Robust Optimal Strategies in Markov Decision Problems

Gal Oren*and Eilon Solan†

January 5, 2014

Abstract

An optimal strategy in a Markov decision problem is *robust* if it is optimal in every decision problem (not necessarily stationary) that is close to the original problem. We prove that when the state and action spaces are finite, an optimal strategy is robust if and only if it is the unique optimal strategy.

Keywords: Markov decision problems, optimal strategy, robustness.

1 Introduction

Sequential decision problems are used by practitioners in various areas. To properly analyze such a problem, and to find an optimal strategy, one needs to provide the stage payoff, the transitions, and the discount factor after every possible history. In many cases, it is impossible or too costly to know the data of the problem precisely, and then only estimation of the data is available to the decision maker. Thus one cannot analyze the actual decision problem, but only an approximating one, and an optimal strategy in the model that is studied might be suboptimal in the actual one. We say that an optimal strategy in a given decision problem $\mathcal{P}$ is *robust* if it remains optimal in every decision problem that is close to $\mathcal{P}$.

In many cases, to simplify the analysis one approximates a nonstationary decision problem by a stationary one. This way one smooths out fluctuations in stage payoff, transitions, or discounting that are due to predictable or unpredictable changes in the environment. For example, consider a lake that is used for fishing, and suppose that the increase rate of the population is stationary and known. The fishermen can then predict the evolution of the fish population as a function of the amount of fish they extract, yet the fish prices as well as the interest rate in the market fluctuate randomly. If these last two variables are approximately stationary, then any robust optimal strategy in the fictitious stationary problem is also optimal in the actual one.

*The School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. e-mail: gal123oren@gmail.com.

†Corresponding author. The School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. e-mail: eilons@post.tau.ac.il. Telephone: (972)-3-6409635.

In this note we prove that an optimal strategy in a stationary decision problem with finitely many states and actions is robust if and only if it is the unique optimal strategy, and by virtue of an example we show that this result does not extend to decision problems with countably many states or actions. This result highlights the extent to which an optimal strategy remains optimal when the data of the decision problem is known only approximately.

2 The Model and the Main Result

A decision problem is defined over a finite set of states S. In each state $s \in S$ the decision maker has a finite collection of available actions, denoted by A_s. Denote $A = (A_s)_{s\in S}$. We let $\Xi := \{(s,a)\colon s \in S, a \in A_s\}$ be the set of all pairs of state and an action available at that state. Set $H_{S,A} := \cup_{n=0}^{\infty} (\Xi^n \times S)$ be the space of all finite histories, and $H^+_{S,A} := \cup_{n=1}^{\infty} \Xi^n$ be the space of all histories including the current action. The notation $h^n = (s^1, a^1, \ldots, s^n) \in \Xi^{n-1} \times S$ will always denote a possible history at stage n. Given an infinite play $h^\infty = (s^1, a^1, s^2, a^2, \ldots) \in \Xi^\infty$ and $n \in \mathbb{N}$, we denote by $h^n = (s^1, a^1, \ldots, s^n) \in \Xi^{n-1} \times S$ the possible history at stage n that h^∞ induces. For every finite set S denote by $\Delta(S)$ the set of probability distributions over S.

Definition 1 *A* decision problem *over the set of states S and the sets of actions $A = (A_s)_{s\in S}$ is a triplet $\mathcal{P} = (q, r, \lambda)$ where*

- $q : H^+_{S,A} \to \Delta(S)$ *is a transition function;*
- $r : H^+_{S,A} \to \mathbb{R}$ *is a bounded payoff function; and*
- $\lambda : H_{S,A} \to (0,1)$ *is a discounting function.*

A decision problem evolves as follows. The initial state $s^1 \in S$ is given. At each stage $n \geq 1$, given the history so far $h^n = (s^1, a^1, \cdots, s^n) \in H_{S,A}$, the decision maker chooses an action $a^n \in A_{s^n}$, receives the stage payoff $r(h^n, a^n)$ and a new state s^{n+1} is chosen according to $q(h^n, a^n)$. The decision maker's goal is to maximize the discounted sum

$$\sum_{m=1}^{\infty} \left(r(h^m, a^m)\lambda(h^m) \prod_{k<m} (1 - \lambda(h^k)) \right). \tag{1}$$

Because the payoff function r is bounded, so is the sum in (1). Note that the transition and stage payoff depend on the current action of the decision maker while the discount factor does not depend on it. Nothing in what follows would have changed if the discount factor depended on the decision maker's current action.

