Minimax-Bayes Reinforcement Learning

Thomas Kleine Buening thomkl@uio.no
University of Oslo

Christos Dimitrakakis christos.dimitrakakis@gmail.com
University of Neuchâtel, University of Oslo, Chalmers University of Technology

Hannes Eriksson hannes.eriksson@zenseact.com
Zenseact, Chalmers University of Technology

Divya Grover divya.grover@chalmers.se
Chalmers University of Technology

Emilio Jorge emilio.jorge@chalmers.se
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

Bayesian methods offer a principled approach for obtaining nearly-optimal adaptive policies in reinforcement learning. However, the selection of the prior distribution may be at least as important as the algorithm used. In this work, we consider the problem of a Bayesian agent interacting with a Markov Decision Process (MDP), where the prior is selected in a minimax fashion in order to ensure robustness and where nature is assumed to select a prior adversarially. We define the zero-sum game as either a maximin utility or a minimax regret game. As nature can choose a prior so that all policies will obtain zero utility, we focus mainly on the minimax regret problem.

Minimax-Bayes decision problems have been discussed extensively previously in the monograph by Berger (1985). There the problem is to find a worst-case prior so as to obtain guarantees in terms of the expected loss in a Bayesian decision procedure. More recently, Grünwald and Dawid (2004) discussed the problem of Bayesian experiment design in this context. Arguably, the reinforcement learning problem in the Bayesian setting is a strict generalisation of experiment design. However, this setting has not received much attention in the past, even though the related concept of maximum entropy have been used in inverse reinforcement learning (Ziebart et al., 2008). In particular, while the notion of a worst-case prior is well-established in simple decision problems, it is unclear whether it can be easily characterised in reinforcement learning settings. In this paper, we first give an overview of basic theoretical concepts, and provide some extensions and modifications of existing minimax theorems to this setting. This helps provide an intuition about what is achievable. These are complemented by experiments in discrete and continuous settings, where we aim to identify both minimax priors and their corresponding best response policies.

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, who 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 approximately minimax solution achievable in many settings, but that they are much more robust than Bayes-adaptive policies under common priors.

2. Setting

A Markov decision process 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, $P : \mathcal{S} \times \mathcal{A} \to \mathbb{\Delta}(\mathcal{S})$ is a transition function, $\rho : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a reward function, and T is a (potentially random) horizon. 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 horizon itself is also finite. Let $\mathcal{M}$ denote the space of MDPs. 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. Together, the history $h_t = (s^t, a^{t-1})$ describes the information that is available to the agent before choosing an action in round t. The agent's utility $\mathscr{U}$ is an additive function of individual rewards $\mathscr{U} \triangleq \sum_{t=1}^{T} r_t$. 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 $\mathscr{U}(\pi, \mu) \triangleq \mathbb{E}_\mu^\pi[\mathscr{U}]$ with maximal utility denoted by $\mathscr{U}^*(\mu) \triangleq \max_{\pi \in \Pi} \mathscr{U}(\pi, \mu)$. Typically the policy is adaptive, so that the agent's actions can depend on what it has observed in the past.

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, we have $\pi(a_t = a \mid h_t) = \pi(a_t = a \mid s_t)$. 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. $\mathscr{U}^*(\mu) = \sup_{\pi \in \Pi} \mathscr{U}(\pi, \mu) = \max_{\pi \in \Pi_1^{\mathrm{D}}} \mathscr{U}(\pi, \mu)$.

Strategies. Typical minimax results rely on the notion of mixed actions called strategies. A strategy σ is a probability measure over policies. If Π is a set of base policies, we denote the set of probability measures over Π by $\mathbb{\Delta}(\Pi)$.

Fact 1 *For any strategy $\sigma \in \mathbb{\Delta}(\Pi^{\mathrm{D}})$ 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.*

2.1 Utility

For a distribution β over MDPs, we define the utility of a particular policy π to be:

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

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

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

denote the Bayes-optimal utility for a belief. We recall the fact that this is a convex function (c.f. DeGroot, 1970). By definition, and due to convexity, the following bounds hold: $\mathscr{U}(\pi, \beta) \leq \mathscr{U}^*(\beta) \leq \int_\mu \mathscr{U}^*(\mu) \, \mathrm{d}\beta(\mu)$, $\forall \pi \in \Pi$. In the

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

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.

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 selects β in an arbitrary way from a set of possible priors B. Then we wish to find a policy π^* achieving:

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

One basic open question is whether the maximum exists. The answer is positive if the game between nature and the agent has a value, i.e. $\mathscr{U}^* = \sup_{\pi \in \Pi} \inf_{\beta \in B} \mathscr{U}(\pi, \beta) = \inf_{\beta \in B} \sup_{\pi \in \Pi} \mathscr{U}(\pi, \beta) = \mathscr{U}_*$. Let π^* and β^* be the maximin policy and minimax prior respectively. If the game has a value then there exists an equalising policy which is optimal in response to some minimax belief β^*, and vice versa. A sufficient condition for this to occur is for $\mathscr{U}^*(\beta)$ to be convex and differentiable everywhere (c.f. Grünwald and Dawid, 2004). In particular, an equalising strategy can always be found when Π is finite.

