Minimax-Bayes Reinforcement Learning

Thomas Kleine Buening[*] University of Oslo

Christos Dimitrakakis[*] University of Neuchatel

Hannes Eriksson[*] Zenseact

Divya Grover[*] Chalmers University of Technology

Emilio Jorge[*] Chalmers University of Technology

Abstract

While the Bayesian decision-theoretic framework offers an elegant solution to the problem of decision making under uncertainty, one question is how to appropriately select the prior distribution. One idea is to employ a worst-case prior. However, this is not as easy to specify in sequential decision making as in simple statistical estimation problems. This paper studies (sometimes approximate) minimax-Bayes solutions for various reinforcement learning problems to gain insights into the properties of the corresponding priors and policies. We find that while the worst-case prior depends on the setting, the corresponding minimax policies are more robust than those that assume a standard (i.e. uniform) prior.

1 Introduction

Reinforcement learning is the problem of an agent learning how to act in an unknown environment through interaction and reinforcement. In the standard setting, the learning agent acts in an unknown Markov Decision Process μ, within some class of MDPs $\mathcal{M}$. The agent observes the state $s_t \in \mathcal{S}$ of the MDP and selects an action $a_t \in \mathcal{A}$ using a policy π. It then observes a reward $r_t \in \mathbb{R}$ and the next state s_{t+1}. The agent's goal is to maximise utility, defined as the sum of rewards to some horizon T, $u = \sum_{t=1}^{T} r_t$, in expectation, i.e. $\mathbb{E}_\mu^\pi(u)$, where $\mathbb{E}_\mu^\pi$ is the expectation under the MDP and policy. Since the true μ is unknown, this optimisation problem is ill-posed. In the Bayesian setting, this conundrum is solved by selecting some *subjective* prior distribution β over MDPs and maximising $\mathbb{E}_\beta^\pi(u) = \int_{\mathcal{M}} \mathbb{E}_\mu^\pi(u) \, \mathrm{d}\beta(\mu)$. Then it remains to compute the optimal adaptive (i.e. history-dependent) policy, something that can be only done approximately in general, due to the fact that the number of adaptive policies increases exponentially with the problem horizon.

The above discussion assumes that the agent has *somehow* chosen a prior. However, it is not clear how such a prior can be selected from first principles, if we have no domain knowledge, but still want to be robust. The minimax-Bayes idea (Berger, 1985) is to assume that nature selects the *worst* possible prior β^* for the agent, but *without* knowledge of the agent's policy. This can be formalised by having nature play the minimising player in a simultaneous-move zero-sum game defined by the expected utility $\mathbb{E}_\beta^\pi(u)$, where the agent (who maximises) chooses π, and nature (who minimises) chooses β. In simple Bayesian decision problems (e.g. linear regression) the minimax-Bayes problem is well-studied and β^* sometimes corresponds to a maximum entropy prior. However, in an interactive setting, results are limited to one-shot experiment design (Grünwald and Dawid, 2004), which shows that maximum entropy priors are not the worst-case priors generally.

In reinforcement learning, which can be seen as a sequential generalisation of one-shot experiment design, this problem has not received much attention in the past. Sometimes, the concept of maximum entropy has been used in reinforcement learning as a penalty term on the policy (e.g. Todorov, 2006; Haarnoja et al., 2018; Eysenbach and Levine, 2021) as well as in the context of inverse reinforcement learning (Ziebart, 2010), but an explicit connection to the minimax-Bayes literature has not been made. In preliminary work, Androulakis and Dimitrakakis (2014) analysed variants of the weighted majority algorithm for finding minimax priors in a restricted version of this setting.

Contributions. In this paper, we study the basic theoretical and algorithmic properties of minimax-Bayes reinforcement learning. This includes (a) characterising the exis-

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

[*] Authors contributed equally to this work.

tence of solutions under different assumptions on the policy and MDP space (b) defining algorithms, together with convergence guarantees when possible, and (c) performing numerical experiments to illustrate the behaviour of (approximate) minimax-Bayes algorithms and contrast them with Bayesian RL algorithms that assume a standard maximum-entropy (e.g. uniform) prior.

The paper is organised as follows. In Section 2, we formally introduce the setting. In Section 3, we introduce regret definitions and prove some basic properties of the regret as well as relations between Bayesian regret and Bayes-optimal regret. Section 4 discusses the existence of a value for the game between a Bayesian agent and Nature, which selects the prior. Section 5 develops algorithms for finding approximately minimax policies in certain policy classes. In particular, we consider (a) finite-horizon Bayes-optimal policies (b) posterior sampling policies, and (c) parametrised adaptive policies. Our results indicate that, not only is an approximate minimax solution achievable in many settings but that they are much more robust than Bayes-adaptive policies under common priors. Finally, Section 7 contains the related work and conclusions.

2 Setting

A Markov Decision Process (MDP) is a tuple $\mu = \langle \mathcal{S}, \mathcal{A}, P, \rho, T \rangle$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition function, $\rho : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a reward function, and T is a (potentially random) horizon. Let $\mathcal{M}$ denote the space of MDPs.

For simplicity, in our theoretical development, we focus on the setting where the agent is acting in a finite state space $\mathcal{S}$ with a finite set of actions $\mathcal{A}$, the reward function ρ is known, and the horizon T is fixed and finite, although many of our results could be more generally applicable. In each round t, the agent observes state $s_t \in \mathcal{S}$, chooses an action $a_t \in \mathcal{S}$ and receives a reward $r_t = \rho(s_t, a_t)$. We write $s^t = (s_1, \dots, s_t)$ and $a^t = (a_1, \dots, a_t)$ for the sequence of states and actions up to round t. Given the reward function, the history $h_t = (s^t, a^{t-1})$ describes the information available to the agent before choosing an action in round t. The agent's utility u is an additive function of individual rewards $u \triangleq \sum_{t=1}^{T} r_t$. The agent is acting in an MDP through a policy $\pi \in \Pi$, where we let Π denote a generic policy space. For a fixed MDP $\mu \in \mathcal{M}$ and policy $\pi \in \Pi$, the expected utility is given by $U(\pi, \mu) \triangleq \mathbb{E}_\mu^\pi[u]$ with maximal utility denoted by $U^*(\mu) \triangleq \max_{\pi \in \Pi} U(\pi, \mu)$.

When the MDP is unknown, as in the reinforcement learning problem, the policy is adaptive and the agent's actions can depend on what it has been observed in the past, as we explain below.

2.1 Policies.

Let $\mathcal{H}$ be the set of all histories. A (stochastic) policy π is a set of probability measures $\{ \pi(\cdot \mid h) \mid h \in \mathcal{H} \}$ on the set of actions $\mathcal{A}$. We denote the set of all behavioural[1] policies by Π^{S}. A policy is *deterministic* if, for each history $h_t = (s^t, a^{t-1})$, there exists an action $a \in \mathcal{A}$ such that $\pi(a_t = a \mid h_t) = 1$. We denote the set of deterministic policies by Π^{D}. A policy is *memoryless* (or reactive) if, for all histories h_t with $s_t = s$, we have $\pi(a_t = a \mid h_t) = \pi(a_t = a \mid s_t = s)$. We denote the set of memoryless (stochastic) policies by Π_1^{S}. The set of memoryless deterministic policies is denoted by Π_1^{D}. Obviously, $\Pi_1^{\mathrm{D}} \subset \Pi^{\mathrm{D}} \subset \Pi^{\mathrm{S}}$ and $\Pi_1^{\mathrm{D}} \subset \Pi_1^{\mathrm{S}} \subset \Pi^{\mathrm{S}}$. Finally, for any MDP μ there exists a deterministic, memoryless policy that is optimal, i.e. $U^*(\mu) = \max_{\pi \in \Pi} U(\pi, \mu) = \max_{\pi \in \Pi_1^{\mathrm{D}}} U(\pi, \mu)$ (see e.g. Puterman, 2014).

Strategies. Typically, minimax results rely on the notion of mixed strategies. Here, we let $\sigma \in \Delta(\Pi)$ denote a probability measure over a set of base policies Π.

Fact 1. For any strategy $\sigma \in \Delta(\Pi^{\mathrm{D}})$ there exists an equivalent stochastic policy $\pi \in \Pi^{\mathrm{S}}$ such that $\sigma(a_t | h_t) = \pi(a_t \mid h_t)$ for all histories h_t with positive probability.

2.2 Utility and Beliefs

In the following, we overload the $U(\pi, \beta)$ to also mean the expected utility of π with respect to a distribution β over MDPs:

$$U(\pi, \beta) \triangleq \mathbb{E}_\beta^\pi[u] = \int_{\mathcal{M}} U(\pi, \mu) \, \mathrm{d}\beta(\mu), \tag{1}$$

under appropriate measurability assumptions.

There are two possible ways to interpret the distribution β, depending on how it is chosen. If β is chosen by the agent selecting π, it corresponds to the subjective belief of the decision maker about which is the most likely MDP *a priori*. Then, $U(\pi, \beta)$ corresponds to the expected utility of a particular policy under this belief. Let

$$U^*(\beta) \triangleq \max_{\pi \in \Pi} U(\pi, \beta)$$

denote the Bayes-optimal utility for a belief. We recall the fact that this is a convex function (c.f. DeGroot, 1970). By definition, the following bounds hold:

$$U(\pi, \beta) \leq U^*(\beta) \leq \int_{\mathcal{M}} U^*(\mu) \, \mathrm{d}\beta(\mu), \qquad \forall \pi \in \Pi,$$

so that $U^*(\beta)$ is convex with respect to β. In the above, the left-hand side is the utility of an arbitrary policy, while the right side can be seen as the expected utility we would obtain if the true MDP was revealed to us.

[1] That is, history-dependent and stochastic policies.

