1 Policy Gradient

Let π_ϕ be a policy of our agent with parameters ϕ. That is in our case a neural network that given an observation returns a probability distribution over actions.
Let $s_0 \sim S()$, $s_{i+1} \sim S(s_i, a_i)$ be a probability distribution over the states in our environment (we assume that the environment is Markovian).
Let $r_i = R(s_i, s_{i+1})$ be a function that gives a state, and the next state returns a reward.

Let us consider the formula below:

$$\mathbb{E}_{\tau \sim P(|\pi_\phi)} \text{reward}(\tau)$$

Where τ is a sequence of states and actions forming a trajectory.
To be more precise

$$\tau = (s_0, a_0, s_1, \ldots, s_n)$$

And $P(|\pi_\phi)$ is the probability distribution of trajectories, given that actions are taken according to π_ϕ

For simplicity let's assume finite state/action and trajectory space. For example, in the case of uncountable state/action space, we would need to replace sums with integrals.

$$\sum_\tau P(\tau|\pi_\phi) \text{reward}(\tau)$$

That is for a given policy π_ϕ the sum above gives us its expected reward. For now, lets assume that

$$\text{reward}(\tau) = \sum_i \gamma^i R(s_i, s_{i+1})$$

As we want to maximize the expected reward achieved by the policy so we want to compute:

$$\nabla_\phi \sum_\tau P(\tau|\pi_\phi)\text{reward}(\tau)$$

What is equal to

$$\sum_\tau \nabla_\phi P(\tau|\pi_\phi)\text{reward}(\tau)$$

We know that

$$\nabla \log(F(x)) = \frac{1}{F(x)}\nabla(F(x))$$

Therefore we can write

$$\sum_\tau \left[\nabla_\phi \log\left(P(\tau|\pi_\phi)\right)\right] \text{reward}(\tau)P(\tau|\pi_\phi)$$

What we can approximate by sampling

$$\mathbb{E}_{\tau\sim P(|\pi_\phi)}\left[\nabla_\phi \log\left(P(\tau|\pi_\phi)\right)\text{reward}(\tau)\right]$$

Let's focus on

$$\nabla_\phi \log\left(P(\tau|\pi_\phi)\right) \text{reward}(\tau)$$

for some trajectory τ and lets unpack the terms

$$P(\tau|\pi_\phi) = S(s_0)\prod_i \left[\pi_\phi(a_i|s_i)S(s_{i+1}|s_i,a_i)\right]$$

$$\text{reward}(\tau) = \sum_i \gamma^i R(s_i,a_i)$$

as $\log(ab) = \log(a) + log(b)$ we have also that

$$\log\left(P(\tau|\pi_\phi)\right) = \log\left(S(s_0)\right) + \sum_i \left[\log(\pi_\phi(a_i|s_i)) + \log(S(s_{i+1}|s_i,a_i))\right]$$

As $\nabla(A+B) = \nabla(A) + \nabla(B)$

$$\nabla_\phi \log\left(P(\tau|\pi_\phi)\right) =$$
$$\nabla_\phi \log\left(S(s_0)\right) + \sum_i \left[\nabla_\phi \log(\pi_\phi(a_i|s_i)) + \nabla_\phi \log(S(s_{i+1}|s_i,a_i))\right]$$

as we cannot influence the probability of state transition when states and actions are already sampled so

$$\nabla_\phi \log\left(P(\tau|\pi_\phi)\right) = \sum_i \nabla_\phi \log(\pi_\phi(a_i|s_i))$$

So we can write

$$\nabla_\phi \log\left(P(\tau|\pi_\phi)\right) \text{reward}(\tau) =$$

$$\left[\sum_i \nabla_\phi \log(\pi_\phi(a_i|s_i))\right]\left[\sum_i \gamma^i R(s_i, a_i)\right]$$

What we can further simplify to

$$\left[\sum_i \gamma^i R(s_i, a_i)\left(\sum_{j=0}^{i} \nabla_\phi \log(\pi_\phi(a_j|s_j))\right)\right]$$

That is we use prob gradients up to the given reward.
Why we can do that, intuitively future should not affect the past but more formally consider a prefix $\hat{\tau} = (s_0, a_1, \ldots, s_i, a_i)$ of a trajectory $\tau = (s_0, a_1, \ldots, s_i, a_i, \ldots s_n)$ our whole formula has the following form

$$\mathbb{E}_{\tau\sim P(|\pi_\phi)} \nabla_\phi \log\left(P(\tau|\pi_\phi)\right)\text{reward}(\tau)$$

and it will sum over all possible futures so for example for the next term, we have that its state will come from $\hat{s_{i+1}} \sim S(|s_i)$ and the following action will come from $\hat{a}_{i+1} \sim \pi_\phi(|\hat{s}_{i+1})$

Therefore we can write

$$\sum_{\hat{s}_{i+1}} \sum_{\hat{a}_{i+1}} S(\hat{s}_{i+1}|s_i)\pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1})\nabla_\phi \log(\pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1}))$$

To be more clear about the part above

- We have fixed a prefix $\hat{\tau} = (s_0, a_1, \ldots, s_i, a_i)$
- We then have looked at

$$\mathbb{E}_{\tau\sim P(|\pi_\phi)} \nabla_\phi \log\left(P(\tau|\pi_\phi)\right)\text{reward}(\tau)$$

- We have focuses on the sum of trajectories τ that have a prefix $\hat{\tau}$
- We have taken a part of this sum that can be written as follows

$$\sum_{\hat{s}_{i+1}} \sum_{\hat{a}_{i+1}} S(\hat{s}_{i+1}|s_i)\pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1})\nabla_\phi \log(\pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1}))$$

Now observe that

$$\sum_{\hat{a}_{i+1}} \pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1}) \nabla_\phi \log(\pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1})) =$$

$$\sum_{\hat{a}_{i+1}} \pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1}) \frac{1}{\pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1})} \nabla_\phi(\pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1})) =$$

$$\sum_{\hat{a}_{i+1}} \nabla_\phi(\pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1})) =$$

$$\nabla_\phi \sum_{\hat{a}_{i+1}} (\pi_\phi(\hat{a}_{i+1}|\hat{s}_{i+1})) =$$

$$\nabla_\phi 1 = 0$$

So we can use just this formula

$$\left[\sum_i \gamma^i R(s_i, a_i) \left(\sum_{j=0}^{i} \nabla_\phi \log(\pi_\phi(a_j|s_j)) \right) \right]$$

We can also rewrite it in reward-to-go format.