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

Note that this is only a best-response policy, and not a solution to the maximin problem (2). In addition, 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 will focus on the problem of minimax *regret*, i.e. the gap between the agent's policy and that of an oracle.

3. Properties of the regret

We generally write $\mathscr{R}(\mathcal{A}, \mathcal{I})$ to mean the regret of some algorithm $\mathcal{A}$ relative to an oracle with information $\mathcal{I}$. For simplicity, we consider a finite set of base MDPs $\mathcal{M}$. 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 $\mathscr{R}(\pi, \mu) \triangleq \mathscr{U}^*(\mu) - \mathscr{U}(\pi, \mu)$.*

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: $\mathscr{R}(\pi, \beta) \triangleq \mathscr{U}^*(\beta) - \mathscr{U}(\pi, \beta) = \sum_\mu \beta(\mu)[\mathscr{U}(\pi^*(\beta), \mu) - \mathscr{U}(\pi, \mu)]$, where $\pi^*(\beta) = \arg\max_\pi \mathscr{U}(\pi, \beta)$.*

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 $\mathscr{L}(\pi, \beta) \triangleq \mathbb{E}_{\mu \sim \beta}[\mathscr{R}(\pi, \mu)] = \sum_\mu \beta(\mu) \mathscr{R}(\pi, \mu) = \sum_\mu \beta(\mu)[\mathscr{U}^*(\mu) - \mathscr{U}(\pi, \mu)]$.*

These definitions of the 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.

Remark 4 *The Bayesian regret of a policy π is greater than the Bayes-optimal regret, i.e. $\mathscr{R}(\pi, \beta) \leq \mathscr{L}(\pi, \beta)$.*

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.

Proof For the discrete case, $\mathscr{R}(\pi, \beta) = \sum_\mu \beta(\mu)[\mathscr{U}(\pi^*(\beta), \mu) - \mathscr{U}(\pi, \mu)] \leq \sum_\mu \beta(\mu)[\mathscr{U}^*(\mu) - \mathscr{U}(\pi, \mu)] = \mathscr{L}(\pi, \beta)$, since $\mathscr{U}(\pi^*(\beta), \mu) < \mathscr{U}^*(\mu)$ by definition of $\mathscr{U}^*(\mu)$. The continuous case follows similarly, under mild technical conditions on measurability.

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 $\mathscr{L}(\pi, \beta) = \mathscr{L}(\pi^*(\beta), \beta) + \mathscr{R}(\pi, \beta)$.

Remark 5 *$\mathscr{R}(\pi, \beta)$ is convex in β.*

Proof From the definition $\mathscr{R}(\pi, \beta) = \mathscr{U}^*(\beta) - \mathbb{E}_{\mu\sim\beta}[\mathscr{U}(\pi, \mu)]$. As $\mathscr{U}^*(\beta)$ is convex in β and $\mathbb{E}_{\mu\sim\beta}[\mathscr{U}(\pi, \mu)]$ 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.

Let us now attempt to see whether zero-sum games defined with respect to the regret always have a value. We would expect this to be the case if the regret was a bilinear function of the policy and prior. However, at least for the Bayes-optimal regret, this is not the case.

Intuitively, the worst-case regret of any policy can be taken over MDPs rather than beliefs, as the Bayes-optimal regret is a convex function. This implies that, for any policy, the maxima of the function lie on beliefs which are degenerate. 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 B denote the set of beliefs and we work under the assumption that the degenerate beliefs are contained in the belief space.

Lemma 6 (Lattimore (2021)) *If for each MDP $\mu \in \mathcal{M}$ there exists an associated Dirac belief $\beta_\mu \in B$, then for any policy π we have $\max_{\mu\in\mathcal{M}} \mathscr{R}(\pi, \mu) = \max_{\beta\in B} \mathscr{R}(\pi, \beta)$.*

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

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

We find a similar result for the Bayesian regret.

