Actor Critic Agent

Marco Marini

May 19, 2020

Abstract

1 TD Error

The TD error is defined as

$$\delta_t = r_t - r_\pi + v_\pi(s_{t+1}) - v_\pi(s_t) \tag{1}$$

2 Policy actor

The policy actor estimate the probabilities $\pi_a(s)$ of choose action a at status s. The function is the softmax of the actions preferences $h_a(s)$

$$\pi(a, s) = \frac{e^{h_a(s)}}{\sum_k e^{h_k(s)}} \tag{2}$$

simplifying the notation with

$$\begin{aligned} \pi(a, s) &= \pi_a \\ h_a(s) &= h_a \end{aligned}$$

The update of policy gradient is

$$\begin{aligned} \nabla \ln \pi_a &= \frac{1}{\pi_a} \frac{\partial}{\partial h_a} \pi_a \\ &= \frac{1}{\pi_a (\sum_k e^{h_k})^2} \left[e^{h_a} \nabla h_a - e^{h_a} \nabla \sum_k e^{h_k} \right] \\ &= \frac{1}{\sum_k e^{h_k}} \left[\nabla h_a - \sum_k \nabla e^{h_k} \right] \\ &= \frac{1}{\sum_k e^{h_k}} \left[\nabla h_a - \sum_k e^{h_k} \nabla h_k \right] \end{aligned}$$

Let be

$$A_i(a) = 1, \Rightarrow i = a$$
$$A_i(a) = 0 \Rightarrow i \neq a$$

then

$$\begin{aligned}\nabla \ln \pi_a &= \frac{1}{\sum_k e^{h_k}} \sum_i \left[A_i(a) - e^{h_i}\right] \nabla h_i \\ &= \sum_i \left[\frac{A_i(a)}{\sum_k e^{h_k}} - \pi_i\right] \nabla h_i\end{aligned}$$

The backwork propagated TD error to the output neural network is

$$\begin{aligned}\delta_{h_a}(t) &= \delta(t)\nabla \ln \pi_a \\ &= \delta(s_t) \sum_i \left[\frac{A_i(a_t)}{\sum_k e^{h_k}} - \pi_i\right]\end{aligned} \tag{3}$$

The updated actor preferences are

$$h_a^*(s_t) = h_a(s_t) + \alpha_h \delta_{h_a}(t) \tag{4}$$

3 Gaussian policy actor

The Gaussian policy actor estimate the probabilities $\pi(a, s)$ of choose a continuous action a at status s as a normal distributed function of two parameters $\mu(s)$ and $\sigma(s)$.

We change the notation to avoid ambiguity between the constant $\pi = 3.14\ldots$ and the policy $\pi(a, s)$:

$$\begin{aligned}\pi(a, s) &= p(a, s) \\ &= \frac{1}{\sigma(s)\sqrt{2\pi}} e^{-\frac{(a-\mu(s))^2}{\sigma(s)^2}}\end{aligned} \tag{5}$$

$$\sigma(s) = e^{h_\sigma(s)} \tag{6}$$

Futhermore we change the notation to simplifying the equation:

$$\begin{aligned}p(a, s) &= p \\ \mu(s) &= \mu \\ \sigma(s) &= \sigma \\ h_\sigma(s) &= h_\sigma\end{aligned}$$

to infere the gradient of logarithm of p

$$\nabla \ln p = \left(\frac{\partial}{\partial \mu} + \frac{\partial}{\partial h_\sigma} \right) \ln p$$

the partial derivative by μ is

$$\begin{aligned}
\frac{\partial}{\partial \mu} \ln p &= \frac{1}{p} \frac{\partial p}{\partial \mu} \\
&= \frac{1}{p\sigma\sqrt{2\pi}} e^{\frac{-(a-\mu)^2}{\sigma^2}} \frac{\partial}{\partial \mu} \left[-\frac{(a-\mu)^2}{\sigma^2} \right] \\
&= -\frac{1}{\sigma^2} [2(x-\mu)(-1))] \\
&= \frac{2}{\sigma^2} (x-\mu) \\
\frac{\partial}{\partial h_\sigma} \ln p &= \frac{1}{p} \frac{\partial p}{\partial \sigma} \frac{\partial \sigma}{\partial h_\sigma} \\
\frac{\partial p}{\partial \sigma} &= \frac{1}{\sigma^2 \sqrt{2\pi}} \left\{ \sigma \frac{\partial}{\partial \sigma} \left[e^{-\frac{(a-\mu)^2}{\sigma^2}} \right] - e^{-\frac{(a-\mu)^2}{\sigma^2}} \right\} \\
&= \frac{p}{\sigma} \left\{ \sigma \frac{\partial}{\partial \sigma} \left[-(x-\mu)^2 \sigma^{-2} \right] - 1 \right\} \\
&= \frac{p}{\sigma} \left[-\sigma(a-\mu)^2(-2\sigma^{-3}) - 1) \right] \\
&= \frac{p}{\sigma} \left[2(a-\mu)^2 \sigma^{-2} - 1 \right] \\
\frac{\partial \sigma}{\partial h_\sigma} &= \sigma \frac{\partial}{\partial h_\sigma} \ln p \\
&= \frac{1}{p} \frac{p}{\sigma} \left[2(a-\mu)^2 \sigma^{-2} - 1 \right] \sigma \\
&= 2(a-\mu)^2 \sigma^{-2} - 1
\end{aligned}$$

