Then ρ_∞, defined by (2), exists and

$$\rho_\infty(Z) = Z_1 + \sigma(\rho_\infty(Z')),$$

where $Z'_t = Z_{t+1} - Z_1, t > 0$ and $\sigma(X) = \Sigma(\mathcal{L}(X))$ for any random variable X (see (2)).

Note that σ is an unconditional coherent risk measure. First we prove the following lemma:

Lemma 1. *(i) Let Z_1 be bounded and let (3) hold. Let $\sigma_{t|\mathcal{F}_{t-1}}(Z_t)$ be defined by (2) and let it be support-bounded for all t. Then $\rho_\infty(Z)$ exists and the convergence in (1) is uniform; in particular, $0 \leq \rho_t(Z_t) - \rho_\infty(Z) \leq \sum_{\tau=t}^\infty \epsilon_\tau$.*
(ii) For any coherent risk measure σ and any real random sequence Z_t uniformly converging a.s. to a random variable $Z^\star$, $\sigma(Z_t) \to \sigma(Z^\star)$.

Proof of Lemma. (i) Clearly, Z_t fulfills the assumptions of Theorem 1, so $\rho_\infty(Z)$ exists. Denote $e_t = \sum_{\tau \geq t} \epsilon_\tau$. For any $t > 0$ and $s > 0$, knowing that $\rho_t(Z_t)$ is non-increasing, we have

$$\begin{aligned}0 &\leq \rho_t(Z_t) - \rho_{t+s}(Z_{t+s}) \leq \rho_t(Z_t) - \rho_{t+s}(Z_t - e_t) \\ &= \rho_t(Z_t) - \rho_t(Z_t - e_t) = e_t.\end{aligned}$$

Recall that e_t is deterministic, therefore $\rho_t(Z_t)$ is Cauchy sequence in sup norm, hence uniformly convergent (L^∞ is Banach space).
(ii) Let $Z_t \to Z^\star$ uniformly. Then there exists a sequence $e_t \to 0$ such that $Z^\star - e_t \leq Z_t \leq Z^\star + e_t$, so, by coherence, $\sigma(Z^\star) - e_t \leq \sigma(Z_t) \leq \sigma(Z^\star) + e_t$ implying $\sigma(Z_t) \to \sigma(Z^\star)$. □

Proof of the Proposition. Thanks to Theorem 1, the limit defining the l.h.s. exists. From the definition and the coherence property

$$\rho_\infty(Z) = \lim_{t\to\infty} \rho_t(Z_t) = Z_1 + \lim_{t\to\infty}(\sigma(S_t))$$

where $S_1 = 0$ and

$$\begin{aligned}S_t =& \sigma_{2|\mathcal{F}_1}(\dots \sigma_{t|\mathcal{F}_{t-1}}(Z'_{t-1}) \dots) \\ =& \Sigma(\mathcal{L}(\Sigma(\dots \Sigma(\mathcal{L}(Z'_{t-1}|\mathcal{F}_{t-1}) \dots)|\mathcal{F}_0)), \quad t \geq 0.\end{aligned}$$

By Lemma 1 (i), S_t converges uniformly to $\rho_\infty(Z')$, so, by (ii) of the same Lemma, $\lim_{t\to\infty}(\sigma(S_t)) = \sigma(\lim_{t\to\infty} S_t) = \sigma(\rho_\infty(Z'))$. □

2. Contractiveness of the Bellman operator

Consider a dynamic programming problem

$$V(s_1) \overset{\mathrm{def}}{=} \sup_{a_1(s_1)\in X(s_1), a_{t+1}\in A, t\geq 1} \varrho_\infty(Y^{s_1,a_1(s_1),a_2,\dots}), \quad s_1 \in \mathcal{S},$$

$$Y_t^{s_1,a_1(s_1),a_2,\dots} = \sum_{\tau=1}^{t} \gamma^{\tau-1} r(s_\tau, a_\tau(s_\tau)),$$