The second view of β is to assume that the MDP is *actually* drawn randomly from the distribution β. If this is known, then the subjective value of a policy is equal to its true expected value. However, it is more interesting to consider the case where nature arbitrarily selects β from a set of possible priors $\mathcal{B}$. Then we wish to find a policy π^* achieving:

$$\max_{\pi \in \Pi} \min_{\beta \in \mathcal{B}} U(\pi, \beta). \tag{2}$$

A minimax solution exists if the game *has a value*, i.e. $\max_{\pi \in \Pi} \min_{\beta \in \mathcal{B}} U(\pi, \beta) = \min_{\beta \in \mathcal{B}} \max_{\pi \in \Pi} U(\pi, \beta)$. Then there exists a maximin policy π^* which is optimal in response to some minimax belief β^*, and vice versa. A sufficient condition for this to occur is for $U^*(\beta)$ to be convex and differentiable everywhere (c.f. Grünwald and Dawid, 2004). In particular, a maximin *strategy* (i.e. a distribution over policies) can always be found when Π is finite. On the other hand, for any fixed prior β, there is always an optimal deterministic policy. Note that this is only a *best-response* policy and not a solution to the maximin problem (2).

Fact 2. For any distribution β over MDPs, there exists a deterministic, history-dependent policy that is optimal, i.e. $U^*(\beta) = \max_{\pi \in \Pi} U(\pi, \beta) = \max_{\pi \in \Pi^D} U(\pi, \beta)$.

Unfortunately, looking at the problem from the point of view of utility maximisation is somewhat problematic. This is because an unrestricted set of priors for nature may lead to absurd solutions: nature could pick a prior so that all rewards are zero, thus trivially achieving minimal utility. For that reason, we actually focus on the problem of minimax *regret*, i.e. the gap between the agent's policy and that of an oracle. We give the appropriate definitions in the next section.

3 Properties of the regret

We generally write $R(\pi, \mathcal{I})$ to mean the regret of some algorithmic policy π relative to an oracle with information $\mathcal{I}$.

Let us start with the regret of a policy relative to an oracle that knows the underlying MDP:

Definition 1 (Regret)**.** The regret of a policy π for an MDP μ is $R(\pi, \mu) \triangleq U^*(\mu) - U(\pi, \mu)$.

Since this regret notion may be too strong, it is also interesting to define the regret of a policy with respect to the oracle that knows β. This allows us to take into account oracles which have less knowledge than the actual MDP.

Definition 2 (Bayes-optimal Regret)**.** This is the regret of a policy π with respect to the Bayes-optimal policy[2] for β: $R(\pi, \beta) \triangleq U^*(\beta) - U(\pi, \beta) = \int_{\mathcal{M}} \mathrm{d}\beta(\mu)[U(\pi^*(\beta), \mu) - U(\pi, \mu)]$, where $\pi^*(\beta) = \arg\max_\pi U(\pi, \beta)$.

[2]Generally this policy will belong to the set of history-dependent policies, but in some cases, it makes sense to restrict them to e.g. a subset of parametrised policies.

This notion of regret tells us how much we lose relative to a computationally unbounded oracle that knows the prior. We can use it to measure the loss both due to a misspecified prior, by fixing $\pi^*(\beta_0)$ to some prior β_0 and examining $R(\pi^*(\beta_0), \beta)$ as the actual prior β varies, and due to computational approximations, by measuring $R(\pi^*_\epsilon(\beta), \beta)$ for policies calculated with some approximate algorithm.

Finally, we may wish to subjectively calculate our expected regret under an oracle that knows the underlying MDP. Since the agent does not know the underlying MDP, it necessarily measures regret under a Bayesian prior.

Definition 3 (Bayesian regret)**.** The Bayesian regret of a policy π under a prior β is $L(\pi, \beta) \triangleq \mathbb{E}_{\mu \sim \beta}[R(\pi, \mu)] = \sum_\mu \beta(\mu) R(\pi, \mu) = \sum_\mu \beta(\mu)[U^*(\mu) - U(\pi, \mu)]$.

These definitions of regret are closely related, as we shall show in the remainder. It will be illuminating to look at the difference between the regret the agent subjectively expects to suffer with respect to some prior distribution β, relative to the regret of the same policy compared to the Bayes-optimal policy for the same prior.

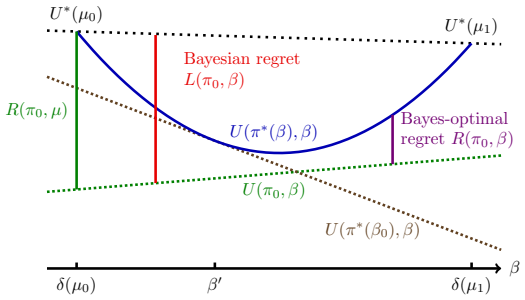


Figure 1: Illustration of the notions of regret for different policies with a belief β over two MDPs μ_1 and μ_2, where $\delta(\mu)$ denotes the Dirac belief on μ. Any *fixed* policy π_0 will have a utility that is a linear function of the belief (green dotted line). The blue curve shows the utility of the Bayes-optimal policy $\pi^*(\beta) = \arg\max_\pi U(\pi, \beta)$. This policy is prior-aware, and hence not fixed, but depends on the prior β. Note that by definition, $U(\pi^*(\beta), \beta)$ is convex. However, if we fix a Bayes-optimal policy for a specific prior β_0, we obtain a tangent $U(\pi^*(\beta_0), \beta)$ to the Bayes-optimal curve at β_0. The Bayesian regret (of π_0) (red line) is the expected regret of a policy compared against an oracle that knows the MDP (black dotted line). The Bayes-optimal regret (of π_0) is the difference in performance to the Bayes-optimal policy (purple line).

Remark 1. *The Bayesian regret of a policy π is greater than the Bayes-optimal regret, i.e. $R(\pi, \beta) \leq L(\pi, \beta)$.*

Proof. Note that $R(\pi, \beta) = \int_{\mathcal{M}} \mathrm{d}\beta(\mu)[U(\pi^*(\beta), \mu) - U(\pi, \mu)] \leq \int_{\mathcal{M}} \mathrm{d}\beta(\mu)[U^*(\mu) - U(\pi, \mu)] = L(\pi, \beta)$, since $U(\pi^*(\beta), \mu) \leq U^*(\mu)$ by definition of $U^*(\mu)$. □

The above also follows from the fact that for any policy π and prior β, the Bayesian regret of π equals the Bayesian regret of the Bayes-optimal policy[3] plus the Bayes-optimal regret of π, that is, $L(\pi, \beta) = L(\pi^*(\beta), \beta) + R(\pi, \beta)$. Geometrically, this follows from the fact that the utility of any fixed policy is lower bounding the convex Bayes-optimal utility curve, as can be seen in Figure 1. The following fact also follows from a simple geometrical argument:

Remark 2. *$R(\pi, \beta)$ is convex in β.*

Proof. By definition of the Bayesian-optimal regret, we have $R(\pi, \beta) = U^*(\beta) - \mathbb{E}_{\mu\sim\beta}[U(\pi, \mu)]$. As $U^*(\beta)$ is convex in β and $\mathbb{E}_{\mu\sim\beta}[U(\pi, \mu)]$ is linear in β, their difference is also convex. □

Of course, the game where nature sees the agent's policy π first before selecting a prior is strictly determined and nature can simply select a single MDP (Dirac distribution) as its best response to π. In this particular case, this follows directly from the convexity of the Bayes-optimal regret.

Following the steps of the proof by Lattimore (2021) for the bandit case, we can show that the maximum regret is attained in Dirac beliefs. Here, we let $\mathcal{B}$ denote the set of beliefs and we work under the assumption that the degenerate beliefs are contained in the belief space.

Lemma 1 (Lattimore (2021))**.** *If for each MDP $\mu \in \mathcal{M}$ there exists an associated Dirac belief $\beta_\mu \in \mathcal{B}$, then for any policy π we have* $\max_{\mu\in\mathcal{M}} R(\pi, \mu) = \max_{\beta\in\mathcal{B}} R(\pi, \beta)$.

This immediately implies that the minimax regret is the same over both beliefs and MDPs:

$$\min_{\pi\in\Pi} \max_{\mu\in\mathcal{M}} R(\pi, \mu) = \min_{\pi\in\Pi} \max_{\beta\in\mathcal{B}} R(\pi, \beta) \tag{3}$$

We find a similar result for the Bayesian regret.

Lemma 2. *If for each MDP $\mu \in \mathcal{M}$ there exists an associated Dirac belief $\beta_\mu \in \mathcal{B}$, then for any π:*

$$\max_{\mu\in\mathcal{M}} R(\pi, \mu) = \max_{\beta\in\mathcal{B}} L(\pi, \beta). \tag{4}$$

Proof. For any $\beta \in \mathcal{B}$, we have

$$\begin{aligned}
\max_{\mu\in\mathcal{M}} R(\pi, \mu) &\geq \max_{\mu\in\mathrm{supp}(\beta)} R(\pi, \mu) \\
&= \max_{\mu\in\mathrm{supp}(\beta)} U(\pi^*(\mu), \mu) - U(\pi, \mu) \\
&\geq \int_{\mathrm{supp}(\beta)} \mathrm{d}\beta(\mu)[U(\pi^*(\mu), \mu) - U(\pi, \mu)] \\
&= L(\pi, \beta).
\end{aligned}$$

Consequently $\max_\mu R(\pi, \mu) \geq \max_\beta L(\pi, \beta)$. Using $\delta(\mathcal{M})$ to denote the set of Dirac beliefs over $\mathcal{M}$, we have: $\max_\beta L(\pi, \beta) \geq \max_{\beta\in\delta(\mathcal{M})} L(\pi, \beta) = \max_{\mu\in\mathcal{M}} R(\pi, \mu)$, due to the fact that $R(\pi, \mu) = L(\pi, \beta_\mu)$ for the singular belief β_μ on MDP μ. As a result, it must hold that $\max_{\mu\in\mathcal{M}} R(\pi, \mu) \geq \max_{\beta\in\mathcal{B}} L(\pi, \beta) \geq \max_{\mu\in\mathcal{M}} R(\pi, \mu)$. □

Lattimore and Szepesvári (2019) show that for the problem of prediction with partial information, the minimax regret equals the minimax Bayesian regret. We show that this also holds in a general setting, as an immediate consequence of Lemma 2.

