CSE6243: Advanced Machine Learning **Fall 2024**

Lecture 23: Policy Gradient and Actor Critic

Lecturer: Bo Dai *Scribes: Malav Patel*

Note: *LaTeX template courtesy of UC Berkeley EECS Department.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

23.1 Recap

In the previous lecture(s) we discussed the Bellman Optimality Equation and the Bellman Expectation Equation. There are two main settings we discussed, those being planning and learning. In the planning stage we are often interested in determining an optimal policy π^*, evaluating a particular policy π (i.e. determine V^π, or doing both (via policy iteration). Let us write a simple two step procedure that describes policy iteration:

Algorithm 1 Policy Iteration

```
Require: initial policy $\pi_1$
for t = 1, ..., T do
    Solve $V^{\pi_t} = R^{\pi_t} + \gamma P^{\pi_t} V^{\pi_t}$ (policy evaluation)
    Update $\pi_{t+1} = \underset{\pi}{arg\,max}\ R^{\pi} + \gamma P^{\pi} V^{\pi_t}$ (policy update)
end for
return $\pi_T$
```

With large enough T we showed that the algorithm converges to V^* and π^*.

23.2 New Content

In the learning setting how does policy iteration change? In this case we no longer have a reward model R or transition matrix P, so the above algorithm must be changed.

23.2.1 Modifying the Policy Update

In the learning setting we are now concerned with a policy update as the solution to the following optimization problem:

$$l(\theta) := \max_{\pi_\theta(\cdot|s)} \mathbb{E}_s \Big[\sum_s \pi_\theta(a|s) R(s,a) + \gamma \int \sum_a P(s'|s,a) \pi_\theta(a|s) V^{\pi_t}(s') ds' \Big] \tag{23.1}$$

From this two questions arise: how do we sample s and how do we do optimization of the above objective (23.1)?

23.2.1.1 How to Optimize the Objective

Using the handy log trick we can compute the gradient as an expectation:

$$\begin{aligned}\nabla\ l(\theta) &= \mathbb{E}_{s,a\sim\pi}\Big[\nabla_\theta \log \pi(a|s)\Big[R(s,a)+\gamma\int P(s'|s,a)V^{\pi_t}ds'\Big]\Big]\\ &= \mathbb{E}_{s,a\sim\pi}[\nabla \log \pi(a|s)Q^{\pi_t}(s,a)]\end{aligned}$$

Where we recognize the term in the nested brackets as the action value function evaluated at (s,a). So, in the learning setting, we can estimate $Q^{\pi_t}(s,a)$. How can we do this estimation? To answer this, recall the Bellman equation for Q:

$$\begin{aligned}Q_\phi^\pi(s,a) &= R(s,a)+\gamma\int P(s'|s,a)\pi(a'|s')Q_\phi^\pi(s',a')ds'da'\\ &= R(s,a)+\gamma\mathbb{E}_{s',a'}\Big[Q_\phi^\pi(s',a')\Big]\end{aligned}$$

From this equation we can define an optimization problem that minimizes the norm of the difference between the LHS and RHS:

$$\begin{aligned}\min_\phi f(\phi) &:= \mathbb{E}_{s,a}\Big[\Big\|Q_\phi^\pi(s,a)-(R(s,a)+\gamma\mathbb{E}_{s',a'}\Big[Q_\phi^\pi(s',a')\Big])\Big\|^2\Big]\\ \nabla f(\phi) &= \mathbb{E}_{s,a}[(Q_\phi^\pi(s,a)-R(s,a)-\gamma\mathbb{E}_{s',a'}[Q_\phi^\pi(s',a')])(\nabla Q_\phi^\pi(s,a)-\mathbb{E}_{s',a'}[\nabla Q_\phi^\pi(s',a')])]\end{aligned}$$

Note that we run into the doubling sampling issue. In order to estimate the gradient of this objective, it is apparent that we will need to run trajectories of states and actions *twice* as seen by the two expectation over (s',a') in the gradient. In practice, this is computationally inefficient as we often have no way of rewinding the environment to resample the state action pair.

In order to solve the double sampling issue we propose the following alternative objective:

$$\min_\phi \max_h\ \mathbb{E}_{s,a}[\|Q_\phi^\pi(s,a)-(R(s,a)+\gamma Q_\phi^\pi(s,a))\|^2]-\mathbb{E}_{s,a,s',a'}[\|h(s,a)-\gamma Q_\phi^\pi(s',a')\|^2]$$

We claim that the minimizer of the new objective is equivalent to that of $f(\phi)$. To see this first note that the optimal function h^* can be found by taking the gradient of the second expectation in function space.

$$\mathbb{E}_{s,a,s',a'}[2(h(s,a)-\gamma Q_\phi^\pi(s',a'))]=0$$

This is solved by setting $h^*(s,a)=\gamma\mathbb{E}_{s',a'|s,a}[Q_\phi^\pi(s',a')]$. Now let us return to the expression for $f(\phi)$:

$$\begin{aligned} f(\phi) &= \mathbb{E}_{s,a}\left[\left\|Q^{\pi}_{\phi}(s,a) - (R(s,a) + \gamma\mathbb{E}_{s',a'}\left[Q^{\pi}_{\phi}(s',a')\right])\right\|^2\right] && (23.2)\\ &= \mathbb{E}_{s,a}\left[\left\|\underbrace{(Q^{\pi}_{\phi}(s,a) - (R(s,a) + \gamma Q^{\pi}_{\phi}(s',a')))}_{a} - \underbrace{(\gamma\mathbb{E}_{s',a'}[Q^{\pi}_{\phi}(s',a')] - \gamma Q^{\pi}_{\phi}(s',a'))}_{b}\right\|^2\right] && (23.3)\\ &= \mathbb{E}_{s,a}[\|a\|^2 - \|b\|^2] && (23.4) \end{aligned}$$

From here it is straightforward to see that the objective is equivalent to the minimax objective defined above.

23.2.1.2 Sampling s

We return to the question of sampling s. This is necessary because we intend to approximate expectations over s in the above equations with Monte Carlo simulation. In this problem, it is sufficient to sample from the stationary distribution of the environment:

$$d^{\pi}(s) = \int \pi(P^{\pi}(s'|s))\mu^0(s)ds'$$

Here $\mu^0(s)$ is the distribution of the initial state, often defined by user. Note that a peculiar complication arises: in the optimization over the parameters above we rely on an estimate of the action value function which depends on the current policy π_t. Upon taking a single gradient step in θ, the policy changes by some amount $\Delta\pi$. As a result, if we solve the full optimization problem (i.e. take many steps of the gradient) our policy may change quite considerably and as a result our gradient estimate which depends on π_t will become less and less accurate. So in essence, at each gradient step, the gradient estimate gets worse and our resulting estimate of the new policy diverges. To prevent this from happening, we opt not to solve the full optimization problem but instead only take one step in the gradient direction before moving to update the action value function. The resulting two step loop resembles the actor-critic update:

Algorithm 2

```
Require: initial policy $\pi_1$
for k = 1, . . . , T do
    $\theta_{k+1} = \theta_k + \eta_k\ \mathbb{E}_{s,a\sim\pi(\cdot|s)}[\nabla \log \pi(a|s) Q^{\pi_k}(s,a)]$ (actor update)
    Update $Q^{\phi_{k+1}}$ (critic update)
end for
return $\pi_T$
```