Actor Critic Agent

Marco Marini

July 12, 2020

Abstract

Description of Actor Critic Agent and definitions of performance indicators

Contents

1 Actor Critic Agent

The Policy Actor Critic Agent is composed by an artificial neural network (ANN) that generates the differential state value of critic component and the policy parameters for action selection of each actor (the action space dimensions) from the state input signals. At each iteration with the environment, the agent computes the average reward r_π and the differential TD error δ_t

$$\delta_t = R_{t+1} - \bar{R} + \hat{v}(S_{t+1}) - \hat{v}(S_t) \tag{1}$$

The differential TD error indicate how different is the differential state value determined by the current reward from the estimated differential state value.

If the error is positive the agent changes the policy reinforcing the current strategy. It increments the probabilities of selected actions and at the same time reduces the probabilities of the other actions.

If the error is negative the agent changes the policy weakening the current strategy. It reduces the probabilities of selected actions and increments the probabilities of the other actions.

The magnitude of changes is proportional to the TD Error by a α_i parameter for each i actor that allows to control the convergence of the algorithm.

The TD error is used to determine the updated output values and train the AAN with these expected outputs.

1.1 Discrete Policy Actor

The policy actor computes the probabilities $\pi(a|s)$ of choose action a at status s. The probabilities are the softmax function of the action preferences $h(s, a, \vec{\theta})$ given by the ANN

$$\pi(a|s, \vec{\theta}) = \frac{e^{h_a(s,a,\vec{\theta})}}{\sum_{b \in A} e^{h(s,b,\vec{\theta})}} \tag{2}$$

To train the ANN the agent updates the $\vec{\theta}$ parameters at each step t

$$\vec{\theta}_{t+1} = \vec{\theta}_t + \alpha \delta_t \nabla \ln \pi(A_t|S_t, \vec{\theta}_t) \tag{3}$$

The policy gradient function is

$$\begin{aligned}
\nabla \ln \pi(A_t|S_t, \vec{\theta}) &= \frac{1}{\pi(A_t|S_t, \vec{\theta})} \nabla \pi(A_t|S_t, \vec{\theta}) \\
\pi(a|s, \vec{\theta}) &= \frac{e^{h_a(s,a,\vec{\theta})}}{D(s, \vec{\theta})} \\
D(s, \vec{\theta}) &= \sum_{b \in A} e^{h(s,b,\vec{\theta})} \\
\nabla \pi(A_t|S_t, \vec{\theta}) &= \frac{1}{D^2(S_t, \vec{\theta})} \left[D(S_t, \vec{\theta}) \nabla e^{h(S_t,A_t,\vec{\theta})} - e^{h(S_t,A_t,\vec{\theta})} \nabla D(S_t, \vec{\theta}) \right] \\
&= \frac{1}{D^2(S_t, \vec{\theta})} \left[D(S_t, \vec{\theta}) e^{h(S_t,A_t,\vec{\theta})} \nabla h(S_t, A_t, \vec{\theta}) - e^{h(S_t,A_t,\vec{\theta})} \nabla D(S_t, \vec{\theta}) \right] \\
&= \frac{\pi(A_t|S_t, \vec{\theta})}{D(S_t, \vec{\theta})} \left[D(S_t, \vec{\theta}) \nabla h(S_t, A_t, \vec{\theta}) - \nabla D(S_t, \vec{\theta}) \right] \\
\nabla D(s, \vec{\theta}) &= \sum_{b \in A} \nabla e^{h(s,b,\vec{\theta})} \\
&= \sum_{b \in A} e^{h(s,b,\vec{\theta})} \nabla h(s, b, \vec{\theta}) \\
\nabla \pi(A_t|S_t, \vec{\theta}) &= \frac{\pi(A_t|S_t, \vec{\theta})}{D(S_t, \vec{\theta})} \left[D(S_t, \vec{\theta}) \nabla h(S_t, A_t, \vec{\theta}) - \sum_{b \in A} e^{h(s,b,\vec{\theta})} \nabla h(s, b, \vec{\theta}) \right]
\end{aligned}$$

