Computer Science Department - Rutgers University **Fall 2021**

Deriving the Policy Gradient

16:198:605

A policy $\pi_\theta(a|x)$ specifies (as a function of parameters θ) the probability of selecting action a when in state x. For the moment, we'll be taking x to be a discrete state and actions a discrete actions, but a lot of this is going to generalize to continuous spaces as well.

Let $p_0(x)$ be the probability of *starting* in state x at time 0.

We have that the *value* of policy π_θ, or expected (infinite, discounted) reward associated with this policy is given by

$$V(\theta) = \sum_x p_0(x) J_\theta(x). \tag{1}$$

where

$$J_\theta(x) = \sum_a \pi_\theta(a|x) \left[r_{x,a} + \beta \sum_y p^a_{x,y} J_\theta(y) \right]. \tag{2}$$

We'd like to find θ^* that maximizes the value, i.e., the policy that produces the largest possible expected value in application. In order to do this, we want to effectively do a sort of gradient ascent -

$$\theta_{\text{new}} \leftarrow \theta_{\text{old}} + \alpha \nabla_\theta V(\theta_{\text{old}}) \tag{3}$$

In order to do such a thing, we need to be able to calculate or at least estimate $\nabla V(\theta)$. To do so, we need to compute or estimate the $\nabla J_\theta(x)$. But the recursive nature of $J_\theta(x)$ makes that difficult.

$$\begin{aligned} \nabla J_\theta(x) &= \sum_a \left[\nabla \pi_\theta(a|x) \right] \left[r_{x,a} + \beta \sum_y p^a_{x,y} J_\theta(y) \right] + \sum_a \pi_\theta(a|x) \nabla \left[r_{x,a} + \beta \sum_y p^a_{x,y} J_\theta(y) \right] \\ &= \sum_a \left[\nabla \pi_\theta(a|x) \right] J_\theta(x,a) + \sum_a \pi_\theta(a|x) \nabla \left[r_{x,a} + \beta \sum_y p^a_{x,y} J_\theta(y) \right]. \end{aligned} \tag{4}$$

Note the use of $J_\theta(x,a)$ notation to indicate the 'expected utility of taking action a in state x, following policy π_θ beyond that point). Simplifying...

$$\nabla J_\theta(x) = \sum_a \left[\nabla \pi_\theta(a|x) \right] J_\theta(x,a) + \beta \sum_a \pi_\theta(a|x) \sum_y p^a_{x,y} \nabla \left[J_\theta(y) \right]. \tag{5}$$

It's convenient to re-arrange the second sum in the following way:

$$\nabla J_\theta(x) = \sum_a \left[\nabla \pi_\theta(a|x) \right] J_\theta(x,a) + \beta \sum_y \nabla \left[J_\theta(y) \right] \sum_a \pi_\theta(a|x) p^a_{x,y}. \tag{6}$$

Note that the summation on the left can be interpreted nicely, as the probability from transitioning from x to y in one step *under the policy* π_θ. Hence,

$$\nabla J_\theta(x) = \sum_a \left[\nabla \pi_\theta(a|x) \right] J_\theta(x,a) + \beta \sum_y \nabla \left[J_\theta(y) \right] P^\theta_1(x \to y). \tag{7}$$

Note in the above - we want to compute $\nabla J_\theta(x)$. We can compute $\nabla \pi_\theta$ based on the structure of the policy. We can *estimate* $J_\theta(x,a)$ from data. Similarly, we can estimate $P^\theta_1(x \to y)$ from data. The only unknown term in the above (relatively speaking) is $\nabla \left[J_\theta(y) \right]$. We can start to nest these, however. Note that for any $x_0, x_1, x_2, \ldots$:

$$\begin{aligned}
\nabla J_\theta(x_0) &= \sum_{a_0} \left[\nabla \pi_\theta(a_0|x_0) \right] J_\theta(x_0, a_0) + \beta \sum_{x_1} \nabla \left[J_\theta(x_1) \right] P_1^\theta(x_0 \to x_1) \\
\nabla J_\theta(x_1) &= \sum_{a_1} \left[\nabla \pi_\theta(a_1|x_1) \right] J_\theta(x_1, a_1) + \beta \sum_{x_2} \nabla \left[J_\theta(x_2) \right] P_1^\theta(x_1 \to x_2) \\
\nabla J_\theta(x_2) &= \sum_{a_2} \left[\nabla \pi_\theta(a_2|x_2) \right] J_\theta(x_2, a_2) + \beta \sum_{x_3} \nabla \left[J_\theta(x_3) \right] P_1^\theta(x_2 \to x_3) \\
&\ldots
\end{aligned} \tag{8}$$

Substituting in, we can push off the unknown gradient-

$$\begin{aligned}
\nabla J_\theta(x_0) &= \sum_{a_0} \left[\nabla \pi_\theta(a_0|x_0) \right] J_\theta(x_0, a_0) + \beta \sum_{x_1} P_1^\theta(x_0 \to x_1) \nabla \left[J_\theta(x_1) \right] \\
&= \sum_{a_0} \left[\nabla \pi_\theta(a_0|x_0) \right] J_\theta(x_0, a_0) \\
&+ \beta \sum_{x_1} P_1^\theta(x_0 \to x_1) \left[\sum_{a_1} \left[\nabla \pi_\theta(a_1|x_1) \right] J_\theta(x_1, a_1) + \beta \sum_{x_2} \nabla \left[J_\theta(x_2) \right] P_1^\theta(x_1 \to x_2) \right] \\
&= \sum_{a_0} \left[\nabla \pi_\theta(a_0|x_0) \right] J_\theta(x_0, a_0) \\
&+ \beta \sum_{x_1} P_1^\theta(x_0 \to x_1) \sum_{a_1} \left[\nabla \pi_\theta(a_1|x_1) \right] J_\theta(x_1, a_1) \\
&+ \beta^2 \sum_{x_1} \sum_{x_2} P_1^\theta(x_0 \to x_1) P_1^\theta(x_1 \to x_2) \nabla \left[J_\theta(x_2) \right]
\end{aligned} \tag{9}$$