The backward propagated TD errors to the output network layer are

$$\delta_\mu(t) = \frac{2}{\sigma^2}(x-\mu)\delta(t) \tag{7}$$

$$\delta_{h_\sigma}(t) = \left[2\frac{(x-\mu)^2}{\sigma^2} - 1 \right] \delta(t) \tag{8}$$

The updated actor parameters are:

$$\mu^*(s_t) = \mu(s_t) + \alpha_\mu \delta_\mu(t) \tag{9}$$

$$h^*_\sigma(s_t) = h_\sigma(s_t) + \alpha_{h_\sigma} \delta_{h_\sigma}(t) \tag{10}$$

4 Performance Indicators

The agent has a lot of iper parameters to tune and optimize the learning rate.

In this section we define performance indicators to tune such parameters.

4.1 Critic Indicator

The critic computes the updated value of current state by appling the the bootstrap equation:

$$v^*(s_t) = v(s_{t+1}) + r_t - r_\pi$$

The ratio of MSE after and before the learning activity indicates the quality of such activity.

$$\begin{aligned} J_v(s_t) &= [v^*(s_t) - v(s_t)]^2 \\ J'_v(s_t) &= [v^*(s_t) - v'(s_t)]^2 \\ K_v(s_t) &= \frac{J'_v(s_t)}{J_v(s_t)} \end{aligned} \tag{11}$$

A ratio $K_v(s_t) \geq 1$ means a step-size parameter α too high and a ratio $K_v(s_t) \ll 1$ means a step-size parameter too low with very poor capacity of learning.

Because the $J(s_t)$ should approach to 0 in optimal conditions, we should take into consideration only the steps that have a $J(s_t) > \varepsilon$.

In learning session we can evaluate the value of perfomance indicator and adjust the step-size parameter accordingly. We want that a fraction p of all the steps have a K_v indicator less than 1, we calculate $K_{v,p}$ the p centile of K_v and correct the step-size parameter by a factor

$$\eta_v = \frac{1}{K_{v,p}} \tag{12}$$

4.2 Policy Actor Indicators

The actor computes the updated preferernces of current state by adding a step-size parameter to gradient and TD error

$$h^*_a(s_t) = h_a(s_t) + \alpha_h \delta_{h_a}(t)$$

To avoid comuptation overflow the preferences are constratints to a limited range e.g. $(-7, +7)$. The changes of preferences should also be limited to a fraction of the range $(-\varepsilon_h, \varepsilon_h)$, so we can measure the squared distance of changes of preferences:

$$\begin{aligned} J_h(s_t) &= \sum_a [h^*_a(s_t) - h_a(s_t)]^2 \\ &= \alpha_h^2 \sum_a \delta_{h_a}^2 \end{aligned} \tag{13}$$

A $J_h(s_t) \geq \varepsilon_h^2$ means a α_h parameter value too high and $J_h(s_t) \ll \varepsilon_h^2$ means a α_h parameter value too small, producing an uneffective changes on preferences.

