Minimax-Bayes Reinforcement Learning

Thomas Kleine Buening*,1 Christos Dimitrakakis*,1,2,3 Hannes Eriksson*,3,4 Divya Grover*,3 Emilio Jorge*,3

*Equal contribution [1]University of Oslo [2]Université de Neuchâtel [3]Chalmers university of Technology [4]Zenseact

Problem setup

- MDP $\mu = (S, A, P, R, T) \in \mathcal{M}$
- Utility $\mathscr{U} = \sum_{t=1}^{T} r_t$

For a fixed MDP $\mu \in \mathcal{M}$, we define the

- Utility $\mathscr{U}(\pi, \mu) = E_\mu^\pi[\mathscr{U}]$
- Optimal Utility $\mathscr{U}^*(\mu) = \max_\pi \mathscr{U}(\pi, \mu)$

For a distribution β over MDPs, we define the

- Utility $\mathscr{U}(\pi, \beta) = E_\beta^\pi[\mathscr{U}] = \int_{\mathcal{M}} \mathscr{U}(\pi, \mu) d\beta(\mu)$
- Bayes-optimal utility $\mathscr{U}^*(\beta) = \sup_\pi \mathscr{U}(\pi, \beta)$

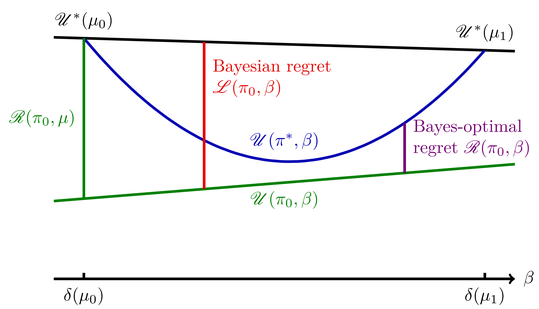
Interpretation of β:

1. The agent's subjective belief about which MDP is the most likely a priori.
2. The MDP is actually drawn randomly from distribution β.

Suppose Nature selects β arbitrarily or adversarially, then we are interested in finding

$$\max_\pi \min_\beta \mathscr{U}(\pi, \beta).$$

However, for an unrestricted set of priors, Nature could pick a prior such that all rewards are zero, thus trivially achieving minimal utility. Instead we consider regret.



Notions of regret

For a fixed MDP $\mu \in \mathcal{M}$, we define

$$\mathscr{R}(\pi, \mu) = \mathscr{U}^*(\mu) - \mathscr{U}(\pi, \mu)$$

For a prior β, we define the

- Bayes-optimal regret
$$\mathscr{R}(\pi, \beta) = \mathscr{U}^*(\beta) - \mathscr{U}(\pi, \beta)$$
- Bayesian regret (comparing against an oracle)
$$\mathscr{L}(\pi, \beta) = \mathrm{E}_{\mu\sim\beta}[\mathscr{R}(\pi, \mu)] = \int_{\mathcal{M}} \mathscr{U}^*(\mu) - \mathscr{U}(\pi, \mu) d\beta(\mu)$$

Of course, we always have $\mathscr{R}(\pi, \beta) \leq \mathscr{L}(\pi, \beta)$.

Minimax game against Nature

We define minimax games with respect to the Bayes-optimal regret $\min_\pi \max_\beta \mathscr{R}(\pi, \beta)$ and Bayesian regret $\min_\pi \max_\beta \mathscr{L}(\pi, \beta)$.

Corollary (value of the game)

The minimax game with resepct to the utility and the Bayesian regret have a value, i.e. it holds that

$$\max_\pi \min_\beta \mathscr{U}(\pi, \beta) = \min_\beta \max_\pi \mathscr{U}(\pi, \beta), \quad \min_\pi \max_\beta \mathscr{L}(\pi, \beta) = \max_\beta \min_\pi \mathscr{L}(\pi, \beta).$$

Lemma

The minimax game with respect to the Bayes-optimal regret may not have a value, i.e.,

$$\min_\pi \max_\beta \mathscr{R}(\pi, \beta) < \max_\beta \min_\pi \mathscr{R}(\pi, \beta).$$

Lemma (Bayesian regret of the Bayes-optimal policy)

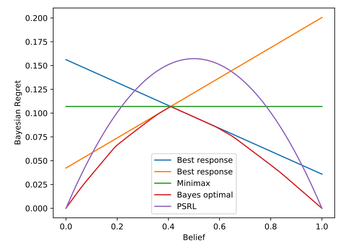
The worst-case Bayesian regret of the Bayes-optimal policy equals the minimax Bayesian regret, i.e.

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

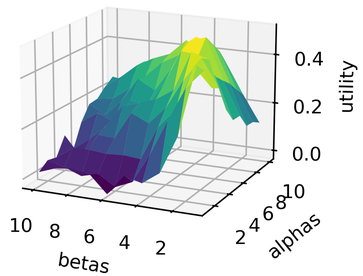
Finding minimax solutions

In the paper we describe two methods for obtaining minimax solutions based on using gradient descent ascent (Lin et al. 2020) or approximate cutting plane (Bertsimas and Vempala. 2020) that are applicable in certain settings.

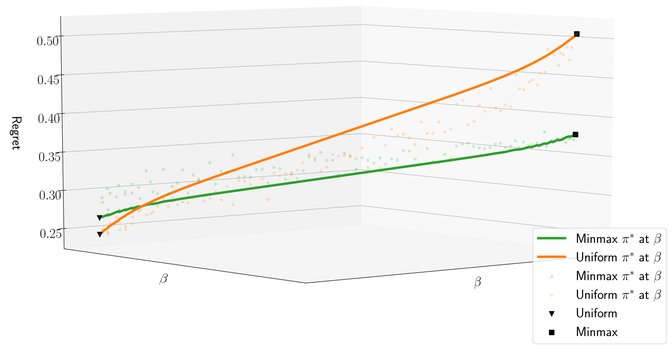
Experiments



(a) Regret for a two MDP task for a variety of policies.



(b) Bayesian regret of the Bayes-optimal policy in two armed Bernoulli bandit task. The first arms prior is fixed to $\mathcal{Beta}(4, 2)$ while the other is given by the values on the x- and y-axis.



(c) Displays a t-SNE embedding for (approximately) minimax and uniform beliefs with the Bayesian regret $\mathscr{L}$ for the minimax policy and optimal adaptive policy for the uniform belief. The policy associated with the uniform belief incurs larger Bayesian regret for beliefs further away from the uniform belief compared to the minimax policy.

European Workshop on Reinforcement Learning 2022