Corollary 1. *If for each MDP $\mu \in \mathcal{M}$ there exists an associated Dirac belief $\beta_\mu \in \mathcal{B}$, then for any π:*

$$\min_{\pi\in\Pi} \max_{\mu\in\mathcal{M}} R(\pi, \mu) = \min_{\pi\in\Pi} \max_{\beta\in\mathcal{B}} L(\pi, \beta) \tag{5}$$

Equations (3) and (5) can be made intuitive through a simple geometric argument. Due to the linearity of the expected regret with respect to the belief for any fixed policy, the best response for nature always includes singular beliefs.

4 Minimax theorems

The above results merely make precise the intuition that when playing second, nature does not need to randomise: it can simply pick the worst-case MDP for the policy we have chosen. However, we typically want to model a worst-case setting by assuming nature picks its distribution without knowing which policy the decision maker will pick. For that reason, it is important to investigate whether the normal form game against nature, where nature and the agent play without seeing each other's move, has a value. We would expect this to be the case if the regret was a bilinear function of the policy and prior. Consequently, the answer is positive with respect to both the Bayesian regret and the utility in the finite setting. However, this is not the case for the Bayes-optimal regret.

Corollary 2. *For a finite set of MDPs in a finite state-action space, with a known reward function and a finite horizon, the utility and Bayesian regret satisfy:*

$$\min_{\beta\in\mathcal{B}} \max_{\pi\in\Pi} U(\pi, \beta) = \max_{\pi\in\Pi} \min_{\beta\in\mathcal{B}} U(\pi, \beta), \tag{6}$$

$$\max_{\beta\in\mathcal{B}} \min_{\pi\in\Pi} L(\pi, \beta) = \min_{\pi\in\Pi} \max_{\beta\in\mathcal{B}} L(\pi, \beta) \tag{7}$$

Proof. First note that, due to Fact 1, the stochastic policy π can always be written as a distribution σ over deterministic behavioural policies $d \in \Pi^{\mathrm{D}}$ so that $U(\pi, \beta) = \sum_\mu \sum_d \beta(\mu) U(d, \mu)\sigma(d)$. The result follows from the standard minimax theorem. Similarly for regret, we use $L(\pi, \beta) = \sum_\mu \sum_d \beta(\mu) R(d, \mu)\sigma(d)$. □

The same does not hold for the Bayes-optimal regret, since for arbitrary policy spaces the agent's Bayes-optimal policy

[3] This is equal to the difference between the Bayes-optimal value and the upper bound.

has zero Bayes-optimal regret, as it is aware of the prior distribution. However, the minimax value is generally greater than zero.

Lemma 3. *The game $R(\pi, \beta)$ does not have a value when $\mathcal{M}$ contains at least two MDPs μ, μ' whose optimal policy sets have an empty intersection.*

Proof. For $\pi \in \Pi^{\mathrm{D}}$, we have $\max_\beta \min_\pi R(\pi, \beta) = 0$, so that $\min_\pi \max_\beta R(\pi, \beta) \geq \max_\beta \min_\pi R(\pi, \beta) = 0$. From (3), it then follows that $\min_\pi \max_\mu R(\pi, \mu) = \min_\pi \max_\beta R(\pi, \beta) \geq \max_\beta \min_\pi R(\pi, \beta) = 0$. It remains to show that $\min_\pi \max_\mu R(\pi, \mu) > 0$. Assume the contrary. Then there is some policy π^* for which $\max_\mu R(\pi^*, \mu) = 0$. However, there exists at least one μ' whose optimal policy does not coincide with π^*, hence $R(\pi^*, \mu') > 0$. □

Finally, it is interesting to consider the Bayesian regret of the Bayes-optimal policy. For the worst-case Bayesian regret of the Bayes-optimal policy, we find that it is equal to the minimax Bayesian regret.

Lemma 4. *For finite $\mathcal{M}$, the worst-case Bayesian regret of the Bayes-optimal policy equals the minimax Bayesian regret, i.e.*

$$\max_{\beta \in \mathcal{B}} L(\pi^*(\beta), \beta) = \max_{\beta \in \mathcal{B}} \min_{\pi \in \Pi} L(\pi, \beta) = \min_{\pi \in \Pi} \max_{\beta \in \mathcal{B}} L(\pi, \beta).$$

Proof. By definition of the Bayes-optimal policy, we have $U(\pi^*(\beta), \beta) = \max_\pi U(\pi, \beta)$. Thus,

$$\begin{aligned}
\max_\beta L(\pi^*(\beta), \beta) &= \max_\beta \sum_\mu \beta(\mu)[U^*(\mu) - U(\pi^*(\beta), \mu)] \\
&= \max_\beta \min_\pi \sum_\mu \beta(\mu)[U^*(\mu) - U(\pi, \mu)] \\
&= \max_\beta \min_\pi L(\pi, \beta).
\end{aligned}$$

While the above holds for arbitrary $\mathcal{M}$, for the second equality we need to use Corollary 2, which states that the game has a value when $\mathcal{M}$ is finite, so that $\max_\beta \min_\pi L(\pi, \beta) = \min_\pi \max_\beta L(\pi, \beta)$. □

It is important to emphasise that this does not imply that $\pi^*(\beta^*)$ is a minimax policy, but merely that its value at the worst-case belief β^* is equal to the value of the game. As we shall see in Section 6.2, in settings with a finite number of policies, β^* is located at a vertex with at least two best response policies π^*, where the minimax policy must be a mixture between those.

Open questions. This concludes our preliminary discussion of minimax values for Bayesian games on MDPs. While it is clear that standard minimax theorems apply in the discrete case when we consider stochastic policies, it is an open question whether those can be extended to a more general setting. In particular, do the utility and Bayesian regret game have a value with an uncountable family of priors such as the Dirichlet-product prior? It is also an open question whether a value for the game exists when we are restricted to deterministic policies in some cases. We conjecture that this is generally not the case. For example in discrete, finite horizon problems, the set of policies pure deterministic policies is finite, and so it is unlikely that one of them is maximin. We explore these questions experimentally, after we first develop some algorithms in the following section.

5 Algorithms

In this section, we attempt to answer some of the above questions empirically. In particular, does there exist an equilibrium for bandit problems, where the Bayes-optimal policy can be efficiently approximated through Gittins indices? What about settings where we must restrict the policy space to parametrised or tree policies? Does solving the minimax problem approximately lead to robust policies? Are the worst-case priors we obtain through optimisation actually preferable in some way to standard priors such as the uniform one? For example, do they lead to more robust policies?

For the infinite horizon case, we cannot consider the Bayes-optimal regret, as it requires us to compute the Bayes-optimal policy. However, we can always target the Bayesian regret, which is an upper bound on the Bayes-optimal regret. (And since the former is usually the same as the minimax regret, it gives us a minimax policy).

Section 5.1 describes a stochastic gradient descent-ascent algorithm for finding an approximate minimax regret pair. For the finite horizon case, we can obtain the Bayes-optimal response to any prior distribution. More specifically, when the set of possible MDPs is finite, and we have an optimal policy oracle, we can employ a cutting plane algorithm, described in Section 5.2. This allows us to obtain the set of all best response policies to the worst-case prior, and hence the minimax policy.

5.1 Gradient descent ascent

We want to calculate the minimax pair (π^*, β^*) for the Bayesian regret. This can be done through gradient descent-ascent (GDA) (Lin et al., 2020), which alternates performing a gradient step for the prior and performing a gradient step for the policy. We show convergence guarantees for GDA in the finite MDP setting, for certain parametrisations of the policy. To calculate the minimax solution for the Bayesian regret, we need the gradient with

respect to the policy and the prior.

$$\nabla_\pi L(\pi,\beta) = -\int_{\mathcal{M}} d\beta(\mu)\nabla_\pi U(\pi,\mu) \tag{8}$$

$$\nabla_\beta L(\pi,\beta) = \int_{\mathcal{M}} R(\pi,\mu)\nabla_\beta d\beta(\mu). \tag{9}$$

Intuitively, Algorithm 1 works as follows: First, we sample M MDPs from the current prior β_{t-1}. We use those to do a policy gradient step and obtain a new policy π_t using standard policy gradient algorithms, as well as a gradient step in the prior space to obtain a new prior β_t. Since each gradient may not be exact, we use $G_\pi(\pi,\beta)$ and $G_\beta(\pi,\beta)$ to denote the approximate gradient with respect to the policy and prior respectively. Appendix A describes how we obtain those in detail. Since gradient steps may lead us outside the feasible prior space $\mathcal{B}$, we use a projection $\mathcal{P}_\mathcal{B}$ to ensure we have a valid prior distribution. Finally, we return a randomly selected policy-prior pair from the ones generated during the algorithm's run.

Algorithm 1 Stochastic GDA

Input policy π_0, belief β_0, learning rates (η_π, η_β) and stochastic gradient estimators G_π, G_β for $\nabla_\pi L, \nabla_\beta L$
for $t = 1, \dots, T$ **do**
  Get directions $g_\beta = \frac{1}{M}\sum_i G_\beta^{(i)}(\pi_{t-1},\beta_{t-1})$ and $g_\pi = \frac{1}{M}\sum_i G_\pi^{(i)}(\pi_{t-1},\beta_{t-1})$ using M i.i.d samples
  $\pi_t \leftarrow \pi_{t-1} - \eta_\pi g_\pi$
  $\beta_t \leftarrow \mathcal{P}_\mathcal{B}\Big(\beta_{t-1} + \eta_\beta g_\beta\Big)$
end for
Output β^*, π^* uniformly at random from $\{(\beta_1,\pi_1),\dots,(\beta_T,\pi_T)\}$

5.1.1 Convergence guarantees for finite set of MDPs

In the MDP setting with n MDPs, we have $\mathcal{B}$ as the probability simplex which has the diameter $D = \sqrt{2}$. Additionally, the gradient

$$\nabla_\beta L(\pi,\beta) = \sum_i^n R(\pi,\mu_i)\nabla_\beta P(\mu_i|\beta) \tag{10}$$

$$\nabla_{\beta_i} L(\pi,\beta) = R(\pi,\mu_i) \tag{11}$$

is constant and therefore convex.