Lemma 7 *If for each MDP $\mu \in \mathcal{M}$ there exists an associated Dirac belief $\beta_\mu \in B$, then for any π:*

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

Proof For any β, we have

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

Consequently $\max_\mu \mathscr{R}(\pi, \mu) \geq \max_\beta \mathscr{L}(\pi, \beta)$. Once more, $\max_\beta \mathscr{L}(\pi, \beta) \geq \max_{\beta\in\delta(\mathcal{M})} \mathscr{L}(\pi, \beta) = \max_{\mu\in\mathcal{M}} \mathscr{R}(\pi, \mu)$, due to the fact that $\mathscr{R}(\pi, \mu) = \mathscr{L}(\pi, \beta_\mu)$ for the singular belief β_μ on MDP μ. As a result, $\max_{\mu\in\mathcal{M}} \mathscr{R}(\pi, \mu) \geq \max_{\beta\in B} \mathscr{L}(\pi, \beta) \geq \max_{\mu\in\mathcal{M}} \mathscr{R}(\pi, \mu)$.

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

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 7.

Corollary 8 *If for each MDP $\mu \in \mathcal{M}$ there exists an associated Dirac belief $\beta_\mu \in B$, then for any π:*

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

Equations (3) and (5) can be made intuitive through a simple geometry 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've chosen. However, we typically want to model a worst-case setting by assuming nature picks its distribution without knowing which policy the DM 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 can answer this in the positive with respect to both the Bayesian regret and the utility in the finite setting.

Corollary 9 *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:*

$$\min_{\beta} \max_{\pi} \mathscr{U}(\pi, \beta) = \max_{\pi} \min_{\beta} \mathscr{U}(\pi, \beta), \qquad \max_{\beta} \min_{\pi} \mathscr{L}(\pi, \beta) = \min_{\pi} \max_{\beta} \mathscr{L}(\pi, \beta) \tag{6}$$

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 $\mathscr{U}(\pi, \beta) = \sum_\mu \sum_d \beta(\mu) \mathscr{U}(d, \mu) \sigma(d)$. The result follows from the standard minimax theorem. Similarly for the regret, we use $\mathscr{L}(\pi, \beta) = \sum_\mu \sum_d \beta(\mu) \mathscr{R}(d, \mu) \sigma(d)$.

The same does not hold for the Bayes-optimal regret, since for arbitrary policy spaces the agent's Bayes-optimal policy has zero Bayes-optimal regret, as it is aware of the prior distribution. However, the minimax value is generally greater than zero. Since $\mathscr{R}(\pi, \beta)$ is convex in β, we do not obtain the standard bilinear form, and the game may not have a value.

Lemma 10 *The game $\mathscr{R}(\pi, \beta)$ 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 $\pi \in \Pi^{\mathrm{D}}$, $\max_\beta \min_\pi \mathscr{R}(\pi, \beta) = 0$. Consequently, $\min_\pi \max_\beta \mathscr{R}(\pi, \beta) \geq \max_\beta \min_\pi \mathscr{R}(\pi, \beta) = 0$. From (3), we obtain $\min_\pi \max_\mu \mathscr{R}(\pi, \mu) = \min_\pi \max_\beta \mathscr{R}(\pi, \beta) \geq \max_\beta \min_\pi \mathscr{R}(\pi, \beta) = 0$. It remains to show that $\min_\pi \max_\mu \mathscr{R}(\pi, \mu) > 0$. Assume the contrary. Then there is some policy π^* for which $\max_\mu \mathscr{R}(\pi^*, \mu) = 0$. However, there exists at least one μ' whose optimal policy does not coincide with π^*, hence $\mathscr{R}(\pi^*, \mu') > 0$.

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 11 *The worst-case Bayesian regret of the Bayes-optimal policy equals the minimax Bayesian regret, i.e.*

$$\max_{\beta} \mathscr{L}(\pi^*(\beta), \beta) = \max_{\beta} \min_{\pi} \mathscr{L}(\pi, \beta) = \min_{\pi} \max_{\beta} \mathscr{L}(\pi, \beta).$$

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

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

Finally, $\max_\beta \min_\pi \mathscr{L}(\pi, \beta) = \min_\pi \max_\beta \mathscr{L}(\pi, \beta)$ by merit of Corollary 9.

We should clarify 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, 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 strategy 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, discrete, finite horizon problems, as the set of policies is then finite, meaning that no pure deterministic policy may be equalising. 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, 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 methods

We want to calculate the minimax pair (π^*, β^*) through an alternating gradient algorithm. The algorithms are essentially the same for both maximin utility and the minimax Bayesian regret. More particularly, the Bayesian regret gradient is obtained as follows:

$$\nabla_\pi \mathscr{L}(\pi, \beta) = -\int d\beta(\mu) \nabla_\pi \mathscr{U}(\pi, \mu) \qquad \nabla_\beta \mathscr{L}(\pi, \beta) = \int_{\mathcal{M}} \mathscr{R}(\pi, \mu) \nabla_\beta d\beta(\mu). \tag{7}$$

To solve the minimax problem, we consider both an descent-ascent algorithm (GDA) (Lin et al., 2020) and GDMax (Jin et al., 2020, Alg. 2), which performs a gradient step for the prior, and full optimization for the policy. GDMax is used in combination with a Gittins index policy in the bandit case, and with a myopic BAMDP-policy in the MDP case. GDA is used with a parametrised history-dependent policy. For Bayesian regret $\mathscr{L}$ and finite MDP setting, convergence guarantees exist for GDA, but not for parametrised belief spaces. To be doubly sure of the numerical stability of the gradient algorithms, we compare them with a cutting-plane method.

