Safe Exploration in Markov Decision Processes

Teodor Mihai Moldovan MOLDOVAN@CS.BERKELEY.EDU
Pieter Abbeel PABBEEL@CS.BERKELEY.EDU
University of California at Berkeley, CA 94720-1758, USA

Abstract

In environments with uncertain dynamics exploration is necessary to learn how to perform well. Existing reinforcement learning algorithms provide strong exploration guarantees, but they tend to rely on an ergodicity assumption. The essence of ergodicity is that any state is eventually reachable from any other state by following a suitable policy. This assumption allows for exploration algorithms that operate by simply favoring states that have rarely been visited before. For most physical systems this assumption is impractical as the systems would break before any reasonable exploration has taken place, i.e., most physical systems don't satisfy the ergodicity assumption. In this paper we address the need for *safe* exploration methods in Markov decision processes. We first propose a general formulation of safety through ergodicity. We show that imposing safety by restricting attention to the resulting set of guaranteed safe policies is NP-hard. We then present an efficient algorithm for guaranteed safe, but potentially suboptimal, exploration. At the core is an optimization formulation in which the constraints restrict attention to a subset of the guaranteed safe policies and the objective favors exploration policies. Our framework is compatible with the majority of previously proposed exploration methods, which rely on an exploration bonus. Our experiments, which include a Martian terrain exploration problem, show that our method is able to explore better than classical exploration methods.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

1. Introduction

When humans learn to control a system, they naturally account for what we think of as safety. For example, when a novice pilot learns how to fly an RC helicopter, they will slowly spin up the blades until the helicopter barely lifts off, then quickly put it back down. They will repeat this a few times, slowly starting to bring the helicopter a little bit off the ground. When doing so they would try out the cyclic (roll and pitch) and rudder (yaw) control, while—until they have become more skilled—at all times staying low enough that simply shutting it down would still have it land safely. When a driver wants to become skilled at driving on snow, they might first slowly drive the car to a wide open space where they could start pushing their limits. When we are skiing downhill, we are careful about not going down a slope into a valley where there is no lift to take us back up.

One would hope that exploration algorithms for physical systems would be able to account for safety and have similar behavior naturally emerge. Unfortunately most existing exploration algorithms completely ignore safety issues. More precisely phrased, most existing algorithms have strong exploration guarantees, but to achieve these guarantees they assume ergodicity of the Markov decision process (MDP) in which the exploration takes place. An MDP is *ergodic* if any state is reachable from any other state by following a suitable policy. This assumption does not hold true in the exploration examples presented above as each of these systems could break during (non-safe) exploration.

Our first important contribution is a definition of safety, which, at its core, requires restricting attention to policies that preserve ergodicity with some well controlled probability. Imposing safety is, unfortunately, NP-hard in general. Our second important contribution is an approximation scheme leading to guaranteed safe, but potentially sub-optimal, exploration.[1] A third contribution is the consideration of

[1]Note that existing (unsafe) exploration algorithms are

uncertainty in the dynamics model that is correlated over states. While usually the assumption is that uncertainty in different parameters is independent—as this makes problem more tractable computationally—being able to learn about state-action pairs before visiting them is critical for safety.

Our experiments illustrate that our method indeed achieves safe exploration, in contrast to plain exploration methods. They also show that our algorithm is almost as computationally efficient as planning in a known MDP—but then, as every step leads to an update in knowledge about the MDP, this computation is to be repeated after every step. Our approach is able to safely explore grid worlds of size up to 50 100. Our method can make safe any type of exploration that relies on exploration bonuses, which is the case for most existing exploration algorithms, including, for example, the methods proposed in (Brafman & Tennenholtz, 2001; Kolter & Ng, 2009). In this article we do not focus on the exploration objective and use existing ones.

Safe exploration has been the focus of a large number of articles. (Gillula & Tomlin, 2011; Aswani & Bouffard, 2012) propose safe exploration methods for linear systems with bounded disturbances based on model predictive control and reachability analysis. They define safety in terms of safe regions of the state space, which, we will show, is not always appropriate in the context of MDPs. The safe exploration for MDP methods proposed by (Geramifard et al., 2011; Hans et al., 2008) gauge safety based on the best best estimate of the transition measure but they ignore the level of uncertainty in this estimate. As we will show, this is not sufficient to provably guarantee safety.

Provably efficient exploration is a recurring theme in reinforcement learning (Strehl & Littman, 2005; Li et al., 2008; Brafman & Tennenholtz, 2001; Kearns & Singh, 2002; Kolter & Ng, 2009). Most methods, however, tend to rely on the assumption of ergodicity which rarely holds in interesting practical examples; consequently, these methods are rarely applicable for physical systems. The issue of provably guaranteed safety, or risk aversion, under uncertainty in the MDP parameters has also been studied in the reinforcement literature. In (Nilim & El Ghaoui, 2005) they propose a robust MDP control method assuming the transition frequencies are drawn from an orthogonal convex set by an adversary. Unfortunately, it seems impossible to use their method to constrain some safety objective while optimizing a different exploration objective.

also sub-optimal, in that they are not guaranteed to complete exploration in the minimal number of time steps.

In (Delage & Mannor, 2007) they present a safe exploration algorithm for the special case of Gaussian distributed ambiguity in the reward and state-action-state transition probabilities, but their safety guarantees are only accurate if the ambiguity in the transition model is small.

2. Notation and Assumptions

Due to space constraints, we will not give a general introduction to Markov decision processes (MDPs). For an introduction to MDPs we refer the readers to (Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996).

We use capital letters to denote random variables; for example, the total reward is: $V := \sum_{t=0}^{\infty} R_{S_t,A_t}$. We represent the policies and the initial state distributions by probability measures. Usually the measure π will correspond to a policy and the measure $s := \delta(s)$, which puts measure only in state s, will correspond to starting in state s. With this notation, the usual value recursion, assuming a known transition measure, p, reads:

$$E^p_{s,\pi}[V] = \sum_{a,s'} \pi_{s,a} \left(E[R]_{s,a} + p_{s,a,s'} E^p_{s',\pi}[V] \right).$$

We specify the transition measure as a superscript of the expectation operator rather than a subscript for typographical convenience; in this case, and in general, the positioning of indexes as subscripts or superscripts adds no extra significance. We will let the transition measure p sometimes sum to less than one, that is $\sum_{s'} p_{s,a,s'} \leq 1$. The missing mass is implicitly assigned to transitioning to an absorbing "end" state, which, for example, allows us to model γ discounting by simply using γp as a transition measure.

We model ambiguous dynamics in a Bayesian way, allowing the transition measure to also be a random variable. When this is the case, we will use P to denote the, now random, transition measure. The *belief*, which we will denote by β, is our Bayesian probability measure over possible dynamics, governing P and R. Therefore, the expected return under the belief and policy π, starting from state s, is $E_\beta E^P_{s,\pi}[V]$. We allow beliefs under which transition measures and rewards are arbitrarily correlated. In fact, such correlations are usually necessary to allow for safe exploration. For compactness we will often use lower case letters to denote the expectation of their upper case counterparts. Specifically, we will use the notations $p := E_\beta[P]$ and $r := E_\beta[R]$ throughout.

3. Problem formulation

3.1. Exploration Objective

Exploration methods, as those proposed in (Brafman & Tennenholtz, 2001; Kolter & Ng, 2009), operate by finding optimal policies in constructed MDPs with exploration bonuses. The R-MAX algorithm, for example, constructs an MDP based on the discounted expected transition measure and rewards under the belief, and adds a deterministic exploration bonus equal to the maximum possible reward in the MDP, $\xi^{\beta}_{s,a} = r_{\max}$, to any transitions that are not sufficiently well known. Our method allows adding safety constraints to any such exploration methods. Henceforth, we will restrict attention to such exploration methods, which can be formalized as optimization problems of the form:

$$\text{maximize}_{\ \pi_o}\ E^{\gamma p}_{s_0,\pi_o} \sum_{t=0}^{\infty} \left(r_{S_t,A_t} + \xi^{\beta}_{S_t,A_t} \right). \qquad (1)$$

3.2. Safety Constraint

The issue of safety is closely related to *ergodicity*. Almost all proposed exploration techniques presume ergodicity; authors present it as a harmless technical assumption but it rarely holds in interesting practical problems. Whenever this happens, their efficient exploration guarantees cease to hold, often leading to very inefficient policies. Informally, an environment is ergodic if any mistake can be forgiven eventually. More specifically, a belief over MDPs is ergodic if and only if any state is reachable from any other state via some policy or, equivalently, if and only if:

$$\forall s, s', \exists\ \pi_r \text{ such that } E_{\beta} E^{P}_{s,\pi_r}[B_{s'}] = 1, \qquad (2)$$

where $B_{s'}$ is an indicator random variable of the event that the system reaches state s' at least once: $B_{s'} = 1\{\exists t < \infty \text{ such that } S_t = s'\} = \min\left(1, \sum_t 1_{S_t = s'}\right)$.

Unfortunately, many environments are not ergodic. For example, our robot helicopter learning to fly cannot recover on its own after crashing. Ensuring almost sure ergodicity is too restrictive for most environments as, typically, there always is a very small, but non-zero, chance of encountering that particularly unlucky sequence of events that breaks the system. Our idea is to restrict the space of eligible policies to those that preserve ergodicity with some user-specified probability, δ, called the *safety level*. We name these policies *δ-safe*. Safe exploration now amounts to choosing the best exploration policy from this set of safe policies.

Informally, if we stopped a δ-safe policy π_o at any time T, we would be able to return from that point to the home state s_0 with probability δ by deploying a return policy π_r. Executing only δ-safe policies in the case of a robot helicopter learning to fly will guarantee that the helicopter is able to land safely with probability δ whenever we decide to end the experiment. In this example, T is the time when the helicopter is recalled (perhaps because fuel is running low), so we will call T the *recall time*. Formally, an outbound policy π_o is δ-safe with respect to a home state s_0 and a stopping time T if and only if:

$$\exists \pi_r \text{ such that } E_{\beta} E^{P}_{s_0,\pi_o}\left[E^{P}_{S_T,\pi_r}[B_{s_0}] \right] \geq \delta. \qquad (3)$$

Note that, based on Equation (2), any policy is δ-safe for any δ if the MDP is ergodic with probability one under the belief. For convenience we will assume that the recall time, T, is exponentially distributed with parameter $1 - \gamma$, but our method also applies when the recall time equals some deterministic horizon. Unfortunately, expressing the set of δ-safe policies is NP-hard in general, as implied by the following theorem proven in the appendix.

Theorem 1. *In general, it is NP-hard to decide whether there exist δ-safe policies with respect to a home state, s_0, and a stopping time, T, for some belief, β.*

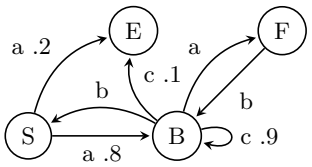


Figure 1. Starting from state S, the policy (aababab...) is safe at a safety level of .8. However, the policy (acccc...) is not safe since it will end up in the sink state E with probability 1. State-action Sa and state B can neither be considered safe nor unsafe, since both policies use them.

3.3. Safety Counter-Examples

We conclude this section with counter-examples to three other, perhaps at first sight more intuitive, definitions of safety. First, we could have tried to define safety in terms of sets of safe states or state-actions. That is, we might think that making the non-safe states and actions unavailable to the planner (or simply inaccessible) is enough to guarantee safety. Figure 1 shows an MDP where the same state-action is used both by a safe and by an unsafe policy. The idea behind this counter-example is that safety depends not only on the states visited, but also on the number of visits, thus, on the policy. This shows that safety should be defined in terms of safe policies, not in terms of safe states or state-actions.

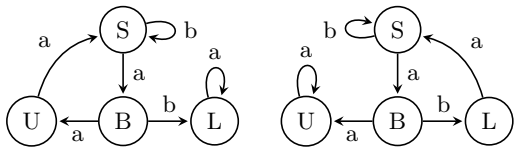


Figure 2. Under out belief the two MDPs above both have probability .5. It is intuitively unsafe to go from the start state S to B since we wouldn't know whether the way back is via U or L, even though we know for sure that a return policy exists.

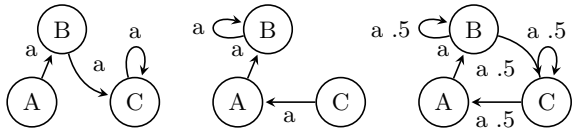


Figure 3. The two MDPs on the left both have probability .5. Under this belief, starting from state A, policy (aaa...) is unsafe. However, under the mean transition measure, represented by the MDP on the right, the policy is safe.

Second, we might think that it is perhaps enough to ensure that there exists a return policy for each potential sample MDP from the belief, but not impose that it be the same for all samples. That is, we might think that condition 3 is too strong and, instead, it would be enough to have:

$$E_\beta 1\{\exists \pi_r : E^P_{s_0,\pi_o} E^P_{S_T,\pi_r}[B_{s_0}] = 1\} \geq \delta.$$

Figure 2 shows an MDP where this condition holds, yet all policies are naturally unsafe.

Third, we might think that it is sufficient to simply use the expected transition measure when defining safety, as in the equation below. Figure 3 shows that this is not the case; the expected transition measure is not a sufficient statistic for safety.

$$\exists \pi_r \text{ such that } E^p_{s_0,\pi_o}\left[E^p_{S_T,\pi_r}[B_{s_0}]\right] \geq \delta.$$

4. Guaranteed Safe, Potentially Sub-optimal Exploration

Although imposing the safety constraint in Equation (3) is NP-hard, as shown in Theorem 1, we can efficiently constrain a lower bound on the safety objective, so the safety condition is still provably satisfied. Doing so could lead to sub-optimal exploration since the set of policies we are optimizing over has shrunk. However, we should keep in mind that the exploration objectives represent approximate solutions to other NP-hard problems, so optimality has already been forfeited in existing (non-safe) approaches to start out with. Algorithm 1 summarizes the procedure and the experiments presented in the next section show that, in practice, when the ergodicity assumptions are violated, safe exploration is much more efficient than plain exploration.

Algorithm 1 Safe exploration algorithm

Require: prior belief β, discount γ, safety level δ.
Require: function ξ : belief $\to$ exploration bonus
  $M, N \leftarrow$ new MDP objects
  repeat
    $s_0, \varphi \leftarrow$ current state and observations
    update belief β with information φ
    $\xi^\beta_{s,a} \leftarrow \xi(\beta)$ (exploration bonus based on β)
    $\sigma^\beta_{s,a} \leftarrow \sum_{s'} E_\beta[\min(0, P_{s,a,s'} - E_\beta[P_{s,a,s'}])]$
    M.transition measure $\leftarrow E_\beta[P](1 - 1_{s=s_0})$
    M.reward function $\leftarrow 1_{s=s_0} + (1 - 1_{s=s_0})\sigma^\beta_{s,a}$
    $\pi^r, v \leftarrow M$.solve()
    N.transition measure $\leftarrow \gamma E_\beta[P]$
    N.reward function $\leftarrow E_\beta[R_{s,a}] + \xi^\beta_{s,a}$
    N.constraint reward func. $\leftarrow (1-\gamma)v_s + \gamma\sigma^\beta_{s,a}$
    N.constraint lower bound $\leftarrow \delta$
    $\pi^o, v^\xi, v^\sigma \leftarrow N$.solve under constraint()
    $q^\sigma_{s,a} \leftarrow (1-\gamma)v_s + \gamma\sigma^\beta_{s,a} + \sum_{s'} p_{s,a,s'} v^\sigma_{s'}$
    $a \leftarrow \text{argmax}_{\{\pi^o_{s_0,a} > 0\}}\, q^\sigma_{s_0,a}$ (de-randomize policy)
    take action a in environment
  until $\xi^\beta = 0$, so there is nothing left to explore

Putting together the exploration objective defined in Equation (1) and the safety objective defined in Equation (3) allows us to formulate safe exploration at level δ as a constrained optimization problem:

$$\begin{aligned} \text{maximize}_{\ \pi_o,\pi_r} \quad & E^{\gamma p}_{s_0,\pi_o} \sum_t \left(r_{S_t,A_t} + \xi^\beta_{S_t,A_t}\right) \\ \text{such that:} \quad & E_\beta E^P_{s_0,\pi_o}\left[E^P_{S_T,\pi_r}[B_{s_0}]\right] \geq \delta. \end{aligned}$$

The exploration objective is already conveniently formulated as the expected reward in an MDP with transition measure γp, so we will not modify it. On the other hand, the safety constraint is difficult to deal with as is. Ideally, we would like the safety constraint to also equal some expected reward in an MDP. We will see that, in fact, it takes two MDPs to express the safety constraint.

First, we express the inner term, $E^P_{S_T,\pi_r}[B_{s_0}]$, as the expected reward in an MDP. We can replicate the behaviour of B_{s_0}, that is counting only the first time state s_0 is reached, by using a new transition measure, $P \cdot (1 - 1_{s=s_0})$ under which, once s_0 is reached, any further actions lead immediately to the implicit "end"

state. Formally, we express this by the identity:

$$E^P_{S_T,\pi_r}[B_{s_0}] = E^{P\cdot(1-1_{s=s_0})}_{S_T,\pi_r} \sum_{t=0}^{\infty} 1_{S_t=s_0}.$$

We now focus on the outer term, $E^P_{s_0,\pi_o}\left[E^P_{S_T,\pi_r}[B_{s_0}]\right]$. Since the recall time, T, is exponentially distributed with parameter $1-\gamma$, we can view S_T as the final state in a γ-discounted MDP starting at state s_0, following policy π_o. In this MDP, the inner term plays the role of a terminal reward. To put the problem in a standard form, we convert this terminal reward to a step-wise reward by multiplying it by $1-\gamma$.

$$E^P_{s_0,\pi_o}\left[E^P_{S_T,\pi_r}[B_{s_0}]\right] = E^{\gamma P}_{s_0,\pi_o} \sum_{t=0}^{\infty} (1-\gamma)\left[E^P_{S_t,\pi_r}[B_{s_0}]\right].$$

At this point, we have expressed the safety constraint in the MDP formalism, but the transition measures of these MDPs, $P(1-1_{s=s_0})$ and γP, are still random. If we could replace these random transition measures with their expectations under the belief β that would significantly simplify the safety constraint. It turns out we can do this, at the expense of making the constraint more stringent. Our tool for doing so is Theorem 2 presented below, but proven in the appendix. It shows that we can replace a belief over MDPs by a single MDP with the expected transition measure, featuring an appropriate reward correction such that, under any policy, the value of this MDP is a lower bound on the expected value under the belief.

Theorem 2. *Let β be a belief such that for any policy, π, and any starting state, s, the total expected reward in any MDP drawn from the belief is between 0 and 1; i.e. $0 \le E^P_{s,\pi}[V] \le 1$, β-almost surely. Then the following bound holds for any policy, π, and any starting state, s:*

$$E_\beta E^P_{s,\pi} \sum_{t=0}^{\infty} R_{S_t,A_t} \ge E^p_{s,\pi} \sum_{t=0}^{\infty} \left(E_\beta[R_{S_t,A_t}] + \sigma^\beta_{S_t,A_t}\right)$$
$$\text{where} \quad \sigma^\beta_{s,a} := \sum_{s'} E_\beta\left[\min(0, P_{s,a,s'} - E_\beta[P_{s,a,s'}])\right].$$

We first apply Theorem 2 to the outer term, yielding the following bound:

$$E^P_{s_0,\pi_o}\left[E^P_{S_T,\pi_r}[B_{s_0}]\right] = E^{\gamma P}_{s_0,\pi_o} \sum_{t=0}^{\infty} (1-\gamma)\left[E^P_{S_t,\pi_r}[B_{s_0}]\right]$$
$$\ge E^{\gamma p}_{s_0,\pi_o} \sum_{t=0}^{\infty} \left((1-\gamma) E_\beta E^P_{S_t,\pi_r}[B_{s_0}] + \gamma\sigma^\beta_{S_t,A_t}\right).$$

We, then, apply it again to the inner term:

$$E_\beta E^P_{s,\pi_r}[B_{s_0}] = E^{P\cdot(1-1_{s=s_0})}_{s,\pi_r} \sum_{t=0}^{\infty} 1_{S_t=s_0} \ge \tag{4}$$
$$\ge E^{p\cdot(1-1_{s=s_0})}_{s,\pi_r} \sum_{t=0}^{\infty} \left(1_{S_t=s_0} + (1-1_{S_t=s_0})\sigma^\beta_{S_t,A_t}\right).$$

Combining the last two results allows us to replace the NP-hard safety constraint with a stricter, but now tractable, constraint. The resulting optimization problem corresponds to the guaranteed safe, but potentially sub-optimal exploration problem:

$$\text{maximize}_{\ \pi_o,\pi_r} \quad E^{\gamma p}_{s_0,\pi_o} \sum_t \left(r_{S_t,A_t} + \xi^\beta_{S_t,A_t}\right) \tag{5}$$
$$\text{s.t.:} \quad E^{\gamma p}_{s_0,\pi_o} \sum_{t=0}^{\infty} \left((1-\gamma)v_{S_t} + \gamma\sigma^\beta_{S_t,A_t}\right) \ge \delta \quad \text{and}$$
$$v_s = E^{p\cdot(1-1_{s=s_0})}_{s,\pi_r} \sum_{t=0}^{\infty} \left(1_{S_t=s_0} + (1-1_{S_t=s_0})\sigma^\beta_{S_t,A_t}\right).$$

The term v_s represents our lower bound for the inner term per Equation (4), and is simply the value function of the MDP corresponding to the inner term; i.e. the MDP with transition measure $p(1-1_{s=s_0})$ and reward function $1_{s=s_0} + (1-1_{s=s_0})\sigma^\beta_{s,a}$, under policy π_r. Since the return policy, π_r, does not appear anywhere else, we can split the safe exploration problem we obtained in Equation (5) into two steps:

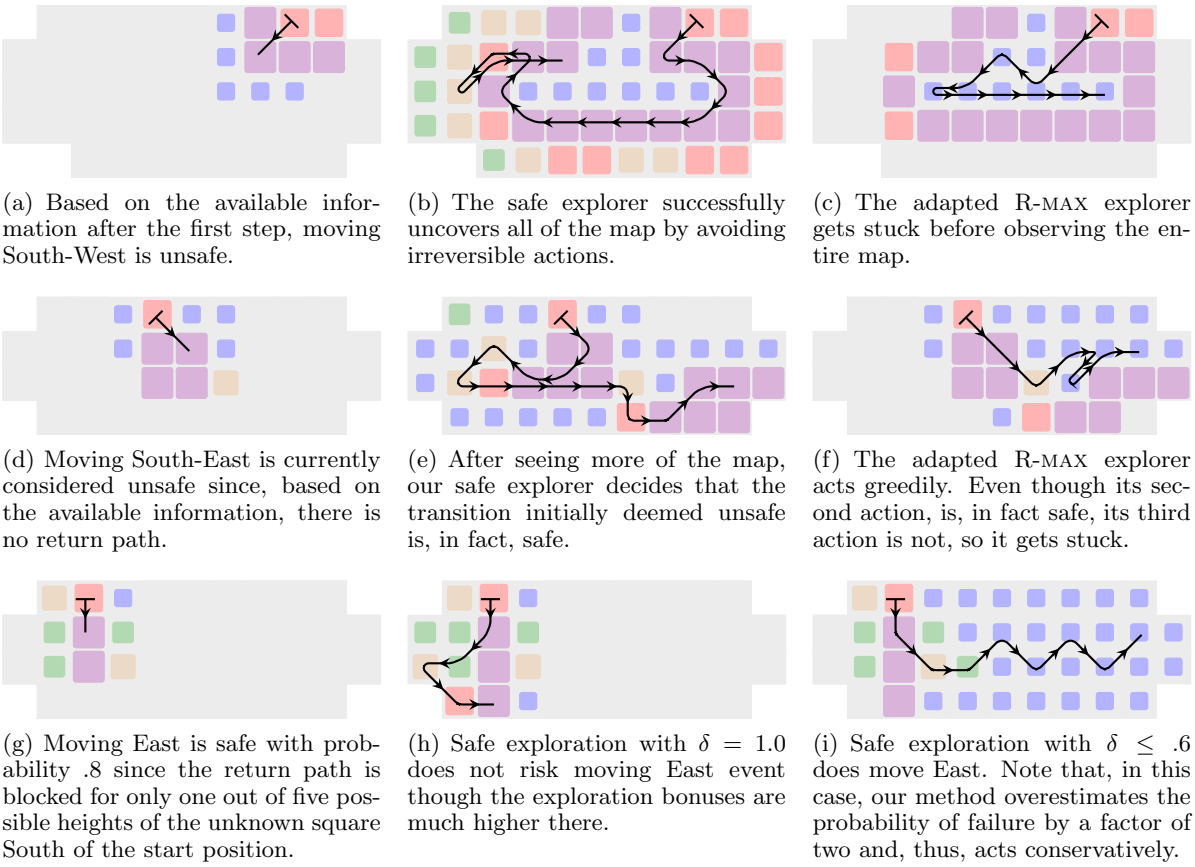
Step one: find the optimal return policy π^*_r, and corresponding value function v^*_s, by solving the standard MDP problem below:

$$E^{p\cdot(1-1_{s=s_0})}_{s,\pi_r} \sum_{t=0}^{\infty} \left(1_{S_t=s_0} + (1-1_{S_t=s_0})\sigma^\beta_{S_t,A_t}\right).$$

Step two: find the optimal exploration policy π^*_o under the strict safety constraint, by solving the constrained MDP problem below:

$$\text{maximize}_{\ \pi_o} \quad E^{\gamma p}_{s_0,\pi_o} \sum_t \left(r_{S_t,A_t} + \xi^\beta_{S_t,A_t}\right)$$
$$\text{s.t.:} \quad E^{\gamma p}_{s_0,\pi_o} \sum_{t=0}^{\infty} \left((1-\gamma)v^*_{S_t} + \gamma\sigma^\beta_{S_t,A_t}\right) \ge \delta.$$

The first step amounts to solving a standard MDP problem while the second step amounts to solving a constrained MDP problem. As shown by (Altman, 1999), both can be solved efficiently either by linear programming, or by value-iteration. In our experiments we used the LP formulation with the state-action occupation measure as optimization variable.



(a) Based on the available information after the first step, moving South-West is unsafe.

(b) The safe explorer successfully uncovers all of the map by avoiding irreversible actions.

(c) The adapted R-MAX explorer gets stuck before observing the entire map.

(d) Moving South-East is currently considered unsafe since, based on the available information, there is no return path.

(e) After seeing more of the map, our safe explorer decides that the transition initially deemed unsafe is, in fact, safe.

(f) The adapted R-MAX explorer acts greedily. Even though its second action, is, in fact safe, its third action is not, so it gets stuck.

(g) Moving East is safe with probability .8 since the return path is blocked for only one out of five possible heights of the unknown square South of the start position.

(h) Safe exploration with $\delta = 1.0$ does not risk moving East event though the exploration bonuses are much higher there.

(i) Safe exploration with $\delta \leq .6$ does move East. Note that, in this case, our method overestimates the probability of failure by a factor of two and, thus, acts conservatively.

Figure 4. Exploration experiments in simple grid worlds. See text for full details. Square sizes are proportional to corresponding state heights between 1 and 5. The large, violet squares have a height of 5, while the small, blue squares have a height of 1. Gray spaces represent states that have not yet been observed. Each row corresponds to the same grid world. The first column shows the belief after the first exploration step, while the second and third columns show the entire trajectory followed by different explorers.

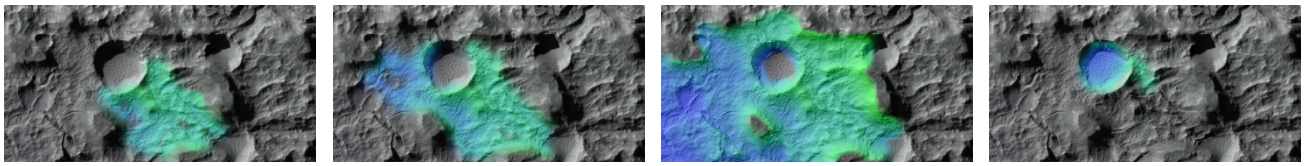
Solutions to the constrained MDP problem will usually be stochastic policies, and, in our experiments, we found that following them sometimes leads to random walks which explore inefficiently. We addressed the issue by de-randomizing the exploration policies in favor of safety. That is, whenever the stochastic policy proposes multiple actions with non-zero measure, we choose the one among them that optimizes the safety objective.

