Big Blank Title for the Project

Author

7th April 2020

1 Softmax PPO Closed Form Update

We will consider direct functional representation with tabular parameterization, i.e. $\pi \equiv p^\pi$ is essentially an S × A table satisfying the constraints

$$\begin{aligned} \sum_a p^\pi(a|s) &= 1, \qquad \forall s \in \mathcal{S} \\ p^\pi(a|s) &\geq 0, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}. \end{aligned}$$

Our goal is to find the closed form solution to the following optimization problem

$$\pi_{t+1} = \arg\max_{\pi \in \Pi} \left[\sum_s d^{\pi_t}(s) \sum_a p^{\pi_t}(a|s) \left(A^{\pi_t}(s,a) + \frac{1}{\eta} \right) \log \frac{p^\pi(s,a)}{p^{\pi_t}(s,a)} \right], \tag{1}$$

subject to the above constraints on p^π.

We begin by formulating this problem using Lagrange multipliers λ_s, $\lambda_{s,a}$ for all states s and actions a:

$$\begin{aligned} \mathcal{L}(p^\pi, \lambda_s, \lambda_{s,a}) = & \sum_s d^{\pi_t}(s) \sum_a p^{\pi_t}(a|s) \left(A^{\pi_t}(s,a) + \frac{1}{\eta} \right) \log \frac{p^\pi(a|s)}{p^{\pi_t}(a|s)} \\ & - \sum_{s,a} \lambda_{s,a} p^\pi(a|s) - \sum_s \lambda_s \left(\sum_a p^\pi(a|s) - 1 \right). \end{aligned} \tag{2}$$

KKT conditions for this problem are:

$$\begin{aligned} \nabla_{p^\pi(x,b)} \mathcal{L}(p^\pi, \lambda_s, \lambda_{s,a}) &= 0, \qquad \forall x \in \mathcal{S}, b \in \mathcal{A} && (3) \\ \sum_a p^\pi(a|s) &= 1, \qquad \forall s \in \mathcal{S} && (4) \\ p^\pi(a|s) &\geq 0, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A} && (5) \\ \lambda_s &\geq 0, \qquad \forall s \in \mathcal{S} && (6) \\ \lambda_s \left(\sum_a p^\pi(a|s) - 1 \right) &= 0 \qquad \forall s \in \mathcal{S} && (7) \\ \lambda_{s,a} p^\pi(a|s) &= 0, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}. && (8) \end{aligned}$$

Let us now try to solve this system. Solving the first equation for arbitrary state-action pair (x,b), gives us:

$$\begin{aligned} \nabla_{p^\pi(b|x)} \mathcal{L}(p^\pi, \lambda_s, \lambda_{s,a}) &= d^{\pi_t}(x) p^{\pi_t}(b|x) \left(A^{\pi_t}(x,b) + \frac{1}{\eta} \right) \frac{1}{p^\pi(b|x)} - \lambda_{x,b} - \lambda_x = 0 \\ \Rightarrow \qquad p^\pi(b|x) &= \frac{d^{\pi_t}(x) p^{\pi_t}(b|x)(1 + \eta A^{\pi_t}(x,b))}{\eta(\lambda_x + \lambda_{x,b})}. \end{aligned} \tag{9}$$

Let us set

$$\lambda_{s,a} = 0, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{10}$$

Then combining Eq. 9 with the second KKT condition gives us

$$\lambda_s = \frac{1}{\eta} \sum_a d^{\pi_t}(s) p^{\pi_t}(a|s)(1 + \eta A^{\pi_t}(s,a)). \tag{11}$$

Therefore, with the additional assumption $d^{\pi_t}(s) > 0$, $p^\pi(a|s)$ becomes

$$p^\pi(a|s) = \frac{p^{\pi_t}(a|s)(1 + \eta A^{\pi_t}(s,a))}{\sum_b p^{\pi_t}(b|s)(1 + \eta A^{\pi_t}(s,b))}. \tag{12}$$

Note that $d^{\pi_t}(s), p^{\pi_t}(a|s) \geq 0$ for any state-action pair, since they are proper measures. Therefore, all that remains is to ensure that

$$1 + \eta A^{\pi_t}(s,a) \geq 0$$

to satisfy the third and fourth KKT conditions. But how to do that? One straightforward way is to define $p^\pi(a|s) = 0$ whenever $1 + \eta A^{\pi_t}(s,a) \leq 0$, and accordingly re-define λ_s. Therefore, this gives us the final solution to our original optimization problem (Eq. 1):

$$\pi_{t+1} = p^\pi(s,a) = \frac{p^{\pi_t}(a|s) \max(1 + \eta A^{\pi_t}(s,a), 0)}{\sum_b p^{\pi_t}(b|s) \max(1 + \eta A^{\pi_t}(s,b), 0)}. \tag{13}$$

This leaves one last problem: Is it always true that given any state s, there exists atleast one action a, such that $1 + \eta A^{\pi_t}(s,a) \geq 0$? Because otherwise, we would fail to satisfy the second KKT condition.