Let be

$$\begin{gathered}
I(a, b) = 1, \Rightarrow a = b \\
I(a, b) = 0 \Rightarrow a \neq b
\end{gathered}$$

then

$$\sum_{b \in A} I(At, b) \nabla h(St, b), \vec{\theta}) = \nabla h(St, At), \vec{\theta})$$

$$\begin{aligned}
\nabla \pi(A_t|S_t, \vec{\theta}) &= \frac{\pi(A_t|S_t, \vec{\theta})}{D(S_t, \vec{\theta})} \sum_{b \in A} \left[D(S_t, \vec{\theta}) I(A_t, b) - e^{h(S_t, b, \vec{\theta})} \right] \nabla h(S_t, b, \vec{\theta}) \\
\nabla \ln \pi(A_t|S_t, \vec{\theta}) &= \sum_{b \in A} \left[I(A_t, b) - \pi(S_t, b, \vec{\theta}) \right] \nabla h(S_t, b, \vec{\theta}) \\
\delta_h(S_t, A_t, b, \theta_t) &= \delta_t \left[I(A_t, b) - \pi(S_t, b, \vec{\theta}) \right] \\
\vec{\theta}_{t+1} &= \vec{\theta}_t + \sum_{b \in A} \alpha \delta_h(S_t, A_t, b, \theta_t) \nabla h(S_t, b, \vec{\theta}_t)
\end{aligned}$$

The same result is obtained by training the network with labels

$$h^*(S_t, A_t, b) = h(S_t, b, \vec{\theta}_t) + \alpha \delta_h(S_t, A_t, b, \vec{\theta}_t) \tag{4}$$

1.2 Gaussian policy actor

The Gaussian policy actor computes the probabilities $\pi(a|s)$ of choose a continuous action a at status s as a normal distributed function of two parameters $\mu(s, \vec{\theta}_\mu)$ and $\sigma(s, \vec{\theta}_h)$.

$$\pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h) = \frac{1}{\sigma(s, \vec{\theta}_h)\sqrt{2\pi}} e^{-\frac{(a-\mu(s,\vec{\theta}_\mu))^2}{\sigma(s,\vec{\theta}_h)^2}} \tag{5}$$

$$\sigma(s, \vec{\theta}_h) = e^{h(s, \vec{\theta}_h)} \tag{6}$$

To train the ANN the agent updates the $\vec{\theta}_\mu, \vec{\theta}_h$ parameters at each step t

$$\vec{\theta}_{\mu t+1} = \vec{\theta}_{\mu t} + \alpha_\mu \delta_t \nabla \ln \pi(A_t|S_t, \vec{\theta}_{\mu t}, \vec{\theta}_{ht}) \tag{7}$$

$$\vec{\theta}_{h+1} = \vec{\theta}_{ht} + \alpha_h \delta_t \nabla \ln \pi(A_t|S_t, \vec{\theta}_{\mu t}, \vec{\theta}_{ht}) \tag{8}$$

The policy gradient functions for $\vec{\theta}_\mu$ is

$$\begin{aligned}
\nabla \ln \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h) &= \pi^{-1}(a|s, \vec{\theta}_\mu, \vec{\theta}_h) \nabla \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h) \\
&= \pi^{-1}(a|s, \vec{\theta}_\mu, \vec{\theta}_h) \frac{\partial \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h)}{\partial \mu(s, \vec{\theta}_\mu)} \nabla \mu(s, \vec{\theta}_\mu) \\
\frac{\partial \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h)}{\partial \mu(s, \vec{\theta}_\mu)} &= -\frac{1}{\sigma^3(s, \vec{\theta}_h)\sqrt{2\pi}} e^{-\frac{(a-\mu(s,\vec{\theta}_\mu))^2}{\sigma^2(s,\vec{\theta}_h)}} \frac{\partial}{\partial \mu(s, \vec{\theta}_\mu)} (a - \mu(s, \vec{\theta}_\mu))^2 \\
&= -\pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h)(-2)(a - \mu(s, \vec{\theta}_\mu))\sigma^{-2}(s, \vec{\theta}_h) \\
\nabla \ln \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h) &= \frac{2[a - \mu(s, \vec{\theta}_\mu)]}{\sigma^2(s, \vec{\theta}_h)} \nabla \mu(s, \vec{\theta}_\mu)
\end{aligned}$$

