Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Stochastic Safest and Shortest Path Problems

Florent Teichteil-Königsbuch
florent.teichteil@onera.fr
Onera — The French Aerospace Lab
F-31055, Toulouse, France

Abstract

Optimal solutions to Stochastic Shortest Path Problems (SSPs) usually require that there exists at least one policy that reaches the goal with probability 1 from the initial state. This condition is very strong and prevents from solving many interesting problems, for instance where all possible policies reach some dead-end states with a positive probability. We introduce a more general and richer dual optimization criterion, which minimizes the average (undiscounted) cost of only paths leading to the goal among all policies that maximize the probability to reach the goal. We present policy update equations in the form of dynamic programming for this new dual criterion, which are different from the standard Bellman equations. We demonstrate that our equations converge in infinite horizon without any condition on the structure of the problem or on its policies, which actually extends the class of SSPs that can be solved. We experimentally show that our dual criterion provides well-founded solutions to SSPs that can not be solved by the standard criterion, and that using a discount factor with the latter certainly provides solution policies but which are not optimal considering our well-founded criterion.

Introduction

Research on probabilistic planning has essentially focused on either maximizing the probability to reach the goal from an initial state (Kolobov et al. 2011; Teichteil-Königsbuch, Kuter, and Infantes 2010; Puterman 1994) or minimizing the average accumulated costs if there exists a policy that reach the goal with probability 1, named *proper* policy (Bertsekas and Tsitsiklis 1996; Bonet and Geffner 2005; 2003; Kolobov, Mausam, and Weld 2010; Kolobov et al. 2011; Yoon et al. 2010). Yet, to the best of our knowledge, no approaches optimize both the probability to reach the goal, and minimize the average accumulated costs at the same time in a proper theoretical framework. Moreover, if the maximum probability to reach the goal is strictly less than 1 for a given problem, i.e. if there does not exist proper policies, it is possible to minimize the average *discounted* accumulated costs (Teichteil-Königsbuch, Vidal, and Infantes 2011), but a drawback of this approach is that costs of paths that do not reach the goal are uselessly (and regrettably) taken into account when optimizing average accumulated costs.

In this paper, we first propose a new infinite-horizon dual optimization criterion, which selects policies that minimize the average (undiscounted) accumulated costs of only paths that reach the goal among all policies that maximize the probability to reach the goal. This dual criterion is often considered as an important evaluation metrics (Younes et al. 2005), but, to the best of our knowledge, no theoretical nor practical means exist to optimize these metrics hand-in-hand. We provide an illustrative example with both goal and dead-end states, which highlight the benefits of our dual criterion with regards to previous approaches. Second, we propose update equations for evaluating this dual criterion for any stationary policy, and prove that these equations always converge to finite-values as the reasoning horizon tends to $+\infty$, without any assumption on the structure of the problem considered or its policies (contrary to previous approaches).

However, in practice, constructing optimal stationary policies for this dual criterion appears to be especially difficult in the general case, i.e. with positive or negative costs. Thus, we provide optimality equations for our dual criterion in the case where all costs are positive. These equations are different from the standard Bellman equations, but: (i) their time complexity is also polynomial in the number of states and actions of the problem, and (ii) they provide the same optimal policies as SSPs for problems where SSP assumptions hold. Finally, we experimentally demonstrate on the basis of various benchmarks, that existing approaches, which optimize either the probability to reach the goal or the average accumulated costs over all reachable paths (not only the ones that reach goal), do not need to provide optimal policies in the sense of our dual criterion.

Goal-oriented Markov Decision Processes. We consider probabilistic planning problems defined as goal-oriented Markov Decision Processes (MDPs), which are tuples $\langle S, A, T, c, G\rangle$ such that (Bertsekas and Tsitsiklis 1996): S is the finite set of states ; G is the finite set of goal states ; A is the finite set of actions ; $T : S \times A \times S \to [0; 1]$ is a transition function such that, for any $(s, a, s') \in S \times A \times S$ and time step $t \in \mathbb{N}$, $T(s, a, s') = Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$; $c : S \times A \times S \to \mathbb{R}$ is the cost function such that, for any $(s, a, s') \in S \times A \times S$, $c(s, a, s')$ is the cost paid when going from state s to state s' and executing action a. We do not

assume positive costs in the general case. We assume that any goal state $g \in G$ is absorbing ($T(g,a,g) = 1, \forall a \in A$), and pays no cost ($c(g,a,g) = 0, \forall a \in A$). We note $app : S \rightarrow 2^A$ the function that gives the set of actions applicable in a given state. A solution of a goal-oriented MDP is a policy $\pi : S \rightarrow A$ that optimizes a given criterion, usually the probability to reach the goal from any initial state, or the average accumulated costs paid from any initial state.

Stochastic Shortest Path Problems. Efficient methods for solving goal-oriented MDPs are available if two assumptions are met (Bertsekas and Tsitsiklis 1996): (i) there exists at least one policy π that reaches the goal with probability 1, named proper policy, and (ii) all improper policies accumulate infinite expected cost. Assumption (ii) means that all cycles in the transition graph, which do not lead to goal states, have positive costs. Problems that meet these assumptions are called *Stochastic Shortest Path* problems (SSPs). Methods for solving SSPs compute the fixed point C^* of the following Bellman equation, which is the optimal achievable accumulated cost averaged over all paths starting in any initial state s, named *total cost criterion* or *cost-to-go* function:

$$C^*(s) = \min_{a \in app(s)} \sum_{s' \in S} T(s,a,s') \left(c(s,a,s') + C^*(s') \right) \quad (1)$$

If the assumptions of SSPs do not hold, for instance in the presence of dead-end states[1] reachable with a positive probability by executing all possible policies, the previous equation may have no solution. Yet, it can be slightly modified by multiplying $C^*(s')$ by a fixed discount factor $0 < \gamma < 1$, giving rise to the *discounted cost criterion* C^*_γ, which is proved to always have a solution (Puterman 1994). Some works (e.g. (Teichteil-Königsbuch, Vidal, and Infantes 2011)) proposed efficient methods to optimize goal MDPs in presence of dead-ends, using the discounted cost criterion, but we will show in the next that such approaches may be not appropriate in some cases with complex cost structures.

