Average Reward Risk-Sensitive Actor-Critic Algorithm

Pierpaolo Necchi

June 12, 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*

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

where d_π is the stationary distribution of the Markov process induced by policy π.

In the risk-neutral setting, the agent aims to find an *average optimal* policy

$$\pi_* = \arg\sup_{\pi} \rho_\pi \tag{2}$$

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*

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

The term $V_\pi(s)$ is usually referred to as the *bias* value, or the *relative* value, since it represents the relative difference in total reward gained starting from a state s as opposed to a generic state. ρ_π serves as a baseline that allows to avoid divergence in the value function definition.

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

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

The action-value function satisfies the following Bellman equation

$$\rho_\pi + Q_\pi(s,a) = \mathcal{R}(s,a) + \int_{\mathbb{S}} \mathcal{P}(s,a,s') V_\pi(s') ds' \tag{6}$$

Using equality (5), we obtain the Bellman equation for the state-value function

$$\rho_\pi + V_\pi(s) = \int_{\mathbb{A}} \pi(s,a) \left[\mathcal{R}(s,a) + \int_{\mathbb{S}} \mathcal{P}(s,a,s') V_\pi(s') ds' \right] da \tag{7}$$

Denoting by T_a (resp. T_π) the transition operator for action a (resp. for policy π)

$$\begin{aligned} T_a V(s) &= \int_{\mathbb{S}} \mathcal{P}(s,a,s') V_\pi(s') \\ T_\pi V(s) &= \int_{\mathbb{A}} \pi(s,a) \int_{\mathbb{S}} \mathcal{P}(s,a,s') V_\pi(s') \end{aligned} \tag{8}$$

The Bellman equations can be rewritten in the shorter form

$$\begin{aligned} \rho_\pi + V_\pi(s) &= \mathcal{R}_\pi(s) + T_\pi V_\pi(s) \\ \rho_\pi + Q_\pi(s,a) &= \mathcal{R}(s,a) + T_a V_\pi(s) \end{aligned} \tag{9}$$

In the discrete case, where the transition operator correspond to matrices, these Bellman equations become linear systems that can be solved to obtain the value functions.

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)**.** *The long-run variance Λ_π under policy π*

$$\begin{aligned}\Lambda_\pi &= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi\left[\sum_{t=0}^{T-1}(R_{t+1}-\rho_\pi)^2\right] \\ &= \mathbb{E}_{\substack{S\sim d_\pi \\ A\sim\pi}}\left[(\mathcal{R}(S,A)-\rho_\pi)^2\right] \\ &= \int_{\mathbb{A}} d_\pi(s)\int_{\mathbb{A}}\pi(s,a)\left(\mathcal{R}(s,a)-\rho_\pi\right)^2 dads\end{aligned} \tag{10}$$

The long-run variance can be decomposed as follows

$$\Lambda_\pi = \eta_\pi - \rho_\pi^2 \tag{11}$$

where η_π is the average square reward per step

Definition 3.2 (Average Square Reward)**.**

$$\begin{aligned}\eta_\pi &= \mathbb{E}_{\substack{S\sim d_\pi \\ A\sim\pi}}\left[\mathcal{R}(S,A)^2\right] \\ &= \int_{\mathbb{S}} d_\pi(s)\int_{\mathbb{A}}\pi(s,a)\mathcal{R}(s,a)^2\end{aligned} \tag{12}$$

As before, we introduce the residual state-value and action-value functions associated with the square reward under policy π

Definition 3.3 (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*

$$U_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\left(R_{t+1}^2-\eta_\pi\right)\middle| S_0 = s\right] \tag{13}$$

Definition 3.4 (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*

$$W_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\left(R_{t+1}^2-\eta_\pi\right)\middle| S_0 = s, A_0 = a\right] \tag{14}$$

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

$$\eta_\pi + W_\pi(s,a) = \mathcal{R}(s,a)^2 + \int_{\mathbb{S}} \mathcal{P}(s,a,s')U_\pi(s')ds' \tag{15}$$

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

$$U_\pi(s) = \int_{\mathbb{A}} \pi(s,a)W_\pi(s,a) \tag{16}$$

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

$$\begin{cases} \max_\pi \rho_\pi \\ \text{subject to } \Lambda_\pi \leq \alpha \end{cases} \tag{17}$$

for a given $\alpha > 0$. Using the Lagrangian relaxation procedure, we can recast (17) to the following uncostrained problem

$$\max_\lambda \min_\pi L(\pi,\lambda) = -\rho_\pi + \lambda(\Lambda_\theta - \alpha) \tag{18}$$

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

Let us consider a family of parametrized policies π_θ, with $\theta \in \Theta \subseteq \mathbb{R}^{D_\theta}$. The optimization problem then becomes

$$\max_\lambda \min_\theta L(\theta,\lambda) = -\rho(\theta) + \lambda(\Lambda(\theta) - \alpha) \tag{19}$$

Using a policy gradient approach, the policy parameters are updated following the gradient ascent direction

$$\nabla_\theta L(\theta,\lambda) = -\nabla_\theta \rho(\theta) + \lambda \nabla_\theta \Lambda(\theta) \tag{20}$$

while the Lagrange multiplier is updated following the gradient descent direction

$$\nabla_\lambda L(\theta,\lambda) = \Lambda(\theta) - \alpha \tag{21}$$

Since

$$\nabla_\theta \Lambda(\theta) = \nabla_\theta \eta(\theta) - 2\rho(\theta)\nabla_\theta \rho(\theta) \tag{22}$$

it is enough to compute $\nabla_\theta \eta(\theta)$ and $\nabla_\theta \rho(\theta)$. These quantities are provided by the policy gradient theorem