The policy gradient functions for $\vec{\theta}_h$ is

$$\begin{aligned}
\nabla \ln \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h) &= \pi^{-1}(a|s, \vec{\theta}_\mu, \vec{\theta}_h) \nabla \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h) \\
&= \pi^{-1}(a|s, \vec{\theta}_\mu, \vec{\theta}_h) \frac{\partial \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h)}{\partial h(s, \vec{\theta}_h)} \nabla h(s, \vec{\theta}_h) \\
\frac{\partial \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h)}{\partial h(s, \vec{\theta}_h)} &= \frac{\partial \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h)}{\partial \sigma(s, \vec{\theta}_h)} \frac{\partial \sigma(s, \vec{\theta}_h)}{\partial h(s, \vec{\theta}_h)} \\
\frac{\partial \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h)}{\partial \sigma(s, \vec{\theta}_h)} &= \frac{\left\{ \sigma(s, \vec{\theta}_h) \frac{\partial}{\partial \sigma(s, \vec{\theta}_h)} \left[e^{-\frac{(a-\mu(s, \vec{\theta}_\mu))^2}{\sigma^2(s, \vec{\theta}_h)}} \right] - e^{-\frac{(a-\mu(s, \vec{\theta}_\mu))^2}{\sigma^2(s, \vec{\theta}_h)}} \right\}}{\sigma^2(s, \vec{\theta}_h) \sqrt{2\pi}} \\
&= \frac{e^{-\frac{(a-\mu(s, \vec{\theta}_\mu))^2}{\sigma^2(s, \vec{\theta}_h)}} \left\{ -\sigma(s, \vec{\theta}_h)(a - \mu(s, \vec{\theta}_\mu))^2 \frac{\partial}{\partial \sigma(s, \vec{\theta}_h)} \left[\sigma^{-2}(s, \vec{\theta}_h) \right] - 1 \right\}}{\sigma^2(s, \vec{\theta}_h) \sqrt{2\pi}} \\
&= \frac{\pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h)}{\sigma(s, \vec{\theta}_h)} \left\{ 2 \frac{(a - \mu(s, \vec{\theta}_\mu))^2}{\sigma^2(s, \vec{\theta}_h)} - 1 \right\} \\
\frac{\partial \sigma(s, \vec{\theta}_h)}{\partial h(s, \vec{\theta}_h)} &= e^{h(s, \vec{\theta}_h)} \\
&= \sigma(s, \vec{\theta}_h) \\
\nabla \ln \pi(a|s, \vec{\theta}_\mu, \vec{\theta}_h) &= 2 \frac{(a - \mu(s, \vec{\theta}_\mu))^2}{\sigma^2(s, \vec{\theta}_h)} - 1
\end{aligned}$$

Let be

$$\begin{aligned}
\delta_\mu(s, a, \vec{\theta}_\mu, \vec{\theta}_h) &= \delta_t 2 \frac{a - \mu(s, \vec{\theta}_\mu)}{\sigma^2(s, \vec{\theta}_h)} \\
\delta_h(s, a, \vec{\theta}_\mu, \vec{\theta}_h) &= \delta_t \left[2 \frac{(a - \mu(s, \vec{\theta}_\mu))^2}{\sigma^2(s, \vec{\theta}_h)} - 1 \right]
\end{aligned}$$

