

A general framework for reinforcement learning

Csaba Szepesvári^{1,2}

¹Bolyai Institute of Mathematics, Univ. of Szeged,
Szeged 6720, Aradi vrt. tere 1., HUNGARY

Tel./Fax: (+36-62) 326 246, e-mail: szepcs@math.u-szeged.hu

²Department of Photophysics, Institute of Isotopes of the
Hungarian Academy of Sciences,
Budapest 1525, P.O. Box 77. HUNGARY

Abstract: *In this article we propose a general framework for sequential decision making. The framework is based on the observation that the derivation of the optimal behaviour under various decision criteria follows the same pattern: the cost of policies can be decomposed into the successive application of an operator that defines the related dynamic programming algorithm and this operator describes completely the structure of the decision problem. We take this mapping (the so called one step lookahead (OLA) cost mapping) as our starting point. This enables the unified treatment of various decision criteria (e.g. the expected value criterion or the worst-case criterion). The main result of this article says that under minimal conditions optimal stationary policies are greedy w.r.t. the optimal cost function and vice versa. Based on this result we feel that former results on reinforcement learning can be transferred to other decision criteria provided that the decision criterion is decomposable by an appropriate mapping.*

1 Introduction

Reinforcement learning (RL) refers to a class of learning tasks and algorithms in which the learning system learns an associative mapping $\pi : X \rightarrow A$ from the state to the action space, by minimizing a scalar evaluation (reinforcement) of the cost of its behavior received from the environment. Most RL work assume the use of the *expected* total discounted cost criterion (see e.g. [2, 7, 1]). A lot of alternative decision criteria have been suggested in decision theory [6, 4, 5] to get a more sophisticated consideration of *risk* but most recent results on RL do not concern with this subject. In this paper we propose a general framework for delayed RL that enables the unified treatment of a broad class of decision criteria. The proposed framework is based on the notion of the one step lookahead (OLA) cost operator: The decision criterion must be defined by the successive application of the OLA cost operator that acts on cost functions. We assume that the operator is monotone, non-negative (maps non-negative functions to non-negative functions) and continuous on the left. These properties are satisfied by the operators corresponding to usual cost criteria. For example, the expected and worst-case cost criteria fit the framework. We show that under these properties optimal policies are greedy w.r.t. the optimal cost function and vice versa. Based on this result assuming finite action space one can show that optimal stationary policies exist. In this article we focus only on the development of the theory, the proofs of most of the theorems are presented elsewhere.

2 The OLA mapping

2.1 Notations

Let $B(S)$ denote the set of bounded, real-valued functions over the set S . Let $\mathbf{0}_S \in B(S)$ denote the identically zero function: $\mathbf{0}_S(s) = 0$ for all $s \in S$. For convenience $\mathbf{0}_S$ will be abbreviated to $\mathbf{0}$. We will interpret $\leq$ (or $\geq$) on $B(S)$ in the usual way: $f \leq g$ means that $f(s) \leq g(s)$ for all $s \in S$. $\mathcal{P}(X)$ will denote the power set of X : $\mathcal{P}(X) = \{Y \mid Y \subseteq X\}$.

2.2 Definitions

The basis of our framework is a class of sequential optimal control problems defined in terms of a discrete-time dynamical system with a *state space* X . Let $P : X \times A \rightarrow \mathcal{P}(X)$ be an arbitrary mapping – the *transition rule*, where A is the *action space*. The *decision process*, given by a triple (X, A, P) is as follows: it consists of successive stages indexed by non-negative integers. A stage consists of a state and an action. Assume that at stage t the process is in the state x_t and the decision maker chooses the action a_t . Then the next state of the process satisfies $x_{t+1} \in P(x_t, a_t)$. The rule applied by the decision maker to select the actions at successive stages is called the *policy* of the decision maker. It is casual to assume that the policy depends only on the history of the decision process. If the rule depends only on the last state of the process, the policy is called a *stationary policy*.¹ The objective of decision making is to reduce the cost associated with the decision process. A cost function is a function over X ; it determines the cost that will incur when the process starts in state x . Let $\mathcal{Q} : B(X) \rightarrow B(X \times A)$ be an arbitrary mapping. We will refer to $\mathcal{Q}$ as the *one step lookahead cost mapping* (OLA cost mapping). Mapping $\mathcal{Q}$ defines the cost of choice, i.e., $(\mathcal{Q}f)(x, a)$ is the cost that will incur if the process starts in state x , the first decision is a , and the process stops after one stage and we incur a final cost $f(y)$ provided that the final state is y . The cost of an action sequence is defined by the successive application of the mapping $\mathcal{Q}$: The cost of the action sequence $\{a_1, a_2, \dots, a_t, \dots\}$ is defined as the limit of the following function sequence:

$$v_0(x) = 0, \quad x \in X$$

$$v_{n+1}(x) = (\mathcal{Q}v_n)(x, a_{n+1}), \quad x \in X.$$

Mapping $\mathcal{Q}$ usually depends on the state transition mapping, P . The most well-known example of $\mathcal{Q}$ corresponds to the expected total cost criterion.