Stochastic *Safest* and Shortest Path Problems

The traditional criterion used in SSPs is not well-founded when there is no proper policy. Indeed, in this case, it may diverge because it may sum an infinite number of costs over the paths that do not reach the goal. If discounted, it converges but: 1) it may produce policies that do not maximize the probability to reach the goal (because costs of paths that do not reach the goal may attract the policy), and 2) it is even not optimal considering the costs averaged only over the paths that reach the goal. We think that the only proper way to optimize the probability to reach the goal on one hand, and the accumulated costs averaged over only the paths that reach the goal on the other hand, is to separate these two concurrent criteria in two different, but parallel, evaluation and optimization schemes.

Goal-probability and goal-cost functions. For a given state $s \in S$, policy $\pi \in A^S$, and $n \in \mathbb{N}$, we note $P^{G,\pi}_n(s)$ the probability of reaching the goal G in at most n time steps by executing π from s. This function is named *goal-probability function* in at most n time steps (steps-to-go). In the finite-horizon case, π is a series of policies $(\pi_0, \cdots, \pi_{H-1}), H \in \mathbb{N}$, where π_k is the policy executed at step-to-go k. We also note $C^{G,\pi}_n(s)$ the (undiscounted) costs accumulated by executing π from s, averaged only over the paths that reach the goal G with a positive probability. We name it *goal-cost function* in at most n time steps (steps-to-go). Importantly, this latter function is different from the value function traditionally used in MDPs, since the latter is averaged over all paths starting in s (not only the ones reaching the goal).

Infinite horizon dual optimization criterion. Interestingly, we will prove in this paper that $P^{G,\pi}_n$ and $C^{G,\pi}_n$ both converge to finite values as horizon H (or time steps n) tend to $+\infty$, for any goal-oriented MDPs, stationary policy π, and *without any condition* on the MDP structure. Note that this powerful property is specific to our goal-probability and goal-cost functions; it does not hold for standard MDP criteria, for which convergence is usually conditioned on characteristics of the underlying controlled Markov chain, or on some discount factor, as discussed before.

Based on goal-probability and goal-cost metrics, we define *Stochastic Safest and Shortest Path* problems (S^3Ps for short), which are goal-oriented MDPs where, for all $s \in S$, we aim at finding a policy $\pi^*(s)$ that *minimizes the accumulated costs averaged over the paths that reach the goal from s, among all policies that maximize the probability to reach the goal*:

$$\pi^*(s) \in \underset{\pi : \forall s' \in S, \pi(s') \in \text{argmax}_{\pi' \in A^S} P^{G,\pi'}_\infty(s')}{\text{argmin}} C^{G,\pi}_\infty(s) \quad (2)$$

S^3Ps include the former traditional *Stochastic Shortest Path problems* (S^2Ps for short): $S^2P \subset S^3P$. It is worth noting that S^3Ps also include the single-criterion recently proposed by (Kolobov et al. 2011), where the authors proposed the largest known class of goal-oriented MDP problems, named GSSPs, to optimize either average accumulated costs among only proper policies but with general rewards, or the goal-probability function. But GSSPs do not allow for dual optimization of these two criteria, contrary to us (remind that we also deal with general costs — or rewards). Then, the class of goal MDPs for which optimal solutions exist (i.e. are well-founded) is now extended further: $S^2P \subset GSSP \subset S^3P$.

Illustrative example. Obviously, eq. 2 shows that S^3P optimal policies are the same as SSP optimal policies if there exists a policy reaching the goal with probability 1 (assumption (i) of SSPs) and if such SSP optimal policies are well-founded (equivalent to assumption (ii)). The later assumption may be violated for two different reasons: either the SSP cost-to-go criterion has an infinite value in some initial states, or it has finite values but these values can not be obtained from Bellman equation (eq. 1).

Figure 1 illustrates a goal-oriented MDP for which neither assumptions (i) nor (ii) of SSPs are met, but for which the cost-to-go function used in SSPs is finite (i.e. it can not be obtained using eq. 1). There are 4 possible policies, depending on whether action a_1 or a_2 or a_3 or a_I are chosen in the

[1] A *dead-end* state is a state from which no path can reach the goal with a positive probability, whatever the policy executed.

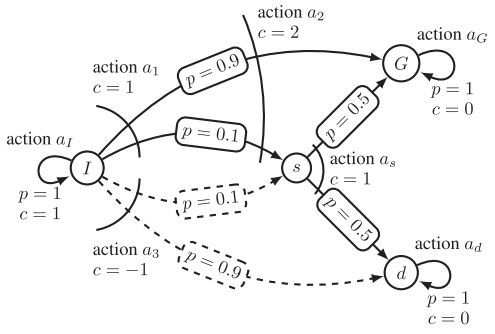

Figure 1: S^{3}P without proper policy. I is the initial state, G the goal, d a dead-end state, and s an intermediate state.

initial state I. Let name them respectively π_1, π_2, π_3 and π_4. Starting in I, policies π_1 and π_2 lead to G (resp. d) with the same probability $0.9 + 0.1 \times 0.5 = 0.95$ (resp. 0.05). Policy π_3 leads to G (resp. d) with probability $0.1 \times 0.5 = 0.05$ (resp. 0.95). Action a_I is absorbing and does not lead at all to G. Thus, the maximum goal-probability policies are π_1 and π_2. However, minimizing the standard cost-to-go criterion will choose policy π_3. Indeed, using eq. 1, the cost-to-go value of s is 1, so that the cost-to-go value of π_1 in I is $C^{\pi_1}(I) = 0.9\times(1+0)+0.1\times(1+1) = 1.1$, the one of π_2 is $C^{\pi_2}(I) = 0.9\times(2+0)+0.1\times(2+1) = 2.1$, and the one of π_3 is $C^{\pi_3}(I) = 0.1\times(-1+1)+0.9\times(-1+0) = -0.9$. Policy π_4 has an infinite value and will be discarded. It means that the standard cost-to-go criterion used in SSP, in this case, does not select the maximum goal-probability policy.