Definition 2 *A* strategy *is a function σ that assigns an action in A_{s^n} to every history $h^n = (s^1, a^1, \cdots, s^n) \in H_{S,A}$.*

Denote by $\mathcal{S}$ the set of all strategies.

Every strategy σ, together with the initial state s^1, induces a probability distribution $\mathbf{P}_{s^1,\sigma}$ over the space of plays Ξ^∞ endowed with the σ-algebra generated by all finite cylinders. Denote by $\mathbf{E}_{s^1,\sigma}$ the corresponding expectation operator. The *expected payoff* induced by a strategy σ is

$$\gamma(s^1,\sigma) := \mathbf{E}_{s^1,\sigma}\left[\sum_{m=1}^{\infty} r(h^m,a^m)\lambda(h^m)\prod_{k<m}(1-\lambda(h^k))\right].$$

We will also be interested in the expected payoff given a history $h^n \in H_{S,A}$, which is

$$\gamma(h^n,\sigma) := \mathbf{E}_{s^1,\sigma}\left[\sum_{m=n}^{\infty} r(h^m,a^m)\lambda(h^m)\prod_{k<m}(1-\lambda(h^k)) \mid h^n\right].$$

This is the expected total payoff assuming that the history h^n occurred. Note that the payoff $\gamma(h^n,\sigma)$ is discounted to stage n.

The value is the maximal payoff that the decision maker can guarantee.

Definition 3 *The* value *of the decision problem $\mathcal{P}$ at the history $h^n \in H_{S,A}$ is*

$$v(h^n) := \sup_{\sigma\in\mathcal{S}} \gamma(h^n,\sigma).$$

A strategy σ^* that attains the supremum in the definition of $v(h^n)$ is called an optimal strategy at h^n. *A strategy is* optimal *if it is optimal at all finite histories in $H_{S,A}$.*

Because the sets S and A are finite, the space of all strategies is compact in the product topology (when S and A are endowed with the discrete topology). Moreover, the payoff function $\sigma \mapsto \gamma(s^1,\sigma)$ is continuous in this topology for every $s^1 \in S$. It follows that an optimal strategy exists. Since the payoff function r is bounded, our model is equivalent to a Markov decision problem with positive bounded payoffs, and therefore by Theorem 7.2.5 in [3] the set of optimal strategies can be characterized using Bellman's Principle of Optimality.

Theorem 1 *The strategy σ^* is optimal if and only if for every history $h^n \in H_{S,A}$ the action $\sigma^*(h^n)$ attains the maximum in*

$$\max_{a\in A_{s^n}}\left(\lambda(h^n)r(h^n,a) + (1-\lambda(h^n))\sum_{s\in S} q(h^n,a)[s]v(h^n,a,s)\right). \tag{2}$$

As mentioned before, in applications we usually do not know the exact payoffs, transitions, and discounting function. The L_∞ metric between decision problems is one way to measure the discrepancy between an actual decision problem and an approximating one.