Note that the last sum can be re-arranged, so that $\sum_{x_1} P_1^\theta(x_0 \to x_1) P_1^\theta(x_1 \to x_2) = P_2^\theta(x_0 \to x_2)$, the probability of going from x_0 to x_2 in two steps, under policy π_θ.

$$\begin{aligned}
\nabla J_\theta(x_0) &= \sum_{a_0} \left[\nabla \pi_\theta(a_0|x_0) \right] J_\theta(x_0, a_0) \\
&+ \beta \sum_{x_1} P_1^\theta(x_0 \to x_1) \sum_{a_1} \left[\nabla \pi_\theta(a_1|x_1) \right] J_\theta(x_1, a_1) \\
&+ \beta^2 \sum_{x_2} P_2^\theta(x_0 \to x_2) \nabla \left[J_\theta(x_2) \right]
\end{aligned} \tag{10}$$

We can continue on this way, pushing hte unknown $\nabla J_\theta(x_i)$ further and further into the future:

$$\begin{aligned}
\nabla J_\theta(x_0) &= \sum_{a_0} \left[\nabla \pi_\theta(a_0|x_0) \right] J_\theta(x_0, a_0) \\
&+ \beta \sum_{x_1} P_1^\theta(x_0 \to x_1) \sum_{a_1} \left[\nabla \pi_\theta(a_1|x_1) \right] J_\theta(x_1, a_1) \\
&+ \beta^2 \sum_{x_2} P_2^\theta(x_0 \to x_2) \sum_{a_2} \left[\nabla \pi_\theta(a_2|x_2) \right] J_\theta(x_2, a_2) \\
&+ \beta^3 \sum_{x_3} P_3^\theta(x_0 \to x_3) \sum_{a_3} \left[\nabla \pi_\theta(a_3|x_3) \right] J_\theta(x_3, a_3) \\
&+ \ldots
\end{aligned} \tag{11}$$

or

$$\nabla J_\theta(x_0) = \sum_{k=0}^{\infty} \beta^k \sum_{x_k} P_k^\theta(x_0 \to x_k) \sum_{a_k} \left[\nabla \pi_\theta(a_k|x_k) \right] J_\theta(x_k, a_k) \tag{12}$$

We can tie this all together ultimately in the following way:

$$\nabla V(\theta) = \sum_{x_0} p_0(x_0) \nabla J_\theta(x_0) = \sum_{k=0}^{\infty} \beta^k \sum_{x_k} \sum_{x_0} p_0(x_0) P_k^\theta(x_0 \to x_k) \sum_{a_k} \left[\nabla \pi_\theta(a_k|x_k) \right] J_\theta(x_k, a_k) \tag{13}$$

Noting that $\sum_{x_0} p_0(x_0) P_k^\theta(x_0 \to x_k)$ is simply the probability that the k-th state X_k is x_k, when the steps are executed under policy π_θ,

$$\nabla V(\theta) = \sum_{k=0}^{\infty} \beta^k \sum_{x_k} \mathbb{P}_\theta \left(X_k = x_k \right) \sum_{a_k} \left[\nabla \pi_\theta(a_k|x_k) \right] J_\theta(x_k, a_k) \tag{14}$$

It's worth noting in the above that at this point, everything is technically known, or can be estimated from data collected over simulation. (We would imagine truncating the summation for instance if β^k becomes sufficiently small, or if the game 'end'.) The issue to some extent is the sum over the states x_k and actions a_k. To that end, the above is frequently rewritten in the following way:

$$\begin{aligned}
\nabla V(\theta) &= \sum_{k=0}^{\infty} \beta^k \sum_{x_k} \mathbb{P}_\theta \left(X_k = x_k \right) \sum_{a_k} \pi_\theta(a_k|x_k) \left[\frac{\nabla \pi_\theta(a_k|x_k)}{\pi_\theta(a_k|x_k)} \right] J_\theta(x_k, a_k) \\
&= \sum_{k=0}^{\infty} \beta^k \sum_{x_k} \mathbb{P}_\theta \left(X_k = x_k \right) \sum_{a_k} \pi_\theta(a_k|x_k) \nabla \ln \left[\pi_\theta(a_k|x_k) \right] J_\theta(x_k, a_k) \\
&= \sum_{k=0}^{\infty} \beta^k \sum_{x_k, a_k} \mathbb{P}_\theta \left(X_k = x_k, A_k = a_k \right) \nabla \ln \left[\pi_\theta(a_k|x_k) \right] J_\theta(x_k, a_k) \\
&= \sum_{k=0}^{\infty} \beta^k \mathbb{E} \left[\nabla \ln \left[\pi_\theta(A_k|X_k) \right] J_\theta(X_k, A_k) \right]
\end{aligned} \tag{15}$$

Note in the above, we now have an expression for ∇V in which everything is either computable ($\nabla \pi_\theta$), or estimable from data ($J_\theta(X_k, A_k)$). Implementation of Policy Gradient strategies all take the approach of various estimation schemes of J_θ in order to try to accurately estimate the above from the data and update the policy accordingly.