Yet, if we first optimize the goal-probability function as defined before, we will select policies π_1 and π_2. By executing policy π_1 (resp. π_2) from I, there are only two paths to the goal: one with probability 0.9 and total cost 1 (resp. 2), the other with probability 0.05 and total cost 2 (resp. 3). The accumulated probability of these paths is 0.95, and can be viewed as a normalization constant of the accumulated cost averaged over these 2 paths. Thus, the goal-cost function of π_1 as defined in this paper, is: $1/0.95 \times (0.9 \times 1 + 0.05 \times 2) = 1/0.95 \simeq 1.05$. The one of π_2 is: $1/0.95\times(0.9\times2+0.05\times3) = 1.95/0.95 \simeq 1.85$. Thus, π_1 is the optimal S^{3}P policy, whose goal-cost function is equal to 1.05. The standard cost-to-go function used in SSPs has a higher value for π_1, because it also averages over the path to the dead-end state d: $0.9 \times 1 + 0.05 \times 2 + 0.05 \times 2 = 1.1$.

Finally, imagine that immediate cost of action a_d is 1 (instead of 0). In this case, the cost-to-go value used in SSPs diverges to $+\infty$ from all states, i.e. it is not well-founded. Yet, the reader can check that previous calculations for S^{3}P policies are still valid and yield the same goal-probability and goal-cost functions, and optimal policy π_1. One may ask whether discounting the cost-to-go function, so that it has now a finite value, will provide the same optimal policy as for S^{3}Ps. As proved later in the experiment section of this paper, the answer is negative for problems with complex cost structures.

Evaluating finite-horizon policies for S^{3}Ps

We present in this section a theorem for evaluating finite-horizon policies, which is fundamental to study mathematical properties of the goal-probability and goal-cost functions. We could prove that $P_n^{G,\pi}$ can be computed via a translation of the original MDP into an MDP where all rewards are equal to 0 except the ones of direct transitions to the goal that are equal to 1 (Kolobov et al. 2011). However, the goal-cost update equation presented below is not equivalent to standard Bellman evaluation equations for MDPs, because costs are averaged only over paths that reach the goal, but not over all paths (as in standard MDPs).

Theorem 1. ***(Policy evaluation equations for finite-horizon S^{3}Ps).*** *Let $H \in \mathbb{N}$ be the finite horizon of the problem. For any step-to-go $1 \leqslant n < H$, any history-dependent policy $\pi = (\pi_0, \cdots, \pi_{H-1})$ and any state $s \in S$:*

$$P_n^{G,\pi}(s) = \textstyle\sum_{s' \in S} T(s, \pi_{H-n}(s), s') P_{n-1}^{G,\pi}(s'), \textit{with:}$$
$$P_0^{G,\pi}(s) = 0, \forall s \in S \setminus G, \textit{and } P_0^{G,\pi}(g) = 1, \forall g \in G \quad (3)$$

If $P_n^{G,\pi}(s) > 0$, $C_n^{G,\pi}(s)$ is well-defined, and satisfies:

$$C_n^{G,\pi}(s) = \frac{1}{P_n^{G,\pi}(s)} \textstyle\sum_{s' \in S} T(s, \pi_{H-n}(s), s') P_{n-1}^{G,\pi}(s') \times \left[c(s, \pi_{H-n}(s), s') + C_{n-1}^{G,\pi}(s') \right], \textit{with:}$$
$$C_o^{G,\pi}(s) = 0, \forall s \in S \quad (4)$$

Proof. Equation 3 is easy to obtained with a reasoning similar, but much simpler, to the one used to demonstrate equation 4. Thus we will only present the demonstration of the latter. Let $\Phi_n^{G,\pi}(s)$ be the set of paths that reach the goal G in at most n time steps by executing policy π from a state s. For any $\phi \in \Phi_n^{G,\pi}(s)$, we note $|\phi|$ the length of ϕ until it reaches the goal, $\phi(i)$ the i^{th} state visited in the path for $0 \leqslant i \leqslant |\phi|$, and ϕ_i the sub-path of ϕ starting in $\phi(i)$.

Calculation of $C_n^{G,\pi}$ is averaged using a conditional probability distribution, conditioned on the only trajectories that reach the goal. Thus, it requires to calculate first the update equation of the following conditional probability ($n \geqslant 1$), where we note $\omega_{s,n}^{G,\pi}$ the event "execution of π from state s will yield a path that will reach the goal in at most n steps-to-go", only meaningful if $P_n^{G,\pi}(s) > 0$ (we have $Pr(\omega_{s,n}^{G,\pi}) = P_n^{G,\pi}(s)$, and probability conditionings on π are implicit):

$$\begin{aligned}
p_{\phi,n}^{G,\pi} &= Pr(\text{“executing } \phi\text{”} \mid \omega_{\phi(0),n}^{G,\pi}) \\
&= \frac{Pr(\omega_{\phi(0),n}^{G,\pi} \mid \text{“executing } \phi\text{”}) Pr(\text{“executing } \phi\text{”})}{Pr(\omega_{\phi(0),n}^{G,\pi})} \\
&= \frac{Pr(\omega_{\phi(1),n-1}^{G,\pi} \mid \text{“ex. } \phi_1\text{”}) Pr(\text{“}\phi(0) \text{ to } \phi(1)\text{”}) Pr(\text{“ex. } \phi_1\text{”})}{P_n^{G,\pi}(\phi(0))} \\
&= T(\phi(0), \pi_{H-n}(\phi(0)), \phi(1)) \frac{P_{n-1}^{G,\pi}(\phi(1))}{P_n^{G,\pi}(\phi(0))} \times \\
&\qquad \underbrace{\frac{Pr(\omega_{\phi(1),n-1}^{G,\pi} \mid \phi_1) Pr(\text{“ex. } \phi_1\text{”})}{P_{n-1}^{G,\pi}(\phi(1))}}_{Pr\left(\text{“ex. } \phi_1\text{”} \middle| \omega_{\phi(1),n-1}^{G,\pi}\right) = p_{\phi_1,n-1}^{G,\pi}}
\end{aligned}$$