For completeness, stochastic GDA (see e.g. (Lin et al., 2020)) is described in Algorithm 1. We assume access to gradient oracles $G_\pi(\pi, \beta, \xi_i)$ and $G_\beta(\pi, \beta, \xi_i)$ for the Bayesian regret $\mathscr{L}$. Note that no guarantees exist for general non-convex non-concave Bayesian regret $\mathscr{L}$, as is the case for Dirichlet belief and parametric policies.

Algorithm 1 Stochastic GDA

Input Initial policy, belief $(\pi_0 \in \Pi, \beta_0 \in \mathcal{B})$, learning rates (η_π, η_β) and stochastic oracles G_π, G_β for $\frac{\delta \mathscr{L}}{\delta \pi}, \frac{\delta \mathscr{L}}{\delta \beta}$
for $t = 1, \ldots, T$ **do**
  Get average gradients $g_\beta = \frac{1}{M} \sum G_\beta(\pi_{t-1}, \beta_{t-1}, \xi_i)$ and $g_\pi = \frac{1}{M} \sum G_\pi(\pi_{t-1}, \beta_{t-1}, \xi_i)$ using M i.i.d samples
  $\pi_t \leftarrow \mathcal{P}_\Pi\Big(\pi_{t-1} - \eta_\pi g_\pi\Big)$
  $\beta_t \leftarrow \mathcal{P}_\mathcal{B}\Big(\beta_{t-1} + \eta_\beta g_\beta\Big)$
end for
Output β^*, π^* uniformly at random from $\{(\beta_1, \pi_1), \ldots, (\beta_T, \pi_T)\}$

Here, $\mathcal{P}_\mathcal{X}$ denotes the projection operator onto set $\mathcal{X}$. For finite MDP setting, Theorem 4.9 in (Lin et al., 2020) holds, but nothing more can be said due to non convexity of the policy space.

5.2 Cutting planes

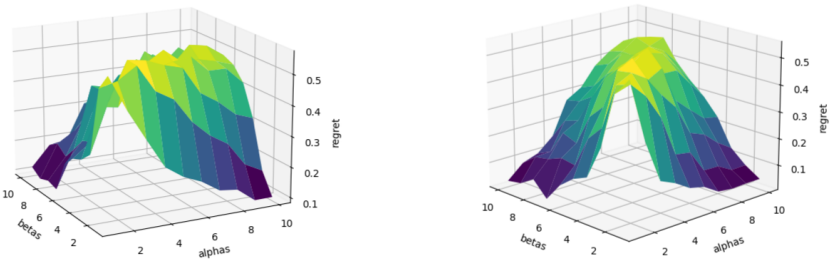
In this section we demonstrate an efficient method for localising the minimax pair (π^*, β^*) for beliefs over a finite set of MDPS, given that an oracle for obtaining the Bayes-optimal policy for a given belief is available. This could for example be obtained in finite horizon tasks with a small horizon. These methods are also applicable when the policy space is a constrained subset of all possible policies. An issue that arises when you only find the ϵ-optimal policy in the policy space is that it is no longer necessarily convex, even if the local optima and their convex combinations will differ by at most ϵ. If this is not a concern, this method can be used anyway and we obtain a locally ϵ-optimal solution. A discussion about what can be said for cases where only an ϵ-optimal policy can be found, and where it is important not to discard the minimax solution, is included in Appendix C.

We use the approximate centroid cutting plane algorithm from Bertsimas and Vempala (2004). Each policy π has a corresponding plane[4] $\mathscr{L}(\pi, \beta)$ over β which is a upper bound of the Bayesian regret of the Bayes-optimal policy. Selecting $\pi^*(\beta)$ allows the use of the plane to discard the halfplane given by the descent direction of the Bayesian regret plane, as this is a upper bound of the regret for those values of β while also being lower than the Bayesian regret in the current β. Selecting a new approximate centroid as the next β to query guarantees fast convergence in the volume of the plausible set of beliefs. With high probability the volume will decay with a factor 2/3 for each step as in Bertsimas and Vempala (2004). As such, less than 1% of the original belief space remains after just 12 steps of the algorithm.