Definition 4 *Let $\mathcal{P} = (q, r, \lambda)$ and $\widehat{\mathcal{P}} = (\widehat{q}, \widehat{r}, \widehat{\lambda})$ be two decision problems defined over the same set of states S and sets of actions $A = (A_s)_{s \in S}$. The* distance *between $\mathcal{P}$ and $\widehat{\mathcal{P}}$ is*

$$d(\mathcal{P}, \widehat{\mathcal{P}}) := \max\{ \sup_{h^n \in H_{S,A}} |\lambda(h^n) - \widehat{\lambda}(h^n)|, \sup_{(h^n, a^n) \in H^+_{S,A}} \|q(h^n, a^n) - \widehat{q}(h^n, a^n)\|_\infty, \sup_{(h^n, a^n) \in H^+_{S,A}} |r(h^n, a^n) - \widehat{r}(h^n, a^n)| \}.$$

In this definition, $\| \cdot \|_\infty$ is the supremum norm: $\|x\|_\infty = \max_{i=1}^d |x_i|$ for every vector $x = (x_i)_{i=1}^d \in \mathbb{R}^d$.

The value function is continuous with respect to the distance $d(\cdot, \cdot)$. We now define the concept of a robust optimal strategy, which is the main interest of this note.

Definition 5 *An optimal strategy σ^* is* robust *in the decision problem $\mathcal{P}$ if it is optimal in every decision problem $\widehat{\mathcal{P}}$ that is close to $\mathcal{P}$: there is $\delta > 0$ such that σ^* is optimal in every decision problem $\widehat{\mathcal{P}}$ that satisfies $d(\mathcal{P}, \widehat{\mathcal{P}}) < \delta$.*

As the next example shows, an optimal strategy, even when it is unique, need not be robust.

Example 1 *Suppose that $S = \mathbb{N} = \{1, 2, 3, \ldots\}$ and $A_s = \{\alpha, \beta\}$ for every $s \in S$. Let $\lambda_0 \in (0, 1)$ be arbitrary and consider the decision problem $\mathcal{P} = (q, r, \lambda)$ defined by*

- *$q(h^n, a)[n + 1] = 1$ for every $h^n \in H_{S,A}$ and $a \in A$: transition is deterministic and independent of the decision maker's action.*
- *$r(h^n, \alpha) = 1$ and $r(h^n, \beta) = 1 - \frac{1}{n}$ for every $h^n \in H_{S,A}$.*
- *$\lambda(h^n) = \lambda_0$ for every $h^n \in H_{S,A}$.*

The unique optimal strategy σ^* is to play α after every history; this strategy guarantees payoff 1. However, since the payoffs $(r(h^n, \beta))_{n \in \mathbb{N}}$ converge to 1 as the history's length increases, for every $\varepsilon > 0$ there is a decision problem $\widehat{\mathcal{P}} = (q, \widehat{r}, \lambda)$ that satisfies (a) $d(\mathcal{P}, \widehat{\mathcal{P}}) < \varepsilon$, (b) $\widehat{r}(h^n, \alpha) = 1$ for every $h^n \in H_{S,A}$, and (c) $\widehat{r}(h^n, \beta) > 1$ for every history $h^n \in H_{S,A}$ for which n is sufficiently large. By Theorem 1 the strategy σ^* is not optimal for $\widehat{\mathcal{P}}$.

In Example 1, the number of states is countable and there are two actions in each state. One can easily construct an analog example with two states and countably many actions.

As we now argue, though in general an optimal strategy may fail to be robust, when the decision problem is stationary there is a simple criterion for determining whether an optimal strategy is robust.

Definition 6 *A decision problem is called* stationary *if the functions q, r, and λ depend on h^n only through s^n; that is, $q(h^n, a) = q(\widehat{h^n}, a)$, $r(h^n, a) = r(\widehat{h^n}, a)$, and $\lambda(h^n) = \lambda(\widehat{h^n})$ for every two histories $h^n = (s^1, a^1, \cdots, s^n) \in H_{S,A}$ and $\widehat{h^n} = (\widehat{s}^1, \widehat{a}^1, \cdots, \widehat{s^n}) \in H_{S,A}$ for which $s^n = \widehat{s^n}$, and every action $a \in A_{s^n}$.*

A strategy σ is *stationary* if $\sigma(h^n)$ depends on h^n only through s^n, that is, $\sigma(h^n) = \sigma(\widehat{h}^{\widehat{n}})$, for every two histories $h^n = (s^1, a^1, \cdots, s^n)$ and $\widehat{h}^{\widehat{n}} = (\widehat{s}^1, \widehat{a}^1, \cdots, \widehat{s}^{\widehat{n}})$ for which $s^n = \widehat{s}^{\widehat{n}}$. By [1], a stationary decision problem has a stationary optimal strategy.

