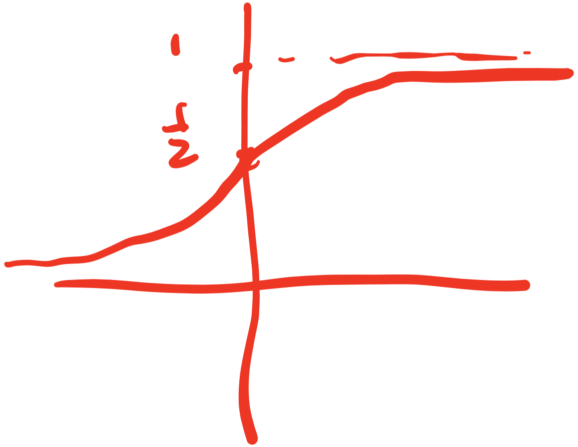
Direct Preference Optimization (DPO)

Recap: Bradley Terry model and reward model (RM) learning

Recap: Bradley Terry model and reward model (RM) learning

The BT model assumes that **humans generate labels** based on the following probablistic model:

$$P(\tau \text{ is prefered over } \tau' \text{ given } x) = \underbrace{\frac{1}{1 + \exp\left(-\left(r^\star(x,\tau) - r^\star(x,\tau')\right)\right)}}_{\sigma\left(r^*(x,\tau) - r^*(x,\tau')\right)}$$



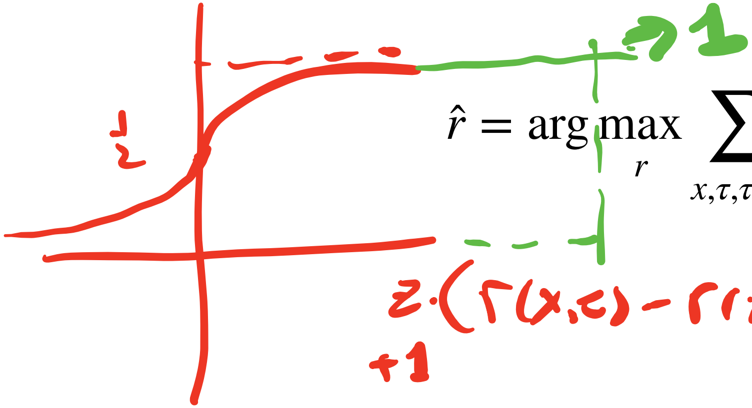
Recap: Bradley Terry model and reward model (RM) learning

The BT model assumes that **humans generate labels** based on the following probablistic model:

$$P(\tau \text{ is prefered over } \tau' \text{ given } x) = \frac{1}{1 + \exp\left(-\left(r^\star(x,\tau) - r^\star(x,\tau')\right)\right)}$$

$D = \{x, \tau, \tau', z\}$ $\quad z = \begin{cases} +1 \leftarrow \tau \text{ is better than } \tau' \\ -1 \end{cases}$

We parameter a reward function (e.g., neural network) $r(x, \tau)$, and learn via MLE / logistic regression



$$\hat{r} = \arg\max_r \sum_{x,\tau,\tau',z} \ln \frac{1}{1 + \exp\left(-z \cdot \underline{\left(r(x,\tau) - r(x,\tau')\right)}\right)}$$

$(x, \tau, \tau', z = 1)$

$\underline{r(x,\tau) - r(x,\tau')}$

$\to +\infty$

$z \cdot (r(x,\tau) - r(x,\tau'))$

$+1$

Recap: KL-reg RL for avoiding reward hacking

β : controls the strength of KL-reg;

$$J(\pi) = \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{\tau\sim\pi(\cdot|x)}\hat{r}(x,\tau) - \beta \mathrm{KL}\left(\pi(\,\cdot\,|x)\middle|\pi_{ref}(\,\cdot\,|x)\right)\right]$$

"stay close" to the SFT policy π_{ref}.