An algorithm can be formulated as following. Let β_t be the approximate centroid (through one of the methods in Bertsimas and Vempala (2004), such as hit-and-run sampling) of the set K_t containing the plausible beliefs that could contain the minimax belief, at step t of the algorithm. C_t is the normal to the Bayes regret plane at β_t and can be obtained : $C_t^T \beta = \mathscr{L}(\pi^*(\beta_t), \beta) = \sum_i \beta_t(\mu_i) \mathscr{R}(\pi^*(\beta_t), \beta = \delta_{\mu_i})$, where each element $C_t^{(i)} = \mathscr{R}(\pi^*(\beta_t), \beta = \delta_{\mu_i})$.

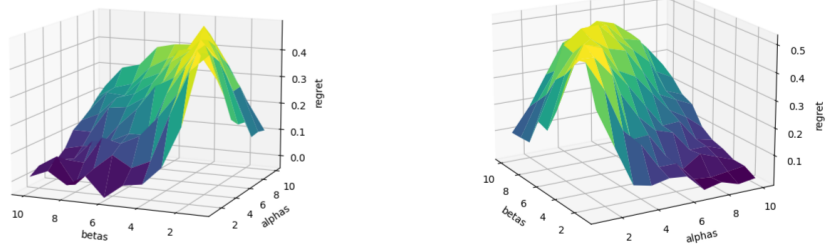
Input: Initial belief set of constraints K_0, Optimal Policy oracle, Policy evaluation oracle, $t = 0$;
for t=0:T-1 **do**
  Obtain $\beta_t \approx E_{K_t}[x]$
  Obtain optimal policy $\pi^*_{\beta_t}$ and $C_t^T \beta = \mathscr{L}(\beta, \pi_{\beta_t})$.
  $K_{t+1} = \{K_t \cap C_t^T(\beta - \beta_t) > 0\}$
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 $\pi^*(\beta^*)$.

4. Due to the Bayesian regret being an expectation over MDPs and therefore is linear.



(a) $\mathcal{Beta}(1, 1)$

(b) $\mathcal{Beta}(3, 3)$



(c) $\mathcal{Beta}(4, 2)$

(d) $\mathcal{Beta}(2, 4)$

Figure 1: 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.

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 characterise 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 the posterior sampling. The final experiment is for the general case of discrete MDPs and parametric adaptive policies, where a value may not exist.

6.1 Illustrations of Worst-Case Priors for Bernoulli Bandits

We wish to analyse 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 discounted rewards Gittins (Gittins, 1979; Gittins et al., 2011) showed that an index policy, the so-called Gittins index, does in fact yield a Bayes-optimal policy. For such K-armed Bernoulli bandits $\theta = (\theta_1, \dots, \theta_K)$ with $\theta_k \in [0, 1]$, we consider Beta product priors such that $\beta(\theta) = \prod_{k=1}^{K} \mathcal{Beta}(a_k, b_k)\{\theta_k\}$. 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 $\mathcal{Beta}(a_1, b_1)$ and the second arm's prior $\mathcal{Beta}(a_2, b_2)$ is set to different values. Figure 1 shows the Bayesian regret for different fixed priors for arm 1 and varying prior for arm 2.

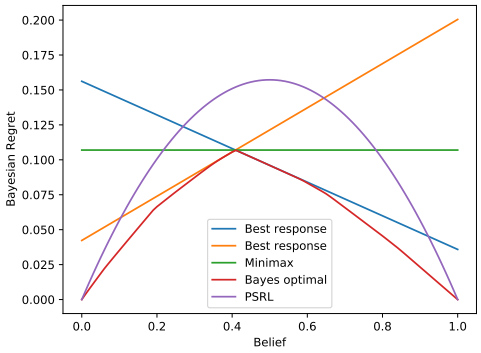
We observe that high Bayesian regret is typically suffered when the second prior's mean approximately matches the mean of the first arm's prior, i.e. $\mathbb{E}[\mathcal{Beta}(a_1, b_1)] = \mathbb{E}[\mathcal{Beta}(a_2, b_2)]$. Moreover, it seems that maximal Bayesian regret is achieved at a completely symmetric prior, i.e. $\mathcal{Beta}(a_1, b_1) = \mathcal{Beta}(a_2, b_2)$, 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 suspected 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 $\mathit{Beta}(0.8, 0.8)$ for 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.

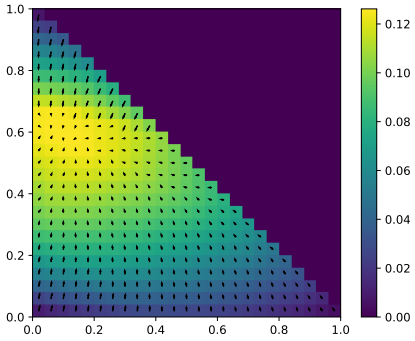
6.2 Finite set of MDPs

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 otherwise. We let the tasks be on a finite horizon $H = 5$ to allow exact computation of the optimal policies and Bayesian regret. Additionally we use $\gamma = 0.9$.

