Information-Theoretic Bounds for Reinforcement Learning based on Duality

Raghav Bongole
Dept. of Information Science and Engineering, KTH Royal Institute of Technology
Supervisors: Prof. Mikael Skoglund (KTH), Prof. Tobias Oechtering (KTH)

Motivation & Research Questions

- Many real-world problems involve agents acting in an **unknown environment**.
- The goal is to find a **robust policy** that maximizes rewards across all environments.
- This leads to the study of **minimax regret**.
- **Example: Stock Portfolio Optimization**
  - Optimize a portfolio of 5 stocks over 10 years.
  - Choices: Buy, sell, or keep stocks annually.
  - Stock values fluctuate with some distribution from a set of distributions.
  - Goal: Maximize worst-case expected portfolio value after 10 years.
- **Research questions:**
  - Given a class of problems, what minimax regret is achievable?
  - Can we use existing Bayesian regret bounds to bound the minimax regret? Spoiler: Yes! using **minimax duality/ the minimax theorem.**

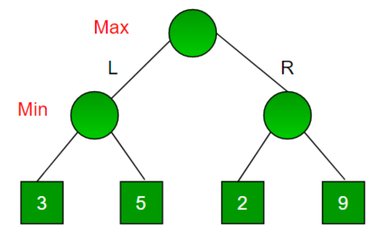

Figure 1: A minimax game

Figure 2: Minimax upper bound

Background

- **Markov Decision Processes (MDPs):** A mathematical model for decision-making in a stochastic environment.

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, y, r, T)$$

such that $p(Y_{t+1}|S_t, A_t, \Theta), y(Y_{t+1}|S_t, \Theta)$ and $r(Y_t, A_t, \Theta)$ is defined, where Θ is the environment parameter.

- **Utility:** Measures the expected cumulative reward from a policy.

$$U_{\mathcal{M}}(\pi, \theta) = \mathbb{E}\left[\sum_{t=1}^{T} r(Y_t, \pi_t(S_t))\right]$$

- **Regret:** Difference between the optimal and achieved utility.

$$\mathfrak{R}_{\mathcal{M}}(\pi, \theta) = U^*_{\mathcal{M}}(\theta) - U_{\mathcal{M}}(\pi, \theta)$$

- **Minimax Regret:** Minimizes the worst-case regret.

$$\mathfrak{M}_{\mathcal{M}} = \min_{\mathbb{P}_\Pi} \max_{\theta} \mathbb{E}_\Pi[\mathfrak{R}_{\mathcal{M}}]$$

- **Worst Case Minimum Bayesian Regret:** Minimum Bayesian Regret under worst case distribution.

$$\mathfrak{F}^*_{\mathcal{M}} = \max_{\mathbb{P}_\theta} \min_{\pi} \mathbb{E}[\mathfrak{R}_{\mathcal{M}}]$$

- **Minimax theorem:** Provides conditions ensuring:

$$\min_x \max_y f(x, y) = \max_y \min_x f(x, y).$$

Method and Results

We derive a **minimax theorem** that provides conditions under which the minimax regret equals the worst-case minimum Bayesian regret:

$$\mathfrak{M}_{\mathcal{M}} = \mathfrak{F}^*_{\mathcal{M}}$$

- A recent work provides upper bounds on the **Minimum Bayesian Regret (MBR)**[1]:

$$\mathfrak{F}_{\mathcal{M}} \leq K_1(\mathbb{P}_\Theta)$$

$\implies$ Worst-case MBR is bounded by:

$$\mathfrak{F}^*_{\mathcal{M}} \leq K_2$$

Under minimax duality conditions, minimax regret can be bounded[2]:

$$\mathfrak{M}_{\mathcal{M}} \leq K_2$$

$\implies$ Minimax regret can be controlled using Bayesian regret bounds.

Minimax regret upper bound for Multi-Armed Bandit problems:

$$\mathfrak{M}_{\mathcal{M}} \leq \sup_{\mathbb{P}_\Theta \in \Delta(\mathcal{O})} \sqrt{\frac{1}{2}|A|H(A^\star)T)}$$

$$\implies \mathfrak{M}_{\mathcal{M}} \leq O\left(\sqrt{|\mathcal{A}| \log |\mathcal{A}| T}\right)$$

References

[1] An information-theoretic analysis of bayesian reinforcement learning
Gouverneur, A., Rodríguez-Gálvez, B., Oechtering, T.J., Skoglund, M.
IEEE Allerton Conference, 2022

[2] Information-Theoretic Minimax Regret Bounds for Reinforcement Learning based on Duality
Bongole, R., Gouverneur, A., Rodríguez-Gálvez, B., Oechtering, T.J., Skoglund, M.
arXiv preprint arXiv:2410.16013, 2024

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM