Minimax-Bayes Reinforcement Learning

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

*Equal contribution [1]University of Oslo [2]University of Neuchatel [3]Chalmers University of Technology [4]Zenseact

Problem Formulation

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

For a fixed MDP $\mu \in \mathcal{M}$, the **expected utility** of policy π is given by

$$U(\pi, \mu) = E_\mu^\pi[U]$$

and the **optimal utility** as $U^*(\mu) = \max_\pi U(\pi, \mu)$.

For a **distribution** β over MDPs, the **expected utility** of a policy π is:

$$U(\pi, \beta) = E_\beta^\pi[U] = \int_{\mathcal{M}} U(\pi, \mu) d\beta(\mu).$$

Its maximum is called the **Bayes-optimal utility**:

$$U^*(\beta) = \sup_\pi U(\pi, \beta).$$

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 the **maximin policy**:

$$\max_\pi \min_\beta U(\pi, \beta).$$

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

Notions of Regret

For a fixed **MDP** $\mu \in \mathcal{M}$, we define the **regret** of policy π as

$$R(\pi, \mu) = U^*(\mu) - U(\pi, \mu).$$

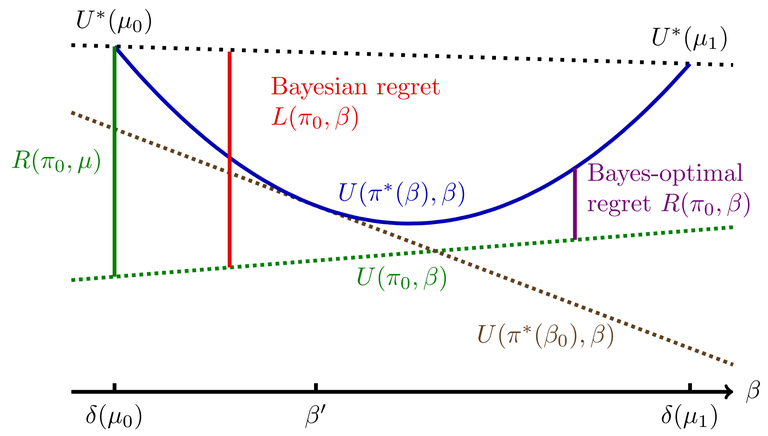
For a fixed **prior** β, we define the **Bayes-optimal regret**

$$R(\pi, \beta) = U^*(\beta) - U(\pi, \beta),$$

and the **Bayesian regret** $L(\pi, \beta) = \mathrm{E}_{\mu\sim\beta}[R(\pi, \mu)]$.

Note that $R(\pi, \beta) \leq L(\pi, \beta)$.

Illustration



Minimax Game against Nature

We define the minimax game against nature with respect to the **Bayes-optimal regret**: $\min_\pi \max_\beta R(\pi, \beta)$, and the **Bayesian regret**: $\min_\pi \max_\beta L(\pi, \beta)$.

Corollary (value of the game)

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

$$\max_\pi \min_\beta U(\pi, \beta) = \min_\beta \max_\pi U(\pi, \beta)$$
$$\min_\pi \max_\beta L(\pi, \beta) = \max_\beta \min_\pi L(\pi, \beta).$$

Lemma

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

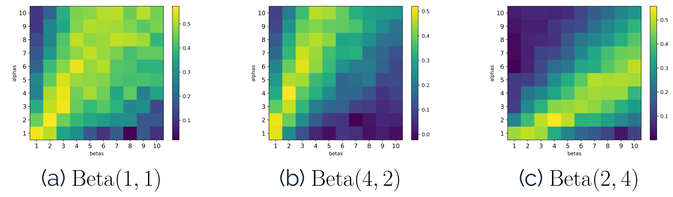
$$\min_\pi \max_\beta R(\pi, \beta) < \max_\beta \min_\pi R(\pi, \beta).$$

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.

$$\max_\beta L\left(\pi^*(\beta), \beta\right) = \min_\pi \max_\beta L(\pi, \beta).$$

Experiments



(a) Beta$(1, 1)$ (b) Beta$(4, 2)$ (c) 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.

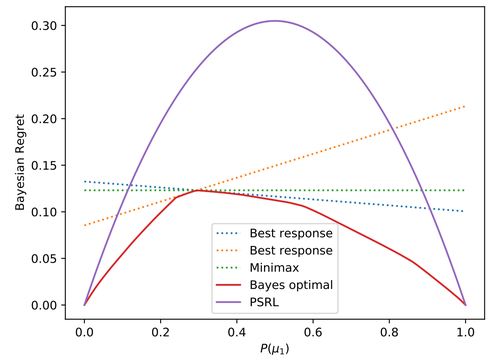


Figure 2. The Bayesian regret of different policies. The dashed lines show the value of three adaptive policies optimal for the maximin-regret prior. Two of them are best responses, which are also optimal on either side of the maximin point.

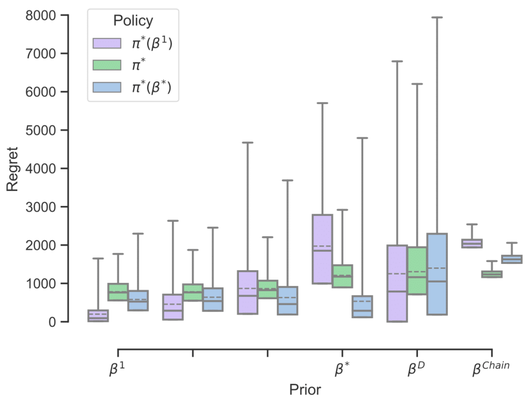


Figure 3. β^D is approximately uniform over deterministic MDPs. β^{Chain} is a delta distribution over the Chain MDP. The MDPs in between β^1 (Uniform) and β^* (Maximin) are interpolated.