Deep Reinforcement Learning

Aditya Rastogi

April 8, 2020

Abstract

Till now we have studied tabular methods in reinforcement learning and we have also understood various model free as well as model based policy and control methods including TD, Monte-Carlo and Dynamic Programming methods. We have also learned how we can use function approximators to scale RL based solutions to large and continuous spaces. In this paper, we will discuss how we can integrate deep learning based methods with novel ideas in reinforcement learning to further scale up RL algorithms.

Contents

1 Policy Gradients

Let's start with the definition of a trajectory. We begin our focus with terminating episodes. A trajectory is a tuple of states and actions that an agent visits from the start till the end of an episode. Hence, the probability of some trajectory τ will be:

$$\begin{aligned} p(\tau) &= p(s_1, a_1, ..., s_T, a_T) \\ &= p(s_1) \prod_{t=1}^{T} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t) \end{aligned}$$

Now, if the policy $\pi(a_t|s_t)$ depends on some parameters θ just like we had in function approximation methods, then $p_\theta(\tau)$ can be written like this:

$$p_\theta(\tau) = p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t) \tag{1}$$

1.1 The goal

Throughout the agent's trajectory, its goal is to maximize its reward. But the trajectory itself is stochastic and hence we focus on maximizing its expected reward. We write the expected reward as follows:

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right] \tag{2}$$

As we stated above, our goal is to maximize $J(\theta)$ and the knobs we can adjust in order to do so are θ. So, our goal is to obtain some θ^* such that the following holds:

$$\theta^* = \arg\max_\theta J(\theta) \tag{3}$$

In order to maximize $J(\theta)$, we would determine $\nabla_\theta J(\theta)$, and would apply any gradient based optimization algorithm like gradient ascent to do parameter updates. Let's proceed by expanding (2) a bit more.

$$\begin{aligned} J(\theta) &= \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right] \\ &= \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[r(\tau) \right] \\ &= \int p_\theta(\tau) r(\tau) d\tau \end{aligned}$$

where $r(\tau)$ is:

$$r(\tau) = \sum_{t=1}^{T} r(s_t, a_t) \tag{4}$$

Therefore, $\nabla_\theta J(\theta)$ would be:

$$\nabla_\theta J(\theta) = \int \nabla_\theta p_\theta(\tau) r(\tau) d\tau \tag{5}$$

$$\nabla_\theta J(\theta) = \int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau) r(\tau) d\tau$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau) r(\tau) \right] \tag{6}$$

So, we need to find $\nabla_\theta p_\theta(\tau)$. To proceed, let's play with (1) a bit.

$$p_\theta(\tau) = p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

It will be easier to get $\nabla_\theta p_\theta(\tau)$ if all the terms are in sum form rather than we dealing with these terms in product. So, let's take log both sides to make things easier.

$$\log p_\theta(\tau) = \log p(s_1) + \sum_{t=1}^{T} \log \pi_\theta(a_t|s_t) + \sum_{t=1}^{T} \log p(s_{t+1}|s_t, a_t)$$

Now, taking gradient with respect to θ both sides, we get:

$$\nabla_\theta \log p_\theta(\tau) = \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \tag{7}$$

$$\frac{\nabla_\theta p_\theta(\tau)}{p_\theta(\tau)} = \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)$$

$$\nabla_\theta p_\theta(\tau) = p_\theta(\tau) * \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \tag{8}$$

Putting (8) in (5), we get:

$$\begin{aligned} \nabla_\theta J(\theta) &= \int \left[p_\theta(\tau) * \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right] r(\tau) d\tau \\ &= \int p_\theta(\tau) \left[\left\{ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right\} r(\tau) \right] d\tau \end{aligned}$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left\{ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right\} r(\tau) \right] \tag{9}$$

Putting (4) in (9), we get:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left\{ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right\} \left\{ \sum_{t=1}^{T} r(s_t, a_t) \right\} \right] \tag{10}$$

1.2 Naive REINFORCE Algorithm

The reinforce algorithm is pretty straightforward. We initialize the policy network with random parameters. Then at each iteration, we sample a trajectory by using this policy network to interact with the environment.

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left[\left\{ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \right\} \left\{ \sum_{t=1}^{T} r(s_{i,t}, a_{i,t}) \right\} \right] \tag{11}$$

Then we use eqn. (11) to calculate $\nabla_\theta J(\theta)$ using which we make the following update:

$$\theta \leftarrow \theta + \alpha * \nabla_\theta J(\theta) \tag{12}$$

Note that the above equation is adding the term $\alpha * \nabla_\theta J(\theta)$ to θ because the learning algorithm is gradient ascent, as we want to maximize $J(\theta)$.

