Maximum Expected Hitting Cost and Informativeness of Rewards

Falcon Z. Dai dai@ttic.edu
Matthew R. Walter mwalter@ttic.edu

Toyota Technological Institute at Chicago

Motivation

We want to inquire how different rewards make solving a reinforcement learning (RL) problem easier or harder in the average reward setting.

- [JOA10] proposed a complexity measure of Markov decision processes (MDP) called diameter but it depends only on the *transitions*. We review and replace it with a **reward-sensitive** quantity called *maximum expected hitting cost* (MEHC).
- What do we mean by reward informativeness? We can look at so-called Π-*equivalent* rewards and compare their MEHCs.
- Potential-based reward shaping (PBRS) [NHR99] provides a way to construct Π-equivalent rewards. Can we characterize this set of equivalent rewards? Yes for a large class of MDPs.

Highlights

- We propose a complexity parameter of MDPs called *maximum expected hitting cost* and show that it refines diameter and thus regret bounds in previous works.
- We show that potential-based reward shaping can change the MEHC of an MDP and thus the regret bound. This results in a set of MDPs equivalent with different learning difficulties as measured by regret.
- We show that MEHCs of rewards related by PBRS differ by a factor of at most two in a large class of MDPs.

Preliminaries

Finite MDP

A *Markov decision process* is defined by the tuple $M = (\mathcal{S}, \mathcal{A}, p, r, r_{\text{max}})$, where S is the state space, A is the action space, p is the transition probability $p : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, r is the reward function $r : \mathcal{S} \times \mathcal{A} \to \mathcal{P}([0, r_{\text{max}}])$ with mean rewards $\bar{r}(s,a) \coloneqq \mathbb{E}[r(s,a)]$. Together with an algorithm $\mathfrak{L}$, we get a stochastic process $(s_t, a_t, r_t)_{t\ge 0}$.

Average reward (gain) and regret

The *accumulated reward* of algorithm $\mathfrak{L}$ after T time steps in MDP M starting in state s is a random variable $R(M, \mathfrak{L}, s, T) \coloneqq \sum_{t=1}^{T} r_t$.

Furthermore, we define the *average reward* or *gain* as $\rho(M, \mathfrak{L}, s) \coloneqq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[R(M, \mathfrak{L}, s, T)\right]$.

This can be maximized by some *stationary* policy and we define the *optimal average reward* of M, which we assume to be *independent* of initial state, as $\rho^*(M) \coloneqq \max_{\pi:S\to A} \rho(M, \pi, s)$.

We will compete with the expected accumulative reward of an optimal stationary policy *on its trajectory*, and define the *regret* of an learning algorithm $\mathfrak{L}$ starting in state s after T time steps as

$$\Delta(M, \mathfrak{L}, s, T) \coloneqq T\rho^*(M) - R(M, \mathfrak{L}, s, T).$$

Diameter and maximum expected hitting cost

Suppose in the stochastic process induced by following a policy π in MDP M, the time to hit state s' starting at state s is $h_{s\to s'}(M,\pi)$. We define the *diameter* of M [JOA10] to be

$$D(M) \coloneqq \max_{s,s'\in\mathcal{S}} \min_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}\left[h_{s\to s'}(M,\pi)\right].$$

We define the *maximum expected hitting cost* of M to be

$$\kappa(M) \coloneqq \max_{s,s'\in\mathcal{S}} \min_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}\left[\sum_{t=0}^{h_{s\to s'}(M,\pi)-1} r_{\text{max}} - r_t\right].$$

Observe that MEHC is a smaller parameter, that is, $\kappa(M) \le r_{\text{max}} D(M)$, since for any s, s', π, we have $r_{\text{max}} - r_t \le r_{\text{max}}$.

Π-equivalent rewards

These rewards assign the same average rewards to the same policies, i.e. $\rho(M_1, \pi, s) = \rho(M_2, \pi, s)$ where M_1 and M_2 differ only in their rewards.

Potential-based reward shaping

Given a potential $\varphi : \mathcal{S} \to \mathbb{R}$, define $r_t^\varphi \coloneqq r_t - \varphi(s_t) + \varphi(s_{t+1})$.

Extended MDP

After visiting state-action (s,a) for $N(s,a)$-many times, we can establish that a confidence interval for both its mean reward $\bar{r}(s,a)$ and its transition $p(\cdot|s,a)$.

$$B_\delta(s,a) \coloneqq \left\{r' \in \mathbb{R} : |r' - \hat{r}(s,a)| \le r_{\text{max}}\, b(\delta, N(s,a))\right\} \cap [0, r_{\text{max}}]$$

and the statistically plausible transitions are

$$C_\delta(s,a) \coloneqq \left\{p' \in \mathcal{P}(\mathcal{S}) : ||p'(\cdot) - \hat{p}(\cdot|s,a)||_1 \le b(\delta, N(s,a))\right\}.$$

We define an *extended MDP* $M^+ \coloneqq (\mathcal{S}, \mathcal{A}^+, p^+, r^+)$, where the action space $\mathcal{A}^+$ is a union over state-specific actions

$$\mathcal{A}_s^+ \coloneqq \{(a, p', r') : a \in \mathcal{A}, p' \in C_\delta(s,a), r' \in B_\delta(s,a)\}.$$

For transition and rewards,

$$p^+\big(\cdot\,|s,(a,p',r')\big) \coloneqq p'(\cdot) \qquad r^+\big(s,(a,p',r')\big) \coloneqq r'.$$

Results

Lemma (MEHC upper bounds the span of values)

Assuming that the actual MDP M is in the extended MDP M^+, i.e. $\bar{r}(s,a) \in B_\delta(s,a)$ and $p(\cdot|s,a) \in C_\delta(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, we have

$$\max_s u_i(s) - \min_{s'} u_i(s') \le \kappa(M)$$

where $u_i(s)$ is the i-step optimal undiscounted value of state s.