5. Experiments

5.1. Grid World

Our first experiment models a terrain exploration problem where the agent has limited sensing capabilities. We consider a simple rectangular grid world, where every state has a height H_s. From our Bayesian standpoint these heights are independent, uniformly distributed categorical random variables on the set $\{1, 2, 3, 4, 5\}$. At any time the agent can attempt to move to any immediately neighboring state. Such move will succeed with probability one if the height of the destination state is no more than one level above the current state; otherwise, the agent remains in the current state with probability one. In other words, the agent can always go down cliffs, but is unable to climb up if they are too steep. Whenever the agent enters a new state it can see the exact heights of all immediately surrounding states. We present this grid world experiment to build intuition and to provide an easily reproducible result. Figure 4 shows a number of examples where our exploration method results in intuitively safe behavior, while plain exploration methods lead to clearly unsafe, suboptimal behavior.

Our exploration scheme, which we call *adapted R-MAX*, is a modified version of R-max exploration (Brafman & Tennenholtz, 2001), where the exploration bonus of moving between two states is now proportional to the



(a) Safe exploration with $\delta =$ .98 leads to a model entropy reduction of 7680.

(b) Safe exploration with $\delta =$ .90 leads to a model entropy reduction of 12660.

(c) Safe exploration with $\delta =$ .70 leads to a model entropy reduction of 35975.

(d) Regular (unsafe, $\delta = 0$) exploration leads to a model entropy reduction of 3214.

Figure 5. Simulated safe exploration on a 2km by 1km area of Mars at -30.6 degree latitude and 202.2 degrees longitude, for 15000 time steps, at different safety levels. See text for full details. The color saturation is inversely proportional to the standard deviation of the height map under the posterior belief. Full coloration represents a standard deviation of 1cm or less. We report the difference between the entropies of the height model under the prior and the posterior beliefs as a measure of performance. Images: NASA/JPL/University of Arizona.

number of neighboring unknown states that would be uncovered as a result of the move, to account for the remote observation model. The safety costs for this exploration setup, as prescribed by Theorem 2 are:

$$\sigma^{\beta}_{s,a} = -2E_{\beta}[P_{s,a}](1 - E_{\beta}[P_{s,a}]) = -2\mathrm{Var}_{\beta}[P_{s,a}]$$

where $P_{s,a} := 1_{H_{s+a} \leq H_s + 1}$ is the probability that attempted move a succeeds in state s and the belief β describes the distribution of the heights of unseen states. In practice we found that this correction is a factor of two larger than would be sufficient to give a tight safety bound.

A somewhat counter intuitive result is that adding safety constraints to the exploration objective will, in fact, improve the fraction of squares explored in randomly generated grid worlds. The reason why plain exploration performs so poorly is that the ergodicity assumptions are violated, so efficiency guarantees no longer hold. Figure 6 in the appendix summarizes our exploration performance results.

5.2. Martian Terrain

For our second experiment, we model the problem of autonomously exploring the surface of Mars by a rover such as the *Mars Science Laboratory (MSL)* (Lockwood, 2006). The MSL is designed to be remote controlled from Earth but communication suffers a latency of 16.6 minutes. At top speed, it could traverse about 20m before receiving new instructions, so it needs to be able to navigate autonomously. In the future, when such rovers become faster and cheaper to deploy, the ability to plan their paths autonomously will become even more important. The MSL is designed to a static stability of 45 degrees, but would only be able to climb slopes up to 5 degrees without slipping (MSL, 2007). Digital terrain models for parts of the surface of Mars are available from the *High Resolution Imaging Science Experiment (HiRISE)* at a scale of 1.00 meter/pixel and accurate to about a quarter of a meter. The MSL would be able to obtain much more accurate terrain models locally by stereo vision.

The state-action space of our model MDP is the same as in the previous experiment, with each state corresponding to a square area of 20 by 20 meters on the surface. We allow only transitions at slopes between -45 and 5 degrees. The heights, H_s, are now assumed to be independent Gaussian random variables. Under the prior belief, informed by the HiRISE data, the expected heights and their variances are:

$$E_{\beta}[H] = D^{20}[g \circ h] \quad \text{and}$$
$$\mathrm{Var}_{\beta}[H] = D^{20}[g \circ (h - g \circ h)^2] + v_0$$

where h are the HiRISE measurements, g is a Gaussian filter with $\sigma = 5$ meters, "$\circ$" represents image convolution, D^{20} is the sub-sampling operator and $v_0 = 2^{-4}\mathrm{m}^2$ is our estimate of the variance of HiRISE measurements. We model remote sensing by assuming that the MSL can obtain Gaussian noisy measurements of the height at a distance d away with variance $v(d) = 10^{-6}(d + 1\mathrm{m})^2$.

To account for this remote sensing model we use a first order approximation of the entropy of H as an exploration bonus:

$$\xi^{\beta}_{s,a} = \sum_{s'} \mathrm{Var}_{\beta}[H_{s'}]/v(d_{s,s'}).$$

Figure 5 shows our simulated exploration results for a 2km by 1km area at -30.6 degrees latitude and 202.2 degrees longitude (PSP, 2008). Safe exploration at level 1.0 is no longer possible, but, even at a conservative safety level of .98, our method covers more ground than the regular (unsafe) exploration method which promptly get stuck in a crater. Imposing the safety constraint naively, with respect to the expected

Table 1. Per-step planning times for the 50×100 grid world used in the Mars exploration experiments, with $\gamma = .999$.

Problem setting	Planning time (s)
Safe exploration at .98	5.86 ± 1.47
Safe exploration at .90	10.94 ± 7.14
Safe exploration at .70	4.57 ± 3.19
Naive constraint at .98	2.55 ± 0.42
Regular (unsafe) exploration	1.62 ± 0.26

transition measure, as argued against at the end of Section 3.3, performs as poorly as unsafe exploration even if the constraint is set at .98.

