Chapter 12 – REINFORCE

Author: Gianmarco Scarano

gianmarcoscarano@gmail.com

1. Introduction

Recalling the policy gradient theorem (infinite setting), we said that we can estimate this through this object over here:

$$\nabla_\theta J(\pi_\theta) = \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s,a\sim\mathbb{P}_h^{\pi_\theta}} \nabla_\theta \ln \pi_\theta(a\,|\,s) \cdot Q^{\pi_\theta}(s,a)$$
$$= \frac{1}{1-\gamma} \mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \left[\nabla_\theta \ln \pi_\theta(a\,|\,s) \cdot Q^{\pi_\theta}(s,a) \right]$$

The problem, now, is to get $Q^{\pi_\theta}(s,a)$, but this looks like a familiar problem, where we can use the return G as an unbiased estimate of Q (MC).

Here's where REINFORCE comes in:

2. REINFORCE

Initialize policy parameters θ arbitrarily
for each episode $\{s_1, a_1, r_2, \cdots, s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$ **do**
  for $t = 1$ to $T-1$ **do**
    $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) G_t$
  endfor
endfor
return θ

This is good for us, since it's an unbiased estimation of Q, but this has a high variance problem, which we could avoid by introducing baselines (function of the state).

2.1 Baseline

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s,a\sim d_\mu^{\pi_\theta}} \left[\nabla_\theta \ln \pi_\theta(a\,|\,s) \cdot \left(Q^{\pi_\theta}(s,a) - b(s) \right) \right]$$

We are influencing policy gradient in some way but we are not adding bias. Let's see:

Through some expansions, explicating of the expectation as well as the gradient and through simple mathematic manipulations, we can infer at the end that $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \nabla_\theta \log_n \pi_\theta(a|s) \cdot b(s) = 0$, so the baselines do not introduce bias.

Since the baselines have to be action-independent, a common choice is the Value Function. Our $b(s)$ then becomes:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s,a\sim d^{\pi_\theta}} \left[\nabla_\theta \ln \pi_\theta(a\,|\,s) \left(Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) \right) \right]$$

Where $Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$ is called **ADVANTAGE FUNCTION**, which we denote as $A^{\pi_{\theta_t}}(s,a)$.

This advantage function tells us how good an action is compared to the average value of the state.

Namely, $Q^{\pi_\theta}(s,a)$ is the value of an action in the state, while $V^{\pi_\theta}(s)$ is the AVERAGE value of the state.

2.2 REINFORCE with Baseline

Initialize policy parameter θ, baseline b
for iteration=$1, 2, \cdots$ **do**
  Collect a set of trajectories by executing the current policy
  At each timestep t in each trajectory τ^i, compute
    Return $G_t^i = \sum_{t'=t}^{T-1} r_{t'}^i$, and
    Advantage estimate $\hat{A}_t^i = G_t^i - b(s_t)$.
  Re-fit the baseline, by minimizing $\sum_i \sum_t ||b(s_t) - G_t^i||^2$,
  Update the policy, using a policy gradient estimate $\hat{g}$,
    Which is a sum of terms $\nabla_\theta \log \pi(a_t|s_t, \theta)\hat{A}_t$.
    (Plug $\hat{g}$ into SGD or ADAM)
endfor

We're still using the return and collecting MC samples

Actually, if we can access the true value function, the Temporal Difference (TD) error is an unbiased estimate of the advantage function.

For this reason, we can use the Temporal Difference (TD) error to compute the policy gradient.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta (s, a) \cdot \delta^{\pi_\theta}]$$

$$\delta^{\pi_\theta} = r + \gamma \cdot V^{\pi_\theta}(s') - V^{\pi_\theta}(s)$$

Introduces bias, but it's fine as we know from the Temporal Difference lessons. This, though, leads us to the notion of critic.

2.3 Reducing variance with Critic

Let's remember that we had a very major problem when computing the estimate G: high variance.

We can then, use an estimate V/Q by using a Critic which is also parameterized.

$$Q_w(s, a) \approx Q^{\pi_\theta}(s, a)$$
$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a)\ Q_w(s, a)]$$
$$\Delta\theta = \alpha \nabla_\theta \log \pi_\theta(s, a)\ Q_w(s, a)$$

We can select any blend between TD and MC estimators for Q_w

Actor-Critic Algorithms use 1-step TD

- **MC Policy gradient:**
  - Here the target is the return G.
  - $Q_w(s, a) = (G_t - V_v(s_t))$
- **Actor-Critic**
  - Here the target is a Temporal Difference target and relies on bootstrapping
  - We can also choose different timescales $(n - step)$ or $TD(\lambda)$ with forward/backward view
  - $Q_w(s, a) = (r + \gamma \cdot V_v(s_{t+1}) - V_v(s_t))$
  - Could be also applied using LFA / Eligibility traces

Policy Gradient Summary

$$\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) G_t] && \text{REINFORCE} \\
&= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a)\ Q^w(s, a)] && \text{Q Actor-Critic} \\
&= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a)\ A^w(s, a)] && \text{Advantage Actor-Critic} \\
&= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a)\ \delta] && \text{TD Actor-Critic} \\
&= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a)\ \delta e] && \text{TD}(\lambda) \text{ Actor-Critic}
\end{aligned}$$

Critic does policy evaluation to estimate Q, V or A using bootstrapping (*if it uses MC we do not call it a critic*)