Lemma 5. *If the policy π is parameterised as a softmax over actions, independently for each h_t and the horizon T is fixed. Then $L(\pi,\beta)$ is $T^2(|\mathcal{A}|+1)$-smooth and $L(\cdot,\beta)$ is T^2-Lipschitz*

With these properties, and a batch size $M = 1$, the requirements of Theorem 4.9 of Lin et al. (2020) are fulfilled and Algorithm 1 will find a $\epsilon-$stationary point in terms of Moreau envelopes, given appropriate step sizes, with an iteration complexity of

$$\mathcal{O}\left(|\mathcal{A}|^3 T^6 \left(\frac{\left(T^4+\sigma^2\right)\widehat{\Delta}_\Phi}{\epsilon^6} + \frac{\widehat{\Delta}_0}{\epsilon^4}\right)\max\left\{1,\frac{\sigma^2}{\epsilon^2}\right\}\right), \tag{12}$$

as long as $\mathbb{E}_G\left[\|G(\pi,\beta) - \nabla L(\pi,\beta)\|^2\right] \le \sigma^2$. Note that no guarantees exist for general non-convex non-concave Bayesian regret L, as is the case for Dirichlet beliefs and parametric policies.

Here the stationarity is defined as $||\nabla\Phi_{1/2l}(\pi)||_2 \le \epsilon$ as in Lin et al. (2020). We have $\Phi(\cdot) = \max_{\beta\in\mathcal{B}} L(\cdot,\beta)$ and $\Phi_\lambda(\pi) = \min_{w\in\Pi}\Phi(w) + (1/2\lambda)||\omega-\pi||_2^2$ is the Moreau envelope of Φ. Finally we obtain $\widehat{\Delta}_\Phi = \Phi_{1/2l}(\pi_0) - \min_\pi \Phi_{1/2l}(\pi)$ and $\widehat{\Delta}_0 = \Phi(\pi_0) - L(\pi_0,\beta_0)$.

5.2 Cutting planes

In this section we demonstrate an efficient method for localising the minimax pair (π^*,β^*) for beliefs over a finite set of MDPs, given that an oracle for the Bayes-optimal policy for a given belief is available. This could for example be obtained in finite horizon tasks with a sufficiently small horizon such that a tree-policy is tractable. An example of this can be found in (Duff, 2002, Section 1.5).

We use the approximate centroid cutting plane algorithm from Bertsimas and Vempala (2004), which can be seen as a high dimensional extension of the bisection algorithm. The goal here is to find a way to repeatedly obtain a plane where we can reject one side of the half-plane, quickly shrinking the plausible set of beliefs. Each policy π has a corresponding regret plane[4] $L(\pi,\beta)$ over β. Since $L(\pi^*(\beta),\beta)) \le \max_{\beta\in\mathcal{B}} L(\pi^*(\beta),\beta))$, any $\beta : L(\pi^*(\beta'),\beta')) > L(\pi^*(\beta'),\beta))$ can not be the minimax β and can be discarded. This is the same as discarding the half-plane given by the descent direction of the Bayesian regret plane. An illustration of this principle in two dimensions can be found in Figure 2.

Selecting a new approximate centroid as the next β to query guarantees fast convergence in the volume of the plausible set of beliefs given the following lemma.

Lemma 6 (Lemma 5 Bertsimas and Vempala (2004))**.** *Each cut in Algorithm 2 will reduce the volume of the set K_t by at least 1/3 with high probability.*

The full procedure is described in Algorithm 2. Here β_t is the approximate centroid (through one of the methods in Bertsimas and Vempala (2004), such as hit-and-run sampling) of the set K_t. K_t contains the plausible beliefs that could be the minimax belief, at step t of the algorithm. The cut is given by C_t which is the normal to the Bayes regret

[4] Due to the Bayesian regret being an expectation over MDPs and hence is linear.

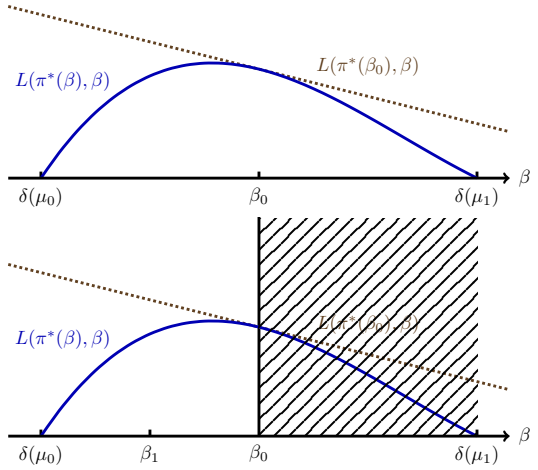


Figure 2: Illustration of cutting plane algorithm for two dimensions. The top image illustrates the Bayesian regret plane obtained for queried belief β_0 while the bottom image shows how the cut obtained by the plane discards the right side of the belief space and a new queried belief β_1 is obtained.

plane at β_t where each element $C_t^{(i)} = R(\pi^*(\beta_t), \beta = \delta_{\mu_i})$.

Algorithm 2 Cutting plane algorithm for finding minimax belief

Input: Initial belief set of constraints K_0, Optimal Policy oracle, Policy evaluation oracle, $t = 0$;
for $t \in 0, \dots, T-1$ **do**
  Obtain $\beta_t \approx \mathbb{E}_{K_t}[x]$
  Obtain optimal policy $\pi^*_{\beta_t}$ and $C_t^{(i)} = R(\pi^*(\beta_t), \beta = \delta_{\mu_i})$.
  $K_{t+1} = K_t \cap \{\beta : C_t^T(\beta - \beta_t) > 0\}$
end for
Return $\beta^* \in K_T$ that has $\frac{\text{VOL}(K_T)}{\text{VOL}(K_0)} < \left(\frac{2}{3}\right)^T$ with high probability and corresponding $\pi^*(\beta^*)$.

This method is also applicable when the policy space is a set of ϵ-optimal policies $\Pi^\epsilon \subset \Pi$, i.e. such that $\max_{\pi\in\Pi^\epsilon} U(\pi,\beta) \geq \max_{\pi\in\Pi} U(\pi,\beta) - \epsilon$ for any $\beta \in \mathcal{B}$. It is natural to look at such a policy space, because policies obtained through look-ahead tree search or neural network may be adaptive, but they can only be ϵ-optimal in general.

Lemma 7. *If* $\max_{\pi\in\Pi^\epsilon} L(\pi,\beta) \leq \max_{\pi\in\Pi} L(\pi,\beta) + \epsilon$ *for all* $\beta \in \mathcal{B}$ *then*

$$\min_{\pi\in\Pi} L(\pi, \beta^{\epsilon,*}) \geq \max_{\beta\in\mathcal{B}} \min_{\pi\in\Pi} L(\pi,\beta) - \epsilon \tag{13}$$

where $\beta^{\epsilon,*} = \arg\max_{\beta\in\mathcal{B}} \min_{\pi\in\Pi^\epsilon} L(\pi,\beta)$.

Additionally, if $\min_{\pi\in\Pi} L(\pi,\beta)$ *is c-concave in* β *then* $||\beta^{\epsilon,*} - \beta^*||_2 < \sqrt{\epsilon/c}$.

A proof is provided in the appendix.

6 Experiments

We perform three experiments to see how minimax priors differ from common uniform priors, and examine the relative robustness of the corresponding policies. The first characterises worst-case priors for Bernoulli bandits. The second experiment is on finite MDP sets with a finite horizon. Here we verify the feasibility of the cutting plane algorithm for finding minimax solutions. We also illustrate the regret of posterior sampling. The final experiment is for the general case of discrete MDPs and parametric adaptive policies, where a value may not exist.[5]

6.1 Illustrations of Worst-Case Priors for Bernoulli Bandits

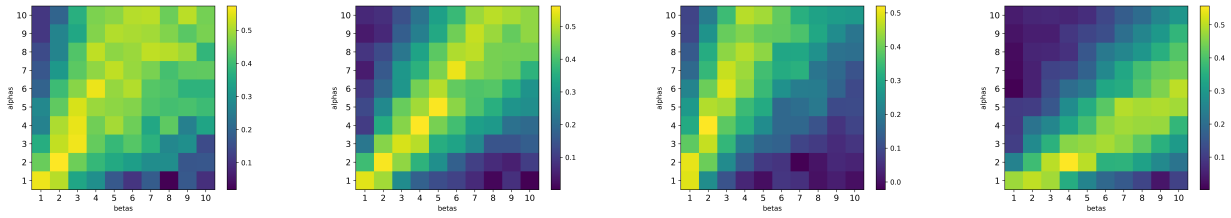
We are interested in analysing the worst-case priors when the Bayesian agent is responding to nature's prior with a Bayes-optimal policy. In general, computing the Bayes-optimal policy is intractable. However, for Bernoulli bandits with infinite horizon and geometrically discounted rewards, so that the utility is defined as $u = \sum_t \gamma^t r_t$, Gittins (Gittins, 1979; Gittins et al., 2011) showed that an index policy, the so-called Gittins index, yields a Bayes-optimal policy.