Now, noting $c(\phi)$ the accumulated cost over a path ϕ, we have for any state s such that $P_n^{G,\pi}(s) > 0$:

$$
\begin{aligned}
C_n^{G,\pi}(s) &= \textstyle\sum_{\phi \in \Phi_n^{G,\pi}(s)} \; Pr(\text{“executing } \phi\text{”} \mid \omega_{s,n}^{G,\pi})\, c(\phi) \\
&= \textstyle\sum_{\phi \in \Phi_n^{G,\pi}(s)} \; p_{\phi,n}^{G,\pi} \times (c(s, \pi_{H-n}(s), \phi(1)) + c(\phi_1)) \\
&= \textstyle\sum_{\phi \in \Phi_n^{G,\pi}(s)} \; T(s, \pi_{H-n}(s), \phi(1)) \frac{P_{n-1}^{G,\pi}(\phi(1))}{P_n^{G,\pi}(s)} \times \\
&\qquad p_{\phi_1,n-1}^{G,\pi} \times (c(s, \pi_{H-n}(s), \phi(1)) + c(\phi_1)) \\
&= \frac{1}{P_n^{G,\pi}(s)} \textstyle\sum_{s' \in S} \; T(s, \pi_{H-n}(s), s') P_{n-1}^{G,\pi}(s') \times \\
&\qquad \Big[c(s, \pi_{H-n}(s), s') + \underbrace{\textstyle\sum_{\phi_1 \in \Phi_{n-1}^{G,\pi}(s')} \; p_{\phi_1,n-1}^{G,\pi}\, c(\phi_1)}_{C_{n-1}^{G,\pi}(s')} \Big]
\end{aligned}
$$

The last calculation step is due to the fact that $\sum_{\phi_1 \in \Phi_{n-1}^{G,\pi}(s')} \; p_{\phi_1,n-1}^{G,\pi} = 1$. □

Division by $P_n^{G,\pi}(s)$ in eq. 4 may be surprising, but it is a normalization constant of the mean defining the goal-cost function: indeed, the sum of $T(s, \pi_{H-n}(s), s') P_{n-1}^{G,\pi}(s')$ over successor states s' in eq. 4 is actually equal to $P_n^{G,\pi}(s)$.

Solving infinite horizon S^{3}Ps

This work was primarily motivated by infinite horizon goal-oriented MDP problems, for which SSP assumptions do not hold. This is actually the key to understand the intuition behind our dual criterion and why it converges to an infinite-horizon fixed point. The goal-probability function converges because: 1) states for which no paths lead to the goal have a constant 0 value after each update; 2) other states will eventually end up in G (with a monotonically increasing probability) or in one of the former states, ensuring convergence of the goal-probability function. The goal-cost function converges because: 1) it is not defined for states where no paths lead to the goal (as intended); 2) costs of other states are accumulated only along (and averaged only among) paths that reach the goal, whose transient probabilities converge to zero when the length of paths tends to $+\infty$, and paying no cost after reaching the goal. In comparison, the criterion used in SSPs accumulates costs also along paths that do not reach the goal (if any), whose costs may diverge to $\pm\infty$ in the general case, suppressing convergence guarantees.

The mathematical foundations of this intuition are actually quite complex, and rely on the following lemma, which proves that *the transition operator restricted to a stable subset of states in $S \setminus G$ that reach the goal with a positive probability for a given stationary policy, is a contraction.* This lemma can be seen as a non-trival generalization of the contraction property of proper policies' transition operator used in SSPs (Bertsekas and Tsitsiklis 1996), which had to be a contractive mapping over the entire state space.

Lemma 1. *Let $\mathcal{M}$ be a general goal-oriented MDP, π a stationary policy, T^π the transition matrix for policy π, and for all $n \in \mathbb{N}$, $\mathcal{X}_n^\pi = \{s \in S \setminus G : P_n^{G,\pi}(s) > 0\}$. Then: (i) for all $s \in S$, $P_n^{G,\pi}(s)$ converges to a finite value as n tends to $+\infty$; (ii) there exists $\mathcal{X}^\pi \subset S$ such that $\mathcal{X}_n^\pi \subset \mathcal{X}^\pi$ for all $n \in \mathbb{N}$ and T^π is a contraction over $\mathcal{X}^\pi$.*