MEHC replaces diameter and leads to tighter problem-dependent regret bounds on `UCRL2` (and other algorithms), $\tilde{O}(\kappa S\sqrt{AT})$.

Theorem (MEHC under PBRS)

Given an MDP M with finite maximum expected hitting cost $\kappa(M) < \infty$ and an unsaturated optimal average reward $\rho^*(M) < r_{\text{max}}$, the maximum expected hitting cost of any PBRS-parametrized MDP M^φ is bounded by a multiplicative factor of two

$$\frac{1}{2}\kappa(M) \le \kappa(M^\varphi) \le 2\kappa(M).$$

Toy example

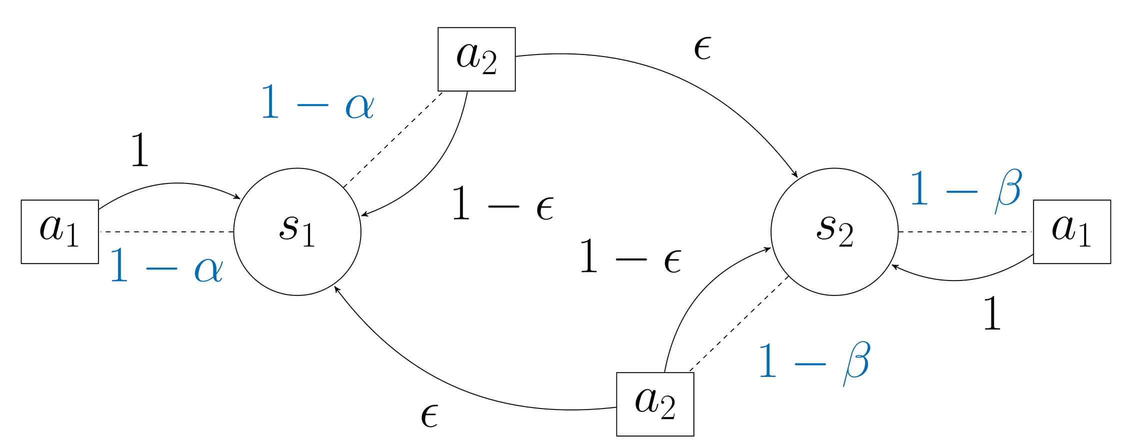

Figure 1: Circular nodes represent states and square nodes represent actions. The solid edges are labeled by the transition probabilities and the dashed edges are labeled by the mean rewards. Furthermore, $r_{\text{max}} = 1$. For concreteness, one can consider setting $\alpha = 0.11, \beta = 0.1, \epsilon = 0.05$. a_1 is the "stay" action and a_2, the "sometimes switch" action.

Obviously it is best to go to s_2 and then stay. However, taking a_2 at state s_1 usually looks as bad as taking a_1. We can differentiate the actions better by shaping with a potential of $\varphi(s_1) \coloneqq 0$ and $\varphi(s_2) \coloneqq {}^{(\alpha-\beta)}/{}_{2\epsilon}$. The shaped mean rewards become,

$$\bar{r^\varphi}(s_1, a_2) = 1-\alpha-\varphi(s_1)+\epsilon\varphi(s_2)+(1-\epsilon)\varphi(s_1) = 1-{}^{(\alpha+\beta)}/{}_2 > 1-\alpha = \bar{r^\varphi}(s_1, a_1)$$

and

$$\bar{r^\varphi}(s_2, a_2) = 1-\beta-\varphi(s_2)+\epsilon\varphi(s_1)+(1-\epsilon)\varphi(s_2) = 1-{}^{(\alpha+\beta)}/{}_2 < 1-\beta = \bar{r^\varphi}(s_2, a_1).$$

The maximum expected hitting cost becomes smaller

$$\begin{aligned}\kappa(M^\varphi) &= \max\left\{\alpha, \beta, \varphi(s_1) - \varphi(s_2) + \frac{\alpha}{\epsilon}, \varphi(s_2) - \varphi(s_1) + \frac{\beta}{\epsilon}\right\} \\ &= \max\left\{\alpha, \beta, \frac{\alpha+\beta}{2\epsilon}, \frac{\alpha+\beta}{2\epsilon}\right\} \\ &= \frac{\alpha+\beta}{2\epsilon} < \frac{\alpha}{\epsilon} = \kappa(M).\end{aligned}$$

Open questions

- Many different reward functions can motivate the same near-optimal behaviors. How can we find helpful potentials, for example in the context of inverse reinforcement learning (IRL)? Or to construct them from verbal instructions or demonstrations?

References

[FLP19] Ronan Fruit, Alessandro Lazaric, and Matteo Pirotta. Exploration-exploitation in reinforcement learning. 2019.

[GLD00] Robert Givan, Sonia Leach, and Thomas Dean. Bounded-parameter markov decision processes. *Artificial Intelligence*, 122(1-2):71--109, 2000.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563--1600, 2010.

[KBP13] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238--1274, 2013.

[NHR99] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278--287, 1999.

[Wie03] Eric Wiewiora. Potential-based shaping and q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205--208, 2003.