The $\vec{\theta}_\mu, \vec{\theta}_h$ parameters are updated by

$$\begin{aligned}
\vec{\theta}_{\mu t+1} &= \vec{\theta}_{\mu t} + \alpha_\mu \delta_\mu(S_t, A_t, \vec{\theta}_{\mu t}, \vec{\theta}_{ht}) \\
\vec{\theta}_{ht+1} &= \vec{\theta}_{ht} + \alpha_h \delta_h(S_t, A_t, \vec{\theta}_{\mu t}, \vec{\theta}_{ht})
\end{aligned}$$

The same result is obtained by training the network with labels

$$\mu^*(S_t) = \mu(S_t, \vec{\theta}_{\mu t}) + \alpha_\mu \delta_\mu(S_t, A_t, \vec{\theta}_{\mu t}, \vec{\theta}_{ht}) \tag{9}$$

$$h^*(S_t) = h(S_t, \vec{\theta}_{\mu t}) + \alpha_h \delta_h(S_t, A_t, \vec{\theta}_{\mu t}, \vec{\theta}_{ht}) \tag{10}$$

1.3 Learning process

Initially the agent do not have any experience and the differential state value generated by the ANN starts from a random initial estimation. The same is for the behavior generated by the actors where the preferences are random initialized.

Since the very first interactions the agent generates wrong differential state value estimation with high value of TD Error. Consequently the agent easily changes the policy behavior, trying different policies with the exploration of both the policy and state space. It tries to improve the average reward, the value state estimation at the same time selecting better policies.

Different factors influence the learning process. External agent factors like the number of states, actions and policies not yet explored tent to generate high error rate. The error increases as the agent experiences new environment states or try new policies, but as the agent faces states already visited it can exploit the past experiences reducing the error and improving the state value estimation and the policy.

Instead the agent internal factor, such as the ANN approximation, generates errors that can be reduced only to a low-bound limit value.

The ANN learning parameters can influence both in the quality of approximated functions and in the speed of adaptation.

High value of step-size parameter η increases the correction of weights increasing consequently the speed of adaptation, but can easily generate corrective value too big bringing the network to not converging behaviors. On the other hand small value of step-size parameter weakens the correction of weights decreasing the speed of adaptation.

Also the actor parameters α_i, which modulate the policy changes from the critic, influence the improvement capacity. High value of α_i determine strong policy changes even in the face of small errors, making it difficult a fine correction, the other way around very small value of α_a bring very little variations on policy without improvement.

If the actor parameters that govern the policy have a limited range of values we may choose an $\alpha_i = \alpha_i^*$ so that the average of corrections is ε_i a fraction of the limited range.

For the discrete policy actors we calculate the average of squared preferences variations $\Delta^2 h_i$

$$\begin{aligned}
h^*_{ia}(t) - h_{ia}(t) &= \Delta h_{ia}(t) \\
&= \alpha_i \delta_{h_{ia}}(t) \\
\Delta^2 h_i(t) &= \frac{1}{m} \sum_a^m \Delta h_{ia}(t) \\
\Delta^2 h_i &= \frac{1}{n} \sum_t^n \Delta^2 h_i(t) \\
&= \frac{1}{n} \frac{1}{m} \sum_t^n \sum_a^m \Delta^2 h_{ia}(t) \\
&= \alpha_i^2 \frac{1}{n} \frac{1}{m} \sum_t^n \sum_a^m \delta^2_{h_{ia}}(t) \\
\Delta^2 h_i &= \varepsilon_i^2 \\
\alpha^*_i &= \frac{\varepsilon_i}{\sqrt{\frac{1}{n} \frac{1}{m} \sum_t^n \sum_a^m \delta^2 h_{ia}(t)}}
\end{aligned}$$

if α_i is the value used in the session then

$$\alpha^*_i = \frac{\varepsilon_i}{\sqrt{\frac{1}{n} \frac{1}{m} \sum_t^n \sum_a^m \Delta^2 h_{ia}(t)}} \alpha_i \tag{11}$$

