Actor-Critic Model and the KGRE-Rec Problem

Based on our MDP Formulation, the goal is to learn a stochastic policy π that maximizes the expected cumulative reward, conditioned on a particular initial user node in the knowledge graph.

$$J(\theta) = \mathbb{E}_{\pi}[\sum_{t=0}^{T-1} \gamma^t R_{t+1} | u]$$

Let's define J(θ) as the expected sum of rewards the agent obtains in every timestep, going from t=0 to t=T, following a policy π(θ).

$$J(\theta) = \mathbb{E}[\sum_{t=0}^{T} R(s_t, a_t) | \pi(\theta))]$$

If we define the episode (finite-horizon case) as a trajectory going from t=0 to t=T, then J(θ) is the sum over all trajectories, of the probability that is selected according to θ, times the reward obtained on this trajectory . This leads us to the following expression for J(θ) -

$$J(\theta) = \mathbb{E}[\sum_{t=0}^{T} R(s_t, a_t) | \pi(\theta))] = \sum_{\tau} P(\tau|\theta) R(\tau)$$

Goal: Find the set of parameters θ, which maximizes J(θ)

We shall use calculus, and a sleight-of-hand trick to compute the derivative of J(θ), in order to maximize J(θ)

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_{\tau} P(\tau|\theta) R(\tau) = \sum_{\tau} \nabla_\theta P(\tau|\theta) R(\tau) = \sum_{\tau} P(\tau|\theta) \frac{\nabla_\theta P(\tau|\theta)}{P(\tau|\theta)} R(\tau)$$

$$\nabla_\theta J(\theta) = \sum_{\tau} P(\tau|\theta) \nabla_\theta \log P(\tau|\theta) R(\tau)$$

Since it is practically impossible to compute every trajectory , we will use a reduced number (only m) trajectories. On selection of the particular m trajectories, the probability term in the sum vanishes, since the trajectories have now been selected and there is no longer any uncertainty in that regard. Also, the expectation becomes an arithmetic average over the m trajectories.

$$\nabla_\theta J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^i|\theta) R(\tau^i)$$

Now, let's simplify (or complicate?) the expression inside the logarithm. We notice that the probability that a trajectory is followed is equal to the product of probabilities that each step of the trajectory is reached when following a stochastic policy $\pi(\theta)$

$$P(\tau|\theta) = \prod_{t=0}^{T} P(s_{t+1}|s_t, a_t)\pi_\theta(a_t|s_t)$$

Next, we turn the logarithm of products into a sum of logarithms

$$\nabla_\theta \log P(\tau|\theta) = \nabla_\theta \log[\prod_{t=0}^{T} P(s_{t+1}|s_t, a_t)\pi_\theta(a_t|s_t)]$$

$$\nabla_\theta \log P(\tau|\theta) = \nabla_\theta \sum_{t=0}^{T} \log P(s_{t+1}|s_t, a_t) + \nabla_\theta \sum_{t=0}^{T} \log \pi_\theta(a_t|s_t)$$

The first term on the RHS vanishes, as $P(s_{t+1}|s_t, a_t)$ does not depend on θ

$$\nabla_\theta \log P(\tau|\theta) = \nabla_\theta \sum_{t=0}^{T} \log \pi_\theta(a_t|s_t)$$

To conclude, the final expression for the gradient of J(θ) is as follows -

$$\nabla_\theta J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) R(\tau^i)$$

What does the above expression really mean? It's calculating the variations of the policy π with the set of parameters, θ. These variations are (amplified by $R(^i)$, i.e. shifted towards the trajectory with a higher magnitude of rewards.

What's wrong with this expression? Consider a trajectory wherein the rewards at early steps are negative, and at later steps are positive, such that the total reward is zero. In this case, this trajectory won't contribute to the gradient and the network won't learn any new values for the parameters.

To fix the above issue, we use the following modified reward function, which calculates discounted rewards starting from the *current state* to the terminal state.

$$R_t = \sum_{j=t}^{T} \gamma^j r_j$$

So, the gradient now becomes -

$$\nabla_\theta J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) R_t$$

To reduce the variance, and to make subsequent gradient updates stable (and smaller), the cumulative reward is reduced by a baseline, b_t

$$\nabla_\theta J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)(R_t - b_t)$$

A good choice for the baseline b_t could be the value function, $V(s_t)$, which returns the value of the state s_t - and R_t is the reward as a result of taking action at at step t, which is very similar to $Q(s_t,a_t)$, i.e. the state-action value given action at taken at step st. So, we end up with:

$$\nabla_\theta J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)(Q(s_t, a_t) - V(s_t))$$

where, π***(a\|s) is the actor, and Q(s_t,a_t) - V(s_t), the critic.***