$D = \{x, \tau, \tau', z\}$

$\tau, \tau' \sim \pi_{ref}(\cdot|x)$

Recap: KL-reg RL for avoiding reward hacking

β : controls the strength of KL-reg;

$$J(\pi) = \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{\tau\sim\pi(\cdot|x)}\hat{r}(x,\tau) - \beta\mathsf{KL}\left(\pi(\,\cdot\,|x)\middle|\pi_{ref}(\,\cdot\,|x)\right)\right]$$

"stay close" to the SFT policy π_{ref}.

ChatGPT uses PPO to optimize $J(\pi)$….

When models are large…

RM + PPO can be hard to optimize…

At least need to maintain 4 big models in GPU RAM (RM, π, V, π_{ref}…)

Question today:

Can we combine the two stages together and learn policy directly?

Outline

1. KL-reg RL revisit and its closed-form solution
2. Reparametrization trick — modeling RM difference using policy directly
3. DPO Algorithm

First thing…

We will directly operate at the trajectory level, i.e., a trajectory is an action

Given prompt x, and an "action" (a trajectory) $\tau = \{y_0, y_1, \ldots, y_{H-1}\}$, what's the likelihood of the "action" under the policy π_θ?

$\pi_\theta(\tau \mid x)$

First thing…

We will directly operate at the trajectory level, i.e., a trajectory is an action

Given prompt x, and an "action" (a trajectory) $\tau = \{y_0, y_1, \ldots, y_{H-1}\}$, what's the likelihood of the "action" under the policy π_θ?

$$\pi_\theta(\tau \mid x) = \prod_{h=0}^{H-1} \pi_\theta(y_h \mid x, y_{<h})$$

history

$\{s_h, a_h\}_{h=0}^{H-1}$

First thing…

We will directly operate at the trajectory level, i.e., a trajectory is an action

Given prompt x, and an "action" (a trajectory) $\tau = \{y_0, y_1, \ldots, y_{H-1}\}$, what's the likelihood of the "action" under the policy π_θ?

$$\pi_\theta(\tau \mid x) = \prod_{h=0}^{H-1} \underline{\pi_\theta(y_h \mid x, y_{<h})}$$

Likelihood of predicting y_h given the past..

KL-reg RL objective

$$J(\pi) = \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{\tau\sim\pi(\cdot|x)}\hat{r}(x,\tau) - \beta \mathrm{KL}\left(\pi(\cdot|x)\middle|\pi_{ref}(\cdot|x)\right)\right]$$

What's the $\arg\max_{\pi} J(\pi)$?

KL-reg RL objective

$$J(\pi) = \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{\tau\sim\pi(\cdot|x)}\hat{r}(x,\tau) - \beta \mathrm{KL}\left(\pi(\cdot|x)\middle|\pi_{ref}(\cdot|x)\right)\right]$$

$\pi(\tau|x)$

What's the $\arg\max_{\pi} J(\pi)$?

Consider on a (x,τ) pair, what is $\partial J(\pi)/\partial\pi(\tau|x)$? $=0$ solve for $\pi(\tau|x)$

$$J(\pi) = \sum_{\tau} \pi(\tau|x)\cdot\hat{r}(x,\tau) - \beta\sum_{\tau}\pi(\tau|x)\left(\ln\pi(\tau|x) - \ln\pi_{ref}(\tau|x)\right)$$

$$\frac{dJ}{d\pi(\tau|x)} = \hat{r}(x,\tau) - \beta\left[\left(\ln\pi(\tau|x) - \ln\pi_{ref}(\tau|x)\right) + 1\right] = 0$$

KL-reg RL objective

$$J(\pi) = \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{\tau\sim\pi(\cdot|x)}\hat{r}(x,\tau) - \beta \mathrm{KL}\left(\pi(\cdot\,|\,x)\,\middle|\,\pi_{ref}(\cdot\,|\,x)\right)\right]$$

$\pi(\tau|x) \geq 0$

$\sum_\tau \pi(\tau|x) = 1$

What's the $\arg\max_\pi J(\pi)$?

