Average Reward Risk-Sensitive Actor-Critic Algorithm

Pierpaolo Necchi

May 31, 2016

1 Average Reward

Most of the research in RL has studied a problem formulation where agents maximize the cumulative sum of rewards. However, this approach cannot handle infinite horizon tasks, where there are no absorbing goal states, without discounting future rewards. Traditionally, discounting has served two purposes. In some domains, such as economics, discounting can be used to represent interest earned on rewards, so that an action that generates an immediate reward will be preferred over one that generates the same reward some steps into the future. However, the typical domains studied in RL, such as robotics or games, do not fall in this category. In fact, many RL tasks have absorbing goal states, where the aim of the agent is to get to a given goal state as quickly as possible. Such tasks can be solved using undiscounted RL methods. Clearly, discounting is only really necessary in cyclical tasks, where the cumulative reward sum can be unbounded. More natural long-term measure of optimality exists for such cyclical tasks, based on maximizing the average reward per action [1].

2 Average Reward Control Problem

In the average reward setting, the goal of the agent is to find a policy that maximizes the expected reward per step.

Definition 2.1 (Average Reward)**.** *The average reward ρ_π associated to a policy π is defined as*

$$\begin{aligned}\rho_\pi &= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} R_{t+1}\right] \\ &= \int_{\mathbb{S}} d_\pi(s) \int_{\mathbb{A}} \pi(s,a)\mathcal{R}(s,a)dads \\ &= \mathbb{E}_{\substack{S\sim d_\pi \\ A\sim\pi(S,\cdot)}} [\mathcal{R}(S,A)]\end{aligned} \tag{1}$$

where d_π is the stationary distribution induced by policy π.

In the risk-neutral setting, the optimal policy solves the following optimization problem

$$\rho_* = \sup_{\pi} \rho(\pi) \tag{2}$$

In this setting, we introduce the *average adjusted* value and action-value functions.

Definition 2.2 (Average Adjusted State-Value Function)**.** *The average adjusted state-value function $V_\pi : \mathbb{S} \to \mathbb{R}$ is the expected residual return that can be obtained starting from a state and following policy π*

$$V_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} (R_{t+1} - \rho_\pi) \,\middle|\, S_0 = s \right] \tag{3}$$

Definition 2.3 (Average Adjusted Action-Value Function)**.** *The average adjusted action-value function $Q_\pi : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ is the expected residual return that can be obtained starting from a state, taking an action and then following policy π*

$$Q_\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} (R_{t+1} - \rho_\pi) \,\middle|\, S_0 = s, A_0 = a \right] \tag{4}$$

We have the following relation between the state-value function and the action-value function

$$V_\pi(s) = \int_{\mathbb{A}} \pi(s,a) Q_\pi(s,a) \tag{5}$$

These functions satisfy the following Bellman equations

$$\begin{aligned} \rho_\pi + V_\pi(s) &= \mathcal{R}_\pi(s) + T_\pi V_\pi(s) \\ \rho_\pi + Q_\pi(s,a) &= \mathcal{R}(s,a) + T_a V_\pi(s) \end{aligned} \tag{6}$$

where T_a denotes the transition operator for action a while T_π denotes the transition operator for the policy π, i.e.

$$\begin{aligned} T_a V(s) &= \int_{\mathbb{S}} \mathcal{P}(s,a,s') V_\pi(s') \\ T_\pi V(s) &= \int_{\mathbb{A}} \pi(s,a) \int_{\mathbb{S}} \mathcal{P}(s,a,s') V_\pi(s') \end{aligned} \tag{7}$$

3 Risk-Sensitive Average Reward Control Problem

In many application, in addition to maximining the average reward, the agent may want to control risk by minimizing some measure of viariability in rewards. In [2], the authors consider the long-run variance of π, defined as

Definition 3.1 (Long-Run Variance)**.** *Given a policy π, the long-run variance Λ_π is*

$$\begin{aligned}\Lambda_\pi &= \lim_{T\to\infty}\frac{1}{T}\mathbb{E}_\pi\left[\sum_{t=0}^{T-1}(R_{t+1}-\rho_\pi)^2\right]\\ &= \int_{\mathbb{A}} d_\pi(s)\int_{\mathbb{A}}\pi(s,a)\left(\mathcal{R}(s,a)-\rho_\pi\right)^2 dads\\ &= \mathbb{E}_{\substack{S\sim d_\pi\\ A\sim\pi(S,\cdot)}}\left[\left(\mathcal{R}(S,A)-\rho_\pi\right)^2\right]\end{aligned} \tag{8}$$

We can split the long-run variance in the following way

$$\Lambda_\pi = \eta_\pi - \rho_\pi^2 \tag{9}$$

where η_π is the expected square reward per step

$$\begin{aligned}\eta_\pi &= \int_{\mathbb{S}} d_\pi(s)\int_{\mathbb{A}}\pi(s,a)\mathcal{R}(s,a)^2\\ &= \mathbb{E}_{\substack{S\sim d_\pi\\ A\sim\pi(S,\cdot)}}\left[\mathcal{R}(S,A)^2\right]\end{aligned} \tag{10}$$

We introduce the state-value and action-value functions associated with the square reward under policy π

Definition 3.2 (Average Adjusted Square State-Value Function)**.** *The average adjusted square state-value function $U_\pi : \mathbb{S}\to\mathbb{R}$ is the expected square residual return that can be obtained starting from a state and following policy π*

$$U_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\left(R_{t+1}^2-\eta_\pi\right)\middle| S_0 = s\right] \tag{11}$$

Definition 3.3 (Average Adjusted Square Action-Value Function)**.** *The average adjusted square action-value function $Q_\pi : \mathbb{S}\times\mathbb{A}\to\mathbb{R}$ is the expected residual square return that can be obtained starting from a state, taking an action and then following policy π*

$$W_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\left(R_{t+1}^2-\eta_\pi\right)\middle| S_0 = s, A_0 = a\right] \tag{12}$$

The average adjusted square action-value function satisfies the following Bellman equation

$$\eta_\pi + W_\pi(s,a) = \mathcal{R}(s,a)^2 + \int_{\mathbb{S}}\mathcal{P}(s,a,s')U_\pi(s')ds' \tag{13}$$

Moreover, the average adjusted square state-value function is related to the action-value function by the following

$$U_\pi(s) = \int_{\mathbb{A}}\pi(s,a)W_\pi(s,a) \tag{14}$$

Using this equality, we can write an analogous Bellman equation for the square state-value function. In the risk-sensitive setting, the agent wants to find a policy that solves the following optimization problem

$$\begin{cases} \max_\pi \rho_\pi \\ \text{subject to } \Lambda_\pi \leq \alpha \end{cases} \tag{15}$$

for a given $\alpha > 0$. Using the Lagrangian relaxation procedure, we can recast (15) to the following uncostrained problem

$$\max_\lambda \min_\pi L(\theta, \lambda) = -\rho_\pi + \lambda(\Lambda_\theta - \alpha) \tag{16}$$

Let us observe that the discussion can be easily extended to other risk-sensitive performance measures, such as the standard mean-variance criterion or the Sharpe ratio.

4 Risk-Sensitive Policy Gradient

5 Risk-Sensitive Actor-Critic Algorithm

References

[1] S. Mahadevan, *Average reward reinforcement learning: Foundations, algorithms, and empirical results*, Machine learning, 22 (1996), pp. 159–195.

[2] L. Prashanth and M. Ghavamzadeh, *Actor-critic algorithms for risk-sensitive reinforcement learning.*, arXiv preprint arXiv:1403.6530, (2014).