CS4789

AUSTIN WU

Date: May 28, 2025.

Problem 1:

a:

Since we have fixed $A = A^{\pi_{\theta_t}}(s, a)$, we know that

$$r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}$$

Thus each term is

$$r(\theta)A, (1-\epsilon)A, (1+\epsilon)A$$

If the first term is the min, then the gradient becomes

$$\begin{aligned}\nabla_\theta[r(\theta)]A &= A\nabla_\theta[\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}] \\ &= A\,\nabla_\theta r(\theta) \\ &= A\,\nabla_\theta\Big[\frac{\pi_\theta(a \mid s)}{\pi_{\theta_t}(a \mid s)}\Big] \\ &= A\,\frac{\pi_\theta(a \mid s)}{\pi_{\theta_t}(a \mid s)}\,\nabla_\theta \log \pi_\theta(a \mid s) \\ &= A\,r(\theta)\,\nabla_\theta \log \pi_\theta(a \mid s).\end{aligned}$$

For the second term, it becomes

$$\nabla_\theta[(1-\epsilon)A] = 0$$

And for the third term, it becomes

$$\nabla_\theta[(1+\epsilon)A] = 0$$

Since they are constant with respect to theta.

b:

When $r(\theta)$ drifts above $1+\epsilon$ with $A > 0$ or below $1-\epsilon$ with $A < 0$, we choose the clipped bound instead of the main function. This causes the gradient to get set to 0, This implies that $r(\theta)$ can't drift above or below these bounds. This means that the polciy ratio cannot increase more than its bounded $[1-\epsilon, 1+\epsilon]$ keeping policy updates close to π_{θ_t}.

c:

Suppose instead we use

$$\hat{l}_{\text{final}}(\theta) = \sum_{s,a} \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_t}}(a|s), 1-\epsilon, 1+\epsilon\right) \cdot A^{\pi_{\theta_t}}(s,a)$$

Instead of having a minimization.
Then if $A > 0$ and $r < 1-\epsilon$ the function without the min would always be clipped by $(1-\epsilon)A$ for any state action pair which has gradient 0. This implies that if r falls below the $1-\epsilon$ lower bound, there is no ability for the gradient to be in a direction to increase it.
This contradicts the function with the min because if it falls below the $1-\epsilon$ bound the gradient pushes $r(\theta)$ up so $min(rA, (1-\epsilon)A) = rA$. This then creates the gradient $A\, r(\theta)\, \nabla_\theta \log \pi_\theta(a \mid s)$ from the previous part, allowing a corrective gradient to bring it back up.
Similarly, if $A < 0$ and $r > 1+\epsilon$ the function without the min would always be clipped by $(1+\epsilon)A$ for any state action pair with gradient 0. This implies that if r falls above the $1+\epsilon$ upper bound, there is no ability for the gradient to be in a direction to decrease it.
This similarly contradicts the function with the min because if it rises above the $1+\epsilon$ bound the gradient pushes $r(\theta)$ down so $min(rA, (1+\epsilon)A) = rA$. This then creates the gradient $A\, r(\theta)\, \nabla_\theta \log \pi_\theta(a \mid s)$ from the previous part, allowing a corrective gradient to bring it back down.

Problem 2:

a:

Set $\pi_{\text{ref}} = \begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}$. Then $\frac{1}{3} \leq \frac{1}{3} \leq \frac{1}{3}$. Now define $\pi_0 = \begin{bmatrix}0.3 & 0.2 & 0.5\end{bmatrix}$. Then $0.2 \leq 0.3 \leq 0.5$. Thus we get that

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathop{\mathbb{E}}_{x,y_w,y_l \in \mathcal{D}}\left[\log\sigma(\beta \log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)})\right]$$

Simplifies to

$$= -\mathop{\mathbb{E}}_{x,y_w,y_l \in \mathcal{D}}\left[\log\sigma(\beta \log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)})\right]$$

Since $\pi_{\text{ref}}(y|x)$ is always equal to $\frac{1}{3}$. Thus if we sample with single prefrence ordering $y_1 \leq y_2$ we get when going from π_0 to the next iteration

$$= -\mathop{\mathbb{E}}_{x,y_2,y_1 \in \mathcal{D}}\left[\log\sigma(\beta \log\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)})\right]$$

$$= -\mathop{\mathbb{E}}_{x,y_2,y_1 \in \mathcal{D}}\left[\log\sigma(\beta \log\frac{3}{2})\right]$$

Thus if we create polices π_1, π_2 that have a good loss function, we must maximize the ratio of $\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)}$. This also implies that polices with equivalent ratios have the same loss function since we are only sampling once. Thus after update the some possible answers are arbitrarly setting $\pi_1(y_3) = \frac{1}{2}$. Then say

$$\frac{\pi_1(y_2)}{\pi_1(y_2)} = 1.75 = r$$

Thus since probability spaces must stay normalized we are constrained to

$$\pi_1(y_1) + 1.75\pi_1(y_1) = \frac{1}{2} \implies 2.75\pi_1(y_1) = \frac{1}{2} \implies \pi_1(y_1) = \frac{2}{11} \approx 0.18, \pi_1(y_2) = \frac{22}{69} \approx 0.31$$