Consider on a (x, τ) pair, what is $\partial J(\pi)/\partial \pi(y\,|\,x)$?

$$\frac{\partial J(\pi)}{\partial \pi(y\,|\,x)} = \hat{r}(x,\tau) - \beta\left(\ln \pi(\tau\,|\,x) - \ln \pi_{ref}(\tau\,|\,x) + 1\right) = 0$$

solve for $\pi(\tau|x)$

$$\exp\left(\ln \pi(\tau|x)\right) = \exp\left(\frac{1}{\beta}\hat{r}(x,\tau) + \ln \pi_{ref}(\tau|x) - \cancel{1}\right)$$

KL-reg RL objective

$$J(\pi) = \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{\tau\sim\pi(\cdot|x)}\hat{r}(x,\tau) - \beta \mathrm{KL}\left(\pi(\cdot\,|\,x)\,\middle|\,\pi_{ref}(\cdot\,|\,x)\right)\right]$$

What's the $\arg\max_{\pi} J(\pi)$?

Consider on a (x, τ) pair, what is $\partial J(\pi)/\partial\pi(y\,|\,x)$?

$$\frac{\partial J(\pi)}{\partial \pi(y\,|\,x)} = \hat{r}(x,\tau) - \beta\left(\ln \pi(\tau\,|\,x) - \ln \pi_{ref}(\tau\,|\,x) + 1\right)$$

$$\pi(\tau\,|\,x) \propto \pi_{ref}(\tau\,|\,x)\exp\left(\hat{r}(x,\tau)/\beta\right)$$

0

$\beta = +\infty$

$\beta \to 0^+$ $\quad \pi(\tau|x) \to \arg\max_{\tau} \hat{r}(x,\tau)$

KL-reg RL objective

$$J(\pi) = \mathbb{E}_{x \sim \nu}\left[\mathbb{E}_{\tau \sim \pi(\cdot|x)} \hat{r}(x,\tau) - \beta \mathrm{KL}\left(\pi(\cdot \mid x) \middle| \pi_{ref}(\cdot \mid x)\right)\right]$$

What's the $\arg\max_{\pi} J(\pi)$?

Consider on a (x, τ) pair, what is $\partial J(\pi) / \partial \pi(y \mid x)$?

$$\frac{\partial J(\pi)}{\partial \pi(y \mid x)} = \hat{r}(x,\tau) - \beta\left(\ln \pi(\tau \mid x) - \ln \pi_{ref}(\tau \mid x) + 1\right)$$

$$\pi(\tau \mid x) \propto \pi_{ref}(\tau \mid x) \exp\left(\hat{r}(x,\tau)/\beta\right)$$

Normalization

$$\pi(\tau \mid x) = \pi_{ref}(\tau \mid x) \exp\left(\frac{\hat{r}(x,\tau)}{\beta}\right) / Z(x), \text{ where } Z(x) = \mathbb{E}_{\tau \sim \pi_{ref}(\cdot|x)} \exp(\hat{r}(x,\tau)/\beta)$$

$$= \sum_{\text{all } \tau} \pi_{ref}(\tau|x) \exp(\hat{r}(x,\tau)/\beta)$$

$$\sum_{\tau} \pi(\tau|x) = 1$$

KL-reg RL objective

$$J(\pi) = \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{\tau\sim\pi(\cdot|x)}\hat{r}(x,\tau) - \beta \mathrm{KL}\left(\pi(\cdot\,|\,x)\,\middle|\,\pi_{ref}(\cdot\,|\,x)\right)\right]$$

$\hat{\pi} \leftarrow \arg\max_{\pi} J(\pi)$

In sum, the optimal policy is:

$$\hat{\pi}(\tau\,|\,x) = \frac{\pi_{ref}(\tau\,|\,x)\cdot\exp\left(\frac{\hat{r}(x,\tau)}{\beta}\right)}{Z(x)}$$

$\Rightarrow \sum_{\tau} \hat{\pi}(\tau|x) = 1$

1. When $\beta \to 0$: $\hat{\pi}(\tau|x) \to \arg\max_{\tau} \hat{r}(x,\tau)$

2. When $\beta \to \infty$: $\hat{\pi}(\tau|x) \to \pi_{ref}(\tau|x)$

Outline

1. KL-reg RL revisit and its closed-form solution

2. Reparametrization trick — modeling **RM difference** using policy directly

3. DPO Algorithm

Can we parameterize RM using policies?