In Figure 2a the Bayesian regret for a two MDP task can be found. This helps visualise how the Bayes-optimal policy is piecewise linear function consisting of the minimum of a set of locally optimal policies. We also compare with the Bayesian regret of the PSRL policy (Strens, 2000) which samples 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. Figure 2b 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 an additional experiment



(a) There are two Bayes-optimal policies at the minimax point that can be mixed to obtain a minimax policy. The PSRL policy's Bayesian regret is necessarily quadratic.



(b) 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.

Figure 2: Visualisation of Bayesian regret for (a) two and (b) three finite-horizon MDPs.

(Appendix E, Table 5), we compare of the worst case Bayesian regret of the minimax policy and of the Bayes optimal policy for the uniform belief.

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 minmax 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.

Finding the Minimax Prior. The metrics of interest while studying the minimax belief β^* are the Bayesian regret of the minimax policy and the optimal adaptive policy for the uniform belief, the diameter of the prior β given by, $D(\beta) \triangleq \int_{\mathcal{M}} D(\mu)\,\mathrm{d}\beta(\mu) = \int_{\mathcal{M}} \mathrm{d}\beta(\mu) \max_{s \neq s' \in \mathcal{S}} \min_{\pi, \mathcal{S} \to \mathcal{A}} \mathbb{E}[\mathcal{T}(s' \mid \mu, \pi, s)]$, and the average Wasserstein distance of the prior β, $W(\beta) \triangleq \int_{\mathcal{M}} \frac{1}{|\mathcal{S}||\mathcal{A}|} \sum_{s,a} |F_\mu(s' \mid s, a) - F_{\hat{\mu}}(s' \mid s, a)|\,\mathrm{d}\beta(\mu)$.

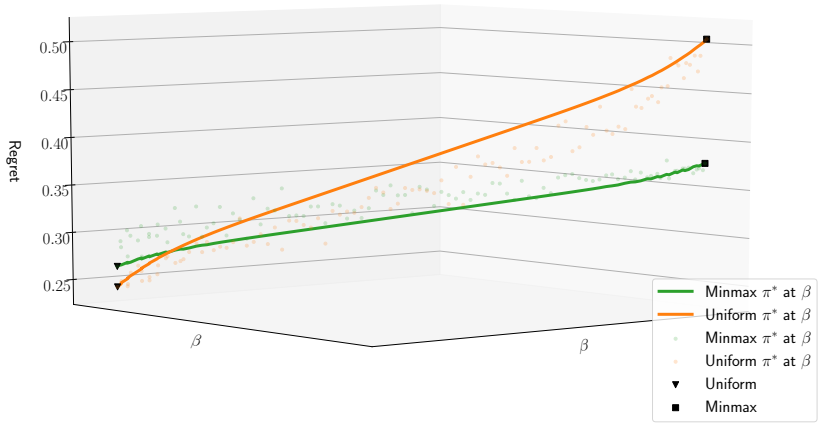


Figure 3: The following plot illustrates a t-SNE embedding of the belief space, evaluated for beliefs around the minimax belief β^* and the uniform belief β^1. The height represents the Bayesian regret $\mathscr{L}$, evaluated for each of the two optimal policies, minimax π^* and the optimal adaptive policy for the uniform belief.

	$D(\beta)$	$\max D(\mu)$	$W(\beta)$	Minimax $\mathscr{L}$	Uniform $\mathscr{L}$
Uniform β^1	3.76 ± 6.32	$1.52 * 10^3$	0.25 ± 0.07	0.28 ± 0.12	0.24 ± 0.08
Minimax β^*	$159.65 \pm 1.07 * 10^5$	$1.07 * 10^8$	0.11 ± 0.04	0.37 ± 0.02	0.51 ± 0.11

Table 1: Statistics for the uniform belief β^1 and the minimax belief β^* obtained using GAD, collected from 10^6 sampled MDPs from each belief. We show the average and maximum diameter, the average Wasserstein distance and the Bayesian regret of the two respective policies.

Figure 3 illustrates that the Bayesian regret of the minimax policy is less sensitive to changes in belief compared to the optimal policy for the uniform belief. This is verified quantitatively in Table 1. The approximate minimax solution results in MDPs with larger diameter, indicating the MDPs in general are more difficult to traverse, while the regret of the minimax policy is nearly constant. Additional results and discussion for the infinite MDP setting is available in Appendix D.

7. Conclusion

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. 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. However, the computation of minimax-Bayes policies has not been previously considered. We find that not only this appears to be feasible, but also that such policies can be significantly more robust than those based on standard uninformative priors.

Acknowledgments

This work was partially supported by the Research Council of Norway under grant 302203. This work was also supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. We are grateful for their support.

