Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling (ICAPS 2019)

Robust and Adaptive Planning under Model Uncertainty

Apoorva Sharma,[1] James Harrison,[2] Matthew Tsao,[3] Marco Pavone[1]
Department of [1]Aeronautics and Astronautics, [2]Mechanical Engineering, [3]Electrical Engineering
Stanford University, Stanford, CA 94305
{apoorva,jharrison,mwtsao,pavone}@stanford.edu

Abstract

Planning under model uncertainty is a fundamental problem across many applications of decision making and learning. In this paper, we propose the Robust Adaptive Monte Carlo Planning (RAMCP) algorithm, which allows computation of risk-sensitive Bayes-adaptive policies that optimally trade off exploration, exploitation, and robustness. RAMCP formulates the risk-sensitive planning problem as a two-player zero-sum game, in which an adversary perturbs the agent's belief over the models. We introduce two versions of the RAMCP algorithm. The first, RAMCP-F, converges to an optimal risk-sensitive policy without having to rebuild the search tree as the underlying belief over models is perturbed. The second version, RAMCP-I, improves computational efficiency at the cost of losing theoretical guarantees, but is shown to yield empirical results comparable to RAMCP-F. RAMCP is demonstrated on an n-pull multi-armed bandit problem, as well as a patient treatment scenario.

Introduction

In many sequential decision making domains, from personalized medicine to human-robot interaction, the underlying dynamics are well understood save for some latent parameters, which might vary between episodes of interaction. For example, we might have models for the evolution of a disease of a patient under various treatments, but they may depend on unobserved, patient-specific physiological parameters. Faced with a new patient, the agent must learn over the course of a *single* episode of interaction, as it simultaneously tries to identify the underlying parameters while maximizing its objective of improving the patient's health.

Such challenges are commonplace, yet not well addressed by standard episodic reinforcement learning, which assumes no prior knowledge and learns over the course of repeated interaction on the same system. These problems are better addressed by a Bayesian approach to reinforcement learning, in which the agent can leverage prior knowledge in the form of a *belief distribution* over a family of likely models, which can be updated via Bayes' rule as the agent interacts with the system (Ghavamzadeh et al. 2015). Incorporating this prior knowledge enables effective learning within a single episode of interaction, and also allows the the agent to consider how to balance identification of the latent parameters with maximizing the objective.

While leveraging a prior distribution over models is powerful, coming up with accurate priors remains a challenge. In the contexts of human interaction scenarios like patient treatment, these distributions might be obtained experimentally from past interactions, or may be chosen heuristically by a domain expert. Therefore, this prior over models is likely to be inaccurate, and thus it is paramount that the agent plan in a manner that is robust to incorrect priors, especially in safety-critical settings.

Contributions The contributions of this work are threefold. First, we present (to our knowledge) the first mathematical framework to incorporate robustness to priors in the context of Bayesian RL. Second, we present Robust Adaptive Monte Carlo Planning (RAMCP), an sampling-based tree search algorithm for online planning in this framework for discrete MDPs and discrete priors over models. The approach fundamentally consists of an adversarial weighting step on top of standard risk-neutral tree-search approaches to Bayesian RL such as BAMCP (Guez, Silver, and Dayan 2013). In particular, we introduce two version of the RAMCP algorithm, which we refer to as RAMCP-F and RAMCP-I. For the first, we prove that this adversarial optimization does not oscillate, and indeed converges to the optimal solution. RAMCP-I, on the other hand, sacrifices convergence guarantees for empirical performance. Finally, we demonstrate the algorithms through numerical experiments including a patient treatment scenario, and compare the performance of the two versions of the RAMCP algorithm.

Background

This work considers adding robustness to model-based Bayesian reinforcement learning. In this setting, we wish to control an agent in a system defined by the Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, R, H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $T(s'|s,a)$ is the transition function, $R(s,a,s')$ is the stage-wise reward function, and H is the problem horizon. We assume that the exact transition dynamics T depend on a parameter θ, and denote this dependence as T_θ. We assume that the agent knows this MDP, but is uncertain about the true setting of the parameter θ, and instead maintains a belief distribution over this parameter. This structured representation of the agent's knowledge (and lack of knowledge) of the planning problem allows considering the objectives of exploration and robustness in addition to the standard MDP objective of maximizing reward.

The Explore/Exploit Dilemma As the agent acts in the true MDP, the observed state transitions will provide information about the true underlying system parameters. An agent might *explore*, i.e. choose actions with the aim of reducing uncertainty over θ, or *exploit*, choosing actions to maximize cumulative reward given its current estimates of θ. The Bayesian setting allows optimally making this tradeoff.

Writing the observed transitions so far, or the *history* in an environment at time t as $h_t = (s_0, a_0, \ldots, s_t)$ and given a prior distribution over the model parameters b_{prior}, we can define optimal behavior in the Bayesian setting. Let $\mathcal{H}$ denote the set of possible histories for a given MDP. Then, we will write the set of stochastic history-dependent polices $\pi : \mathcal{H} \times \mathcal{A} \to [0, 1]$ as Π. Let

$$V(h_t,\pi) = \mathbb{E}_{\pi,b}\left[\sum_{\tau=t}^{H-1} R(s_\tau, a_\tau, s_{\tau+1}) \mid (s_0, a_0, \ldots, s_t) = h_t\right]$$

denote the value function associated with policy π, for history h, and with model distribution b. A history-dependent policy π^* is said to be Bayes-optimal with respect to the prior b_{prior} if it has associated value function $V(\{s\}, \pi^*) = \sup_{\pi \in \Pi} V(\{s\}, \pi)$ (Martin 1967).

The problem is a special case of a Partially Observable Markov Decision Processes (POMDP), where the hidden portion of the state (here, the parameters θ) is fixed over the course of an episode. As with all POMDPs, this problem can be cast into a belief-state MDP by augmenting the state at time t with the posterior belief (or, equivalently, the history up to that point h_t). This is the Bayes-Adpative Markov Decision Process (BAMDP) formulation, and the optimal policy in this MDP is the Bayes-optimal policy (Duff 2002). In general, optimizing these policies is computationally difficult. Information-state techniques (as in, e.g., Gittins indices for bandit problems (Gittins, Glazebrook, and Weber 2011)) are typically intractable due to a continuously growing information-state space (Duff 2002). Offline global value approximation approaches, typically based on offline POMDP solution methods, scale poorly to large state spaces (Ghavamzadeh et al. 2015). Online approaches (Wang et al. 2005), (Guez, Silver, and Dayan 2013), (Chen et al. 2016) use tree search with heuristics to either simplify the problem or guide the search.