For K-armed Bernoulli bandits $\theta = (\theta_1, \dots, \theta_K)$ with $\theta_k \in [0,1]$, we then consider Beta product priors such that $\beta(\theta) = \prod_{k=1}^K \mathcal{B}eta(a_k, b_k)\{\theta_k\}$. To illustrate how the Bayes-expected regret of the Bayes-optimal policy changes with respect to the prior, we consider a two-armed Bernoulli bandit, where the first arm's prior is fixed to some distribution $\mathcal{B}eta(a_1, b_1)$ and the second arm's prior $\mathcal{B}eta(a_2, b_2)$ is set to different values. Figure 3 shows the Bayesian regret for different fixed priors for arm 1 and varying prior for arm 2.

We observe that high Bayesian regret is typically suffered when the second prior's mean approximately matches the mean of the first arm's prior, i.e. $\mathbb{E}[\mathcal{B}eta(a_1, b_1)] = \mathbb{E}[\mathcal{B}eta(a_2, b_2)]$. Moreover, it seems that maximal Bayesian regret is achieved at a completely symmetric prior, i.e. $\mathcal{B}eta(a_1, b_1) = \mathcal{B}eta(a_2, b_2)$, irrespective of how the first arm's prior is chosen. More generally, we can observe that lower values of a and b yield higher Bayesian regret, making the intuition precise that the Bayes-optimal policy suffers higher Bayesian regret when the prior provides less information. Based on this, a worst-case prior can be conjectured to make arms maximally indistinguishable a priori; as one may expect.

We also allowed all priors to vary to discover the actual worst-case prior. We found this depends heavily on the discount factor γ and the number of arms K. For $K = 2$ and $\gamma = 0.9$, we found it is approximately $\mathcal{B}eta(0.8, 0.8)$ for

[5]The code is made available at `https://github.com/minimaxBRL/minimax-bayes-rl`.



(a) $\mathcal{B}eta(1,1)$ (b) $\mathcal{B}eta(3,3)$ (c) $\mathcal{B}eta(4,2)$ (d) $\mathcal{B}eta(2,4)$

Figure 3: The Bayesian regret of the Bayes-optimal policy in two-armed Bernoulli bandits, where the first arm's prior is fixed. The x- and y-axis denote the parameters of the second arm's prior.

both arms. In general, the worst-case prior is symmetric with parameters increasing in the number of arms and the discount factor, i.e. moving towards short-tailed priors.

6.2 Finite Set of MDPs

In this section, we study the properties of minimax problems where we have a belief over a finite set of MDPs. The transition matrix is randomly sampled from an exponential distribution before being normalised. The agent starts in state 1, and the reward is 1 for taking the first action in state N, and zero elsewhere. We use a finite horizon $T = 5$ to allow exact computation of the optimal policies and Bayesian regret. Additionally we use $\gamma = 1$.

Figure 4 show the Bayesian regret for a two-MDP task. This helps us visualise that the Bayes-optimal value is a piecewise linear function consisting of the minimum over locally optimal policies. We also compare with the Bayesian regret of the PSRL policy (Strens, 2000), which for every episode acts optimally with respect to a sampled MDP from the belief. The quadratic curve for PSRL is due to the fact that we allow the policy to change with the belief.

In additional experiments in Appendix C, we study the Bayesian regret landscape for a three MDP setup (see Figure 6). We also compare the worst case Bayesian regret of the minimax policy and of the Bayes optimal policy for the uniform belief for a few different setups with 16 different MDPs in Table 1 and can see that the minimax policy significantly outperforms the uniform best response policy.

6.3 Infinite Set of MDPs

In the following experiments, we study priors over an infinite space of MDPs. The main prior of interest is Dirichlet product-priors. We use the minimax policy gradient algorithm to simultaneously update the parameters of the belief β and the parameters of the policy π. We choose a history-dependent policy parametrisation using a softmax rule. In these experiments we study MDPs with 5 states and two actions. Further, we consider problems with horizon $T = 1000$.

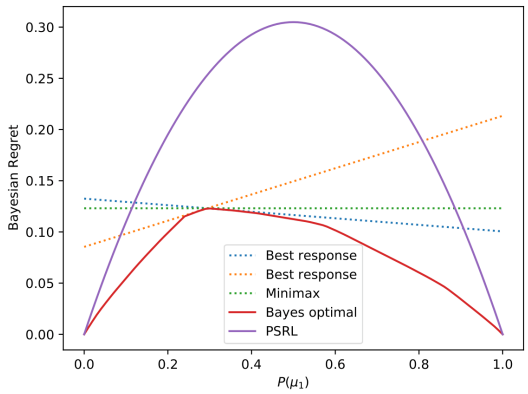


Figure 4: This figure shows the Bayesian regret of different policies. The dashed lines show the value of three adaptive policies optimal for the maximin-regret prior. Two of them are best responses, which are also optimal on either side of the maximin point. The minimax-regret policy is shown in green, and it has a uniform regret no matter what the actual prior is. The solid lines show policies which have knowledge of the MDP prior: the Bayes-optimal policy and the best PSRL policy for that specific prior. Their dependency on the prior makes their regret a concave function.

In Figure 5 we investigate the performance of the minimax policy π^* compared to the baseline *best response* adaptive policies, $\pi^*(\beta^1), \pi^*(\beta^*)$, to the uniform prior β^1 and the maximin prior β^*, respectively. The three policies are evaluated on six different priors. These are, the uniform prior β^1, the maximin prior β^*, two priors interpolated between the uniform and the maximin prior, a uniform prior over deterministic MDPs β^D and a delta distribution over the parameters of the Chain environment (Strens, 2000), β^{Chain}.

In this setting we can only expect to find approximate minimax solutions. Thus, there is no guarantee the obtained minimax solution is globally robust to changes in belief. However, in Figure 5 we observe the minimax policy π^* to be the most robust taking all priors into account.

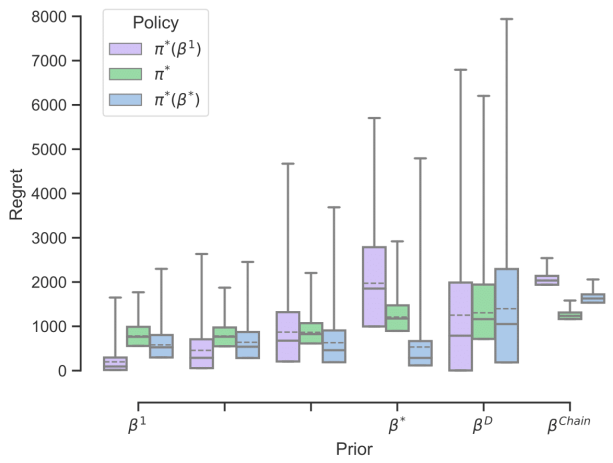


Figure 5: β^D is approximately uniform over deterministic MDPs. β^{Chain} is a delta distribution over the Chain MDP. The MDPs in between β^1 (Uniform) and β^* (Maximin) are interpolated. The mean is depicted with a dashed line, the solid line is the median and the upper whisker is the 99.9% percentile.

7 Discussion and Conclusion

Related work We studied the problem of minimax-Bayes reinforcement learning. Although minimax-Bayes problems are well-known in statistical inference (c.f. Berger, 1985), they have received little attention in sequential problems. Older work such as Arrow et al. (1949) is interested in minimax and Bayes optimal solutions to decision making tasks but without combining them. Similarly, Hodges Jr and Lehmann (1952) relaxes the property of minimax risk to restricted Bayes solutions where the maximal risk is bounded while also changing the objective to an interpolation between the expected and maximal risk. While this is work in the same spirit as ours it is fundamentally different. Grünwald and Dawid (2004) studied the problem of one-shot experiment design prior to estimation. In the partial monitoring setting, Lattimore and Szepesvári (2019) made connections between the Bayesian minimax regret and the minimax regret.

There have been a variety of work interested in using meta learning to create Bayes-(adaptive) optimal agents such as Hochreiter et al. (2001); Wang et al. (2016); Mikulik et al. (2020); Zintgraf et al. (2021). They use recurrent neural networks to encode an episode's history so as to adapt optimally in a new episode in a new MDP. As they are interested in optimising for specific MDP distribution, β is considered fixed and they solve $\max_\pi \mathbb{E}_{\mu\sim\beta} U(\mu, \pi)$ without studying β's impact on the utility or regret.

Work on Bayesian robust reinforcement learning (Derman et al., 2020; Petrik and Russel, 2019) is related in the manner that they search for policies that are robust against interference from nature. The difference is that they wish to find policies that are good against the worst MDP from the set of MDPs that are plausible with respect to a specific posterior, rather than against an adversarial prior.

Conclusion In this work we study the computation of minimax-Bayes policies, which have not been previously considered. We also include conditions for when the solutions can be guaranteed to be found efficiently. Experimentally we find that these policies not only appear to be feasible, but also that such policies can be significantly more robust than those based on standard uninformative priors. Finally, we make exposition of many important properties of minimax-Bayes solutions for reinforcement learning to make a basis for future work in this area.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, the Swedish research council grant on "Learning, Privacy and the Limits of Computation" and the Norwegian research council grant on "Algorithms and Models for Socially Beneficial AI". We are grateful for their support. Many thanks to Emmanouel Androulakis, whose Master thesis developed MWA algorithms for this problem, and to Tor Lattimore for discussions about minimax properties in the Bayesian setting.

References

E. G. Androulakis and C. Dimitrakakis. Generalised entropy mdps and minimax regret. *arXiv preprint arXiv:1412.3276*, 2014.

K. J. Arrow, D. Blackwell, and M. A. Girshick. Bayes and minimax solutions of sequential decision problems. *Econometrica, Journal of the Econometric Society*, pages 213–244, 1949.

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

D. Bertsimas and S. Vempala. Solving convex programs by random walks. *J. ACM*, 51(4):540–556, jul 2004. ISSN 0004-5411. doi: 10.1145/1008731.1008733. URL https://doi.org/10.1145/1008731.1008733.