References

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

Dimitris Bertsimas and Santosh 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.

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

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

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

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

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

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR, 2020.

Tor Lattimore. Personal Communication, March 2021.

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

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

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

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Appendix A. Ommitted proofs

Proof of Lemma 6. For any β

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

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

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

Appendix B. Convex sets in isotropic position

Definition 12 *A convex set K is in isotropic position if $\mathbb{E}_K[x] = 0$ and $\mathbb{E}_K[xx^T] = I$.*

As long as K is a fully-dimensional convex set, K can be converted into isotropic position with a affine transformation where $K' = y : y = B(x - z), x \in K$ where $A = E_K[(x - z)(x - z)^T]$, z is the centroid of K and $B^2 = A^{-1}$. The existence of B is given by the fact that K is fully dimensional and that A is positive-definite.

Lemma 13 *(Lemma 6.14, Bertsimas and Vempala (2004) formulated as in Bubeck et al. (2015)). Let $\mathcal{K}$ be a convex set in isotropic position. Then for any $w \in \mathbb{R}^n, w \neq 0, z \in \mathbb{R}^n$, we have:*

$$\mathrm{Vol}\left(\mathcal{K} \cap \left\{x \in \mathbb{R}^n : (x - z)^\top w \geq 0\right\}\right) \geq \left(\frac{1}{e} - \|z\|_2\right) \mathrm{Vol}(\mathcal{K}).$$

Appendix C. ϵ-optimal cutting planes

If only an oracle for obtaining an $\epsilon-$optimal policy can be obtained then the properties are weakened if we wish to guarantee that we do not remove the minimax belief. For simplicity, we still assume that the value functions can be accurately calculated[5]. Let

$$C^T\beta = \mathscr{L}(\pi^{\epsilon-\max}_{\beta_t}, \beta) = \sum_i \beta(\mu_i)\mathscr{R}(\pi^{\epsilon-\max}_{\beta_t}, \beta = \delta_{\mu_i}). \tag{8}$$

We wish to cut $c^T(\beta - \beta_t) > \frac{\epsilon}{||C||} > \delta, c = \frac{C}{||C||}$ We can see the cutting plane as the following plane $c^T(\beta - \beta_t + c\frac{2\epsilon}{||C||}) > 0$. This can be interpreted as a cutting plane passing through $\beta_t - c\frac{\epsilon}{||C||}$. Since Lemma 13 requires isotropic position the set needs to be transformed. In the transformed set this point is $-Bc\frac{\epsilon}{||C||}$ away from the centroid.

Given that $||B||\epsilon$ is sufficiently small and $||C||$ sufficiently large we can guarantee a desired reduction in volume of the mass of K_t. If $||C||$ is small we have an approximately equalizing policy and we obtain $f(\beta^*) \geq f(\beta_t) + C^T(\beta_t - \beta^*) \implies 0 \leq f(\beta_t) - f(\beta^*) \leq ||C||||\beta_t - \beta^*|| \leq ||C||\sqrt{2}$.

Input: Initial belief set of constraints K_0, $\epsilon-$optimal policy oracle, Policy evaluation oracle, $t = 0$;
for t=0:T-1 **do**
  Sample $\{b_i\}_i^k$ from K_t
  Calculate $\hat{A} = 1/k \sum b_i b_i^T, \hat{B} = \sqrt{\hat{A}^{-1}}, \beta_t = 1/k \sum_i b_i$
  $\epsilon_t = \frac{\delta}{4||B||}$
  Obtain ϵ_t-optimal policy π_{β_t} and $C_t^T\beta = \mathscr{L}(\beta, \pi_{\beta_t})$.
  if $|C_t| \leq \delta$ **then**
    Break
  end if
  $c_t = C_t/||C_t||$
  $K_{t+1} = \{K_t \cap c_t^T(\beta - \beta_t + c\frac{\epsilon_t}{||C||}) > 0\}$
end for
Return $\sqrt{2}\delta$ optimal policy or $\beta^* \in K_T$ with VOL$(K_T) < (9/10)^t$.

If enough samples k are taken from K_t, $\hat{B}$ can be seen as accurate approximations of B, although the exact number needed is left as an open question. For approximating β_t there is a linear amount of terms needed, see Bertsimas and Vempala (2004). Then, we obtain a reduction of volume with a factor of $1/e - ||\frac{Bc\epsilon}{||C||}|| \geq 1/10$ such that after T steps we obtain a reduction in volume of $(\frac{9}{10})^T$. An optimal selection of ϵ can be calculated based on the computational complexity of the oracle such that fewer but larger cuts might be preferred. An issue is that B grows the set shrinks. As such, it might not be feasible to take a large amount of steps.

5. Usually, this is easier than finding the optimal policy. Extending to the approximate case would be a question of multiplying ϵ by a scalar in the algorithm.