Robustness to Incorrect Priors While encoding model uncertainty through a prior over model parameters enables optimally balancing exploration and exploitation, it is likely that these priors may be inaccurate. Previous work in policy optimization for BAMDPs has focused on optimizing performance in expectation, and thus does not offer any notion of robustness to misspecified priors (Guez, Silver, and Dayan 2013). Robust MDPs are posed as MDPs with uncertainty sets over state transitions, and approaches to this problem aim to optimize the worst-case performance over all possible transition models (Nilim and El Ghaoui 2005). However, this minimax approach does not consider the belief associated with a transition model, and thus is typically over-conservative and can not optimally balance exploration and exploitation.

Tools from risk theory can be used to achieve tunable, distribution-dependent conservatism. Given a reward random variable Z, a *risk metric* is a function $\rho(Z)$ that maps the uncertain reward to a real scalar, which encodes a preference model over uncertain outcomes where higher values of $\rho(Z)$ are preferred. A key concept in risk theory is that of a *coherent* risk metric. These metrics satisfy axioms originally proposed in (Artzner et al. 1999), which ensure a notion of rationality in risk assessment. We refer the reader to (Majumdar and Pavone 2017) for a more thorough discussion of why coherent risk metrics are a useful tool in decision making. While expectation and worst-case are two possible coherent risk metrics, the set of CRMs is a rich class of metrics including the Conditional Value at Risk (CVaR) metric popular in mathematical finance (Majumdar and Pavone 2017). The rationality of coherent risk metrics contrasts with standard approaches (especially in stochastic control (Glover and Doyle 1988)) such as mean-variance and exponential risk metrics, for which there are simple examples of clearly absurd decision-making (Rabin and Thaler 2001).

Applying risk metrics to the reward of a single MDP, yields an optimization problem with strong connections to the Robust MDP formulation, with the choice of risk metric in the risk-sensitive MDP corresponding to a particular choice of uncertainty set in the Robust MDP formulation (Chow et al. 2015) (Osogami 2012). In contrast, in this work we apply tools from risk theory to the Bayesian setting in which we have a prior over MDPs to balance all three objectives of robustness, exploration, and exploitation in a coherent mathematical framework. To our knowledge, this work represents the first approach that considers all three of these objectives in the context of sequential decision making.

Problem Statement

We aim to compute a history dependent policy π which is optimal according to a coherent risk metric over model uncertainty. To make this objective more concrete, let $\tau = (s_0, a_0, \ldots, s_{H-1}, a_{H-1}, s_H)$ be particular trajectory realization. Note that the probability of a given trajectory depends on both the choice of policy π and the transition dynamics of the MDP, T_θ. The cumulative reward of a given trajectory $J(\tau)$ can be calculated by summing the stage-wise rewards

$$J(\tau) = \sum_{t=0}^{H-1} R(s_t, a_t, s_{t+1}). \tag{1}$$

Note that since the distribution over τ is governed by the stochasticity in each environment as well as the uncertainty over environments, the distribution of the total cost of a trajectory $J(\tau)$ is as well.

In this work, we focus our attention to risk-sensitivity with respect to the randomness from model uncertainty only. Concretely, we can write the objective as

$$\Pi^* = \arg\max_{\pi} \ \rho\left(\mathbb{E}\left[J(\tau)|\theta, \pi\right]\right), \tag{2}$$

where the risk metric is with respect to the random variable induced by the distribution over models. Note that Π^* denotes the set of optimal policies.

In this work, we consider a finite collection of M possible parameter settings, $\Theta = \{\theta_i\}_{i=1}^M$, and write our prior belief over Θ as the vector b_{prior}. While limiting ourselves to discrete distributions over model parameters is somewhat restrictive, these simplifications are common in, for example, sequential Monte Carlo. Indeed, computing continuous posterior distributions exactly is often intractable, making this discrete approximation necessary (Guez et al. 2014).

The above objective differs from that typically used in the risk-sensitive reinforcement learning literature, which applies the risk metric ρ directly to the total reward random variable $J(\tau)$ (Tamar et al. 2017). In contrast, we first marginalize out the effects of stochasticity via the expectation, and then apply ρ to the multinomial random variable $\mathbb{E}\left[J(\tau)|\theta = \theta_i\right]$ where each outcome corresponds to the expected value of the policy π on model θ_i. The robustness provided by risk-sensitivity protects against modeling errors in distribution. In the BAMDP context, the primary error in distribution is in the model belief: the true distribution over models has all its mass on one model, so effectively any belief will be incorrect. Furthermore, these beliefs will be updated online, and thus are susceptible to noise. Comparatively, we assume that for a particular model, the stochasticity in transition dynamics is well characterized. Thus, we argue that the optimization objective (2) is well-aligned with our goal of enabling robustness to model uncertainty.

Approach

In this section we discuss the high-level approach taken in RAMCP. We begin by reformulating (2) as a two-player, zero-sum game. This game is played between an agent computing optimal history-dependent policies with respect to a belief over models, and an adversary perturbing this belief. Armed with this problem reformulation, we describe Generalized Weakened Fictitious Play (GWFP) (Leslie and Collins 2006), a framework for computing Nash equilibria of two-player zero-sum games (as well as a collection of other game settings). Finally, we outline application of GWFP to (2).

Reformulation as a Zero-Sum Game

Our reformulation of the objective stems from a universal representation theorem which all coherent risk metrics (CRMs) satisfy.