Our main result is the following.

Theorem 2 *An optimal strategy σ^* of a stationary decision problem with finitely many states and actions $\mathcal{P}$ is robust if and only if it is the unique optimal strategy in $\mathcal{P}$.*

As mentioned before, the value function is continuous with respect to the distance $d(\cdot,\cdot)$. To prove Theorem 2 we need the following result, which bounds the derivative of the value function at stationary decision problems. Its proof is similar to that of Theorem 4.3.7 in [2] (see Eq. (4.20) on page 186). In the statement of this theorem and later, whenever $\mathcal{P}$ and $\widehat{\mathcal{P}}$ are two decision problems, we denote their value functions by v and $\widehat{v}$ respectively.

Theorem 3 *Let $\mathcal{P} = (q, r, \lambda)$ be a stationary decision problem and $\widehat{\mathcal{P}} = (\widehat{q}, \widehat{r}, \widehat{\lambda})$ be a decision problem that are both defined over the same set of states S and sets of actions $A = (A_s)_{s\in S}$. Then for every history $h^n \in H_{S,A}$ we have*

$$|v(h^n) - \widehat{v}(h^n)| \leq 3 \max\left\{1, \frac{2\|\widehat{r}\|_\infty}{\min_{s\in S}\lambda(s)}\right\} d(\mathcal{P}, \widehat{\mathcal{P}}).$$

Proof of Theorem 2. Since $\mathcal{P}$ is stationary, we can denote its transitions and payoffs by $(q(s,a), r(s,a))_{s\in S, a\in A_s}$ and the discounting function by $(\lambda(s))_{s\in S}$. Suppose first that σ^* is the unique optimal strategy in $\mathcal{P}$. In particular, σ^* is stationary, and therefore we can denote it by $\sigma^* = (\sigma^*(s))_{s\in S}$. By Theorem 1 this implies in particular that for every $s \in S$ there is a unique action $a_s \in A_s$ that attains the maximum in

$$\max_{a\in A_s}\left(\lambda(s)r(s,a) + (1-\lambda(s))\sum_{s'\in S} q(s,a)[s']v(s,a,s')\right), \tag{3}$$

and, moreover, $\sigma^*(s) = a_s$. We will now show that σ^* is robust.

Set

$$\varepsilon := \min_{s\in S}\left\{v(s) - \max_{a\in A_s\setminus\{a_s\}}\left(\lambda(s)r(s,a) + (1-\lambda(s))\sum_{s'\in S} q(s,a)[s']v(s,a,s')\right)\right\}. \tag{4}$$

Because a_s is the unique maximizer in (3), and because S and $\{A_s, s \in S\}$ are finite sets, it follows that $\varepsilon > 0$. Let $\widehat{\mathcal{P}}$ be a decision problem whose distance from $\mathcal{P}$ is less than δ, where δ satisfies (a) $\delta \leq \frac{\varepsilon}{9}$, (b) $6\delta\frac{\|r\|_\infty+\delta}{\min_{s\in S}\lambda(s)} \leq \frac{\varepsilon}{3}$, and (c) $2\delta(1+\|r\|_\infty(|S|+2)+\delta) < \frac{\varepsilon}{12}$. By Theorem 3 and conditions (a) and (b),

$$|v(h^n) - \widehat{v}(h^n)| \leq 3\max\left\{1, \frac{2\|\widehat{r}\|_\infty}{\min_{s\in S}\lambda(s)}\right\} d(\mathcal{P}, \widehat{\mathcal{P}}) \leq \frac{\varepsilon}{3},$$

for every $h^n \in H_{S,A}$. Fix $h^n \in H_{S,A}$. We will show that a_{s_n} is the unique maximizer in (2) for $\widehat{\mathcal{P}}$, so that σ^* is the unique optimal strategy for the decision problem $\widehat{\mathcal{P}}$. Set