We can correct the α_h parameter multiplying it by a γ_h factor so that the corrected $J_h(s_t)$ is equal to ε_h^2, we have

$$\begin{aligned}
\varepsilon_h^2 &= (\gamma_h \alpha_h)^2 \sum_a \delta_{h_a}^2 \\
&= \gamma_h^2 J_h(s_t) \\
\gamma_h &= \frac{\varepsilon_h}{\sqrt{J_h(s_t)}}
\end{aligned}$$

Asserting we want to have a p fraction of samples with a $J_h(s_t) < \varepsilon_h^2$, we calculate $J_{h,p}$ the p centile of $J_h(s_t)$ and comupute the γ_h

$$\gamma_h = \frac{\varepsilon_h}{\sqrt{J_{h,p}}} \tag{14}$$

The actor than adjusts the network to fit the updated preferences. The same performace indicator defined for the critic is used for each action prference of actor:

$$\begin{aligned}
J_h(s_t) &= \sum_a (h_a^*(s_t) - h_a(s_t))^2 \\
J_h'(s_t) &= \sum_a (h_a^*(s_t) - h_a'(s_t))^2 \\
K_h(s_t) &= \frac{J_h'(s_t)}{J_h(s_t)}
\end{aligned}$$

$$\eta_h = \frac{1}{K_{h,p}} \tag{15}$$

4.3 Gaussian Policy Actor Indicators

The actor computes the updated parameters of current state by adding a step-size parameter to gradient and TD error

$$\begin{aligned}
\delta_\mu &= \frac{2}{\sigma^2}(a - \mu)\delta \\
\delta_{h_\sigma} &= \left[2\frac{(a-\mu)^2}{\sigma^2} - 1 \right] \delta
\end{aligned}$$

The updated gaussian parameters are

$$\begin{aligned}
\mu^*(s_t) &= \mu(s_t) + \alpha_\mu \delta_\mu(s_t) \\
h_\sigma^*(s_t) &= h_\sigma(s_t) + \alpha_{h_\sigma} \delta_{h_\sigma}(s_t)
\end{aligned}$$

We may consider the changes to the gaussian policy parameter limited to a defined range

$$\begin{aligned}
J_\mu(s_t) &< \varepsilon_\mu^2 \\
|\mu^*(s_t) - \mu(s_t)| &< \varepsilon_\mu \\
|\delta_\mu(s_t)| &< \varepsilon_\mu \\
J_{h_\sigma}(s_t) &< \varepsilon_{h_\sigma}^2 \\
[h_\sigma^*(s_t) - h_\sigma(s_t)]^2 &< \varepsilon_{h_\sigma}^2 \\
|\delta_{h_\sigma}(s_t)| &< \varepsilon_{h_\sigma}
\end{aligned}$$

An indicator $J_\mu(s_t) \geq \varepsilon_\mu^2$ means an α_μ parameter value too high, on the other hand an indicator$J_\mu(s_t) \ll \varepsilon_\mu^2$ means an α_μ parameter value too small. An indicator $J_{h_\sigma}(s_t) \geq \varepsilon_{h_\sigma}^2$ means an α_σ parameter value too high and an indicator $J_{h_\sigma}(s_t) \ll \varepsilon_{h_\sigma}^2$ means an α_{h_σ} parameter value too small.

Asserting we want to have a p fraction of samples with a $J_\mu(s_t) < \varepsilon_\mu^2$, we calculate $J_{\mu,p}$ the p centile of $J_\mu(s_t)$ and comupute the γ_μ

$$\begin{aligned}
\varepsilon_\mu^2 &= \gamma_\mu^2 \sigma_\mu^2 \\
&= \gamma_\mu^2 J_\mu(s_t)
\end{aligned}$$

$$\gamma_\mu = \frac{\varepsilon_\mu}{\sqrt{J_{\mu,p}}} \tag{16}$$

In the same way we have

$$\gamma_{h_\sigma} = \frac{\varepsilon_{h\sigma}}{\sqrt{J_{h_\sigma,p}}} \tag{17}$$

For the agent network we have:

$$\begin{aligned}
J_\mu(s_t) &= (\mu^*(\mu_t) - \mu(s_t))^2 \\
J'_\mu(s_t) &= (\mu^*(s_t) - \mu'(s_t))^2 \\
K_\mu(s_t) &= \frac{J'_\mu(s_t)}{J_\mu(s_t)}
\end{aligned}$$

$$\eta_\mu = \frac{1}{K_{\mu,p}} \tag{18}$$

and

$$\begin{aligned}
J_{h_\sigma}(s_t) &= (h_\sigma^*(\mu_t) - h_\sigma(s_t))^2 \\
J'_{h_\sigma}(s_t) &= (h_\sigma^*(s_t) - h'_\sigma(s_t))^2 \\
K_{h_\sigma}(s_t) &= \frac{J'_{h_\sigma}(s_t)}{J_{h_\sigma}(s_t)}
\end{aligned}$$

$$\eta_{h_\sigma} = \frac{1}{K_{h_\sigma,p}} \tag{19}$$