5.3. Computation Time

We implemented our algorithm in Python 2.7.2.7, using Numpy 1.5.1 for dense array manipulation, SciPy 0.9.0 for sparse matrix manipulation and Mosek 6.0.0.119 for linear programming. The discount factor was set to .99 for the grid world experiment and to .999 for Mars exploration. In the latter experiment we also restricted precision to 10^{-6} to avoid numerical instabilities in the LP solver. Table 1 summarizes planning times for our Mars exploration experiments.

6. Discussion

In addition to the safety formulation we discussed in Section 3.2, out framework also supports a number of other safety criteria that we did not discuss due to space constraints:

- Stricter ergodicity ensuring that return is possible within some horizon, H, not just eventually, with probability δ.
- Ensuring that the probability of leaving some predefined safe set of state-actions is lower than $1-\delta$.
- Ensuring that the expected total reward under the belief is higher than δ.

Additionally, any number and combination of these constraints at different δ-levels can be imposed simultaneously.

Acknowledgements

This material is based upon work supported in part by NSF under award IIS-0931463, by ARO under the MAST program, by a Sloan Fellowship, by a gift from Intel, by the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-11-1-0391.

References

MSL Landing Site Selection. User's Guide to Engineering Constraints, 2007. URL http://marsoweb.nas.nasa.gov/landingsites/msl/memoranda/MSL_Eng_User%_Guide_v4.5.1.pdf.

Stratigraphy of Potential Crater Hydrothermal System, 2008. URL http://hirise.lpl.arizona.edu/dtm/dtm.php?ID=PSP_010228_1490.

Altman, Eitan. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

Aswani, Anil and Bouffard, Patrick. Extensions of Learning-Based Model Predictive Control for Real-Time Application to a Quadrotor Helicopter. In *Proc. American Control Conference (ACC) (to appear)*, 2012.

Bertsekas, Dimitri P. and Tsitsiklis, John N. *Neuro-Dynamic Programming*. Athena Scientific, October 1996.

Blondel, Vincent D. and Tsitsiklis, John N. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, September 2000.

Brafman, Ronen I. and Tennenholtz, Moshe. R-MAX - A General Polynomial Time Algorithm for Near-Optimal Reinforcement Learning. In *Journal of Machine Learning Research*, volume 3, pp. 213–231, 2001.

Delage, Erick and Mannor, Shie. Percentile optimization in uncertain Markov decision processes with application to efficient exploration. *ICML; Vol. 227*, pp. 225, 2007.

Geramifard, A, Redding, J, Roy, N, and How, J P. UAV Cooperative Control with Stochastic Risk Models. In *Proceedings of the American Control Conference (ACC)*, San Francisco, CA, 2011.

Gillula, Jeremy H. and Tomlin, Claire J. Guaranteed safe online learning of a bounded system. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2979–2984. IEEE, September 2011.

Hans, A, Schneegaß, D, Schäfer, AM, and Udluft, S. Safe exploration for reinforcement learning. In *ESANN 2008, 16th European Symposium on Artificial Neural Networks*, 2008.

Kearns, Michael and Singh, Satinder. Near-Optimal Reinforcement Learning in Polynomial Time. *Machine Learning*, 49(2):209–232, November 2002.

Kolter, J. Zico and Ng, Andrew Y. Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pp. 1–8, New York, New York, USA, 2009. ACM Press.

Li, Lihong, Littman, Michael L., and Walsh, Thomas J. Knows what it knows: a framework for self-aware learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 568–575, 2008.

Lockwood, Mary Kae. Introduction: Mars Science Laboratory: The Next Generation of Mars Landers. *Journal of Spacecraft and Rockets*, 43(2), 2006.

Nilim, Arnab and El Ghaoui, Laurent. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5):780–798, 2005.

Strehl, Alexander L. and Littman, Michael L. A theoretical analysis of Model-Based Interval Estimation. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pp. 856–863, New York, New York, USA, August 2005. ACM Press.

Sutton, Richard S. and Barto, Andrew G. *Reinforcement learning: an introduction*. MIT Press, 1998.

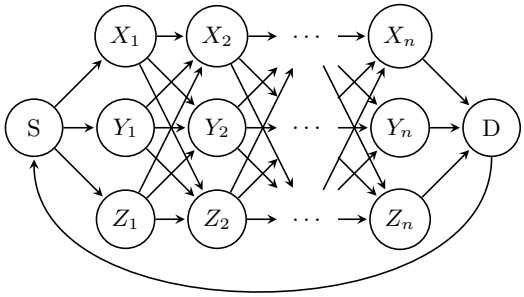


Figure 7. MDP reduction of the 3SAT problem.

Appendix

Proof of Theorem 1.

Proof. We will prove the theorem by reducing the satisfiability problem in conjunctive normal form with three variables (3SAT) to the problem of deciding whether there exists a p-safe policy for a belief that we will construct. The 3SAT problem amounts to deciding whether there exists an assignment to boolean variables $\{U_k\}$ such that the following expression is true:

$$(X_1 \vee Y_1 \vee Z_1) \wedge \cdots \wedge (X_n \vee Y_n \vee Z_n)$$

where each of the variables X_i, Y_i, Z_i equals one of the variables in $\{U_k\}$, possibly negated.

We start by constructing an MDP to represent this problem as shown in Figure 7. In addition to actions corresponding to the outgoing arrows, the agent also has the option of remaining in the same state. A transition from some state to another state will succeed if and only if the boolean variable corresponding to the origin state is true. The boolean variable associated to states S and D are always true. Our belief is the uniform distribution over truth values of the boolean variables $\{U_k\}$.

