The Reward Function in Reinforcement Learning

CMPUT 605 - Theory of RL - Project Presentation

Alireza Masoumian

11 April 2023

Importance of Reward Function

Maximization of the expected value of the cumulative sum of a received reward.

Meaningful

What is the Meaningful Reward?!

Classic Approach

- The Algorithms are <u>too</u> to-the-point.

Rational   Risky   Reward Dependent

What is the General Recipe?!

- How can we transfer the experience between the problems?

Solutions

What is the Meaningful Reward?

What is the General Recipe?

- Consider a large set of rewards within the meta training phase …

  What Unsupervised RL do …

- Design a Self-Consistent General Reward Function …

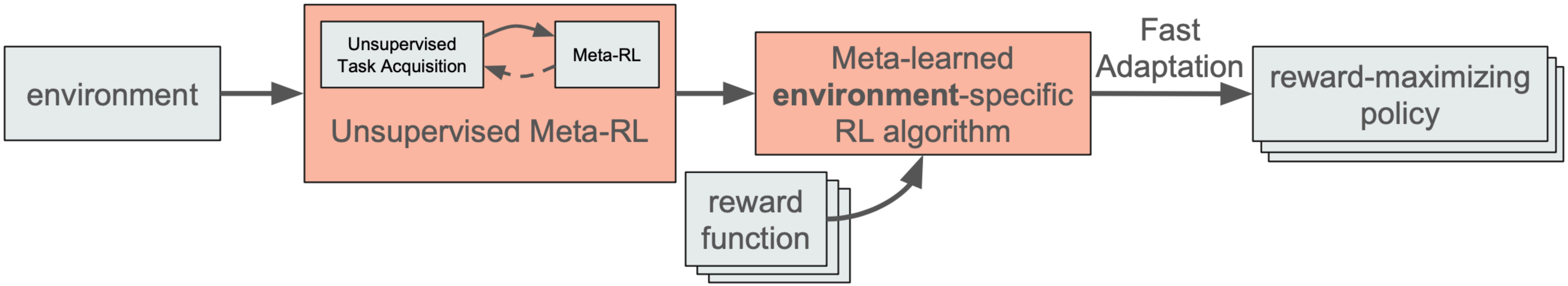
  What I like to do …

An Example for The First Approach

Unsupervised Meta-Learning for Reinforcement Learning

Abhishek Gupta [*1] **Benjamin Eysenbach** [*2] **Chelsea Finn** [3] **Sergey Levine** [1]

An Example for The First Approach



- Considering a set of rewards $r_z(s, a)$ where $z \sim p(z)$.

$$\max_{p(z)} I(\tau; z) = \mathcal{H}[\tau] - \mathcal{H}[\tau \mid z]$$

- The task distribution that provides explorations when it's free and exploits when the task (reward) is given.

An Example for The First Approach

$$f^* \triangleq \arg\max_f \mathbb{E}_{p(r_z)}\left[R(f, r_z)\right]$$

$$\textsc{Regret}(f, p(r_z)) \triangleq \mathbb{E}_{p(r_z)}\left[R(f^*, r_z)\right] - \mathbb{E}_{p(r_z)}\left[R(f, r_z)\right]$$

- We have a Controlled Markov Process $C = (S, A, P, \gamma, \rho)$,

$$\textsc{Regret}_{\text{WC}}(f, C) = \max_{p(r_z)} \textsc{Regret}(f, p(r_z))$$

$$f_C^* \triangleq \arg\min_f \textsc{Regret}_{\text{WC}}(f, C)$$

- Optimal unsupervised meta-learner $F^*(C) = f_C^*$:

$$\mathcal{F}^* \triangleq \arg\min_{\mathcal{F}} \textsc{Regret}_{\text{WC}}(\mathcal{F}(C), C)$$

The Result

- We By optimizing a task proposal distribution that maximizes trajectory-level mutual information, and subsequently performing meta-learning on the proposed tasks, we can acquire the optimal unsupervised meta-learner for trajectory matching tasks.

$$\mathcal{F}^* \triangleq \arg\min_{\mathcal{F}} \textsc{Regret}_{\mathrm{WC}}(\mathcal{F}(C), C)$$

A Confirmation to the Approach

Reward-Free RL is No Harder Than Reward-Aware RL in Linear Markov Decision Processes

Andrew Wagenmaker [1] Yifang Chen [1] Max Simchowitz [2] Simon S. Du [1] Kevin Jamieson [1]

- In contrast to the tabular setting, where we have optimal rate of $\Theta(SA/\epsilon^2)$ in reward-aware and $\Theta(S^2A/\epsilon^2)$ in reward-free.

Second Approach, First Idea

- Give a Self-Consistent General Reward function.
- Based on $C = (S, A, P, \gamma, \rho)$, we can construct a $M = (S, A', R, P, \gamma, \rho)$ such that,

$$A' = \{(a, \hat{s}) : a \in A, \hat{s} \in S\}$$

$$R_{a'}(s_t) = \mathbf{1}[\hat{s} = s_{t+1}] + B_{a'}(s_t)$$

- A bonus reward to motivate exploration.
- It's hard to be memory-less!
- Using the resulting policy as the initialization.

Thank you