Proof. (i) A simple mathematical induction using eq. 3 shows that, for any $s \in S$, $P_n^{G,\pi}(s)$ is increasing with n. As all $P_n^{G,\pi}(s)$ values are bounded by 1, they converge to some $P_\infty^{G,\pi}(s)$ values for all $s \in S$. (ii) This induction also shows that: $\forall n \in \mathbb{N}, \mathcal{X}_n^\pi \subset \mathcal{X}_{n+1}^\pi \subset S$. As S is finite, there exists $\mathcal{X}^\pi \in S$ and $n_0 \in \mathbb{N}$ such that for all $n \in \mathbb{N}$, $\mathcal{X}_n^\pi \subset \mathcal{X}^\pi$ and for all $n \geqslant n_0$, $\mathcal{X}_n^\pi = \mathcal{X}^\pi$. Thus, for all $s \in \mathcal{X}^\pi$ and $n \geqslant n_0$, $P_n^{G,\pi}(s) > 0$. Moreover, as mentioned before, $P_n^{G,\pi}(s)$ increases with n, so that: for all $s \in \mathcal{X}^\pi$, $P_\infty^{G,\pi}(s) = \lim_{n \to +\infty} P_n^{G,\pi}(s) > 0$. Therefore, the probability that any state in $\mathcal{X}^\pi$ is absorbed by the goal state is positive, meaning that $\mathcal{X}^\pi$ is a subset of the transient states of the Markov chain induced by policy π. Let W^π be the sub-matrix of T^π that maps transient states to themselves (transitions between transient states). It has been proved that W^π is a contraction, i.e. $\rho(W^\pi) < 1$, where $\rho(W^\pi)$ is the largest absolute eigenvalue of W^π (see Proposition A.3 in (Puterman 1994)). Now, by reordering transient states in such a way that states in $\mathcal{X}^\pi$ appears first, we can write W^π into the form: $W^\pi = \begin{pmatrix} T^\pi_{|\mathcal{X}^\pi} & A^\pi \\ 0 & B^\pi \end{pmatrix}$. Indeed, if the bottom left sub-matrix were not zero, we could go from a state $\tilde{s}$ not in $\mathcal{X}^\pi \cup G$ to a state in $\mathcal{X}^\pi$ with a positive probability, from which we could then reach the goal state by definition of $\mathcal{X}^\pi$: it means that there would exist $n_1 \in \mathbb{N}$ such that $P_{n_1}^{G,\pi}(\tilde{s}) > 0$, which contradicts the fact that $P_n^{G,\pi}(s) = 0$ for all state $s \notin \mathcal{X}^\pi \cup G$ and $n \in \mathbb{N}$. Finally, thanks to the previous form of W^π, we have: $\rho(T^\pi_{|\mathcal{X}^\pi}) \leqslant \rho(W^\pi) < 1$, i.e. T^π is a contraction over $\mathcal{X}^\pi$. □

Policy evaluation in infinite horizon for S^{3}Ps. Thanks to this helpful lemma, we can now demonstrate the convergence of policy evaluation and policy optimization equations. Like in standard MDPs, we introduce an update operator to prove convergence. For a given $n \in \mathbb{N}^*$ and stationary policy π, we note $\mathcal{L}_n^\pi$ the following operator, defined over functions $J : S \to \mathbb{R}$:

$$
(\mathcal{L}_n^\pi J)(s) = \sum_{s' \in S} T(s, \pi(s), s') [P_{n-1}^{G,\pi}(s') c(s, \pi(s), s') + J(s')]
$$

where $P_{n-1}^{G,\pi}$ is recursively defined as in Theorem 1.

Theorem 2. *Let $\mathcal{M}$ be a general goal-oriented MDP, and π any stationary policy for $\mathcal{M}$. Evaluation equations of Theorem 1 converge to finite values $P_\infty^{G,\pi}(s)$ and $C_\infty^{G,\pi}(s)$ for any $s \in S$ (by convention, $C_n^{G,\pi}(s) = 0$ if $P_n^{G,\pi}(s) = 0, n \in \mathbb{N}$).*

Proof. As shown by Lemma 1, convergence of the goal-probability series of functions is independent from the goal-cost functions. Noting $\mathcal{X}_n^\pi = \{s \in S \setminus G : P_n^{G,\pi}(s) > 0\}$, this lemma also proves that there exists $\mathcal{X}^\pi \subset S$ such that $\mathcal{X}_n^\pi \subset \mathcal{X}^\pi$ for all $n \in \mathbb{N}$ and T^π is a contraction over $\mathcal{X}^\pi$.

We can notice from equation 4 that, for all $n \in \mathbb{N}$, $C_n^{G,\pi}(s) = 0$ for all $s \in G$. For states $s \in S \setminus (G \cup \mathcal{X}^\pi)$, $P_n^{G,\pi}(s) = 0$ for all $n \in \mathbb{N}$ (because $\mathcal{X}_n^\pi \subset \mathcal{X}^\pi$ for all $n \in \mathbb{N}$), so that $P_n^{G,\pi}(s)$ is not defined but constantly equal to zero by convention. Therefore, operator $\mathcal{L}_n^\pi$, restricted to operate over the subspace of functions $\Gamma = \{J : S \to \mathbb{R}\,; J(s) = 0, s \in S \setminus \mathcal{X}^\pi\}$, is equivalent to update equation 4, meaning that the latter converges only and only if $\mathcal{L}_n^\pi$

converges over Γ. We will actually demonstrate that $\mathcal{L}_n^\pi$ is a contraction over Γ ; for all J_1 and J_2 in Γ, and $s \in \mathcal{X}^\pi$, we have:

$$\begin{aligned}|(\mathcal{L}_n^\pi J_1)(s) - (\mathcal{L}_n^\pi J_2)(s)| &\leqslant \max_{s\in\mathcal{X}^\pi} |(\mathcal{L}_n^\pi J_1)(s) - (\mathcal{L}_n^\pi J_2)(s)| \\ &= \max_{s\in\mathcal{X}^\pi} \left|\sum_{s'\in\mathcal{X}^\pi} T^\pi_{|\mathcal{X}^\pi}(s,s')(J_1(s') - J_2(s'))\right| \\ &= \|T^\pi_{|\mathcal{X}^\pi}(J_1 - J_2)\|_{\mathcal{X}^\pi} \leqslant \|T^\pi_{|\mathcal{X}^\pi}\| \cdot \|J_1 - J_2\|_{\mathcal{X}^\pi}\end{aligned}$$

Thus: $\|(\mathcal{L}_n^\pi J_1) - (\mathcal{L}_n^\pi J_2)\|_{\mathcal{X}^\pi} \leqslant \|T^\pi_{|\mathcal{X}^\pi}\| \cdot \|J_1 - J_2\|_{\mathcal{X}^\pi}$ Moreover, by definition of Γ: $\|(\mathcal{L}_n^\pi J_1) - (\mathcal{L}_n^\pi J_2)\|_{S\setminus\mathcal{X}^\pi} = 0$. As $T^\pi_{|\mathcal{X}^\pi}$ is a contraction, $\mathcal{L}_n^\pi$ is a contraction over Γ, for all $n \in \mathbb{N}^*$, and its contraction constant and Γ do not depend on n. Therefore, by (generalized) Banach fixed point theorem, any suite J_n of functions in Γ such that $J_{n+1} = \mathcal{L}_n^\pi J_n$ converges to a unique fixed point $J_\infty = P_\infty^{G,\pi} C_\infty^{G,\pi} \in \Gamma$. □

Policy optimization in infinite horizon for S^{3}Ps. We have just proved that, for any stationary policy $\pi \in A^S$, the goal-probability and goal-cost functions are well-founded (i.e. have finite values) in infinite horizon, and that they can be iteratively computed from equations 3 and 4. Therefore, as the number of states and actions is finite, and thus the number of stationary policies is finite, we can immediately establish the following proposition, which proves that *any* S^{3}P problem in infinite horizon has a solution with finite goal-probability and goal-cost functions.

Proposition 1. *Let $\mathcal{M}$ be a general goal-oriented MDP. (I) There exists an optimal stationary policy π^* that minimizes the infinite-horizon goal-cost function among all policies that maximize the infinite-horizon goal-probability function, ie π^* is solution of eq. 2. (II) Goal-probability P_∞^{G,π^*} and goal-cost C_∞^{G,π^*} functions have finite values.*

This proposition is very general; in particular, it allows practitioners to tackle annoying goal-oriented MDP problems where assumption (i) of SSP holds, but not (ii). Recall that assumption (ii) means that the MDP's transition graph does not contain cycles with non positive costs composed of states not connected to the goal, which ensures that the cost-to-go criterion used in SSPs is well-founded (no such negative-cost cycles) *and* can be optimized using dynamic programming (no such zero-cost cycles). Both subconditions are not assumed in S^{3}Ps, because the goal-cost function is considered only over transient states that reach the goal with a positive probability (i.e. not over cycles composed of states that do not reach the goal). In practice, it means that S^{3}Ps now allow practitioners to solve shortest path problems with negative-cost loops, or also without proper policies.

However, even if Proposition 1 guarantees the existence of optimal stationary policies for every S^{3}P problems, it does not mean that such policies are easily computable in practice. In other terms, there does not necessarily exist practical algorithmic means to optimize P_∞^{G,π^*} and C_∞^{G,π^*} in the general case. The reason is quite tricky and identical to a similar issue that occurs when optimizing the total cost criterion of MDPs in the general case (see (Dawen 1986) and chapters 7 and 10 of (Puterman 1994) for details), or more specifically when optimizing SSPs with zero-cost cycles composed of states that do not reach the goal. In our context, the goal-probability value of stationary policies obtained when convergence is reached, may be surprisingly different from the optimized limit probability. Consider for instance the example depicted in Figure 1: once the optimal goal-probability function is obtained (0.95), applying action a_I from I brings in one (additional) step the same optimal goal-probability as actions a_1 and a_2 in two steps ($1 \times 0.95 = 0.95$), but stationary policy $\pi_4 = (a_I, a_I, \cdots)$ has a zero goal-probability. Thus, as optimizing the goal-probability function does not bring stationary policies whose goal-probability values are equal to the optimized goal-probability function, the goal-cost function (which depends on the goal-probability function, see equation 4) does not need to converge.

Fortunately, we have been able to provide different update equations presented below, which are proved to converge to optimal stationary policies, *provided all transitions from non-goal states have strictly positive costs*. The intuition behind these equations is as follows: once the optimal goal-probability function has converged, the iterative optimization of goal-cost functions *indirectly* selects stationary policies whose goal-probability function equals the optimal one, by rejecting other policies that *necessarily* have higher goal-cost functions. Indeed, by theoretically analysing the following optimization schema, we can see that all policies whose goal-probability is less than the optimal one would have an infinite goal-cost, so that they are automatically discarded by minimizing the goal-cost function. The actual mathematical proof of the following theorem is quite complex and obviously too long to be presented in this paper.

Theorem 3. *Let $\mathcal{M}$ be a goal-oriented MDP such that all transitions from non-goal states have strictly positive costs. Let $P_n^* : S \to [0;1]$ be the series of functions defined as:*

$$P_n^*(s) = \max_{a\in app(s)} \sum_{s'\in S} T(s,a,s') P_{n-1}^*(s'), \textit{with:}$$
$$P_0^*(s) = 0, \forall s \in S \setminus G; P_0^*(g) = 1, \forall g \in G \quad (5)$$

Functions P_n^* converge to a finite-values function P_∞^*.
Let $C_n^* : S \to \mathbb{R}_+$ be the series of functions defined as: $C_n^*(s) = 0$ if $P_\infty^*(s) = 0$, otherwise if $P_\infty^*(s) > 0$:

$$C_n^*(s) = \min_{a\in app(s):\sum_{s'\in S} T(s,a,s')P_\infty^*(s') = P_\infty^*(s)} \frac{1}{P_\infty^*(s)} \times$$
$$\sum_{s'\in S} T(s,a,s') P_\infty^*(s') \left[c(s,a,s') + C_{n-1}^*(s')\right], \textit{with:}$$
$$C_o^*(s) = 0, \forall s \in S \quad (6)$$

Functions C_n^* converge to a finite-values function C_∞^* and any stationary policy π^* obtained from the previous equation when convergence is reached, is optimal for S^{3}Ps.

