# FMA-PG Notes

Shivam and Sharan
September 2021

## 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 $\times$ A table satisfying the constraints

$$\sum_a p^\pi(a|s) = 1, \quad \forall s \in \mathcal{S}$$
$$p^\pi(a|s) \geq 0, \quad \forall s \in \mathcal{S}, \; \forall a \in \mathcal{A}.$$

Our goal is to find the closed form solution to the following optimization problem (from Eq. 6, Sharan et al., 2021):

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

subject to the constraints on $p^\pi$ given above.

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

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

where we abused the notation by using $\lambda_s$ to represent the set $\{\lambda_s\}_{s \in \mathcal{S}}$ and $\lambda_{s,a}$ to represent the set $\{\lambda_{s,a}\}_{s,a \in \mathcal{S} \times \mathcal{A}}$. The KKT conditions (Theorem 12.1, Nocedal and Wright, 2006) for this constrained optimization problem can be written as:

$$\nabla_{p^\pi(x,b)} \mathcal{L}(p^\pi, \lambda_s, \lambda_{s,a}) = 0, \quad \forall x \in \mathcal{S}, \; \forall b \in \mathcal{A} \tag{C1}$$
$$\sum_a p^\pi(a|s) = 1, \quad \forall s \in \mathcal{S} \tag{C2}$$
$$p^\pi(a|s) \geq 0, \quad \forall s \in \mathcal{S}, \; \forall a \in \mathcal{A} \tag{C3}$$
$$\lambda_s \geq 0, \quad \forall s \in \mathcal{S} \tag{C4}$$
$$\lambda_s \left( \sum_a p^\pi(a|s) - 1 \right) = 0, \quad \forall s \in \mathcal{S} \tag{C5}$$
$$\lambda_{s,a} p^\pi(a|s) = 0, \quad \forall s \in \mathcal{S}, \; \forall a \in \mathcal{A}. \tag{C6}$$

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

$$\begin{aligned}
\nabla_{p^\pi(b|x)} \mathcal{L}(p^\pi, \lambda_s, \lambda_{s,a}) &= d^{\pi_t}(x) p^{\pi_t}(b|x) \left( A^{\pi_t}(x,b) + \frac{1}{\eta} \right) \frac{1}{p^\pi(b|x)} - \lambda_{x,b} - \lambda_x = 0 \\
\Rightarrow \qquad p^\pi(b|x) &= \frac{d^{\pi_t}(x) p^{\pi_t}(b|x)(1 + \eta A^{\pi_t}(x,b))}{\eta(\lambda_x + \lambda_{x,b})}.
\end{aligned} \tag{3}$$

Let us set

$$\lambda_{s,a} = 0, \quad \forall s \in \mathcal{S}, \; \forall a \in \mathcal{A}. \tag{4}$$

Combining Eq. 3 with the second KKT condition gives us

$$\lambda_s = \frac{1}{\eta} \sum_a d^{\pi_t}(s) p^{\pi_t}(a|s)(1 + \eta A^{\pi_t}(s,a)). \tag{5}$$

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

$$p^\pi(a|s) = \frac{p^{\pi_t}(a|s)(1 + \eta A^{\pi_t}(s,a))}{\sum_b p^{\pi_t}(b|s)(1 + \eta A^{\pi_t}(s,b))}. \tag{6}$$

Note that $d^{\pi_t}(s), p^{\pi_t}(a|s) \geq 0$ for any state-action pair, since they are proper measures. 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 $1 + \eta A^{\pi_t}(s,a) < 0$, and accordingly re-define $\lambda_s$. This gives us the final solution to our original optimization problem (Eq. 1):

$$\pi_{t+1} = p^\pi(s,a) = \frac{p^{\pi_t}(a|s) \max(1 + \eta A^{\pi_t}(s,a), 0)}{\sum_b p^{\pi_t}(b|s) \max(1 + \eta A^{\pi_t}(s,b), 0)}. \tag{7}$$

**However, it leaves us one last problem to deal with:** Is it always true that given any state $s$, there always 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. Maybe, we can put a condition on $\eta$ in order to fulfill this constraint.

# 2 MDPO Closed Form Update

The paper (Sharan et al., 2021) considers the direct representation along with tabular parameterization of the policy, albeit with a small change in notation as compared to the previous section: $\pi(a|s) \equiv p^\pi(a|s,\theta)$. However, since this notation is more cumbersome, we will stick with our old notation: $\pi(a|s) \equiv p^\pi(a|s)$. The constraints on the parameters $p^\pi(s,a)$ are the same as before: $\sum_a p^\pi(a|s) = 1$, $\forall s \in \mathcal{S}$; and $p^\pi(a|s) \geq 0$, $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$. Our goal, this time, is to solve the following optimization problem (from Eq. 9, Sharan et al., 2021)

$$\pi_{t+1} = \arg\max_{\pi \in \Pi} \left[ \sum_s d^{\pi_t}(s) \sum_a p^{\pi_t}(a|s) \left( Q^{\pi_t}(s,a) \frac{p^\pi(a|s)}{p^{\pi_t}(a|s)} - \frac{1}{\eta} D_\phi(p^\pi(\cdot|s), p^{\pi_t}(\cdot|s)) \right) \right], \tag{8}$$

with the mirror map as the negative entropy (Eq. 5.27, Beck and Teboulle, 2002). This particular choice of the mirror map simplifies the Bregman divergence as follows

$$D_\phi(p^\pi(\cdot|s), p^{\pi_t}(\cdot|s)) = \mathrm{KL}(p^\pi(\cdot|s) \| p^{\pi_t}(\cdot|s)) := \sum_a p^\pi(a|s) \log \frac{p^\pi(a|s)}{p^{\pi_t}(a|s)}. \tag{9}$$