S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

R. Dearden, N. Friedman, and S. Russell. Bayesian q-learning. *Aaai/iaai*, 1998:761–768, 1998.

M. H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, 1970.

E. Derman, D. Mankowitz, T. Mann, and S. Mannor. A bayesian approach to robust reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 648–658. PMLR, 2020.

C. Dimitrakakis and R. Ortner. *Decision making under uncertainty and reinforcement learning*. Springer, 2022.

M. O. Duff. *Optimal Learning Computational Procedures for Bayes-adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts at Amherst, 2002.

B. Eysenbach and S. Levine. Maximum entropy rl (provably) solves some robust rl problems. *arXiv preprint arXiv:2103.06257*, 2021.

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

J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

B. Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific Journal of Mathematics*, 10(4):1257–1261, 1960.

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, 2004.

T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

S. Hochreiter, A. S. Younger, and P. R. Conwell. Learning to learn using gradient descent. In *International conference on artificial neural networks*, pages 87–94. Springer, 2001.

J. L. Hodges Jr and E. L. Lehmann. The use of previous experience in reaching statistical decisions. *The Annals of Mathematical Statistics*, pages 396–407, 1952.

T. Lattimore. Personal Communication, March 2021.

T. Lattimore and C. Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pages 2111–2139. PMLR, 2019.

T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

V. Mikulik, G. Delétang, T. McGrath, T. Genewein, M. Martic, S. Legg, and P. Ortega. Meta-trained agents implement bayes-optimal agents. *Advances in neural information processing systems*, 33:18691–18703, 2020.

M. Petrik and R. H. Russel. Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps. *Advances in neural information processing systems*, 32, 2019.

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

M. Strens. A Bayesian framework for reinforcement learning. In *ICML 2000*, pages 943–950, 2000.

E. Todorov. Linearly-solvable markov decision problems. *Advances in neural information processing systems*, 19, 2006.

J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

B. D. Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

L. Zintgraf, S. Schulze, C. Lu, L. Feng, M. Igl, K. Shiarlis, Y. Gal, K. Hofmann, and S. Whiteson. Varibad: variational bayes-adaptive deep rl via meta-learning. *The Journal of Machine Learning Research*, 22(1):13198–13236, 2021.

A Gradient calculations.

For solving the minimax problem either for the expected utility or the expected regret, we need to calculate the appropriate gradient for both the policy and the prior. The gradients for the expected utility are as follows:

$$\nabla_\pi U(\pi, \beta) = \int_{\mathcal{M}} d\beta(\mu) \nabla_\pi U(\pi, \mu), \qquad \nabla_\beta U(\pi, \beta) = \int_{\mathcal{M}} U(\pi, \mu) \nabla_\beta d\beta(\mu) \tag{14}$$

The Bayesian regret gradient is similarly obtained:

$$\nabla_\pi L(\pi, \beta) = -\int_{\mathcal{M}} d\beta(\mu) \nabla_\pi R(\pi, \mu) \qquad \nabla_\beta L(\pi, \beta) = \int_{\mathcal{M}} R(\pi, \mu) \nabla_\beta d\beta(\mu). \tag{15}$$

Since in the minimax regret scenario, the agent is minimising rather than maximising, the policy update is identical. However, the prior gradient is scaled with respect to the regret rather than the utility. Let us now look at how to calculate those gradients in more detail.

A.1 Policy gradient

Here we look at two classes of policies. The first occurs when there is a finite number of bases (possibly stochastic and behavioural) policies from which the agent chooses one randomly. The second is a class of parametrised stochastic behavioural policies.

Finite policy distributions. For a strategy $\sigma = (\sigma_1, \dots, \sigma_n)$ over a finite set of n policies $\Pi \subset \Pi^{\mathrm{S}}$, we can write

$$U(\sigma, \beta) = \sum_{\pi, \mu} \sigma(\pi) U(\pi, \mu) \beta(\mu). \tag{16}$$

We then obtain

$$\frac{\partial}{\partial \sigma_i} U(\sigma, \beta) = \sum_{\mu} U(\pi_i, \mu) \beta(\mu) \tag{17}$$

We do not use this setting in practice in the paper, but it is an interesting special case.

Stochastic policies. Stochastic policies π in a parametrised policy space $\Pi_W \subset \Pi^{\mathrm{S}}$ can be an arbitrary neural network policy. For a finite set of MDPs, the gradient is:

$$\nabla_\pi U(\pi, \beta) = \sum_{\mu} \nabla_\pi U(\pi, \mu) \beta(\mu). \tag{18}$$

For an infinite set of MDPs, we have

$$\nabla_\pi U(\pi, \beta) = \int_{\mathcal{M}} \nabla_\pi U(\pi, \mu) \, \mathrm{d}\beta(\mu) \approx \frac{1}{M} \sum_{k=1}^{M} \nabla_\pi U(\pi, \mu^{(k)}), \; \mu_k \sim \beta(\mu) \tag{19}$$

So it is only necessary to compute

$$\begin{aligned} \nabla_\pi U(\pi, \mu) &= \sum_h U(h) \, \mathbb{P}_\mu^\pi(h) \sum_t \nabla_\pi \ln \pi(a_t \mid h_t) \\ &= \sum_h U(h) \, \mathbb{P}_\mu^\pi(h) \sum_t \frac{\nabla_\pi \pi(a_t \mid h_t)}{\pi(a_t \mid h_t)}, \end{aligned}$$

where for a given history $h = (s_1, r_1, a_1, \dots, s_T, r_T)$, $h_t = (s_1, r_1, a_1, \dots, s_t, r_t)$. It remains to compute $\nabla_\pi \pi(a_t \mid h_t)$, which can be done automatically using auto-grad software.

However, one particular case is when the policy is parametrised with $\mathbf{w}_a = (w_{a,i})_{i=1}^n$ vectors combined with a statistic $\phi : \mathcal{H} \to \mathbb{R}_+^n$ so that

$$\pi(a_t = a \mid h_t) = \frac{\mathbf{w}_a^\top \phi(h_t)}{\sum_b \mathbf{w}_b^\top \phi(h_t)} = \frac{\sum_i w_{a,i} \phi_i(h_t)}{\sum_b \sum_i w_{b,i} \phi_i(h_t)} \tag{20}$$

$$\frac{\partial}{\partial w_{a,i}}\pi_{\mathbf{w}}(a_t = a \mid h_t) = \frac{\phi_i(h_t)[\sum_{(b,j)\neq(a,i)} w_{b,j}\phi_j(h_t)]}{[\sum_b \sum_j w_{b,j}\phi_j(h_t)]^2}, \qquad \frac{\partial}{\partial w_{b,i}}\pi_{\mathbf{w}}(a_t = a \mid h_t) = -\frac{\phi_i(h_t)\sum_j w_{a,j}\phi_j(h_t)}{[\sum_b \sum_j w_{b,j}\phi_j(h_t)]^2}. \tag{21}$$

With a feature representation $\phi : \mathcal{H} \times \mathcal{A} \to \mathbb{R}^n$ and a softmax policy then

$$\pi(a_t \mid h_t) = \frac{e^{\mathbf{w}^\top \phi(h_t,a_t)}}{\sum_b e^{\mathbf{w}^\top \phi(h_t,b)}}, \qquad \nabla_{\mathbf{w}} \ln \pi(a_t \mid h_t) = \phi(h_t, a_t) - \sum_{a\in\mathcal{A}} \pi(a_t = a \mid h_t)\phi(h_t, a). \tag{22}$$

For the case where $\phi(h_t, a)$ simply partitions the history, so that $\mathbf{w}^\top \phi(h, a) = w_{h,a}$, the above becomes

$$\frac{\partial}{\partial w_{h,a}} \ln \pi(a_t \mid h_t) = \begin{cases} 1 - \pi(a|h), & a_t = a, h_t = h \\ -\pi(a|h), & a_t \neq a, h_t = h \\ 0, & h_t \neq h \end{cases} \tag{23}$$

A.2 Prior gradient.

The steps above were all standard policy gradient steps, which can be implemented with sampled MDPs from the current prior. However, we also need to update the prior distribution with a gradient step. Here we distinguish two cases: a belief over a finite number of MPDs and a Dirichlet belief.

Finite $\mathcal{M}$. Now let us represent the belief as a finite-dimensional vector $\beta = (\beta_i)$ on the simplex. The partial derivative is then:

$$\frac{\partial}{\partial \beta_i} U(\pi, \beta) = \sum_j U(\pi, \mu_j) \frac{\partial}{\partial \beta_i} \beta_j = U(\pi, \mu_j) \tag{24}$$

Dirichlet $\mathcal{M}$. Let us first consider the general case of an infinite MDP space. Then we can approximate the gradient of the expected utility through sampling:

$$\nabla_\beta U(\pi, \beta) = \int_{\mathcal{M}} U(\pi, \mu) \nabla_\beta \ln[\beta(\mu)] d\beta(\mu) \approx \frac{1}{M} \sum_{k=1}^{M} U(\pi, \mu^{(k)}) \nabla_\beta \ln[\beta(\mu^{(k)})], \tag{25}$$

where $\mu^{(k)} \sim \beta$ are samples from the current prior.