Example 2.1 Markovian decision problem with the expected cost criterion. Assume that X is a measurable space and let $M_1(X)$ denote the probability measures over X . Let $\mu : X \times A \rightarrow M_1(X)$ be an arbitrary mapping. We interpret $\mu(x, a)(S)$ as the probability of $x_{t+1} \in S$ provided that $x_t = x$ and $a_t = a$. Let $c : X \times A \rightarrow B(X)$ be a measurable mapping; c is called the *immediate cost mapping* and $c(x, a)(y)$ is the cost that incurs when the process is in state x , action a is executed and the next state is y . The OLA cost mapping corresponding to the total expected cost criterion is the following:

$$(\mathcal{Q}f)(x, a) = \int_X (c(x, a) + \gamma f) d\mu(x, a),$$

where $0 < \gamma \leq 1$ is the discount factor.

We will use this example for the illustration of assumptions and concepts of the framework.

In operations research and decision theory (e.g. [6]) it is well known, that the use of the expected value as a decision criterion is not always reliable and can be misleading. Two

¹It is usual to restrict the set of possible actions from state to state. However, this restriction has no essential effect on subsequent theorems and we disregard it.

examples of this phenomena are the celebrated *St. Petersburg Paradox* and a lottery with zero expected reward, small gain and high stake [6]. Alternative cost criteria are of interest. A family of such criteria can be defined as follows:

Example 2.2 Markovian decision problem: Generalized expectations. *In this case the above OLA mapping is modified as follows:*

$$(\mathcal{Q}f)(x, a) = \left(\int_X (c(x, a) + \gamma f)^p d\mu(x, a) \right)^{\frac{1}{p}},$$

where $0 < p \leq \infty$, $c(x, a)$ and f are assumed to be positive. If $p = \infty$ we get the worst-case cost mapping [4], or the min-max cost criteria. This criterion is intensively used in Game's Theory.

Let us fix an arbitrary OLA mapping, $\mathcal{Q}$. Let $\pi : X \rightarrow A$ be an arbitrary stationary policy (s.p.). The evaluation of policies is based on their evaluation operator:

Definition 2.1 The OLA evaluation operator corresponding to π is $T_\pi : B(X) \rightarrow B(X)$, with

$$(T_\pi f)(x) = (\mathcal{Q}f)(x, \pi(x)),$$

where $f \in B(X)$ and $x \in X$.

Hence $T_\pi v$ is the cost if we use policy π , but we are terminated after one period and incur a final cost $v(x)$ when the final state is x . Greedy behavior will be proved to be optimal. The greedy cost and policy forming operators are defined as follows:

Definition 2.2 The OLA greedy operator is the mapping $G : B(X) \rightarrow B(X)$ that satisfies

$$(Gf)(x) = \inf_{a \in A} (\mathcal{Q}f)(x, a), \quad (1)$$

where $f \in B(X)$ and $x \in X$.

Definition 2.3 The OLA Greedy Policy Forming operator $\Gamma : B(X) \rightarrow \mathcal{P}(A^X)$ is defined by

$$\Gamma f = \{ \pi \in A^X \mid T_\pi f = Gf \}.$$

If A is finite, then the infimum in Equation (1) is reached at some action and thus Γf is non-empty. If A is infinite, then it may be that for some $f \in B(X)$ the infimum is never reached, that is $\Gamma f = \emptyset$.

3 Assumptions and the objective of decision making

Below we postulate 3 assumptions; in the rest of the paper we assume that these assumptions hold. The first assumption expresses the natural requirement that the relation between two "future" cost functions remains the same provided that the first decisions coincide. In other words, this means that if one has two policies that coincide in the first decision then only the remaining decisions determine the ordering of the policies.

Assumption 3.1 We assume that $\mathcal{Q}$ is monotone non-decreasing, that is if $u, v \in B(X)$ and $u \leq v$ then $\mathcal{Q}u \leq \mathcal{Q}v$.