We could also set $\pi_2(y_3) = \frac{1}{4}$. Then since the ratio must be 1.5 implies that

$$\pi_2(y_1) + 1.75\pi_2(y_1) = \frac{3}{4} \implies 2.75\pi_2(y_1) = \frac{3}{4} \implies \pi_2(y_1) = \frac{3}{11} \approx 0.27, \pi_2(y_2) = \frac{21}{44} \approx 0.47$$

Thus $\pi_1 = \begin{bmatrix}\frac{2}{11} & \frac{22}{69} & 0.5\end{bmatrix}, \pi_2 = \begin{bmatrix}\frac{3}{11} & \frac{21}{44} & 0.25\end{bmatrix}$ wich equivlanet loss functions, but π_1 satisifes $y_1 \leq y_2 \leq y_3$ with π_2 does not. Thus our loss functions become

$$\mathcal{L}_{DPO} = \log(\frac{1}{1+e^{-\beta r}})$$

With r being the ratio between elements. Thus the loss with $\beta = 1$ is

$$\mathcal{L}(\pi)_{\text{ref}} = \log(\frac{1}{1+e^{-\beta 1}}) \approx -0.31$$
$$\mathcal{L}(\pi_0) = \log(\frac{1}{1+e^{-\beta 1.5}}) \approx -0.2$$
$$\mathcal{L}(\pi_1) = \log(\frac{1}{1+e^{-\beta 1.75}}) \approx -0.16$$
$$\mathcal{L}(\pi_2) = \log(\frac{1}{1+e^{-\beta 1.75}}) \approx -0.16$$

This means that DPO has a blindness between unobserved comparisons regarding polices.

b:

For $\pi_{\text{ref}}(\cdot|x), \pi_\theta(\cdot|x)$ let

$\pi_{\text{ref}}(y_1\|x)$	$\frac{1}{2}$
$\pi_{\text{ref}}(y_2\|x)$	$\frac{1}{2}$
$\pi_{\text{ref}}(y_3\|x)$	0
$\pi_\theta(y_1\|x)$	0
$\pi_\theta(y_2\|x)$	$\frac{1}{2}$
$\pi_\theta(y_3\|x)$	$\frac{1}{2}$

With $\pi_{\text{ref}}(y_1|x) = \frac{1}{2} \leq \pi_{\text{ref}}(y_2|x) = \frac{1}{2}$ and $\pi_\theta(y_1|x) = 0 \leq \pi_\theta(y_2|x) = \frac{1}{2}$ Then we find that

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathop{\mathbb{E}}_{x,y_w,y_l \in \mathcal{D}}\left[\log \sigma(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)})\right]$$

Now since we are only sampling $(y_w, y_l) = (y_2, y_1)$

$$= -\mathop{\mathbb{E}}_{x,y_w,y_l \in \mathcal{D}}\left[\log \sigma(\beta \log \frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)\pi_\theta(y_l|x)})\right]$$

With notably $\sigma(z) = \frac{1}{1+e^{-z}}$

$$= -\mathop{\mathbb{E}}_{x,y_2,y_1 \in \mathcal{D}}\left[\log \sigma(\beta \log \frac{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y_2|x)\pi_\theta(y_1|x)})\right]$$

Thus plugging in the values we get

$$= -\mathop{\mathbb{E}}_{x,y_2,y_1 \in \mathcal{D}}\left[\log \sigma(\beta \log \frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)})\right]$$
$$= -\mathop{\mathbb{E}}_{x,y_2,y_1 \in \mathcal{D}}\left[\log \sigma(\beta \log \left(\frac{\frac{1}{2}}{0}\right))\right]$$

With asymotic behavior $\beta \log(\frac{\frac{1}{2}}{0}) = +\infty$ thus

$$
\begin{aligned}
&= - \mathop{\mathbb{E}}_{x,y_2,y_1 \in \mathcal{D}} \left[\log \frac{1}{1 + e^{-(+\infty)}} \right] \\
&= - \mathop{\mathbb{E}}_{x,y_2,y_1 \in \mathcal{D}} [\log(1)] \\
&= 0
\end{aligned}
$$

And for KL Difergence we get

$$
\begin{aligned}
D_{KL}(\pi_\theta || \pi_{\text{ref}}) &= \sum_{i=1}^{3} \theta(y_i|x) \log \frac{\pi_\theta(y_i|x)}{\pi_{\text{ref}}(y_i|x)} \\
&= 0 + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{2}} + \frac{1}{2} \log \frac{\frac{1}{2}}{0} \\
&= +\infty
\end{aligned}
$$

Thus we shown that we can pick a policy for $\mathcal{L}_{DPO} = 0$ and $KL(\pi || \pi_{\text{ref}}) = +\infty$