Theorem 1 (Representation Theorem for Coherent Risk Metrics (Artzner et al. 1999))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where Ω is a finite set with cardinality $|\Omega|$, $\mathcal{F}$ is a $\sigma-$algebra over subsets (i.e., $\mathcal{F} = 2^\Omega$), probabilities are assigned according to $\mathbb{P} = (p(1), \ldots, p(|\Omega|))$, and $\mathcal{Z}$ is the space of reward random variables on Ω. Denote by $\mathcal{C}$ the set of valid probability densities:*

$$\mathcal{C} := \left\{ \zeta \in \mathbb{R}^{|\Omega|} \mid \sum_{i=1}^{|\Omega|} p(i)\zeta(i) = 1, \zeta \geq 0 \right\}. \tag{3}$$

Define $p_\zeta \in \mathbb{R}^{|\Omega|}$ as $p_\zeta(i) = p(i)\zeta(i), i = 1, \ldots, |\Omega|$. A risk metric $\rho : \mathcal{Z} \to \mathbb{R}$ with respect to the space $(\Omega, \mathcal{F}, \mathbb{P})$ is a coherent risk metric if and only if there exists a compact convex set $\mathcal{B} \subset \mathcal{C}$ such that for any $Z \in \mathcal{Z}$:

$$\rho(Z) = \min_{\zeta \in \mathcal{B}} \mathbb{E}_{p_\zeta}[Z] = \min_{\zeta \in \mathcal{B}} \sum_{i=1}^{|\Omega|} p(i)\zeta(i)Z(i). \tag{4}$$

This theorem offers an interpretation of CRMs as a worst-case expectation over a set of densities $\mathcal{B}$, often referred to as the *risk envelope*. The particular risk envelope depends on the risk metric chosen, as well as the degree of risk-aversity (which is captured by a parameter for most coherent risk metrics). In this work we focus on *polytopic risk metrics*, for which the envelope $\mathcal{B}$ is a polytope. For this class of metrics, the constraints on the maximization in Equation 4 become linear in the optimization variable ζ, and thus solving for the value of the risk metric becomes a tractable linear programming problem. Note that solving this linear program is at the core of RAMCP. This computational step is why RAMCP is restricted to discrete distributions over models, and can not directly be extended to continuous beliefs. Polytopic risk metrics constitute a broad class of risk metrics, encompassing risk neutrality, mean absolute semi-deviation, spectral risk measures, as well as the CVaR metric often used in financial applications, with the choice of metric determining the form of the polytope (Eichhorn and Römisch 2005).

Through the representation theorem for coherent risk metrics (Equation 4), we can understand Equation 2 as applying an adversarial reweighting ζ to the distribution over models b_{prior}. Let $b_{\text{adv}}(i) = b_{\text{prior}}(i)\zeta(i),\ i = 1, \ldots, M$ represent this reweighted distribution. Thus, (2) may be written

$$\Pi^* = \arg\max_{\pi} \ \min_{\zeta \in \mathcal{B}} \mathbb{E}_{\theta \sim b_{\text{adv}}}\left[\mathbb{E}\left[J(\tau) \mid \theta, \pi\right]\right], \tag{5}$$

where again Π^* denotes the set of optimal policies. We are interested simply in finding a single policy within this set, as opposed to the full set. Note that this takes the form of a two player zero-sum game between the agent (the maximizer) and an adversary (the minimizer). One play of this game corresponds to the following three step sequence:

1. The adversary acts according to its strategy, choosing b_{adv} from the risk envelope.
2. Chance chooses $\theta \sim b_{\text{adv}}$.
3. The agent acts according to its strategy, or policy, $\pi(h)$ in the MDP with dynamics T_θ.

The action of the adversary is to choose a perturbation to the belief distribution from the polytope of valid disturbances that minimizes the expected performance of the agent. The agent seeks to compute an optimal policy under this perturbed belief. The solution to Equation 5 is therefore the optimal Nash equilibrium of the two player game, which we denote as $(b^*_{\text{adv}}, \pi^*)$.

Computing Nash Equilibria

Having formulated (2) as a two-player zero-sum game, we leverage tools developed in algorithmic game theory to efficiently compute Nash equilibria. For two-player, zero-sum games, a Nash equilibrium can be directly computed by solving a linear program of size proportional to the strategy space of each player. In the context of our problem statement, the agent's strategy space is the space of all history dependent policies, and thus solving for a Nash equilibrium directly is computationally intractable. To compute the Nash equilibria of the game defined by (2), we apply iterative techniques which converge to equilibria over repeated simulations of the game.

Fictitious Play (Brown 1949) is a process in which players repeatedly play a game and update their strategies toward the best-response to the average strategy of their opponents. This process has been shown to asymptotically converge to a Nash equilibrium. In (Leslie and Collins 2006), the authors introduced Generalized Weakened Fictitious Play (GWFP), which allows for computation of approximate best-responses but maintains convergence guarantees, and thus has worked well for large-scale extensive form games (Heinrich, Lanctot, and Silver 2015). GWFP converges to Nash equilibria in several classes of games, including two-player zero-sum games such as (5). For the risk-sensitive BAMDP, the GWFP updates to the adversary strategy b and agent strategy π are:

$$b_{k+1} = (1-\alpha_{k+1})b_k + \alpha_{k+1}\mathrm{BR}_\epsilon(\pi_k), \quad (6)$$
$$\pi_{k+1} = (1-\alpha_{k+1})\pi_k + \alpha_{k+1}\mathrm{BR}_\epsilon(b_k) \quad (7)$$

where $\mathrm{BR}_\epsilon(\sigma)$ represents an ϵ-suboptimal best response[1] to strategy σ, and α_{k+1} is an update coefficient chosen such that $\sum_{k=1}^{\infty}\alpha_k = \infty$ and $\lim_{k\to\infty}\alpha_k = 0$. While $\mathrm{BR}_\epsilon(\sigma)$ is typically used to refer to the set of ϵ-best responses, we will use this to refer to a strategy within this set. In this work, we set $\alpha_k = 1/k$ and thus both strategies represent running averages of the best-responses, a property we leverage in our algorithm. Initial values of the belief and policy may be chosen arbitrarily.

The adversarial best response $\mathrm{BR}_\epsilon(\pi_k)$ can be computed by solving the linear program

$$\begin{aligned}&\min_{b,\zeta\in\mathcal{B}} \sum_{i=1}^{M} \hat{V}_{\pi_k}(i) b(i) \\ &\text{s.t. } b_{\text{prior}}(i)\zeta(i) = b(i), \;\; i=1,\dots,M,\end{aligned} \quad (8)$$

where $\mathcal{B}$ is the polytopic risk envelope, and $\hat{V}_{\pi_k}(i)$ is an estimate of $V_{\pi_k}(i) := \mathbb{E}\left[J(\tau) \mid \theta=\theta_i, \pi=\pi_k\right]$. The suboptimality of the solution of this LP is bounded by the error in $\hat{V}_{\pi_k}(i)$.