This assumption is satisfied for the total expected cost (TEC) OLA mapping: let $f, g \in B(X)$ and $f \leq g$. $\mathcal{Q}f \leq \mathcal{Q}g$ means that there should hold

$$\int_X (c(x, a) + \gamma f) d\mu(x, a) \leq \int_X (c(x, a) + \gamma g) d\mu(x, a)$$

for all $(x, a) \in X \times A$. This is equivalent to $\int f d\mu(x, a) \leq \int g d\mu(x, a)$ (since $\gamma > 0$) and this inequality follows from the properties of integration.

Assumption 3.2 $Q0 \geq 0$.

As it is easy to check that for the TEC OLA mapping this assumption means that the average of immediate costs should be non-negative, i.e., there should hold for all $(x, a) \in X \times A$ that

$$\int_X c(x, a) d\mu(x, a) \geq 0.$$

Almost all the applications satisfy this assumption [3].

As a consequence of the above assumptions we have $T_\pi 0 \geq 0$ for all s.p. π and thus $T_\pi^n 0$ is an increasing sequence. Assumption 3.2 is not needed if Q has the contraction property. The following assumption postulates that infinite horizon costs exist for every s.p.

Assumption 3.3 *Assume that $T_\pi^n 0$ converges for every s.p. π . We will denote the limit by v_π and call it the evaluation of π . Further, we assume that v_π is the fixed point of the operator T_π .*

This assumption corresponds to a theorem of most of the conventional constructions. For example, in the case of Markovian decision problems with the TEC criterion the evaluation of a policy is defined as the expectation of

$$\sum_{t=0}^{\infty} \gamma^t c_x(t),$$

where $c_x(t)$ is the cost incurred at stage t provided that we start from state x and apply the policy π . The interpretation of expectation goes as follows: The application of policy π from state x generates a random process and a corresponding probability measure over the set of possible infinite state sequences (this measure is defined by the product of measures $\mu(x, \pi(x))$). The expectation of the total cost is taken according to this measure. The usual way to show that assumption 3.3 holds is that one shows that T_π is a contraction mapping w.r.t. the supremum norm (this is ensured if $0 < \gamma < 1$) and that the evaluation of any policy (the evaluation defined as the expectation of the sum of discounted costs) is indeed the fixed point of its evaluation. Since contraction mappings have exactly one fixed point and the successive application of a contraction mapping on any function converges to the fixed point of the mapping the assumption thus follows.

Definition 3.1 *The optimal cost function $v^* : X \rightarrow \mathbf{R}$ is the function defined by*

$$v^*(x) = \inf_{\pi \in A^X} v_\pi(x). \quad (2)$$

The optimal cost function exists provided that all the cost functions are bounded from below by 0. No policy can do better than v^* . This motivates the following definition:

Definition 3.2 *A s.p. π is called optimal if its evaluation function is the optimal cost function, that is if $v_\pi = v^*$. We denote the set of optimal policies by Π^* .*

4 Optimal policies are greedy

In this section we show that the set of optimal policies, Π^* , is equal to the set of greedy policies w.r.t. the optimal evaluation function. This is proved through a series of propositions. We begin with a simple, but important observation:

Proposition 4.1 *For every $u \in B(X)$ for which $0 \leq u \leq v_\pi$ there holds $\lim_{n \rightarrow \infty} T_\pi^n u = v_\pi$.*

The next lemma gives the first part of the theorem.

Lemma 4.1 *The OLA greedy policies w.r.t. the optimal evaluation function are optimal, i.e.,*

$$\Gamma v^* \subset \Pi^*.$$

The proof exploits the facts that for any s.p. π there holds that $G \leq T_\pi$ and T_π is monotone non-decreasing. The converse of this lemma follows from the generalization of Howard's Policy Improvement (PI) Theorem that states that one-step lookahead improvements are infinite horizon improvements, too. To state this lemma we need the following definition:

Definition 4.1 *We say that the policy π is an OLA improvement w.r.t. the cost function v if $\pi \in \Gamma v$. We say that the policy π is a proper OLA improvement w.r.t. the cost function v if π is an OLA improvement and there exists at least one state, say $x \in X$ such that $(Gv)(x) < v(x)$.*

Howard's PI Theorem says that OLA improvements are global improvements:

Lemma 4.2 (Howard's Policy Improvement) *Let π be an arbitrary s.p. and let $\hat{\pi}$ be an OLA improvement w.r.t. the cost function of π , v_π . Then there holds $v_{\hat{\pi}} \leq v_\pi$. Furthermore, if $\hat{\pi}$ is a proper OLA improvement then for all state x for which $\hat{\pi}$ is a real OLA improvement over π there holds $v_{\hat{\pi}}(x) < v_\pi(x)$.*