Now consider the following simple policy: from D go to S and then stay in S. For any recall time $T > 0$, the recall event will find the agent in state S, so the policy is p-safe for any $p > 0$ if and only if the belief assigns a non-zero measure to MDPs in which state D is accessible form state S, so if any only if there exists at least one boolean assignment for the $\{U_k\}$ such that state D is accessible from S. It is easy to see that, this is the case if and only if the 3SAT formula is satisfied, and this observation completes the reduction.

This result should come as no surprise since similar optimization problems have been shown to be NP-hard in the context of *Partially Observable Markov Decision Processes* (Blondel & Tsitsiklis, 2000).

□

Proof of Theorem 2.

Proof. The result is an immediate consequence of the following Lemma. □

Lemma 3. *Given a belief β and a policy π, there exists a policy dependent reward correction, $\sigma^{\beta,\pi}$, defined below, such that the MDP with transition measure $p := E_\beta P$ and rewards $r + \sigma^{\beta,\pi}$, where $r := E_\beta R$, has the same expected total return as the belief for any initial distribution. Formally:*

$$\forall \rho \quad E_\beta E^P_{\rho,\pi} \sum_{t=0}^{\infty} R_{S_t,A_t} = E^p_{\rho,\pi} \sum_{t=0}^{\infty} \left(r_{s,a} + \sigma^{\beta,\pi}_{s,a}\right)$$

$$\sigma^{\beta,\pi}_{s,a} := \sum_{s'} E_\beta \left[(P_{s,a,s'} - E_\beta[P_{s,a,s'}]) E_{s',\pi,P}[V]\right].$$

Proof. The Markov property under belief β reads:

$$E_\beta E^P_{s,\pi}[V] = \sum_a \pi_{s,a} E_\beta[R_{s,a}] + \\ + \sum_a \pi_{s,a} \sum_{s'} E_\beta[P_{s,a,s'} E^P_{s',\pi}[V]].$$

The Markov property assuming expected transition frequencies and expected rewards with safety penalty is:

$$E^p_{s,\pi}[\bar{V}] = \sum_a \pi_{s,a}(r_{s,a} + \sigma^{\beta,\pi}_{s,a}) + \\ + \sum_a \pi_{s,a} \sum_{s'} p_{s,a,s'} E^p_{s',\pi}[\bar{V}]].$$

Now let $\Delta_s := E_\beta E^P_{s,\pi}[V] - E^p_{s,\pi}[\bar{V}]$. By subtracting the first two equations we get that:

$$\Delta_s = \sum_a \pi_{s,a} \sum_{s'} p_{s,a,s'} \Delta_{s'}.$$

We can see that Δ satisfies the same equation as the value function in an MDP with transition measure p and zero rewards. Since the value function in such an MDP is uniquely defined and identically zero, we conclude $\Delta_s = 0$. □

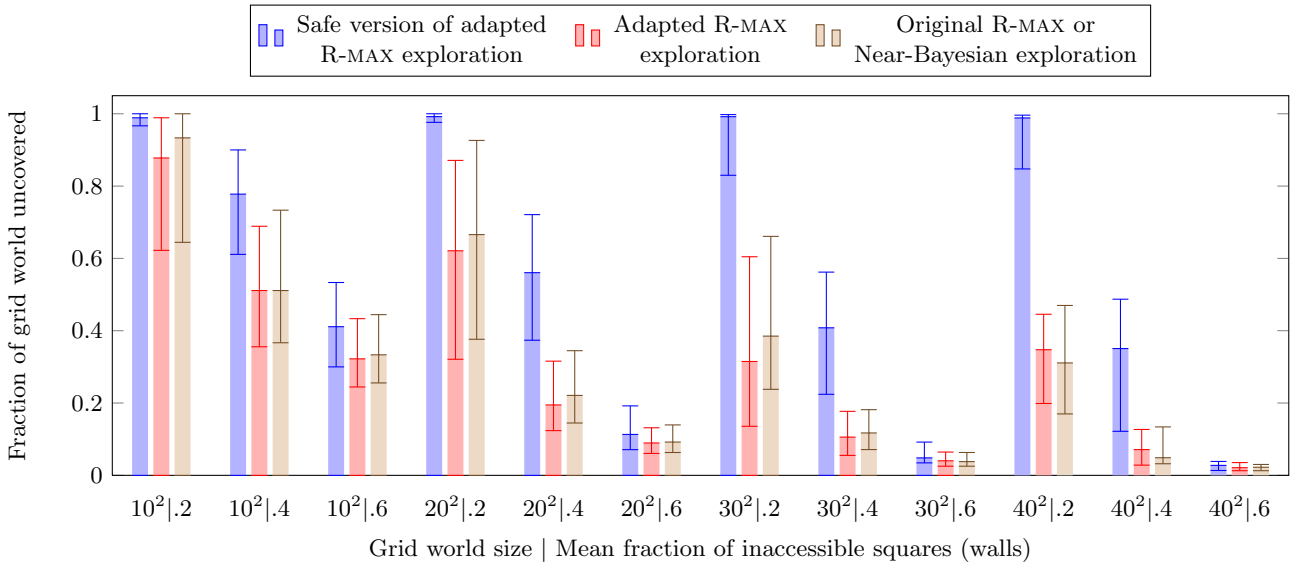


Figure 6. Exploration efficiency comparison. We are showing the median, the upper and the lower quartiles of the fraction of the grid world that was uncovered by different explorers in randomly generated grid worlds. The "amount" of non-ergodicity is controlled by randomly making a fraction of the squares inaccessible (walls). We ran 1000, 500, 100,20 experiments for grids of sizes 10^2, 20^2, 30^2 and 40^2 respectively. We are comparing against our own adapted R-MAX exploration objective, the original R-MAX objective (Brafman & Tennenholtz, 2001) and the Near-Bayesian exploration objective (Kolter & Ng, 2009). The last two behave identically in our grid world environment, since, once a state is visited, all transitions out of that state are precisely revealed.