In sum, the optimal policy of the KL-reg RL objective is:

$$\ln\left(\hat{\pi}(\tau \mid x)\right) = \ln\left(\frac{\pi_{ref}(\tau \mid x) \cdot \exp\left(\frac{\hat{r}(x,\tau)}{\beta}\right)}{Z(x)}\right)$$

Can we parameterize RM using policies?

In sum, the optimal policy of the KL-reg RL objective is:

$$\hat{\pi}(\tau \mid x) = \frac{\pi_{ref}(\tau \mid x) \cdot \exp\left(\frac{\hat{r}(x, \tau)}{\beta}\right)}{Z(x)}$$

$$\ln \hat{\pi}(\tau \mid x) = \ln \pi_{ref}(\tau \mid x) - \ln Z(x) + \frac{\hat{r}(x, \tau)}{\beta}$$

$$\hat{r}(x, \tau) = \beta \ln \frac{\hat{\pi}(\tau \mid x)}{\pi_{ref}(\tau \mid x)} + \beta \cdot \ln Z(x)$$

Can we parameterize RM using policies?

In sum, the optimal policy of the KL-reg RL objective is:

$$\hat{\pi}(\tau \mid x) = \frac{\pi_{ref}(\tau \mid x) \cdot \exp\left(\frac{\hat{r}(x,\tau)}{\beta}\right)}{Z(x)}$$

$$\ln \hat{\pi}(\tau \mid x) = \ln \pi_{ref}(\tau \mid x) - \ln Z(x) + \frac{\hat{r}(x,\tau)}{\beta}$$

$$Z(x) = \mathbb{E}_{\tau \sim \pi_{ref}(\cdot \mid x)} \exp\left(\frac{\hat{r}(x,\tau)}{\beta}\right)$$

$$\hat{r}(x,\tau) = \beta\left(\ln \frac{\hat{\pi}(\tau \mid x)}{\pi_{ref}(\tau \mid x)} + \ln Z(x)\right)$$

$\hat{r}$

Can we parameterize RM using policies?

In sum, the optimal policy of the KL-reg RL objective is:

$$\hat{\pi}(\tau \mid x) = \frac{\pi_{ref}(\tau \mid x) \cdot \exp\left(\frac{\hat{r}(x,\tau)}{\beta}\right)}{Z(x)}$$

$$\ln \hat{\pi}(\tau \mid x) = \ln \pi_{ref}(\tau \mid x) - \ln Z(x) + \frac{\hat{r}(x,\tau)}{\beta}$$

$$\hat{r}(x,\tau) = \beta \left(\ln \frac{\hat{\pi}(\tau \mid x)}{\pi_{ref}(\tau \mid x)} + \ln Z(x) \right)$$

Not done yet, this $Z(x)$ technically contains $\hat{r}$!

Can we parameterize RM using policies?

In sum, the optimal policy of the KL-reg RL objective is:

$$\hat{\pi}(\tau \mid x) = \frac{\pi_{ref}(\tau \mid x) \cdot \exp\left(\frac{\hat{r}(x, \tau)}{\beta}\right)}{Z(x)}$$

$$\ln \hat{\pi}(\tau \mid x) = \ln \pi_{ref}(\tau \mid x) - \ln Z(x) + \frac{\hat{r}(x, \tau)}{\beta}$$

$\hat{r}$

$$\hat{r}(x, \tau) = \beta \left(\ln \frac{\hat{\pi}(\tau \mid x)}{\pi_{ref}(\tau \mid x)} + \ln \underline{Z(x)} \right)$$

Not done yet, this $Z(x)$ technically contains $\hat{r}$!

But $\ln Z(x)$ is a shift that is independent of τ...

$\hat{r}(x, \tau) - \hat{r}(x, \tau')$

Cancelling the normalization constant $Z(x)$ via modeling the difference

$$\hat{r}(x, \tau) = \beta \left(\ln \frac{\hat{\pi}(\tau \mid x)}{\pi_{ref}(\tau \mid x)} + \ln Z(x) \right)$$

Not done yet, this $Z(x)$ technically contains $\hat{r}$!