The proof is based on induction and goes by showing that $T_{\hat{\pi}}^n v \leq v_\pi$ holds for arbitrary n , where v is defined by $v(x) = \min(v_\pi(x), v_{\hat{\pi}}(x))$, $x \in X$. From the above two lemmas the theorem follows easily:

Theorem 4.1 *The set of optimal and greedy policies w.r.t. the optimal cost function, v^* , coincide:*

$$\Pi^* = \Gamma v^*.$$

Proof. We prove that if π is an arbitrary s.p. and $Gv_\pi \neq v_\pi$ then π is not optimal. From this it follows that $\Pi^* \subseteq \Gamma v^*$ and thus by lemma 4.1 $\Pi^* = \Gamma v^*$.

Let π denote an arbitrary optimal policy. Since π is optimal, $v_\pi = v^*$. It should be proved that $Gv_\pi = v_\pi$. We know that $Gv_\pi \leq v_\pi$. Assume that there exists a state x_0 for which $(Gv_\pi)(x_0) < v_\pi(x_0)$. Let $\hat{\pi} \in \Gamma v_\pi$. By the above, policy improvement theorem $v_{\hat{\pi}}(x_0) < v_\pi(x_0)$. This contradicts with the optimality of v_π . Thus $Gv_\pi = v_\pi = T_\pi v_\pi$, and thus $\pi \in \Gamma v_\pi = \Gamma v^*$. $\square$

If there exists at least one optimal policy than a fixed point equation that may be called the Bellman Optimality Equation (BOE) [3] is satisfied:

Corollary 4.1 *Assume that $\Gamma v^* \neq \emptyset$. Then the fixed point equation*

$$v^*(x) = \inf_{a \in A} (Qv^*)(x, a) \quad (3)$$

(or shortly $v^ = Gv^*$) holds.*

Basically two methods can be used to determine an optimal policy:² policy iteration and value iteration. Both methods rely on BOE.

²Linear programming that is also often used for solving MDPs, relies on the fact that Q is linear. If Q is non-linear it can no longer be used and thus we neglect it.

5 Conclusion and future work

We have proposed a general framework for reinforcement learning and/or adaptive control. In the proposed framework various cost criteria are treated in a unified way. We have presented the outline of the main theorems of a generalized reinforcement learning theory. The theorems may have real-world applications as well. The knowledge of these theorems makes it possible, e.g., to think of agents that use a kind of mixed and changing decision criterion: such an agent would adapt its decision criterion based on her experiences: it would use the worst-case cost mapping to avoid dangerous regions while at other places she would use the usual expectation criterion.

Meanwhile some new results were developed. The most important is the characterization of OLA mappings for which there are optimal stationary policies that are optimal in the set of *all* possible policies. Other extensions to these results are that one may substitute the zero terminal reward function by any other finite terminal reward function, the same results hold for non-positive decision problems and for decision problems whose OLA mapping is a contraction. The boundedness condition of the evaluation functions can be weakened, too. Future work should deal with algorithms (iterative learning procedures) arising from this approach. Scaling analysis and convergence speed estimations are required. Partially observable decision processes should also be investigated since they are of great importance from practical aspects.

Taking into account the limited storage space and time constraints calls for asynchronous and partial convergence theorems. Learning from real-time data is another important aspect of learning. It follows that the adaption of stochastic learning methods is also required. It is believed that the proposed framework is useful as it is suitable for the integration of various branches of sequential decision making and thus may eliminate unnecessary repetitions of the proofs of the basic theorems. We feel that the proposed framework also gives a deeper insight in sequential decision making.

References

- [1] A.G. Barto, S.J. Bradtke, and S.P. Singh. *Real-time Learning and Control using Asynchronous Dynamic Programming*. Technical Report 91-57, Computer Science Department, University of Massachusetts, 1991.
- [2] A.G. Barto, R.S. Sutton, and C.W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:834–846, 1983.
- [3] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.
- [4] M. Heger. Consideration of risk in reinforcement learning. 1994. Revised submission to the 11th International Machine Learning Conference ML-94.
- [5] Cs. Szepesvári and András Lőrincz. Behavior of an adaptive self-organizing autonomous agent working with cues and competing concepts. *Adaptive Behavior*, 2(2):131–160, 1994.
- [6] H.A. Taha. *Operations Research: An Introduction*. MacMillen Publishing Company, New York, fourth edition, 1987.
- [7] C.J.C.H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, 1990.