For the Gaussian policy actors we calculate the average of squared parameter variations $\Delta^2 \mu_i$ and $\Delta^2 h_{i\sigma}$

$$\begin{aligned}
\mu^*_i(t) - \mu_i(t) &= \Delta \mu_i(t) \\
&= \alpha_{i\mu} \delta_{i\mu}(t) \\
\Delta^2 \mu_i &= \frac{1}{n} \sum_t^n \Delta^2 \mu_i(t) \\
&= \alpha_{i\mu}^2 \frac{1}{n} \sum_t^n \delta^2_{\mu_i}(t) \\
\Delta^2 \mu_i &= \varepsilon_{i\mu}^2 \\
\alpha^*_{i\mu} &= \frac{\varepsilon_{i\mu}}{\sqrt{\frac{1}{n} \sum_t^n \delta^2 \mu_i(t)}}
\end{aligned}$$

if $\alpha_{i\mu}$ is the value used in the session then

$$\alpha^*_{i\mu} = \frac{\varepsilon_{i\mu}}{\sqrt{\frac{1}{n} \sum_t^n \Delta^2 \mu_i(t)}} \alpha_{i\mu} \tag{12}$$

$$
\begin{aligned}
h^*_{i\sigma}(t) - h_{i\sigma}(t) &= \Delta h_{i\sigma}(t) \\
&= \alpha_{i\sigma}\delta_{ih_\sigma}(t) \\
\Delta^2 h_{i\sigma} &= \frac{1}{n}\sum_t^n \Delta^2 h_{i\sigma}(t) \\
&= \alpha^2_{i\sigma}\frac{1}{n}\sum_t^n \delta^2_{ih_\sigma}(t) \\
\Delta^2 h_{i\sigma} &= \varepsilon^2_{i\sigma} \\
\alpha^*_{i\sigma} &= \frac{\varepsilon_{i\sigma}}{\sqrt{\frac{1}{n}\sum_t^n \delta^2 h_{i\sigma}(t)}}
\end{aligned}
$$

if $\alpha_{i\sigma}$ is the value used in the session then

$$\alpha^*_{i\sigma} = \frac{\varepsilon_{i\sigma}}{\sqrt{\frac{1}{n}\sum_t^n \Delta^2 h_{i\sigma}(t)}}\alpha_{i\sigma} \tag{13}$$

1.4 Performance

The average reward is the main indicators of learning quality, since the agent interacts with the environment the average reward grows up.

Because the process is adaptive the average reward can vary a lot during the interaction with the environment therefore it is necessary to evaluate the trend of the indicator. The logarithmic regression can be used to evaluate the performance

$$r_\pi(t) = \ln(mt + q) \tag{14}$$

An increasing trend indicates the agent is selecting better policies and is getting higher and higher rewards. A decreasing trend indicates the agent is not selecting good policies, this may be caused by low α_a or η parameters.

Another indicator of performance is the squared TD error, as for the average reward it is necessary to evaluated the trend. The error has not negative values that should approach zero when the agent select the local best policy. The exponential regression can be used to evaluate the performance

$$\delta^2(t) = e^{(mt+q)} \tag{15}$$

A decreasing trend indicates the estimation of value states is getting better. An increasing trend indicates the estimation of value states is getting worse it may be caused by the ANN which cannot correct the estimation due to η parameter too high.

Both the critic and actors approximate the value function and policy function with neural network. To monitor the learning activity of ANN we compute the MSE of estimated functions.

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

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

The square error of critic output is

$$\delta^2(t) = [v^*(s_t) - v(s_{t+1})]^2$$

The square error of discrete policy actor outputs are

$$J_h(t) = \sum_i \sum_a^m (h^*_{ia}(s_t) - h_{ia}(s_t))^2$$

and in case of Gaussian actor

$$J_h(t) = \sum_i [\mu^*_i(s_t) - \mu_i(s_t)]^2 + [h^*_{i\sigma}(s_t) - h_{i\sigma}(s_t)]^2$$

where i is index of the actor.