$$s_{t+1} = T(s_t, a_t(s_t), W_{t+1}), \quad t \geq 1.$$

Here,

- $W_t \in \mathcal{X}$, $t \geq 1$, are i.i.d. (we may assume that $\mathcal{X} = (0,1)$ and $W_t \sim U(0.1)$ as above),
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \to \mathcal{S}$ is a jointly measurable mapping, where $\mathcal{S}$ is a complete state space and $\mathcal{A}$ is a (measurable) action space,
- $0 \leq r(\bullet) \leq b_r$, where b_r is deterministic,
- $\varrho_\infty(Z) = -\rho_\infty(-Z)$, where ρ_∞ is a limit nested risk measure (1) defined by filtration $\mathcal{F}_t \overset{\mathrm{def}}{=} \sigma(W_t)$ and a support-bounded coherent risk measure σ as in (2),
- $A = \{a : \mathcal{S} \to \mathcal{A} : a(s) \in X(s)\}$, $X(s) = \{a \in \mathcal{A}, f(a,s) \geq 0\} \neq \emptyset$, $s \in \mathcal{S}$, where $f : \mathcal{A} \times \mathcal{S} \to \mathbb{R}^k$ is jointly measurable,
- $\gamma < 1$ is a discount factor.

Since r is uniformly bounded non-negative and $\gamma < 1$, process $Y^{s_1,a_1(s_1),a_2,\dots}$ has uniformly bounded support (with respect to both $\mathcal{X}$ and the policies) and is non-decreasing. Combined with the assumption of coherent support-bounded conditional risk measure σ, it guarantees existence of ϱ_∞ for any sequence of policies $a_1(s_1), a_2, \ldots$. Hence, the problem is well defined. Moreover, by an application of the Measurable projection theorem, similar to Debut theorem, V is measurable (note that the problem can be expressed as a free maximization with the indicators of X's multiplying the r's, and that the resulting function is jointly measurable).

Proposition 2. *(Bellman Equation)*

$$V(s) = \sup_{a \in X(s)} [r(s,a) + \gamma\varsigma(V(T(s,a,W)))], \quad s \in \mathcal{S}, \tag{4}$$
$$W \sim U(0,1), \text{ where } \varsigma(Z) = -\sigma(-Z).$$

Note that (4) may be also understood as a definition of a risk-averse version of a reinforcement learning problem (see [5]).

Proof. Fix $s_1 \in \mathcal{S}$. Thanks to Proposition 1, Lemma 1 (i), and basic properties of sup, we have

$$V(s_1) = \sup_{a_1(s_1) \in X(s_1)} \left[r(s_1, a_1(s_1)) + \gamma \underbrace{\sup_{a_t \in A, t \geq 2} \varsigma\left(Z^{a_2,a_3,\ldots}\right)}_{=\varsigma(\hat{Z})} \right],$$
$$Z^{a_2,a_3,\ldots} = \varrho_\infty(Y^{s_2,a_2(s_2),a_3,\ldots}),$$
$$\hat{Z} = \sup_{a_2(s_2) \in X(s_2), a_t \in A, t>2} Z^{a_2,a_3,\ldots},$$

where $s_{t+1} = T(s_t, a_t, W_{t+1})$, $t \geq 1$. Note that both Z and $\hat{Z}$ are (possibly) random (but only) $\mathcal{F}_2$-measurable. To prove the "=" under the underbrace, note that $\leq$ follows directly from the monotonicity of ς; as for $\geq$, we have, by the definition of (the finite) sup, that, for each ϵ, there exists feasible $a_2, a_3, \ldots$ such that $\hat{Z} - Z^{a_2,a_3,\ldots} \leq \epsilon$ (note that we maximize separately for each value of s_2), i.e. there exist a uniformly convergent sequence of feasible $Z_1, Z_2, \ldots$ converging to $\hat{Z}$. By Lemma 1 (ii), $\varsigma(Z_k) \to_k \varsigma(\hat{Z})$ which implies $\geq$. □

Next we formulate our main result, which makes many solution techniques applicable to the risk-averse problems.

Theorem 2. *The operator*

$$B : (BV)(s) \overset{def}{=} \sup_{a \in A(s)} [r(s,a) + \gamma\varsigma(V(T(s,a,W))]$$

is a γ contraction w.r.t. sup norm.

Proof. Let $\epsilon > 0$, for any value function V, denote $a_{V,s}$ (a selected) ϵ-optimal solution of $\sup_{a \in A(s)}[r(s,a) + \gamma\varsigma(V(T(s,a,W))]$ and: $Q_{V,s} \in \epsilon - \arg\max_{Q \in \mathcal{M}} \int_0^1 -V(T(s,a_{U,s},w))Q(w)dw$. Then

$$\|BU - BV\|_\infty = \underbrace{\sup_{s \in S_U \overset{def}{=} \{s:(BU)(s)>(BV)(s))\}} [(BU)(s) - (BV)(s)]}_{b_U}$$
$$\vee \underbrace{\sup_{s \in S_V \overset{def}{=} \{s:(BU)(s)\leq(BV)(s))\}} [(BV)(s) - (BU)(s)]}_{b_V}$$