$$D := \delta(1+\|r\|_\infty(|S|+2)+\delta) + \frac{\varepsilon}{3}.$$

By the definition of the distance,

$$\begin{aligned}
&\widehat{\lambda}(h^n)\widehat{r}(h^n,a_{s_n})+(1-\widehat{\lambda}(h^n))\sum_{s'\in S}\widehat{q}(h^n,a_{s_n})[s']\widehat{v}(h^n,a_{s_n},s')\\
&\geq \lambda(s_n)r(s_n,a_{s_n})+(1-\lambda(s_n))\sum_{s'\in S}q(s_n,a_{s_n})[s']v(s')-D\\
&\geq \max_{a\in A_{s_n}\setminus\{a_{s_n}\}}\left(\lambda(s_n)r(s_n,a)+(1-\lambda(s_n))\sum_{s'\in S}q(s_n,a)[s']v(s')\right)-D+\varepsilon. \qquad (5)
\end{aligned}$$

As above, for every $a\in A_{s_n}\setminus\{a_{s_n}\}$ we have

$$\begin{aligned}
&\left(\lambda(s_n)r(s_n,a)+(1-\lambda(s_n))\sum_{s'\in S}q(s_n,a)[s']v(s')\right)\\
&\geq\left(\widehat{\lambda}(h^n)\widehat{r}(h^n,a)+(1-\widehat{\lambda}(h^n))\sum_{s'\in S}\widehat{q}(h^n,a)[s']\widehat{v}(h^n,a,s')\right)-D,
\end{aligned}$$

and therefore it follows that the quantity in (5) is at least

$$\max_{a\in A_{s_n}\setminus\{a_{s_n}\}}\left(\widehat{\lambda}(h^n)\widehat{r}(h^n,a)+(1-\widehat{\lambda}(h^n))\sum_{s'\in S}\widehat{q}(h^n,a)[s']\widehat{v}(h^n,a,s')\right)-2D+\varepsilon,$$

which, by condition (c), is at least

$$\max_{a\in A_{sn}\setminus\{a_{sn}\}}\left(\widehat{\lambda}(h^n)\widehat{r}(h^n,a)+(1-\widehat{\lambda}(h^n))\sum_{s'\in S}\widehat{q}(h^n,a)[s']\widehat{v}(h^n,a,s')\right)+\frac{\varepsilon}{4},$$

so that a_{s_n} is the unique maximizer of

$$\max_{a\in A_s}\left(\widehat{\lambda}(h^n)\widehat{r}(h^n,a)+(1-\widehat{\lambda}(h^n))\sum_{s'\in S}\widehat{q}(h^n,a)[s']\widehat{v}(h^n,a,s')\right).$$

Because this holds for every $h^n\in H_{S,A}$, Theorem 1 implies that the strategy σ^* is indeed optimal in $\widehat{\mathcal{P}}$.

Suppose now that σ^* is not a unique optimal strategy in $\mathcal{P}$. We will show that it is not robust. Indeed, since there are two distinct optimal strategies, Theorem 1 implies that there is a state $s_0\in S$ in which there are two distinct actions, denoted a_{s_0} and $\widehat{a}_{s_0}$, which attain the maximum in

$$\max_{a\in A_{s_0}}\left(\lambda(s_0)r(s_0,a)+(1-\lambda(s_0))\sum_{s'\in S}q(s_0,a)[s']v(s_0,a,s')\right),$$

Assume w.l.o.g that $\sigma^*(s_0)=a_{s_0}$. Fix now $\delta>0$. We will define a (nonstationary) decision problem $\widehat{\mathcal{P}}=(\widehat{q},\widehat{r},\widehat{\lambda})$ that is δ-close to $\mathcal{P}$ and in which σ^* is not optimal. To this end, define