The optimization problem (Eq. 8) then simplifies to

$$\pi_{t+1} = \arg\max_{\pi \in \Pi} \left[ \sum_s d^{\pi_t}(s) \sum_a p^{\pi_t}(a|s) \left( Q^{\pi_t}(s,a) \frac{p^\pi(a|s)}{p^{\pi_t}(a|s)} - \frac{1}{\eta} \sum_{a'} p^\pi(a'|s) \log \frac{p^\pi(a'|s)}{p^{\pi_t}(a'|s)} \right) \right]. \tag{10}$$

Proceeding analogously to the previous section, we use Lagrange multipliers $\lambda_s$, $\lambda_{s,a}$ for all states $s$ and actions $a$ to obtain the function

$$\begin{aligned}\mathcal{L}(p^\pi, \lambda_s, \lambda_{s,a}) = \sum_s d^{\pi_t}(s) \sum_a p^{\pi_t}(a|s) Q^{\pi_t}(s,a) \frac{p^\pi(a|s)}{p^{\pi_t}(a|s)} - \frac{1}{\eta} \sum_s d^{\pi_t}(s) \sum_{a'} p^\pi(a'|s) \log \frac{p^\pi(a'|s)}{p^{\pi_t}(a'|s)} \\ - \sum_{s,a} \lambda_{s,a} p^\pi(a|s) - \sum_s \lambda_s \Bigg( \sum_a p^\pi(a|s) - 1 \Bigg). \qquad (11)\end{aligned}$$

The KKT conditions are exactly the same as before (Eq. C1 to Eq. C6).

Again, we begin by solving the first KKT condition:

$$\begin{aligned}\nabla_{p^\pi(b|x)} \mathcal{L}(p^\pi, \lambda_s, \lambda_{s,a}) &= d^{\pi_t}(x) p^{\pi_t}(b|x) \frac{Q^{\pi_t}(x,b)}{p^{\pi_t}(b|x)} - \frac{d^{\pi_t}(x)}{\eta} \left[ \log \frac{p^\pi(b|x)}{p^{\pi_t}(b|x)} + 1 \right] - \lambda_{x,b} - \lambda_x \\ &= \frac{d^{\pi_t}(x)}{\eta} \left[ \eta Q^{\pi_t}(x,b) - \log \frac{p^\pi(b|x)}{p^{\pi_t}(b|x)} - 1 - \frac{\eta(\lambda_{x,b} + \lambda_x)}{d^{\pi_t}(x)} \right] \\ &= 0 \\ \Rightarrow \qquad \log \frac{p^\pi(b|x)}{p^{\pi_t}(b|x)} &= \eta Q^{\pi_t}(x,b) - \frac{\eta(\lambda_{x,b} + \lambda_x)}{d^{\pi_t}(x)} - 1 \\ \Rightarrow \qquad p^\pi(b|x) &= p^{\pi_t}(b|x) \cdot \exp(\eta Q^{\pi_t}(x,b)) \cdot \exp\left( - \frac{\eta(\lambda_{x,b} + \lambda_x)}{d^{\pi_t}(x)} - 1 \right), \qquad (12)\end{aligned}$$

where in the fourth line, we made the assumption that $d^{\pi_t}(x) > 0$ for all states $x$. We again set

$$\lambda_{s,a} = 0, \quad \forall s \in \mathcal{S},\ \forall a \in \mathcal{A}. \qquad (13)$$

And, we put Eq. 12 in the second KKT condition to get

$$\exp\left( - \frac{\eta \lambda_x}{d^{\pi_t}(x)} - 1 \right) = \left( \sum_b p^{\pi_t}(b|x) \cdot \exp(\eta Q^{\pi_t}(x,b)) \right)^{-1}. \qquad (14)$$

Therefore, we obtain

$$p^\pi(a|s) = \frac{p^{\pi_t}(a|s) \cdot \exp(\eta Q^{\pi_t}(s,a))}{\sum_b p^{\pi_t}(b|s) \cdot \exp(\eta Q^{\pi_t}(s,b))}. \qquad (15)$$

**This leaves one last problem:** Can we ensure that $\lambda_s \geq 0$ for all states $s$? If not, then the fourth KKT condition cannot be satisfied. Maybe, we can set the stepsize $\eta$ in such a way, such that this constraint is always fulfilled.

# References


Beck, A., Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters, 31*(3), 167-175.

Nocedal, J., Wright, S. (2006). Numerical optimization. *Springer Science & Business Media.*

Vaswani, S., Bachem, O., Totaro, S., Mueller, R., Geist, M., Machado, M. C., Castro P. S., Roux, N. L. (2021). A functional mirror ascent view of policy gradient methods with function approximation. *arXiv preprint arXiv:2108.05828.*