The agent's best response $\mathrm{BR}_\epsilon(b_k)$ is a history dependent policy $\pi(h) = \arg\max_a \hat{Q}_{b_k}(h,a)$, where $\hat{Q}_{b_k}(h,a)$ is an estimator of $Q^*_{b_k}(h,a)$, the value of taking action a at history h, then acting optimally when θ is drawn from b_k at the start of the episode. However, computation of this policy is non-trivial, and a naïve approach to value estimation would involve substantial repeated computation that would result in poor performance.

[1] If an opposing player chooses strategy σ, then an ϵ-suboptimal best response to σ is a strategy such that the player obtains a payoff (or cumulative reward) within ϵ of that of an optimal response (which is itself referred to as a best response).

Algorithm 1 RAMCP

function SEARCH(s_0, b_{prior})
  $\hat{V}_\pi(i) \leftarrow 0$ for all $i = 1,\dots,M$
  $k \leftarrow 0$
  $b_{\text{adv}} \leftarrow b_{\text{prior}}$
  while within computational budget **do**
    $k \leftarrow k+1$
    $w \leftarrow M \cdot \mathrm{BR}_\epsilon(\pi)$
    for $i = 1$ **to** M **do**
      $w \leftarrow M \cdot b_{\text{adv}}(i)$
      $V_{\text{br}} \leftarrow$ SIMULATE(s_0, θ_i, w)
      $\hat{V}_\pi(i) \leftarrow \hat{V}_\pi(i) + \frac{1}{k}(V_{\text{br}} - \hat{V}_\pi(i))$
    end for
    { COMPUTEQVALUES(s_0) }
    $b_{\text{adv}} \leftarrow$ solution to linear program (8)
  end while
  return $\pi_{\text{avg}} =$ AVGACTION(h) for all h
end function ▷ Lines in {} for RAMCP-F only.

RAMCP Outline

Carrying out the GWFP process to solve (5) requires estimates $\hat{V}_{\pi_k}(i)$ and $\hat{Q}_{b_k}(h,a)$ at every iteration k. To compute $\hat{V}_{\pi_k}(i)$, we can average the total reward accrued on multiple rollouts of policy π_k on model θ_i, obtaining a Monte Carlo estimate of $\mathbb{E}\left[J(\tau) \mid \theta_i, \pi_k\right]$. Computing $\hat{Q}_{b_k}(h,a)$ is equivalent to approximately solving the BAMDP induced by distribution b_k. Many sampling-based methods exist to estimate the optimal Q function in a BAMDP.

Guez et al. (Guez, Silver, and Dayan 2013) showed that performing Monte-Carlo tree search where the dynamics parameter θ is drawn from b_k at the root of the tree at each iteration can accurately estimate the optimal value function $Q^*_{b_k}(h,a)$. These techniques suggest a naïve approach to solving for the optimal policy: at each iteration of the GWFP process, one could compute $\mathrm{BR}_\epsilon(b_k)$ by running a tree search algorithm on b_k, and compute $\mathrm{BR}_\epsilon(\pi_k)$ by rolling out policy π_k on each model θ_i to get $\hat{V}_{\pi_k}(i)$, and then solving the linear program (Equation 8). This is clearly impractical, as each iteration requires solving a new BAMDP. Furthermore, in order for the GWFP process to converge, the suboptimality of the best responses must go to zero as $k\to\infty$. Thus, each iteration of this naïve implementation would require a growing number of samples, increasing the computational challenges with this approach. Critically, we are able to leverage the structure of the GWFP process to obtain an algorithm that converges to the same result, iterating between performing FP iterations and building the tree. This approach requires growing only one tree, which results in substantial efficiency improvement.

Algorithm 2 Simulate

function SIMULATE(h, θ, w)
  $N(h) \leftarrow N(h) + 1$
  $W(h) \leftarrow W(h) + w$
  $V_{\text{br}} \leftarrow 0$
  if LEN(h) $>= H$ **or** h is terminal **then**
    return V_{br}
  end if
  for all $a \in \mathcal{A}$ **do**
    $N(h,a) \leftarrow N(h,a) + 1$
    $W(h,a) \leftarrow W(h,a) + w$
    $s' \sim T_\theta(s,a)$
    $r \leftarrow R(s,a,s')$
    $V'_{\text{br}} \leftarrow$ SIMULATE(has',θ,w)
    $Q_{ha} \leftarrow r + V'_{\text{br}}$
    [$Q(h,a) \leftarrow Q(h,a) + \frac{1}{N(h,a)}(wQ_{ha} - Q(h,a))$]
    if $a == \arg\max_{a'} Q(h,a')$ **then**
      $V_{\text{br}} \leftarrow Q_{ha}$
      $W_{br}(h,a) \leftarrow W_{br}(h,a) + w$
      [$V(h) \leftarrow V(h) + \frac{1}{N(h)}(wV_{\text{br}} - V(h))$]
    end if
  end for
  return V_{br}
end function ▷ Lines in [] for RAMCP-I only.

Algorithm 3 ComputeQValues

function COMPUTEQVALUES(h)
  if LEN(h) $>= H$ or h is terminal **then**
    $V(h) = 0$
    return 0
  end if
  for all $a \in \mathcal{A}$ **do**
    $w, r, V' \leftarrow [\,], [\,], [\,]$
    for all $s' \in$ CHILDREN(ha) **do**
      APPEND(w, $W(has')$)
      APPEND(r, $R(s,a,s')$)
      APPEND(V', COMPUTEQVALUES(has'))
    end for
    $Q(h,a) \leftarrow \frac{1}{W(h,a)}SUM(w \odot (r + V'))$
  end for
  $V(h) \leftarrow \max_a Q(h,a)$
  return $V(h)$
end function

Algorithm Overview

In this section, we present the RAMCP algorithm, which combines simulation-based search with fictitious play iterations to optimize the risk-sensitive, Bayes-adaptive objective (5). We present two versions of this procedure: RAMCP-F, a slower, but provably asymptotically optimal algorithm, and the more efficient RAMCP-I, for which we do not provide a proof of convergence, but observe good performance empirically. The algorithms share the same overall structure, which is detailed in Algorithm 1. Lines specific to RAMCP-F are enclosed in curly brackets, while those specific to RAMCP-I are enclosed in square brackets.