The proof of this theorem also establishes that the convergence rate of the optimal goal-probability and goal-cost functions depends on a contraction constant, which is equal to the spectral radius of $T^\pi_{|\mathcal{X}^{\pi^*}}$ (see Lemma 1). For a given convergence precision $\epsilon > 0$, eq. 5 and 6 converge in finite time, and the worst-case time complexity of this iterative schema is polynomial in the number of states and actions, like the Bellman equations for standard SSPs. We implemented the optimization schema of Theorem 3 in an algorithm named `GPCI` (Goal-Probability and -Cost Iteration).

Experimental evaluation

The aim of this section is to provide experimental evidence that policies, which minimize the standard accumulated cost criterion of MDPs, are not necessarily optimal S^3P policies for problems where there does not exist policies that reach the goal with probability 1 (proper policies). To check this assumption, we evaluate the goal-probability and goal-cost function of policies optimized for the standard accumulated cost criterion, using evaluation equations of Theorem 1 until convergence (proved in Theorem 2). We then compare the goal-probability and goal-cost functions at a given initial state with the ones optimized by our algorithm `GPCI` (which implements Theorem 3). We tested two optimal algorithms for the standard accumulated cost criterion of MDPs: `VI` (Puterman 1994), which has the same time complexity as `GPCI`, and `LRTDP` (Bonet and Geffner 2003), which is a popular heuristic search algorithm for MDPs. We also compare with a non-optimal but efficient algorithm: `RFF` (Teichteil-Königsbuch, Kuter, and Infantes 2010), which attempts to maximize goal-probability and minimize goal-cost functions without theoretical guarantees.

Interestingly, we will see for some probabilistically hard problems that `GPCI`, which does not rely on heuristic search, is more efficient than all of these algorithms (additionally to being S^3P optimal). Note that past International Planning Competitions (IPCs) partly ranked planners based on their goal-probability and goal-cost functions (Younes et al. 2005), for which we are the first — to the best of our knowledge — to provide practical and theoretically optimal computation means. Results are summarized in Figure 2 for various domains detailed below. For each domain, we present results obtained for a particular problem, because (i) we got exactly the same relative results for all problems of each domain that could be solved by all algorithms, and (ii) the purpose of these tests is obviously not to compare how many problems each algorithm can solve. The largest tested benchmark has 2^{47} states. For `VI` and `GPCI`, we use the knowledge of the initial state to beforehand prune states that are not reachable from the initial state, whatever the policy.

blocksworld and rectangle-tireworld domains. These domains come from the IPC. For both domains, there actually exists proper policies, so that policies optimized for SSPs have finite cost-to-go functions and are thus well-founded. In other terms, we can safely run `VI` and `LRTDP` without discount factor, using the total cost criterion of SSPs (Bertsekas and Tsitsiklis 1996). As expected, we can see that `GPCI`, `VI` and `LRTDP`, all find a policy that reaches the goal with probability 1, and with the same goal-cost function (see Figure 2). `RFF` is efficient in terms of computation time, but its goal-cost function is far from optimal.

triangle-tireworld and exploding-blocksworld domains. These domains also come from the IPC, but unlike the previous ones, there does not exist proper policies (for triangle-tireworld, the maximum goal-probability is very slightly less than 1). Thus, the total cost criterion used in SSPs is not well-founded, so that `VI` and `LRTDP` never converge (checked in our experiments but not presented in the paper). The only other infinite-horizon criterion that can be used with these algorithms so that they converge *without changing the cost structure of the problems*, is the discounted cost criterion, which discounts all costs by a factor $0 < \gamma < 1$ (Puterman 1994; Teichteil-Königsbuch, Vidal, and Infantes 2011). In our experiments, for triangle-tireworld, the maximum goal-probability and minimum goal-cost functions we could get using `VI` and `LRTDP` with the discounted cost-to-go criterion, were obtained with any $\gamma \geqslant 0.95$. For exploding-blocksworld, we had to set $\gamma \geqslant 0.3$. As shown in Figure 2, `VI` and `LRTDP` are able to find the same S^3P optimal policies as `GPCI`: indeed, in these domains, there exists a constant $\alpha > 0$ such that all states that do not reach the goal have the same $\alpha/(1-\gamma)$ accumulated cost, so that there exists a minimum value of γ ensuring that minimizing the cost function (over all paths, even the ones that do not reach the goal) will favor policies that also maximize the goal-probability function. The goal-cost function will then be optimal because all states that do not reach the goal (i.e. they belong to paths that do not lead to the goal) have the same discounted accumulated cost, which is thus neutral for goal-cost optimization. Yet, note that *the minimum discount factor that achieves S^3P optimality can not be known in advance.* Concerning `RFF`, it is not S^3P optimal since it does not even find optimal goal-probabilities.

grid domains. As shown in the previous experiments, the cost structure of IPC domains is too simple to highlight situations where (i) no proper policies exist so that the total cost criterion used in SSPs is unusable, and (ii) the discounted cost-to-go criterion can not provide S^3P optimal policies for any value of γ. For this purpose, we propose the grid domain presented in Figure 3, for which we give two variants `grid-I` and `grid-II`. An agent has to move from an initial state to a goal state using 5 available actions that cost 1 each: `up`, `down`, `right`, `left`, `stay`. In the `grid-I` variant, all doors can close with probability 0.25 and then never open; when doors D1 and D2 (resp. D3 and D4) close, an additional cost of 1 (resp. 3) is added to all future unit moves, so that dead-end states pay different costs depending on where the agent comes from. Clearly, there is a hard tradeoff between maximizing the goal-probability (which requires to choose the less risky and more direct way through doors D3 and D4), and minimizing the accumulated costs (which would require to go through doors D1 and D2 in prevision of doors closing). This is a pitfall for existing approaches that consider only the standard discounted cost-to-go criterion, which they minimize over all paths (even the ones that do not reach the goal). In our experiments, we had to use $\gamma \geqslant 0.99$ to get maximum goal-probability and minimum goal-cost values, but Figure 2 shows that neither `VI` nor `LRTDP` are able to find optimal goal-probability functions found by `GPCI`. Even `RFF` has a better goal-probability function than the former two. As goal-probability functions are worse, goal-cost functions are not comparable because they are averaged over different paths that have a lower probability to reach the goal. In the `grid-II` variant, doors do not close but make the goal disappear (no more reachable) with probability 0.25 when each door is crossed. Additional costs are paid for all future moves like in the first variant. In this variant, paths