But $\ln Z(x)$ is a shift that is independent of τ...

(x, τ, τ')

$\hat{r}(x, \tau) - \hat{r}(x, \tau')$

Cancelling the normalization constant $Z(x)$ via modeling the difference

$$\hat{r}(x, \tau) = \beta \left(\ln \frac{\hat{\pi}(\tau \mid x)}{\pi_{ref}(\tau \mid x)} + \ln Z(x) \right)$$

Not done yet, this $Z(x)$ technically contains $\hat{r}$!

But $\ln Z(x)$ is a shift that is independent of τ…

Given (x, τ, τ'), we just model **reward difference**:

Cancelling the normalization constant $Z(x)$ via modeling the difference

$$\hat{r}(x,\tau) = \beta\left(\ln\frac{\hat{\pi}(\tau\,|\,x)}{\pi_{ref}(\tau\,|\,x)} + \ln Z(x)\right)$$

Not done yet, this $Z(x)$ technically contains $\hat{r}$!

But $\ln Z(x)$ is a shift that is independent of τ...

Given (x, τ, τ'), we just model **reward difference**:

$$\hat{r}(x,\tau) - \hat{r}(x,\tau') = \beta\left(\ln\frac{\hat{\pi}(\tau\,|\,x)}{\pi_{ref}(\tau\,|\,x)} - \ln\frac{\hat{\pi}(\tau'\,|\,x)}{\pi_{ref}(\tau'\,|\,x)}\right)$$

Cancelling the normalization constant $Z(x)$ via modeling the difference

$$\hat{r}(x, \tau) = \beta \left(\ln \frac{\hat{\pi}(\tau \mid x)}{\pi_{ref}(\tau \mid x)} + \ln Z(x) \right)$$

Not done yet, this $Z(x)$ technically contains $\hat{r}$!

But $\ln Z(x)$ is a shift that is independent of τ...

Given (x, τ, τ'), we just model **reward difference**:

$$\hat{r}(x, \tau) - \hat{r}(x, \tau') = \beta \left(\ln \frac{\hat{\pi}(\tau \mid x)}{\pi_{ref}(\tau \mid x)} - \ln \frac{\hat{\pi}(\tau' \mid x)}{\pi_{ref}(\tau' \mid x)} \right)$$

The annoying normalization term gone!

Outline

1. KL-reg RL revisit and its closed-form solution

2. Reparametrization trick — modeling RM difference using policy directly

3. DPO Algorithm

DPO

1. Take any policy π_θ, we can use it to model the reward difference:

$$r_\theta(\tau \mid x) - r_\theta(\tau' \mid x) := \beta\left(\ln \frac{\pi_\theta(\tau \mid x)}{\pi_{ref}(\tau \mid x)} - \ln \frac{\pi_\theta(\tau' \mid x)}{\pi_{ref}(\tau' \mid x)}\right)$$

DPO

1. Take any policy π_θ, we can use it to model the reward difference:

$$r_\theta(\tau|x) - r_\theta(\tau'|x) := \beta\left(\ln\frac{\pi_\theta(\tau|x)}{\pi_{ref}(\tau|x)} - \ln\frac{\pi_\theta(\tau'|x)}{\pi_{ref}(\tau'|x)}\right)$$

2. Now plug this into the MLE loss we had for learning the reward difference:

DPO

1. Take any policy π_θ, we can use it to model the reward difference:

$$r_\theta(\tau|x) - r_\theta(\tau'|x) := \beta\left(\ln\frac{\pi_\theta(\tau|x)}{\pi_{ref}(\tau|x)} - \ln\frac{\pi_\theta(\tau'|x)}{\pi_{ref}(\tau'|x)}\right)$$

2. Now plug this into the MLE loss we had for learning the reward difference:

$$\arg\max_\theta \sum_{x,\tau,\tau',z} \ln \frac{1}{1+\exp\left(-z\cdot\left(r_\theta(x,\tau) - r_\theta(x,\tau')\right)\right)}$$