To compute a risk-sensitive plan for belief b_{prior} from state s_0, the agent calls the SEARCH function. The function iterates between simulating rollouts on different transition models, sampled from an adversarial distribution, and using the improved value estimates from these simulations to improve the agent policy and the adversarial distribution.

Employing ideas from conditional Monte Carlo, the algorithm loops over each model θ_i, and computes a weighting w proportional to the current adversarial belief b_{adv}. This weighting is applied to the statistics recorded in the tree, allowing the algorithm to estimate quantities as if the models were sampled from b_{adv} rather than looped over deterministically.

The algorithm maintains a tree where each node is state or action along a trajectory from s_0. We denote nodes by the history leading to the node. If the node is a state we refer to this history h, and those ending in an action, which we denote as ha. For every node, we store visitation counts $N(h), N(h,a)$, cumulative visitation weights $W(h), W(h,a)$, and value estimates $V(h), Q(h,a)$ for every node in the tree.

The call to SIMULATE(h, θ, w) (Algorithm 2) simulates all possible H-length action sequences starting at h under the a given model θ, and increments the visitation weights along the resulting trajectories by w. The function returns V_{br}, the reward sequence obtained when taking actions were consistent with the greedy policy with respect to the current Q values stored in the tree. By keeping a running average of V_{br} from s_0 for every model θ_i in $\hat{V}_\pi(i)$, $\hat{V}_\pi(i)$ estimates the performance of the agent's average strategy (π_k in GWFP) on model θ_i.

In RAMCP-F, the value estimates in the tree are updated by calling COMPUTEQVALUES (Algorithm 3). This function recurses through the entire simulated tree, and computes value estimates starting at the leaf nodes, via dynamic programming according to the empirical transition probabilities

$$Q(h,a) = \sum_{s'} \hat{p}(s'|h,a)(R(s,a,s') + V(has')) \tag{9}$$

$$V(h) = \max_a Q(h,a) \tag{10}$$

Here, we only sum over visited states, and use $\hat{p}(s'|h,a) = W(has')/W(ha)$ as the empirical transition probability. It can be shown that this probability converges to the transition probability corresponding to a BAMDP with $\theta \sim \bar{b}_{\text{adv}}$, where $\bar{b}_{\text{adv}}$ is the average adversary strategy over the iterations of the algorithm, see the extended version for a proof (Sharma et al. 2019). Since this average strategy is exactly b_k in the GWFP algorithm, this step corresponds to computing $\text{BR}_\epsilon(b_k)$.

Finally, the procedure computes $b_{\text{adv}} = \text{BR}_\epsilon(\pi_k)$ by solving the LP (8) given the current estimates of agent performance in $\hat{V}_\pi$.

Critically, in contrast to the naïve approach, RAMCP avoids performing a separate simulation-based search under a new model distribution b_k at every iteration of of the GWFP. Instead, it performs simulations according to $b_{\text{adv}} = \text{BR}_\epsilon(\pi_k)$ combines the averaging of GWFP with the averaging used in the Monte Carlo estimation of the value estimates. Note that the proposed approach does not compute the running averages π_k and b_k directly. However, since the mixed strategy π_k is what converges to the Nash equilibrium in GWFP, we keep counts $W_{\text{br}}(ha)$, corresponding to whenever action a matches $\text{BR}_\epsilon(b_k)$. The mixed strategy corresponds to sampling actions at history h with probability proportional to $W_{\text{br}}(ha)$. Procedure AVGACTION(h) samples actions in this fashion. Thus, RAMCP-F implicitly carries out the GWFP process, and therefore converges to a Nash equilibrium. More formally, we have the following result:

Theorem 2 (Convergence of RAMCP-F)**.** *Let π_k denote the output of RAMCP-F (Algorithm 1) after k iterations of the outer loop. Then,* $\lim_{k\to\infty} \pi_k \in \Pi^*$ *in probability.*

The proof of this result is given in the appendix of the extended version of the paper (Sharma et al. 2019).

Separating Q value estimation from updating visitation counts through simulation is more convenient for analysis, but requires iterating through the entire tree at each step of the algorithm. Many tree search algorithms, such as UCT (Kocsis and Szepesvári 2006), use stochastic approximation to estimate the Q values via iterative updating, avoiding this recomputation. RAMCP-I is an incremental version of RAMCP-F, which does not call COMPUTEQVALUES, and instead adds the incremental updates to $Q(h, a)$ and $V(h)$ into SIMULATE (see Algorithm 2, lines 32 and 36).

While tools from stochastic approximation can be used to prove convergence when the underlying distribution is stationary, this is not the case in the incremental setting, as the adversarial distribution over models is being recomputed at every iteration of GWFP. We observe that in practice, this incremental version also performs well. Popular techniques for Monte Carlo tree search based on non-uniform sampling such as UCT interact with the constantly updated adversarial distribution in a way that is hard to theoretically characterize. However, the added step of solving the LP to compute the adversarial belief is only a minor increase in computational effort for most reasonably-sized problems. Thus, analysis on the interaction of these tree search methods on the non-stationary tree is a promising but mathematically complex avenue of future work, that may result in algorithms that further increase robustness at little computational cost.

Experiments

We present experimental results for an n-pull multi-armed bandit problem, as well as for a patient treatment scenario. The bandit problem is designed to show the fundamental features of the RAMCP algorithm, while the patient treatment example is a larger scale example motivated by a real-world challenge. In both experiments, we use the CVaR_α risk metric, which at a given α-quantile corresponds to the expectation over the α fraction worst-case outcomes. For $\alpha = 1$, this corresponds to the risk-neutral expectation, and in the limit as $\alpha \to 0^+$, this corresponds to the worst-case metric. For CVaR, the risk polytope $\mathcal{B}$ may be stated in closed form (Majumdar and Pavone 2017).

Multi-armed Bandit

We consider a multi-armed bandit scenario to illustrate several properties of the RAMCP algorithm. Given a finite nnumber of "pulls" of the bandit, and a finite number of reward realizations, we can represent this scenario as an BAMDP. There is one initial state, in which the agent has four possible actions, which each have different stochastic transition models to 6 possible rewarding states. The rewarding states transition deterministically back to the decision state, where the agent must choose another arm to pull.