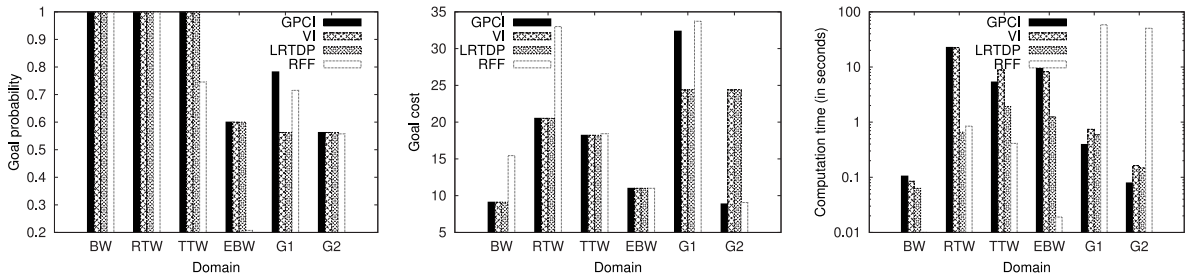

Figure 2: Comparison of goal probability (left plot), goal cost (centre plot), and computation time (right plot), of different algorithms for various domains: `blocksworld` (BW), `rectangle-tireworld` (RTW), `triangle-tireworld` (TTW), `exploding-blocksworld` (EBW), `grid-I` (G1), `grid-II` (G2).

through doors D2 and D1, or D3 and D4, have the same goal-probability values. Figure 2 shows that `VI` and `LRTDP` (and also `RFF`) find the optimal goal-probability function, but they have a very high goal-cost function compared with the optimal one found by `GPCI`. The reason is that `VI` and `LRTDP` take into account all reachable paths during cost minimization, even the ones that do not reach the goal: those paths not reaching the goal have a lower cost if the agent goes through doors D1 and D2 and the treasure disappears. Yet, paths through doors D3 and D4 have the same goal-probability but a lower cost as long as the treasure does not disappear. Moreover, Figure 2 also shows that *`GPCI` has the lowest computation time for the `grid` domains*, because it does not lose time to optimize costs of paths that do not reach the goal (which even `LRTDP` and `RFF` do).

Conclusion

To the best of our knowledge, we provide the first mathematical and algorithmic framework to solve goal-oriented MDPs by optimizing *both* the probability to reach the goal, which does not need to be equal to 1, and the accumulated costs averaged *only* over paths that reach the goal. These metrics are widely used to evaluate planners or policies performances. We experimentally proved that the standard total or discounted cost criteria used in MDPs do not necessarily provide optimal performances for these metrics, contrary to our approach, especially for problems that have a complex cost structure.

The next step will consist in designing efficient heuristic search algorithms for these metrics on the basis of the theoretical material presented in this paper, as well as domain-independent heuristics for the goal-probability and goal-cost functions. We think that this approach is promising, since the rather simple optimal algorithm proposed in this paper for optimizing S^3Ps, `GPCI`, already outperforms heuristic algorithms for SSPs like `LRTDP`, in terms of computation time on some domains with complex cost structure.

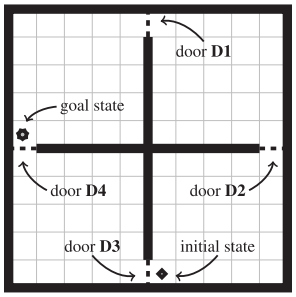

Figure 3: Grid world domain

References

Bertsekas, D. P., and Tsitsiklis, J. N. 1996. *Neuro-dynamic programming*. Athena Scientific.

Bonet, B., and Geffner, H. 2003. Labeled RTDP: Improving the convergence of real-time dynamic programming. In *Proc. ICAPS'03*, 12–21. Trento, Italy: AAAI Press.

Bonet, B., and Geffner, H. 2005. mGPT: A probabilistic planner based on heuristic search. *JAIR* 24:933–944.

Dawen, R. 1986. Finite state dynamic programming with the total reward criterion. *Zeitschrift für Operat. Research* 30(1):A1–A14.

Kolobov, A.; Mausam; Weld, D. S.; and Geffner, H. 2011. Heuristic Search for Generalized Stochastic Shortest Path MDPs. In *Proc. ICAPS'11*.

Kolobov, A.; Mausam; and Weld, D. S. 2010. SixthSense: Fast and Reliable Recognition of Dead Ends in MDPs. In *Proc. AAAI'10*.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: John Wiley & Sons, Inc., 1st edition.

Teichteil-Königsbuch, F.; Kuter, U.; and Infantes, G. 2010. Incremental plan aggregation for generating policies in MDPs. In *Proc. AAMAS'10*, 1231–1238.

Teichteil-Königsbuch, F.; Vidal, V.; and Infantes, G. 2011. Extending classical planning heuristics to probabilistic planning with dead-ends. In *Proc. AAAI'11*.

Yoon, S. W.; Ruml, W.; Benton, J.; and Do, M. B. 2010. Improving determinization in hindsight for on-line probabilistic planning. In *Proc. ICAPS'10*, 209–217.

Younes, H. L. S.; Littman, M. L.; Weissman, D.; and Asmuth, J. 2005. The first probabilistic track of the International Planning Competition. *JAIR* 24:851–887.