Every time the network is trained the label outputs are fed to correct the weights of ANN. The resulting network has therefore a MSE that should be less than the original

$$\delta'^2(t) = [v^*(s_t) - v'(s_{t+1})]^2$$

$$J'_h(t) = \sum_i \sum_a^m (h^*_{ia}(s_t) - h'_{ia}(s_t))^2$$

for the discrete policy agent and

$$J'_h(t) = \sum_i [\mu^*_i(s_t) - \mu'_i(s_t)]^2 + [h^*_{i\sigma}(s_t) - h'_{i\sigma}(s_t)]^2$$

for the Gaussian actors.

The ratio between the total MSE after and before the training activity indicates the quality of such activity.

$$\begin{aligned} J(t) &= \delta^2(t) + J_h(t) \\ J'(t) &= \delta'^2(t) + J'_h(t) \\ K(t) &= \frac{J'(t)}{J(t)} \end{aligned} \tag{16}$$

A ratio $K(t) \geq 1$ means the error after training gets worst due a step-size parameter α too high. A ratio $K(t) = 1$ means no change on error and therefore no improvement. This can be affected by a local minimum reached or a step-size parameter too low with very poor capacity of learning. A ratio $K(t) < 1$ means an improvement of neural network due to correct step-size parameter. A ratio $K(t) = 0$ means a perfect fit of neural network.

We can classify the steps in three class:

C_0 The steps that created a bad approximation with an increased of MSE ($J > \varepsilon \cup K > 1$)

C_1 The steps that create a trivial approximation with a small reduction of MSE ($J > \varepsilon \cup K_0 \leq K \leq 1$ with $K_0 = 0.9$)

C_2 The remaining steps that have a small MSE or that have reduced significantly the MSE

The ideal distribution should have

$$\begin{aligned} C_0 &= 0 \\ C_1 &= 0 \\ C_2 &= 1 \end{aligned} \tag{17}$$

A step parameter η too high generates over correction increasing C_0 and reducing C_2.

An η parameter too small generates under corrections increasing C_1 and reducing C_2.

Because it is difficult to determine the effects of step parameter changes on the result MSE, an empirical way to reduce or increment the parameter is applied, for example increasing or reducing by exponential factors (..., ×0.01, ×0.03, ×0.1, ×0.3, ×3, ×10, ×30, ×100, ...).

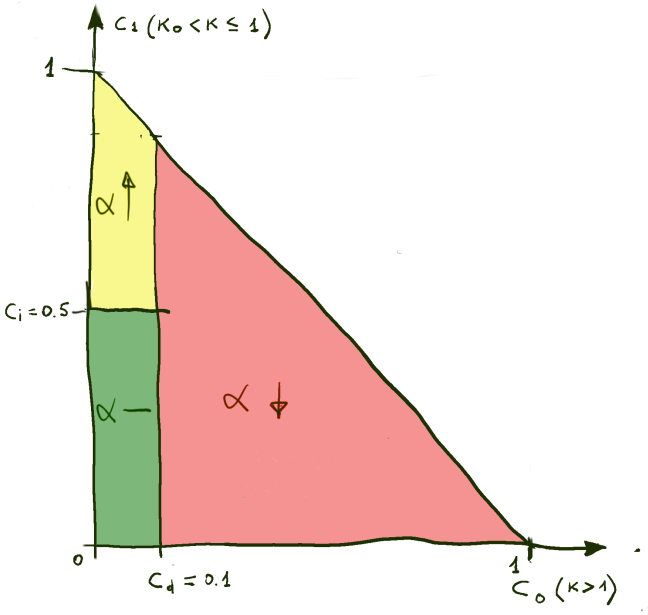

Figure 1: Step classification