We define two possible models, each with different transition probabilities that are known in the planning problem. For illustrative purposes, we chose the transition probabilities to highlight the trade-off between exploitation, exploration and risk. Actions 1 and 2 are *exploratory*: deterministic under each model, and therefore reveal the true model. The actions differ in terms of *risk*: action 1 gives low but similar rewards in both models, while action 2's reward is higher in expectation but varies drastically between the two models. Actions 3 and 4 have more stochastic outcomes under both models, and thus reveal less about the true model. However, these actions serve as choices for *exploitation*, with action 3 offering the highest reward in expectation in θ_1, and action 4 in θ_2. The prior belief is $(0.6, 0.4)$ for θ_1 and θ_2, respectively. An episode consists of two pulls (or two actions) in the environment.

On this small scale example, we test both RAMCP-F and RAMCP-I, and demonstrate that both converge to empirically similar solutions. Figure 1 shows the performance of RAMCP under varying values of α. The top row shows the adversarially perturbed belief over iterations of the outer loop of RAMCP-F. The purple points show the solution to Equation 8. These may switch rapidly, as the adversary will aim to place as little probability mass on the high value models as possible, to minimize the expected value of a given policy. This behavior can be seen for $\alpha = 0.25$. Indeed, this demonstrates the necessity of the averaging in the GWFP process. While the best response beliefs change rapidly, the running average of the adversarially perturbed belief converges, shown in orange. In the second row of Figure 1, the estimated values $\hat{V}_{\pi_k}$ for each θ_i are plotted. The dotted line denotes the expected value of these estimates with respect to the prior belief. As the CVaR quantile α decreases, the expected reward with respect to the prior belief over models decreases, but the performance for the worst-case model improves. This illustrates how by choosing the risk metric used by RAMCP-F (here by tuning the α-quantile of the CVaR metric), the user can obtain policies that meet their standards for robustness. The last row of Figure 1 compares the value under the prior belief as estimated by RAMCP-F against that estimated by RAMCP-I. We see that RAMCP-I converges to similar values as RAMCP-F, suggesting that it is a reasonable approximation to the asymptotically optimal algorithm.

Figure 2 shows the mean performance of the policies obtained through both RAMCP-F and RAMCP-I on the bandit problem. We plot the mean performance of the algorithm as the distribution over underlying models is allowed to shift adversarially away from the prior distribution used for planning,

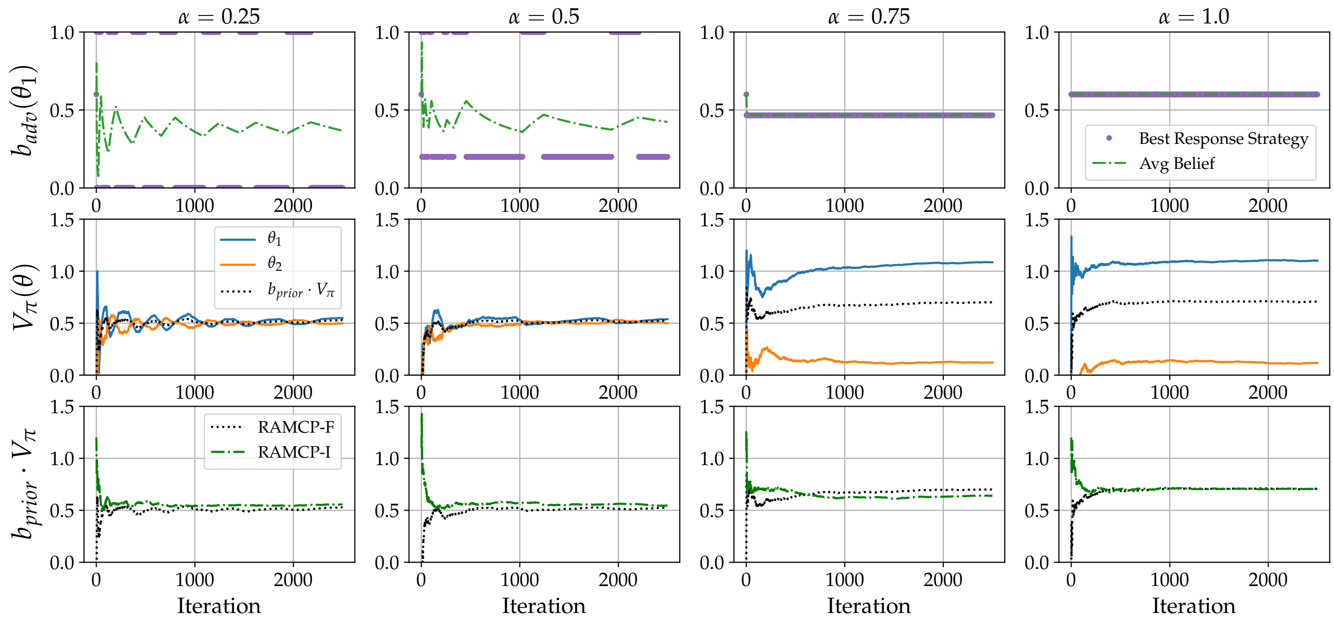

Figure 1: Convergence of the RAMCP-F algorithm for different CVaR quantiles (α). The upper row shows best response belief perturbations (blue) and the running average adversarial belief perturbation (orange) for various CVaR quantiles, α. The second row shows the value function for the policy generated by RAMCP-F operating in an environment with dynamics parameterized by θ_1, and by θ_2 (blue and orange respectively). The dotted line denotes the expected value of the policy under the prior belief over models. The bottom row compares the estimated value under the prior belief as computed by RAMCP-F against that computed by RAMCP-I. We see that the two algorithms converge to similar values.

denoting distance from the prior via the Kullback-Leibler divergence. We see that RAMCP is able to sacrifice expected performance under the provided prior in exchange for robustness to an incorrect prior distribution. Further, we see that both RAMCP-F and RAMCP-I have the same performance, which further supports RAMCP-I as a practically useful modification of RAMCP-F.

Patient Treatment

The problem of developing patient treatment plans is one where being robust yet adaptive to model uncertainty is critical. Each individual patient may respond in different ways to a given treatment strategy, and therefore exploratory actions for model disambiguation are often required. There are many such problems in medicine that have been formulated under the BAMDP or POMDP framework, including choosing drug infusion regimens (Hu, Lovejoy, and Shafer 1996), or HIV treatment plans (Attarian and Tran 2017). With human lives at stake, being robust towards this model uncertainty is also critical, and thus RAMCP offers an attractive solution approach. We demonstrate the effectiveness of RAMCP in such a domain by developing a simplified problem which captures the complexity of such tasks.