DPO

$D = \{x, \tau, \tau', z\}$

DPO optimizes policy π_θ directly using the following loss:

$$\arg\max_{\theta} \sum_{x,\tau,\tau',z} \ln \frac{1}{1 + \exp\left(-z \cdot \beta \underbrace{\left(\ln \frac{\pi_\theta(\tau \mid x)}{\pi_{ref}(\tau \mid x)} - \ln \frac{\pi_\theta(\tau' \mid x)}{\pi_{ref}(\tau' \mid x)}\right)}_{:= r_\theta(\tau, x) - r_\theta(\tau', x)}\right)}$$

The squared loss version of DPO

Optimizing Logistic loss can lead to overfit, we can use square loss (e.g., regression) instead:

$$\arg\max_{\theta} \sum_{x,\tau,\tau',z} \left(\beta \left(\ln \frac{\pi_\theta(\tau \mid x)}{\pi_{ref}(\tau \mid x)} - \ln \frac{\pi_\theta(\tau' \mid x)}{\pi_{ref}(\tau' \mid x)} \right) - z \right)^2$$

$z \in \{+1, -1\}$

τ is better than τ' $(z = +1)$

$$\underline{\ln \frac{\pi_\theta(\tau \mid x)}{\pi_{ref}(\tau \mid x)} - \ln \frac{\pi(\tau' \mid x)}{\pi_{ref}(\tau' \mid x)}} \approx \frac{+1}{\beta}$$

Applying DPO on the openAI gym tasks (next PA)

Q: But these tasks have unknown transition $\rho(\tau) = \prod_h \pi(a_h \mid s)P(s_{h+1} \mid s_h, a_h)$, can we still do DPO?

π

$\tau = \{s_h, a_h\}_{h=0}^{H-1}$

$\ln \frac{\pi(\tau|x)}{\pi_{ref}(\tau|x)}$

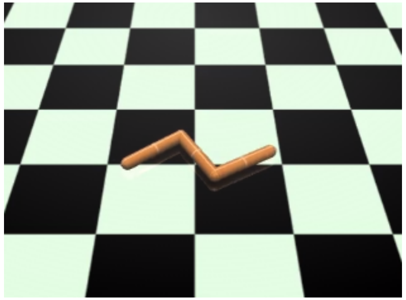
Applying DPO on the openAI gym tasks (next PA)

Q: But these tasks have unknown transition $\rho(\tau) = \prod_h \pi(a_h \mid s) \cancel{P(s_{h+1} \mid s_h, a_h)}$, can we still do DPO?

Note that we only care about trajectory density ratio, so transition cancels out!

$$\ln \frac{\rho_\pi(\tau)}{\rho_{\pi_{ref}}(\tau)} = \ln \prod_h \frac{\pi(a_h \mid s_h)}{\pi_{ref}(a_h \mid s_h)} = \sum_h \ln \frac{\pi(a_h \mid s_h)}{\pi_{ref}(a_h \mid s_h)}$$