For discrete state-action MDPs for a certain number of states and actions, we can use a Dirichlet-product distribution. This means that for each state-action's (s, a) transition distribution, we define a separate Dirichlet distribution $\beta(\mu_{s,a})$ with parameter vector $\alpha_{s,a} \in \mathbb{R}_+^{|S|}$:

$$\beta(\mu) = \prod_{(s,a)} \beta(\mu_{s,a}), \qquad \beta(\mu_{s,a}) = \frac{1}{B(\alpha_{s,a})} \prod_i \mu_{s,a,i}^{\alpha_{s,a,i}-1} \tag{26}$$

where $\mu_{s,a,i} = \mathbb{P}(s_{t+1} = i | s_t = s, a_t = a)$. For the sequel, it is notationally convenient to ignore the s, a subscript and focus only on the next state distribution i

$$\begin{aligned} \frac{\partial}{\partial \alpha_j} \ln \beta(\mu) &= \frac{\partial}{\partial \alpha_j} \ln \left\{ \frac{1}{B(\alpha)} \prod_i \mu_i^{\alpha_i - 1} \right\} \\ &= \frac{\partial}{\partial \alpha_j} \left\{ \ln \frac{1}{B(\alpha)} + \sum_i (\alpha_i - 1) \ln \mu_i \right\} \\ &= \frac{\partial}{\partial \alpha_j} \ln \frac{1}{B(\alpha)} + \ln \mu_j \end{aligned}$$

Note that

$$\begin{aligned} \ln 1/B(\alpha) &= \ln \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \\ &= \ln \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i) \end{aligned}$$

So that

$$\begin{aligned}\frac{\partial}{\partial \alpha_j} \ln 1/B(\alpha) &= \frac{\partial}{\partial \alpha_j} \ln \Gamma(\sum_i \alpha_i) - \frac{\partial}{\partial \alpha_j} \ln \Gamma(\alpha_j) \\ &= \frac{1}{\Gamma(\sum_i \alpha_i)} \frac{\partial}{\partial \alpha_j} \Gamma(\sum_i \alpha_i) - \frac{1}{\Gamma(\alpha_j)} \frac{\partial}{\partial \alpha_j} \Gamma(\alpha_j) \\ &= \psi(\sum_i \alpha_i) - \psi(\alpha_j)\end{aligned}$$

where ψ is the digamma function.

This means that the overall derivative is

$$\begin{aligned}\frac{\partial}{\partial \mu_{s,a,i}} \ln \beta(\mu) &= \frac{\partial}{\partial \mu_{s,a,i}} \ln \prod_{(s',a')} \beta(\mu_{s',a'}) \\ &= \frac{\partial}{\partial \mu_{s,a,i}} \sum_{s',a'} \ln \beta(\mu_{s',a'}) \\ &= \frac{\partial}{\partial \mu_{s,a,i}} \ln \beta(\mu_{s,a}) \\ &= \psi(\sum_j \alpha_{s,a,j}) - \psi(\alpha_{s,a,i}) + \ln(\mu_{s,a,i})\end{aligned}$$

Combinging the above, we get

$$\alpha^{(k)}_{s,a,i} = \alpha^{(k-1)}_{s,a,i} - \delta^{(k)} U(\pi, \mu^{(k)}) \left[\psi(\sum_j \alpha_{s,a,j}) - \psi(\alpha_{s,a,i}) + \ln(\mu^{(k)}_{s,a,i}) \right], \tag{27}$$

where $\delta^{(k)}$ is the step-size.

Reward prior. We can derive a similar update for Beta-distributed rewards, with

$$\alpha^{(k)}_s = \alpha^{(k-1)}_s - \delta^{(k)} U(\pi, \mu^{(k)}) \left[\psi(\alpha_s + \beta_s) - \psi(\alpha_s) + \ln(\rho^{(k)}_s) \right] \tag{28}$$

$$\beta^{(k)}_s = \beta^{(k-1)}_s - \delta^{(k)} U(\pi, \mu^{(k)}) \left[\psi(\alpha_s + \beta_s) - \psi(\beta_s) + \ln(1 - \rho^{(k)}_s) \right]. \tag{29}$$

We can also define the Beta-distribution with alternate parametrisation: $p_s = \alpha_s/(\alpha_s + \beta_s)$, $n_s = \alpha_s + \beta_s$ which implies $\alpha_s = p_s n_s, \beta_s = n_s(1 - p_s)$. We then obtain

$$\frac{\partial}{\partial p_s} \ln \beta(\mu) \tag{30}$$

$$= n_s \frac{\partial}{\partial \alpha_s} \ln \beta(\mu) - n_s \frac{\partial}{\partial \beta_s} \ln \beta(\mu) \tag{31}$$

$$= n_s \left[-\psi(\alpha_s) + \psi(\beta_s) + \ln(\rho^{(k)}_s) - \ln(1 - \rho^{(k)}_s) \right] \tag{32}$$

$$= n_s \left[-\psi(\alpha_s) + \psi(\beta_s) + \ln \left(\frac{\rho^{(k)}_s}{1 - \rho^{(k)}_s} \right) \right] \tag{33}$$

$$\frac{\partial}{\partial n_s} \ln \beta(\mu) = p \frac{\partial}{\partial \alpha_s} \ln \beta(\mu) + (1 - p) \frac{\partial}{\partial \beta_s} \ln \beta(\mu) \tag{34}$$

$$= p \left[-\psi(\alpha_s) + \psi(\beta_s) + \ln(\rho^{(k)}_s) - \ln(1 - \rho^{(k)}_s) \right] + \left[\psi(\alpha_s + \beta_s) - \psi(\beta_s) + \ln(1 - \rho^{(k)}_s) \right] \tag{35}$$

$$= p \left[-\psi(\alpha_s) + \psi(\beta_s) + \ln \left(\frac{\rho^{(k)}_s}{1 - \rho^{(k)}_s} \right) \right] + \left[\psi(\alpha_s + \beta_s) - \psi(\beta_s) + \ln(1 - \rho^{(k)}_s) \right] \tag{36}$$

B Ommitted proofs

Proof of Lemma 1. For any β

$$\max_{\mu\in\mathcal{M}} R(\pi,\mu) \geq \max_{\mu\in\mathrm{supp}(\beta)} R(\pi,\mu) \tag{37}$$
$$= \max_{\mu\in\mathrm{supp}(\beta)} U(\pi^*(\mu),\mu) - U(\pi,\mu) \tag{38}$$
$$\geq \max_{\mu\in\mathrm{supp}(\beta)} U(\pi^*(\beta),\mu) - U(\pi,\mu) \tag{39}$$
$$\geq \sum_{\mu\in\mathrm{supp}(\beta)} \beta(\mu)[U(\pi^*(\beta),\mu) - U(\pi,\mu)] \tag{40}$$
$$= U(\pi^*(\beta),\beta) - U(\pi,\beta) = R(\pi,\beta). \tag{41}$$

Since the above holds for any β, $\max_\mu R(\pi,\mu) \geq \max_\beta R(\pi,\beta)$. Letting $\delta(\mathcal{M})$ denote the degenerate distributions on individual members of $\mathcal{M}$, we have:

$$\max_\beta R(\pi,\beta) \geq \max_{\beta\in\delta(\mathcal{M})} R(\pi,\mu) = \max_{\mu\in\mathcal{M}} R(\pi,\mu)$$

□

Proof of Lemma 5. Let $\pi,\pi',\pi''\in\Pi$. To verify that $L(\pi,\beta)$ is l-smooth we study if

$$||\nabla L(\pi,\beta) - \nabla L(\pi',\beta')|| \leq l||(\pi,\beta)-(\pi',\beta')||. \tag{42}$$

$$||\nabla L(\pi'',\beta'') - \nabla L(\pi',\beta')||_2^2 \tag{43}$$
$$\leq ||(\pi'',\beta'')-(\pi',\beta')||_2^2(\sup_{\pi,\beta}||\nabla^2 L(\pi,\beta)||_2^2) \tag{44}$$
$$= ||(\pi'',\beta'')-(\pi',\beta')||_2^2(\sup_{\pi,\beta}||\nabla_\pi^2 L(\pi,\beta)||_2^2) \tag{45}$$
$$\leq ||(\pi'',\beta'')-(\pi',\beta')||_2^2(\sup_{\pi,\beta}||\nabla_\pi^2 L(\pi,\beta)||_F^2) \tag{46}$$

Here the second transformation is due to the fact that any derivative with respect to β is constant, and therefore the second order derivatives are zero except for ∇_π^2. $||.||_F$ denotes the Frobenius norm.

For stochastic policies π in a parametrised policy space $\Pi_W \subset \Pi^{\mathrm{S}}$, we can write (cf. Dimitrakakis and Ortner (2022)):

$$\nabla_\pi L(\pi,\beta) = \nabla_\pi U(\pi,\beta) = \sum_\beta \nabla_\pi U(\pi,\mu)\beta(\mu). \tag{47}$$

Similarly, we obtain, for the Hessian:

$$\nabla_\pi^2 L(\pi,\beta) = \nabla_\pi^2 U(\pi,\beta) = \sum_\beta \nabla_\pi^2 U(\pi,\mu)\beta(\mu). \tag{48}$$

So it is only necessary to compute

$$\nabla_\pi^2 U(\pi,\mu) = \sum_h U(h)\nabla_\pi(\mathbb{P}_\mu^\pi(h)\sum_t \nabla_\pi \ln\pi(a_t\mid h_t)) \tag{49}$$
$$= \sum_h U(h)(\nabla_\pi(\mathbb{P}_\mu^\pi(h))\sum_t \nabla_\pi \ln\pi(a_t\mid h_t)) + \mathbb{P}_\mu^\pi(h)\sum_t \nabla_\pi^2 \ln\pi(a_t\mid h_t)) \tag{50}$$
$$= \sum_h U(h)(\mathbb{P}_\mu^\pi(h)\sum_t \nabla_\pi \ln\pi(a_t\mid h_t)\sum_t \nabla_\pi \ln\pi(a_t\mid h_t)^T + \mathbb{P}_\mu^\pi(h)\sum_t \nabla_\pi^2 \ln\pi(a_t\mid h_t)) \tag{51}$$
$$= \sum_h U(h)(\mathbb{P}_\mu^\pi(h)\sum_t \nabla_\pi \ln\pi(a_t\mid h_t)\nabla_\pi \ln\pi(a_t\mid h_t)^T + \mathbb{P}_\mu^\pi(h)\sum_t \nabla_\pi^2 \ln\pi(a_t\mid h_t)) \tag{52}$$

where for a given history $h = (s_1, r_1, a_1, \dots, s_T, r_T)$, $h_t = (s_1, r_1, a_1, \dots, s_t, r_t)$.