- $\widehat{q}(h^n, a^n) = q(s^n, a^n)$ for every $h^n = (s^1, a^1, \ldots, s^n) \in H_{S,A}$ and every $a^n \in A_{s^n}$,
- $\widehat{\lambda}(h^n) = \lambda(s^n)$ for every $h^n = (s^1, a^1, \ldots, s^n) \in H_{S,A}$,
- $\widehat{r}(s_0, \widehat{a}_{s_0}) = r(s_0, \widehat{a}_{s_0}) + \delta$, and $\widehat{r}(h^n, a^n) = r(h^n, a^n)$ for every $(h^n, a^n) \in H^+_{S,A} \setminus \{(s_0, \widehat{a}_{s_0})\}$.

That is, the decision problem $\widehat{\mathcal{P}}$ is similar to $\mathcal{P}$, except that in the first stage, the payoff when playing $\widehat{a}_{s_0}$ at s_0 increases by δ. Plainly $d(\mathcal{P}, \widehat{\mathcal{P}}) = \delta$.

Let $\widehat{\sigma}$ be the strategy that coincides with σ^*, except that in the first stage it plays $\widehat{a}_{s_0}$ at state s_0. We argue that $\gamma(\widehat{\sigma}, s_0) > \gamma(\sigma^*, s_0)$, so in particular σ^* is not optimal. Indeed, since from the second stage on the two strategies $\widehat{\sigma}$ and σ^* coincide, and since σ^* is optimal, we have

$$\gamma(h^2, \widehat{\sigma}) = \gamma(h^2, \sigma^*) = v(h^2),$$

for every history $h^2 \in H_{S,A}$. In particular

$$\begin{aligned}
\gamma(s_0, \widehat{\sigma}) &= \widehat{\lambda}(s_0)\widehat{r}(s_0, \widehat{a}_{s_0}) + (1 - \widehat{\lambda}(s_0)) \sum_{s' \in S} \widehat{q}(s_0, \widehat{a}_{s_0})[s'] v(s_0, \widehat{a}_{s_0}, s') \\
&= \lambda(s_0)(r(s_0, \widehat{a}_{s_0}) + \delta) + (1 - \lambda(s_0)) \sum_{s' \in S} q(s_0, \widehat{a}_{s_0})[s'] v(s_0, \widehat{a}_{s_0}, s') \\
&= \gamma(s_0, \sigma^*) + \lambda(s_0)\delta > \gamma(s_0, \sigma^*),
\end{aligned}$$

as desired. Since this is true for every $\delta > 0$, the strategy σ^* is not robust. ∎

The reason that the conclusion of Theorem 2 holds for stationary decision problems and not for general decision problems is that in stationary problems there are finitely many effective histories, so that the minimum in (4) is positive, while in nonstationary problems there are infinitely many histories, so that even when all terms in the minimization (4) are positive, the minimum may not be attained, and the infimum may be 0.

Note that the determination of whether a stationary decision problem has a unique optimal strategy can be done in a polynomial time in $\sum_{s \in S} |A_s|$. Indeed, the calculation of the value function can be done by a linear program, so that the determination of whether there is a unique maximizer to (3), for every $s \in S$, can also be done in a polynomial time.

Theorem 2 leads us to the definition of the following concept. Let the *robustness radius* of $\mathcal{P}$ be the maximal δ such that σ^* is optimal in every decision problem $\widehat{\mathcal{P}}$ whose distance from $\mathcal{P}$ is below δ. Similarly to the previous paragraph, the proof shows that a lower bound on the robustness radius can be calculated in a polynomial time. The determination of the exact robustness radius seems a challenging open problem.

Acknowledgements

Solan acknowledges the support of the Israel Science Foundation, Grant #212/09 and of the Google Inter-university center for Electronic Markets and Auctions.

References

[1] Blackwell D. Discounted Dynamic Programming, *Ann. Math. Stat.*, 36 (1965), 226–235.

[2] Filar J.A. and Vrieze K. *Competitive Markov Decision Processes*, Springer-Verlag, New York, 1997.

[3] Puterman M.L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 2005.