1.3 Towards reducing variance in gradient of expected reward function

Eqn. (10) contains a lot of random variables and the values in the two sums can be vastly different. This results in high variance. Let's have a look at eqn. (10) again and try to find equivalent forms of it with reduced variance.

$$\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left\{ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right\} \left\{ \sum_{t=1}^{T} r(s_t, a_t) \right\} \right] \\
&= \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left\{ \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \right\} \left\{ \sum_{t'=1}^{T} r(s_{t'}, a_{t'}) \right\} \right] \\
&= \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{T} \sum_{t'=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) r(s_{t'}, a_{t'}) \right]
\end{aligned}$$

$$\nabla_\theta J(\theta) = \sum_{t=1}^{T} \sum_{t'=1}^{T} \left[\mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) r(s_{t'}, a_{t'}) \right] \right] \tag{13}$$

$$\nabla_\theta J(\theta) = \sum_{t=1}^{T}\sum_{t'=1}^{T}\left[\mathbb{E}_{(s_t,a_t,s_{t'},a_{t'})}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) r(s_{t'},a_{t'})\right]\right] \tag{14}$$

where (14) follows from

$$\mathbb{E}_{X,Y}\left[\rho(X)\right] = \mathbb{E}_X\left[\rho(X)\right]$$

Proceeding with (14), we get

$$\nabla_\theta J(\theta) = \sum_{t=1}^{T}\sum_{t'=1}^{T}\left[\mathbb{E}_{(s_{t'},a_{t'})}\left[r(s_{t'},a_{t'})\mathbb{E}_{(s_t,a_t)}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)|s_{t'},a_{t'}\right]\right]\right] \tag{15}$$

Consider the inner expectation term:

$$\mathbb{E}_{(s_t,a_t)}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)|s_{t'},a_{t'}\right] = \int\int p(s_t,a_t|s_{t'},a_{t'})\nabla_\theta \log \pi_\theta(a_t|s_t) da_t ds_t \tag{16}$$

Now, we consider $p(s_t,a_t|s_{t'},a_{t'})$,
If $t > t'$, we have

$$p(s_t,a_t|s_{t'},a_{t'}) = \pi_\theta(a_t|s_t) * p(s_t|s_{t'},a_{t'}) \tag{17}$$

Putting (17) in (16), we get for $t > t'$,

$$\begin{aligned}
\mathbb{E}_{(s_t,a_t)}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)|s_{t'},a_{t'}\right] &= \int\int \pi_\theta(a_t|s_t)p(s_t|s_{t'},a_{t'})\nabla_\theta \log \pi_\theta(a_t|s_t) da_t ds_t \\
&= \int p(s_t|s_{t'},a_{t'})\left[\int \pi_\theta(a_t|s_t)\nabla_\theta \log \pi_\theta(a_t|s_t) da_t\right] ds_t \\
&= \int p(s_t|s_{t'},a_{t'})\left[\int \nabla_\theta \pi_\theta(a_t|s_t) da_t\right] ds_t \\
&= \int p(s_t|s_{t'},a_{t'})\left[\nabla_\theta \int \pi_\theta(a_t|s_t) da_t\right] ds_t \\
&= \int p(s_t|s_{t'},a_{t'})\left[\nabla_\theta(1)\right] ds_t \\
&= \int \left[p(s_t|s_{t'},a_{t'}) * 0\right] ds_t \\
&= 0
\end{aligned}$$

Therefore (15) becomes:

$$\nabla_\theta J(\theta) = \sum_{t=1}^{T}\sum_{t'=t}^{T}\left[\mathbb{E}_{(s_{t'},a_{t'})}\left[r(s_{t'},a_{t'})\mathbb{E}_{(s_t,a_t)}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)|s_{t'},a_{t'}\right]\right]\right]$$

And (10) becomes:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{T} \left\{ \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t}^{T} r(s_{t'}, a_{t'}) \right\} \right] \tag{18}$$

Also, we see that

$$\begin{aligned}
\mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{T} \{ \nabla_\theta \log \pi_\theta(a_t|s_t) * c \} \right] &= c * \sum_{t=1}^{T} \left[\mathbb{E}_{(s_t, a_t)} \nabla_\theta \log \pi_\theta(a_t|s_t) \right] \\
&= c * \sum_{t=1}^{T} \left[\int \int p(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) da_t ds_t \right] \\
&= c * \sum_{t=1}^{T} \left[\int p(s_t) \left[\int \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t) da_t \right] ds_t \right] \\
&= c * \sum_{t=1}^{T} \left[\int p(s_t) \left[\int \nabla_\theta \pi_\theta(a_t|s_t) da_t \right] ds_t \right] \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau) b \right] &= b * \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau) \right] \\
&= b * \int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau) d\tau \\
&= b * \int \nabla_\theta p_\theta(\tau) d\tau \\
&= b * \nabla_\theta \left[\int p_\theta(\tau) d\tau \right] \\
&= 0
\end{aligned}$$

Hence, we can also write (18) as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{T} \left[\nabla_\theta \log \pi_\theta(a_t|s_t) \left\{ \left(\sum_{t'=t}^{T} r(s_{t'}, a_{t'}) \right) - b \right\} \right] \right] \tag{19}$$

where b is some constant.