From the setting of a softmax policy and a partitioned history in Eq (23).

$$\frac{\partial}{\partial w_{h,a}} \ln \pi(a_t \mid h_t) = \begin{cases} 1 - \pi(a|h), & a_t = a, h_t = h \\ -\pi(a|h), & a_t \neq a, h_t = h \\ 0, & h_t \neq h \end{cases} \tag{53}$$

$$\frac{\partial\partial}{\partial w_{h,a} \partial w_{h,a'}} \ln \pi(a_t \mid h_t) = \begin{cases} \pi(a|h)(\pi(a|h) - 1), & a = a', h_t = h \\ \pi(a|h)\pi(a'|h), & a \neq a', h_t = h \\ 0, & h_t \neq h. \end{cases} \tag{54}$$

We then get Let $\nabla^2_\pi U(\pi, \mu) = G_1 + G_2$ where

$$G_1 = \sum_h U(h)\, \mathbb{P}^\pi_\mu(h) \sum_t \nabla_\pi \ln \pi(a_t \mid h_t) \nabla_\pi \ln \pi(a_t \mid h_t)^T \tag{55}$$

$$G_2 = \sum_h U(h)\, \mathbb{P}^\pi_\mu(h) \sum_t \nabla^2_\pi \ln \pi(a_t \mid h_t). \tag{56}$$

$$\begin{aligned}
||G_1||_F =& || \sum_h U(h)\, \mathbb{P}^\pi_\mu(h) \sum_t \nabla_\pi \ln \pi(a_t \mid h_t) \nabla_\pi \ln \pi(a_t \mid h_t)^T ||_F && (57)\\
\leq & \max_h |U(h)| || \sum_h \mathbb{P}^\pi_\mu(h) \sum_t \nabla_\pi \ln \pi(a_t \mid h_t) \nabla_\pi \ln \pi(a_t \mid h_t)^T ||_F && (58)\\
\leq & T || \sum_h \mathbb{P}^\pi_\mu(h) \sum_t \nabla_\pi \ln \pi(a_t \mid h_t) \nabla_\pi \ln \pi(a_t \mid h_t)^T ||_F && (59)\\
= & T \sqrt{\sum_{h_t} \sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \left(\mathbb{P}^\pi_\mu(h_t) T \frac{\partial \ln \pi(a_t \mid h_t)}{\partial \omega_{h_t,a}} \frac{\partial \ln \pi(a_t \mid h_t)}{\partial \omega_{h_t,a'}} \right)^2} && (60)\\
\leq & T \sqrt{\sum_{h_t} T^2\, \mathbb{P}^\pi_\mu(h_t)^2 \sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} 1^2} && (61)\\
\leq & T \sqrt{T^2 \sum_{h_t} \mathbb{P}^\pi_\mu(h_t) \sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} 1} && (62)\\
\leq & T \sqrt{T^2 |\mathcal{A}|^2} && (63)\\
\leq & |\mathcal{A}| T^2 && (64)
\end{aligned}$$

Here equation (60) comes from the definition of the Frobenius norm and the fact that every element (h_t, a, a') in the matrix corresponds to $\sum_h \mathbb{I}_{h_t \in h}\, \mathbb{P}^\pi_\mu(h) \frac{\partial \ln \pi(a_t|h_t)}{\partial \omega_{h_t,a}} \frac{\partial \ln \pi(a_t|h_t)}{\partial \omega_{h_t,a'}}$ and that $\mathbb{P}^\pi_\mu(h_t) = \sum_h \mathbb{P}^\pi_\mu(h_t|h)\, \mathbb{P}^\pi_\mu(h) = \sum_h \mathbb{I}_{h_t \in h} 1/T\, \mathbb{P}^\pi_\mu(h)$. Equation (61) follows from the absolute value of equation (53) being bounded by one.

$$||G_2||_F = || \sum_h U(h)\, \mathbb{P}^\pi_\mu(h) \sum_t \nabla^2_\pi \ln \pi(a_t \mid h_t)||_F \tag{65}$$

$$\leq T|| \sum_h \mathbb{P}^\pi_\mu(h) \sum_t \nabla^2_\pi \ln \pi(a_t \mid h_t)||_F \tag{66}$$

$$\leq T|| \sum_{h_t} T\, \mathbb{P}^\pi_\mu(h_t) \nabla^2_\pi \ln \pi(a_t \mid h_t)||_F \tag{67}$$

$$\leq T \sqrt{\sum_{h_t} T^2\, \mathbb{P}^\pi_\mu(h_t)^2 1} \tag{68}$$

$$\leq T \sqrt{T^2 \sum_{h_t} \mathbb{P}^\pi_\mu(h_t) 1} \tag{69}$$

$$\leq T^2 \tag{70}$$

Similarly to the case for G_1, the steps follow the definition of the Frobenius norm, the observation that each element is weighted by $\mathbb{P}^\pi_\mu(h_t)T$, and that the absolute value of the partial derivatives is bounded by 1.

Finally this yields

$$l \leq ||\nabla^2_\pi U(\pi, \mu)||_F \leq ||G_1||_F + ||G_2||_F \leq T^2(|\mathcal{A}| + 1). \tag{71}$$

$L(., \beta)$ is $\mathcal{L}$-Lipschitz if $||\nabla_\pi U(\pi, \mu)||_2 \leq \mathcal{L}$.

$$||\nabla_\pi U(\pi, \mu)||_2 = || \sum_h U(h)\, \mathbb{P}^\pi_\mu(h) \sum_t \nabla_\pi \ln \pi(a_t \mid h_t)||_2 \tag{72}$$

$$\leq || \sum_h U(h)\, \mathbb{P}^\pi_\mu(h) \sum_t \nabla_\pi \ln \pi(a_t \mid h_t)||_F \tag{73}$$

$$\leq T \sqrt{T^2 \sum_{h_t} \mathbb{P}^\pi_\mu(h_t)^2 1^2} \tag{74}$$

$$\leq \max_h(|U(h)|)T. \tag{75}$$

This then gives $\mathcal{L} \leq T^2$.

□

Proof of Lemma 7. Firstly,

$$\min_{\pi \in \Pi} L(\pi, \beta^{\epsilon,*}) \tag{76}$$

$$\geq \min_{\pi \in \Pi^\epsilon} L(\pi, \beta^{\epsilon,*}) - \epsilon \tag{77}$$

$$\geq \min_{\pi \in \Pi^\epsilon} L(\pi, \beta^*) - \epsilon \tag{78}$$

$$\geq \min_{\pi \in \Pi} L(\pi, \beta^*) - \epsilon \tag{79}$$

which completes the first part of the proof.

Secondly from the definition of c-convexity, and the fact that $\nabla_\beta \min_{\pi \in \Pi} L(\pi, \beta^*)^T(\beta - \beta^*)$ must be zero since the gradient must be zero in any direction that does not move out of $\mathcal{B}$, we have

$$\min_{\pi \in \Pi} L(\pi, \beta) \leq \min_{\pi \in \Pi} L(\pi, \beta^*) - c||\beta^* - \beta||^2_2. \tag{80}$$

Rearranging and setting $\beta = \beta^{\epsilon,*}$ finishes the proof. □

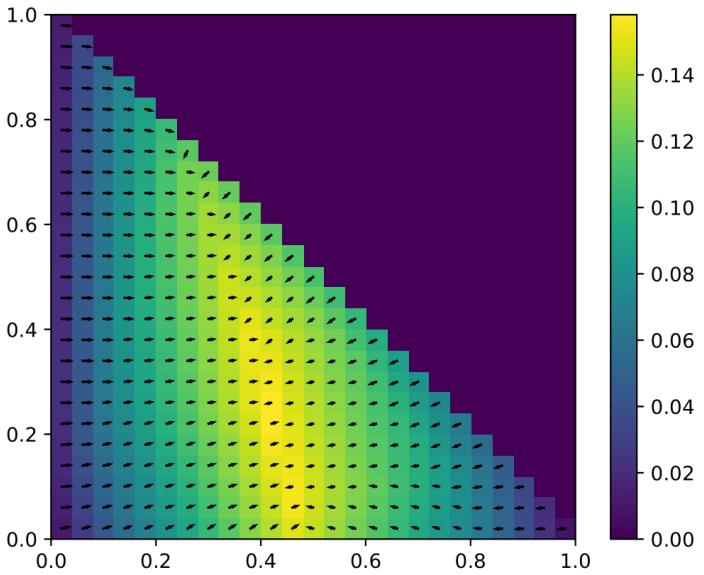


Figure 6: Visualisation of Bayesian regret for three finite-horizon MDPs. The arrows show the gradients of the Bayesian regret for the corresponding Bayes-optimal policy. The axes represent the belief of two of the MDPs while the belief of the final MDP is given by 1-x-y.

Table 1: Comparison of worst-case Bayesian regret for optimal policies at minimax and uniform belief for 16 MDP tasks.

Seed	1	2	3	4	5
Minimax	0.247	0.314	0.348	0.342	0.363
Uniform	0.640	0.554	0.484	0.646	0.850

C Additional results for finite MDPs

In this section we generate MDPs as in the same way as in Section 6.2, with the difference that Table 1 uses $\gamma = 0.9$.

Figure 6 gives an example of what the Bayesian regret landscape looks like for a task with three MDPs. The change in Bayesian regret for the fixed optimal policy of a certain belief is visualised with arrows.

In Table 1 we have some additional results comparing the performance of the uniform-prior and worst-case prior policies. In particular, we generate 5 sets of 16 MDPs. For each set, we calculate the minimax policy and the best response to the uniform prior. We then calculate the worst-case Bayesian regret for each policy. As we can expect, the minimax policy significantly outperforms the uniform best response policy.