Appendix D. Examples of Minimax Solutions

In this section we will cover a few examples of minimax solutions and what the beliefs and policies look like. To begin with, in Table 2 we can see the parameters of the uniform belief β^1, which is the flattened array of the parameters of a Dirichlet distribution. As can be seen, the parameters for all possible states and actions are identical for the uniform belief. In contrast, the following rows include the parameters of the minimax beliefs β^*, obtained from multiple independent experiments. While there is some dispersion among the parameters, we can observe some consistent trend of parameter choices. For instance, the columns for $\theta^3_{s,a}$ to $\theta^7_{s,a}$ are very similar and most of the differences exist for $\theta^0_{s,a}$ and $\theta^2_{s,a}$.

In order to fully appreciate the values of the flattened parameter vectors in Table 2 a refresher on the Dirichlet distribution might be necessary. As a quick summary, the magnitude of the parameters control for the variability of the sampled parameter vectors, whereas the proportion among the parameter control for the location of the samples. For example, in the table we can see $0.01, 2.04$ in columns $\theta^{6-7}_{s,a}$. These parameters would lead to a sampled probability vector with expectation $[0.005, 0.995]$. From this we can deduce that essentially some of the transitions are, most of the time, going to be deterministic. However, not all of the parameter pairs result in deterministic transitions. We can see that $\theta^{4-5}_{s,a}$ are producing probability vectors close to uniform. From these examples, and the result in Table 1 we can see that the minimax beliefs produce MDPs that in general, have higher diameter and are more difficult to traverse.

	$\theta^0_{s,a}$	$\theta^1_{s,a}$	$\theta^2_{s,a}$	$\theta^3_{s,a}$	$\theta^4_{s,a}$	$\theta^5_{s,a}$	$\theta^6_{s,a}$	$\theta^7_{s,a}$
Uniform β^1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Minimax β^*_1	0.47	1.78	0.12	2.12	0.90	1.02	0.01	2.04
Minimax β^*_2	0.01	2.02	0.39	1.81	0.87	1.10	0.01	2.18
Minimax β^*_3	0.53	1.82	0.01	2.15	0.98	0.98	0.01	2.14
Minimax β^*_4	0.25	1.77	0.10	2.28	0.89	1.06	0.01	2.16

Table 2: Examples of uniform β^1 and minimax beliefs β^*.

In Table 3 we can see an example of an adaptive policy optimised for the uniform belief β^1. The policy uses a softmax rule, taking all the $\theta^{0-8}_{s,\cdot}$ parameters into account when making a decision for state s. A stationary policy would have the first eight columns $\theta^{0-7}_{s,a}$ being 0.00, resulting in the policy ignoring the history. In contrast, the minimax policy in Table 4 results in quite a different policy.

	$\theta^0_{s,a}$	$\theta^1_{s,a}$	$\theta^2_{s,a}$	$\theta^3_{s,a}$	$\theta^4_{s,a}$	$\theta^5_{s,a}$	$\theta^6_{s,a}$	$\theta^7_{s,a}$	$\theta^8_{s,a}$
s_0, a_0	-0.26	0.91	0.16	-0.51	-0.10	-0.02	-0.05	-0.07	0.06
s_0, a_1	0.18	-0.70	-0.19	0.69	0.13	-0.05	-0.04	-0.05	-0.02
s_1, a_0	-0.60	-0.21	0.28	-0.38	-0.01	0.48	-0.44	0.03	-0.43
s_1, a_1	0.66	0.22	-0.36	0.20	-0.01	-0.41	0.23	0.03	0.57

Table 3: Example of an optimal adaptive policy for the uniform belief β^1.

	$\theta^0_{s,a}$	$\theta^1_{s,a}$	$\theta^2_{s,a}$	$\theta^3_{s,a}$	$\theta^4_{s,a}$	$\theta^5_{s,a}$	$\theta^6_{s,a}$	$\theta^7_{s,a}$	$\theta^8_{s,a}$
s_0, a_0	0.51	0.61	0.23	-0.03	-0.11	0.11	0.21	0.07	-0.07
s_0, a_1	-0.82	-0.45	-0.30	-0.08	-0.08	-0.33	-0.23	-0.14	-0.06
s_1, a_0	0.02	-0.12	-0.04	-0.39	-0.36	-0.18	-0.22	-0.22	-0.36
s_1, a_1	-0.13	-0.10	0.14	0.19	0.30	-0.28	0.29	0.12	0.08

Table 4: Example of a minimax policy.

Appendix E. Additional results for finite MDPs

Here 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

Table 5: 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

the uniform prior. We then calculate the worst-case Bayesian regret for each policy. As we can expect, the minimax policy does significantly outperform the uniform best response policy.