And equivalently, we can write (6) as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log p_\theta(\tau) \left[r(\tau) - b \right] \right] \tag{20}$$

Now, what value of b should we choose? Since, the above equation holds for any value of b, we choose b such that the variance of the inner term is minimized.

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Here,

$$X = \nabla_\theta \log p_\theta(\tau) \left[r(\tau) - b\right]$$

So, Var(X) becomes,

$$\begin{aligned} Var(X) &= \mathbb{E}\left[\left[\nabla_\theta \log p_\theta(\tau) \left[r(\tau) - b\right]\right]^2\right] - \left[\mathbb{E}\left[\nabla_\theta \log p_\theta(\tau) \left[r(\tau) - b\right]\right]\right]^2 \\ &= \mathbb{E}\left[\left[\nabla_\theta \log p_\theta(\tau) \left[r(\tau) - b\right]\right]^2\right] - \left[\mathbb{E}\left[\nabla_\theta \log p_\theta(\tau) r(\tau)\right]\right]^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial Var(X)}{\partial b} &= \frac{\partial \mathbb{E}\left[\left[\nabla_\theta \log p_\theta(\tau) \left[r(\tau) - b\right]\right]^2\right]}{\partial b} \\ 0 &= \mathbb{E}\left[2 * \left[\nabla_\theta \log p_\theta(\tau) \left\{r(\tau) - b\right\}\right] * \nabla_\theta \log p_\theta(\tau) * (-1)\right] \\ 0 &= \mathbb{E}\left[\left[\nabla_\theta \log p_\theta(\tau) \left\{r(\tau) - b\right\}\right] * \nabla_\theta \log p_\theta(\tau)\right] \\ 0 &= \mathbb{E}\left[\left\{\nabla_\theta \log p_\theta(\tau)\right\}^2 \left\{r(\tau) - b\right\}\right] \\ 0 &= \mathbb{E}\left[\left\{\nabla_\theta \log p_\theta(\tau)\right\}^2 r(\tau)\right] - b * \mathbb{E}\left[\left\{\nabla_\theta \log p_\theta(\tau)\right\}^2\right] \\ b * \mathbb{E}\left[\left\{\nabla_\theta \log p_\theta(\tau)\right\}^2\right] &= \mathbb{E}\left[\left\{\nabla_\theta \log p_\theta(\tau)\right\}^2 r(\tau)\right] \end{aligned}$$

$$b = \frac{\mathbb{E}\left[\left\{\nabla_\theta \log p_\theta(\tau)\right\}^2 r(\tau)\right]}{\mathbb{E}\left[\left\{\nabla_\theta \log p_\theta(\tau)\right\}^2\right]} \tag{21}$$

Putting (4) and (7) in (21), we get

$$b = \frac{\mathbb{E}\left[\left\{\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\right\}^2 \left\{\sum_{t=1}^{T} r(s_t, a_t)\right\}\right]}{\mathbb{E}\left[\left\{\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\right\}^2\right]} \tag{22}$$

How would eqn (19) look like if the reward is received only at the last time step? Say that reward is r_T.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)}\left[\left\{r_T - b\right\}\left\{\nabla_\theta \left[\sum_{t=1}^{T} \log \pi_\theta(a_t|s_t)\right]\right\}\right] \tag{23}$$

where b would be:

$$b = \frac{\mathbb{E}\left[\left\{\nabla_\theta \left[\sum_{t=1}^{T} \log \pi_\theta(a_t|s_t)\right]\right\}^2 r_T\right]}{\mathbb{E}\left[\left\{\nabla_\theta \left[\sum_{t=1}^{T} \log \pi_\theta(a_t|s_t)\right]\right\}^2\right]} \tag{24}$$

If we set

$$\phi = \nabla_\theta \left[\sum_{t=1}^{T} \log \pi_\theta(a_t|s_t)\right] \tag{25}$$

we get

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[(r_T - b)\, \phi \right] \tag{26}$$

and b as

$$b = \frac{\mathbb{E}\left[\phi^2 r_T\right]}{\mathbb{E}\left[\phi^2\right]} \tag{27}$$

And therefore, we can approximate these expectations as:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left[(r_{T,i} - b)\, \phi_i \right] \tag{28}$$

and

$$b \approx \frac{\sum_{i=1}^{N} \left[\phi_i^2 * r_{T,i}\right]}{\sum_{i=1}^{N} \left[\phi_i^2\right]} \tag{29}$$