We consider a model where the state is the patient's health: $s \in \mathcal{S} = \{0, 1, \ldots, 19\}$, where $s = 0$ corresponds to death, and $s = 19$ corresponds to a full health level. The patient starts at $s = 3$. The action space is $\mathcal{A} = \{1, 2, 3\}$, corresponding to different treatment options. We randomly generated 15 possible response profiles to each treatment option. Each response profile assigns a probability mass to a relative change in patient health. Under the different response profiles, this change may be positive or negative, so rapid model identification is important. The exact method used to generate the transition probabilities is provided in the supplementary materials. We consider a prior which assigns a weight of 0.25 to the first response profile, and 0.0536 to the remaining 14.

Figure 3 shows the performance of the RAMCP algorithm on this problem, run for 12500 iterations with a search horizon $H = 4$, 500 times for each α-quantile. As can be seen in the figure, the choice of the CVaR quantile α controls the robustness of the resulting plan to an incorrect prior. The risk-neutral setting ($\alpha = 1.0$) yields the highest expected performance assuming an accurate prior, but its performance can degrade rapidly as the distribution that the model is sampled from changes. Rather than force a practitioner into this tradeoff, RAMCP allows them to encode a preference towards robustness by altering the risk metric. The more robust setting of $\alpha = 0.2$ yields a policy with performance that is less sensitive to the underlying distribution over models, at the cost of a worse reward under the prior distribution. Notably, $\alpha = 0.6$ offers the practitioner an attractive middle ground, offering improved robustness compared to the risk-neutral policy at a small cost to performance on the prior.

We compared the policies obtained by optimizing the coherent CVaR objective through RAMCP against policies optimized with an exponentially weighted objective

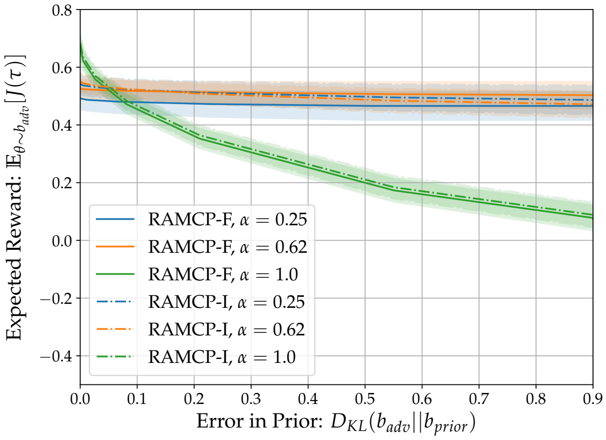

Figure 2: Expected total reward obtained in the bandit problem if the prior is inaccurate. Note that by tuning the CVaR quantile (α) used in the RAMCP algorithm, the resulting policy sacrifices expected performance under the prior to become more robust against distributions over θ that differ from the prior. Error bars denote 90% confidence bounds on the mean total reward. We observe that on this problem, the RAMCP-I algorithm matches the performance of the asymptotically optimal RAMCP-F

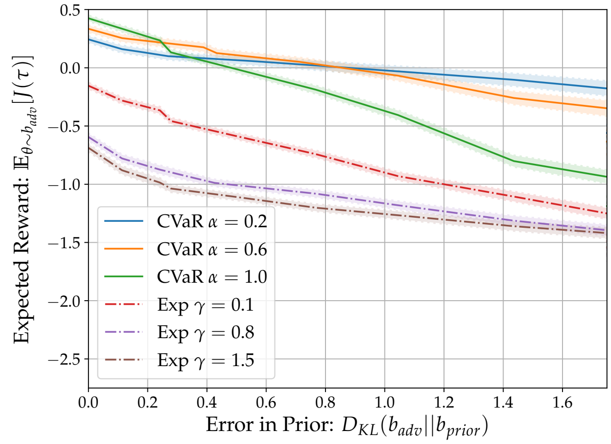

Figure 3: Expected total reward obtained in the patient treatment problem if the prior is inaccurate. As in the bandit problem, tuning the CVaR quantile (α) enables the RAMCP-I algorithm to choose plans that are more robust against distributions over θ that differ from the prior. A naïve approach to add risk sensitivity to the BAMDP via exponential cost shaping yields policies that perform consistently worse, as shown in the dashed-dotted curves. Error bars denote 90% confidence bounds on the mean total reward.

$\tilde{r} = \exp(-\gamma r)$, a common heuristic for incorporating risk-sensitivity into optimization. While reweighting the reward in this way allows using standard BAMDP algorithms and avoids the game-theoretic formulation, the exponential risk formulation is not a coherent risk metric. Furthermore, it does not allow addressing the model uncertainty separately from the transition stochasticity. Experimentally, this is evident in Figure 3. While using exponential reward shaping to add robustness to the BAMDP objective does give resulting policies that are somewhat less sensitive to the underlying prior, the expected performance drops under all possible model distributions.

Related Work

This work aims to optimally trade off between exploration, exploitation, and robustness in the problem of planning under model uncertainty. This problem has been approached in the bandit setting (Galichet, Sebag, and Teytaud 2013), but little work has been done in the MDP setting. Generally, RAMCP can be seen as an extension of the Bayes-Adaptive MDP literature toward safety and robustness. In the discrete MDP case, approaches toward the BAMDP problem are often applications of efficient search techniques for MDPs and POMDPs (Wang et al. 2005) (Guez, Silver, and Dayan 2013). Approaches in the continuous case often use heuristics (Bai, Hsu, and Lee 2013) or leverage tools from adaptive control theory (Slade et al. 2017), (Yu, Liu, and Turk 2017). However, while tools from adaptive control are capable of guaranteeing safety while optimizing performance, they typically do not consider the value of information in performance optimization (Aström and Wittenmark 2013).