Swimmer: continuous controll; goal: move forward fast

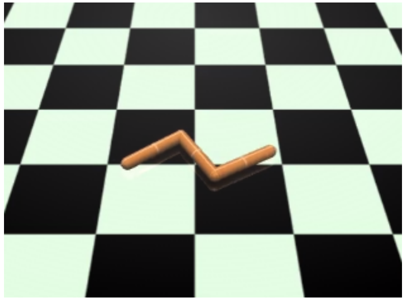


1. Collect pair of trajs using a π_{ref}; label via the ground truth reward

2. Run DPO (squared loss) w/ different β

$\hat{r} - \beta \cdot KL(\pi \mid \pi_{ref})$

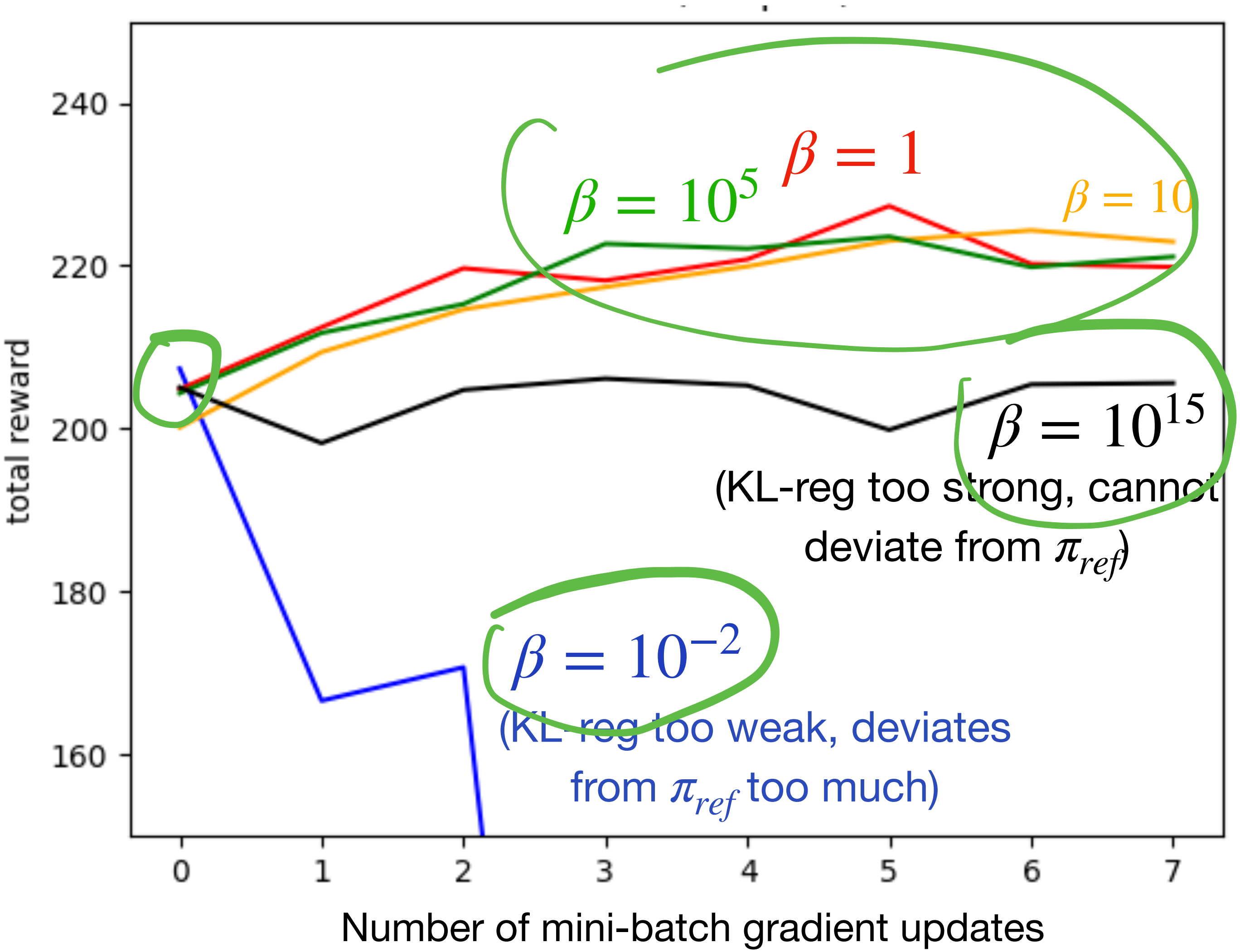
Swimmer: continuous controll; goal: move forward fast



1. Collect pair of trajs using a π_{ref}; label via the ground truth reward

2. Run DPO (squared loss) w/ different β

$\hat{r} - \beta \cdot KL$



Summary

Closed-form solution of the optimal policy of KL-reguarlized RL

Summary

Closed-form solution of the optimal policy of KL-regualrlized RL

DPO reparameterizes the reward difference via policy directly

Summary

Closed-form solution of the optimal policy of KL-regualrized RL

DPO reparameterizes the reward difference via policy directly

Plug the reward difference parameterized by policy into the BT-inspired MLE loss to directly optimize policy