Theorem 4.1 (Policy Gradient)**.**

$$\begin{aligned}\nabla_\theta \rho(\theta) &= \mathbb{E}_{\substack{S\sim d_\pi \\ A\sim\pi}} \left[\nabla_\theta \log \pi_\theta(S,A) Q_\theta(S,A)\right] \\ &= \int_{\mathbb{S}} d_\pi(s) \int_{\mathbb{A}} \pi(s,a) \nabla_\theta \log \pi_\theta(s,a) Q_\theta(s,a) da ds\end{aligned} \tag{23}$$

$$\begin{aligned}\nabla_\theta \eta(\theta) &= \mathbb{E}_{\substack{S\sim d_\pi \\ A\sim\pi}} \left[\nabla_\theta \log \pi_\theta(S,A) W_\theta(S,A)\right] \\ &= \int_{\mathbb{S}} d_\pi(s) \int_{\mathbb{A}} \pi(s,a) \nabla_\theta \log \pi_\theta(s,a) W_\theta(s,a) da ds\end{aligned} \tag{24}$$

Proof. Eq. (23) is the standard policy gradient and its proof can be found in the literature, e.g. [3]. Eq. (24) can be shown in a similar fashion. TODO □

As in the standard risk-neutral case, a state-dependent baseline can be introduced in both gradients without changing the result. In particular, by using the average adjusted value functions as baseline, we can replace the average adjusted action-value functions with the following advantage functions

$$A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s) \tag{25}$$

$$B_\theta(s,a) = W_\theta(s,a) - U_\theta(s) \tag{26}$$

Intuitively, the advantage functions measure how good is to take action a in state s compared to simply following the policy π_θ. The use of a baseline also serves the purpose of reducing the variance of the gradient estimates. Thus, the gradients can be written as

$$\nabla_\theta \rho(\theta) = \mathbb{E}_{\substack{S\sim d_\pi \\ A\sim\pi}} \left[\nabla_\theta \log \pi_\theta(S,A) A_\theta(S,A)\right] \tag{27}$$

$$\nabla_\theta \eta(\theta) = \mathbb{E}_{\substack{S\sim d_\pi \\ A\sim\pi}} \left[\nabla_\theta \log \pi_\theta(S,A) B_\theta(S,A)\right] \tag{28}$$

Results of this type form the basis of many actor-critic algorithms, which employ an approximation of the advantage function to obtain a more accurate estimate of the objective function.

5 Risk-Sensitive Actor-Critic Algorithm

In [2], starting from (27) and (28), the authors derive a risk-sensitive actor-critic algorithm for the average reward setting. In the algorithm proposed, the advantage functions are approximated using a temporal difference (TD) scheme in which a critic maintains a linear approximation of the value functions. More in detail, let δ_n^A and δ_n^B be the TD errors for residual value and square value functions

$$\begin{aligned}\delta_t^A &= R_{t+1} - \widehat{\rho}_{t+1} + \widehat{V}(S_{t+1}) - \widehat{V}(S_t) \\ \delta_t^B &= R_{t+1}^2 - \widehat{\eta}_{t+1} + \widehat{U}(S_{t+1}) - \widehat{U}(S_t)\end{aligned} \tag{29}$$

where $\widehat{V}$, $\widehat{U}$, $\widehat{\rho}$ and $\widehat{\eta}$ are unbiased estimate of V_θ, U_θ, $\rho(\theta)$ and $\eta(\theta)$ respectively. It is easy to show that δ_t^A and δ_t^B are unbiased estimates of the advantage functions.

Proposition 5.1 (Temporal Difference Errors)**.**

$$\begin{aligned} \mathbb{E}_\theta \left[\delta_t^A | S_t = s, A_t = a\right] &= A_\theta(s,a) \\ \mathbb{E}_\theta \left[\delta_t^B | S_t = s, A_t = a\right] &= B_\theta(s,a) \end{aligned} \tag{30}$$

Proof. TODO □

Denoting by $\psi_t = \psi(S_t, A_t) = \nabla_\theta \log \pi(S_t, A_t)$ the compatible feature [3], we can easily obtain an unbiased estimate of the gradients

$$\begin{aligned} \nabla_\theta \rho(\theta) &\approx \psi_t \delta_t^A \\ \nabla_\theta \eta(\theta) &\approx \psi_t \delta_t^B \end{aligned} \tag{31}$$

The value functions are linearly approximated using some veatures vectors $\Phi_V : \mathbb{S} \to \mathbb{R}^{D_V}$ and $\Phi_U : \mathbb{S} \to \mathbb{R}^{D_U}$ as follows

$$\begin{aligned} \widehat{V}(s) &= v^T \Phi_V(s) \\ \widehat{U}(s) &= u^T \Phi_U(s) \end{aligned} \tag{32}$$

Given all these ingredients, the authors propose a three time-scale stochastic approximation algorithm whose pseudo-code is illustrated in 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).

[3] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour, et al., *Policy gradient methods for reinforcement learning with function approximation.*, in NIPS, vol. 99, 1999, pp. 1057–1063.