Robust MDPs (Nilim and El Ghaoui 2005) are a popular tool for planning under model uncertainty, however they have several features which produces undesirable behavior. First, these problems are typically computationally intractable for general uncertainty sets (Bagnell, Ng, and Schneider 2001). Additionally, this framework often leads to extremely conservative policies. Finally, in the BAMDP setting, information gain that simply reweights the probability of elements in an uncertainty set would not change the robust objective, and so the interaction with Bayes-adaptive models is poor. Several of the weaknesses of the robust MDP framework are mitigated by risk-averse approaches to policy optimization (Howard and Matheson 1972), which allow a more naturally tunable notion of conservatism (Chow et al. 2015). However, both the risk-sensitive and robust MDP frameworks do not generally consider that the uncertainty over dynamics models may be reduced as the agent interacts with the environment, still leading to unnecessarily conservative policies that do not consider the value of information gathering actions. We address this problem by introducing notions of risk-sensitivity to BAMDPs.

Discussion & Conclusions

This paper has introduced RAMCP, which marks what we believe to be a first attempt toward efficient methods of computing policies that optimally balance exploration, exploitation, and robustness. This approach has demonstrated good experimental performance on discrete MDPs, typically requiring only minor increases in computation over standard BAMDP solution approaches. There are several clear directions for future work in the effort to bring this tradeoff between safety and performance to more complex domains. First, further analysis into the RAMCP-I formulation of RAMCP are promising, as the algorithm achieves good performance in practice. Second, through the use of value function approximation, the algorithm could be applied to larger, even continuous MDPs. Such approaches have shown good performance for risk-neutral BAMDPs (Guez et al. 2014). Another promising direction for future work is to investigate if the algorithm can be modified to maintain convergence guarantees with more advanced sampling strategies such as UCT (Kocsis and Szepesvári 2006). Finally, this work focused on planning over a discrete distribution over models. Future work may investigate whether sampling strategies such as sequential Monte Carlo can be used to extend RAMCP to continuous beliefs over models parameters.

Acknowledgements Apoorva Sharma was supported in part by the Stanford School of Engineering Fellowship. James Harrison was supported in part by the Stanford Graduate Fellowship and the National Sciences and Engineering Research Council (NSERC). The authors were partially supported by the Office of Naval Research, Young Investigator Program, and by DARPA, Assured Autonomy program.

References

Artzner, P.; Delbaen, F.; Eber, J.-M.; and Heath, D. 1999. Coherent measures of risk. *Mathematical finance*.

Aström, K. J., and Wittenmark, B. 2013. *Adaptive control*. Courier Corporation.

Attarian, A., and Tran, H. 2017. An optimal control approach to structured treatment interruptions for hiv patients: a personalized medicine perspective. *Applied Mathematics*.

Bagnell, J. A.; Ng, A. Y.; and Schneider, J. G. 2001. Solving uncertain markov decision processes. *Neural Information Processing Systems*.

Bai, H.; Hsu, D.; and Lee, W. S. 2013. Planning how to learn. *International Conference on Robotics and Automation*.

Brown, G. W. 1949. Some notes on computation of games solutions. Technical report, RAND Corporation.

Chen, M.; Frazzoli, E.; Hsu, D.; and Lee, W. S. 2016. POMDP-lite for robust robot planning under uncertainty. *International Conference on Robotics and Automation*.

Chow, Y.; Tamar, A.; Mannor, S.; and Pavone, M. 2015. Risk-sensitive and robust decision-making: a CVaR optimization approach. *Neural Information Processing Systems*.

Duff, M. O. 2002. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. Ph.D. Dissertation, University of Massachusetts at Amherst.

Eichhorn, A., and Römisch, W. 2005. Polyhedral risk measures in stochastic programming. *Journal on Optimization*.

Galichet, N.; Sebag, M.; and Teytaud, O. 2013. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. *Asian Conference on Machine Learning*.

Ghavamzadeh, M.; Mannor, S.; Pineau, J.; Tamar, A.; et al. 2015. Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*.

Gittins, J.; Glazebrook, K.; and Weber, R. 2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.

Glover, K., and Doyle, J. C. 1988. State-space formulae for all stabilizing controllers that satisfy an h_∞-norm bound and relations to relations to risk sensitivity. *Systems & control letters*.

Guez, A.; Heess, N.; Silver, D.; and Dayan, P. 2014. Bayes-adaptive simulation-based search with value function approximation. *Neural Information Processing Systems*.

Guez, A.; Silver, D.; and Dayan, P. 2013. Scalable and efficient bayes-adaptive reinforcement learning based on monte-carlo tree search. *Journal of Artificial Intelligence Research*.

Heinrich, J.; Lanctot, M.; and Silver, D. 2015. Fictitious self-play in extensive-form games. *International Conference on Machine Learning*.

Howard, R. A., and Matheson, J. E. 1972. Risk-sensitive markov decision processes. *Management science*.

Hu, C.; Lovejoy, W. S.; and Shafer, S. L. 1996. Comparison of some suboptimal control policies in medical drug therapy. *Operations Research*.

Kocsis, L., and Szepesvári, C. 2006. Bandit based monte-carlo planning. *European Conference on Machine Learning*.

Leslie, D. S., and Collins, E. J. 2006. Generalised weakened fictitious play. *Games and Economic Behavior*.

Majumdar, A., and Pavone, M. 2017. How should a robot assess risk? Towards an axiomatic theory of risk in robotics. *International Symposium on Robotics Research*.

Martin, J. J. 1967. *Bayesian decision problems and Markov chains*. Wiley.

Nilim, A., and El Ghaoui, L. 2005. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*.

Osogami, T. 2012. Robustness and risk-sensitivity in markov decision processes. *Neural Information Processing Systems*.

Rabin, M., and Thaler, R. H. 2001. Anomalies: risk aversion. *Journal of Economic perspectives*.

Sharma, A.; Harrison, J.; Tsao, M.; and Pavone, M. 2019. Robust and adaptive planning under model uncertainty. *arXiv preprint arXiv:1901.02577*.

Slade, P.; Culbertson, P.; Sunberg, Z.; and Kochenderfer, M. J. 2017. Simultaneous active parameter estimation and control using sampling-based Bayesian reinforcement learning. *International Conference on Intelligent Robots and Systems*.

Tamar, A.; Chow, Y.; Ghavamzadeh, M.; and Mannor, S. 2017. Sequential decision making with coherent risk. *Transactions on Automatic Control*.

Wang, T.; Lizotte, D.; Bowling, M.; and Schuurmans, D. 2005. Bayesian sparse sampling for on-line reward optimization. *International Conference on Machine Learning*.

Yu, W.; Liu, C. K.; and Turk, G. 2017. Preparing for the unknown: Learning a universal policy with online system identification. *Robotics: Science and Systems*.