Further we have

$$\begin{aligned}
b_U &\leq \sup_{s \in S_U} [(BU)(s) - r(s,a_{U,s}) - \gamma\varsigma(V(T(s,a_{U,s},W)))] \\
&\leq \sup_{s \in S_U} [r(s,a_{U,s}) - r(s,a_{U,s}) + \gamma\varsigma(U(T(s,a_{U,s},W))) \\
&\quad - \gamma\varsigma(V(T(s,a_{U,s},W)))] + \epsilon \\
&= \gamma \sup_{s \in S_U} [- \sup_{Q \in \mathcal{M}} \int_0^1 -U(T(s,a_{U,s},w))Q(w)dw \\
&\quad + \sup_{Q \in \mathcal{M}} \int_0^1 -V(T(s,a_{U,s},w))Q(w)dw] + \epsilon \\
&\leq \gamma \sup_{s \in S_U} [- \int_0^1 -U(T(s,a_{U,s},w))Q_{V,s}(w)dw \\
&\quad + \int_0^1 -V(T(s,a_{U,s},w))Q_{V,s}(w)dw] + 2\epsilon \\
&\leq \gamma \sup_{s \in S_U} \int_0^1 |U(T(s,a_{U,s},w)) - V(T(s,a_{U,s},w))|Q_{V,s}(w)dw \\
&\quad + 2\epsilon \leq \gamma\|U - V\|_\infty + 2\epsilon
\end{aligned}$$

(the last inequality holds because $Q(w)dw$ is a probability measure). By releasing ϵ and performing a limit transition, we get that $b_U \leq \gamma\|U - V\|_\infty$. By making analogous steps for b_V, we get the Theorem. □

Next we demonstrate the convergence of two well known solution algorithms (see [5]): For any $a \in A$, denote

$$V^a : (V^a)(s) = \varrho_\infty(\sum_{t=1}^\infty \gamma^{t-1} r(s_t, a(s_t))),$$
$$s_{t+1} = T(s_t, a(s_t), W_{t+1}), \quad t \geq 1,$$

and

$$C : (CV)(s) \overset{\text{def}}{=} \arg\max_{a \in A(s)} [r(s,a) + \gamma\varsigma(V(T(s,a,W)))],$$

where the choice of the optimal policy is arbitrary. Let θ be a pre-chosen precision level. Assume C is well defined (i.e. at least one arg max always exists).

Value Iteration

```
choose bounded $V_0 : S \to [0,\infty)$
$n \leftarrow 0$
repeat
    $n \leftarrow n+1$
    $V_n \leftarrow BV_{n-1}$
until $\|V_n - V_{n-1}\|_\infty \leq \theta$
```

Policy Iteration

```
choose $a_0 \in A$
$n \leftarrow 0$
repeat
    $n \leftarrow n+1$
    $a_n \leftarrow CV^{a_{n-1}}$
until $\|V^{a_n} - V^{a_{n-1}}\|_\infty \leq \theta$
```

The following result is a direct consequence of the Banach Fixed Point Theorem (see [2], 1.1) and the equivalence of the value functions and the (classes of) policies.

Theorem 3. *There exists $V_\star$ solving (4). Moreover,*

$$\|V_n - V_\star\|_\infty \leq \frac{\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty,$$
$$\|V^{a_n} - V_\star\|_\infty \leq \frac{\gamma^n}{1-\gamma}\|V^{a_1} - V^{a_0}\|_\infty,$$

for any n.

Corollary 1. *Both the Value iteration algorithm and the Policy iteration algorithm converge at a geometric rate and stop after a finite number of steps.*

Remark 2. Versions using ϵ-optimal solutions may be formulated; however, the ϵ's have to gradually decrease. In particular, it has to be $\epsilon_n \sim \gamma^n$.

Acknowledgements

This work has been supported by grant No. GA19-11062S of the Czech Science Foundation.

References

[1] M. Ang, J. Sun, Q. Yao, On the dual representation of coherent risk measures, Ann. Oper. Res. 262 (1) (2018) 29–46.
[2] A. Granas, J. Dugundji, Fixed Point Theory, vol. 14, Springer, 2003.
[3] O. Kallenberg, Foundations of Modern Probability, 2001.
[4] A. Ruszczyński, Risk-averse dynamic programming for markov decision processes, Math. Program. 125 (2) (2010) 235–261.
[5] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.