RL from human feedback: BT model and REBEL

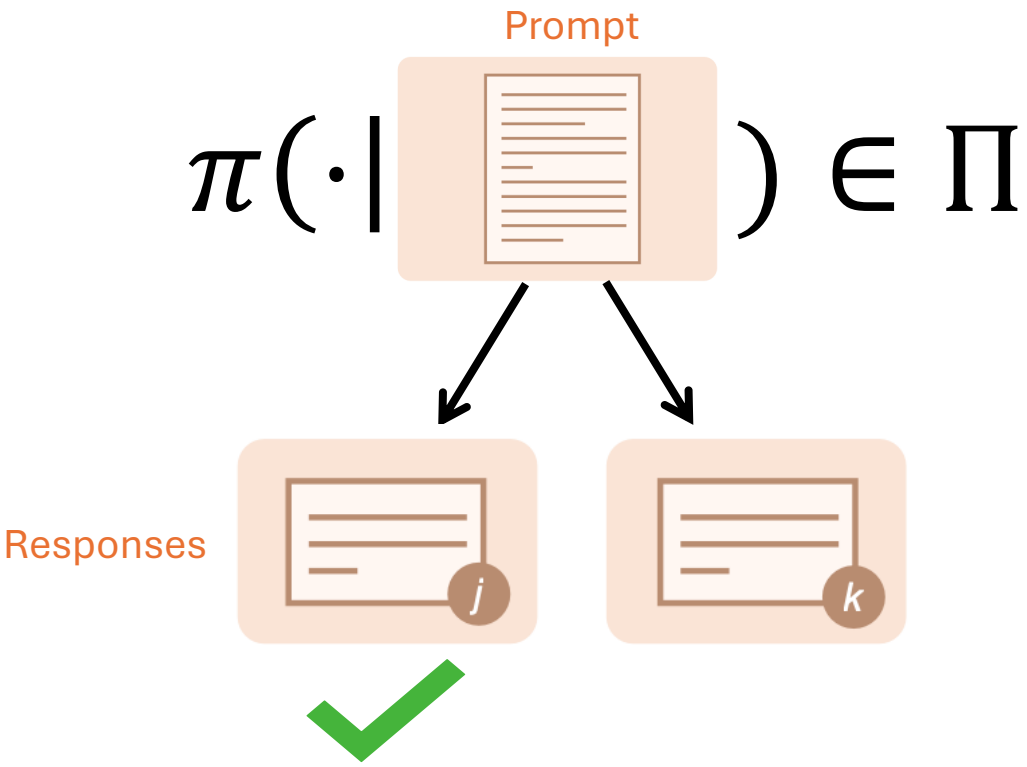
Zhaolin Gao

CS 6789: Foundations of Reinforcement Learning

RL from Human Feedback (RLHF)

RLHF Pipeline

1. Collect preference dataset



2. Learn a reward model

$r($ Prompt , Response $) \in \mathbb{R}$

3. Reinforcement Learning

Optimize LLM s.t. it generates responses with high rewards

$\pi_{t+1} = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} r(x, y)$

Today's Questions

How to learn such a **reward model** from the **human preference data?**

How to design a **RL algorithm** to optimize the LLM given the reward model?

Outline

1. LLM as Policy
2. Bradley-Terry Model as reward
3. RLHF Setting
4. Algorithm: REBEL
5. Guarantee and Proof sketch

LLM as Policy – State and Action

State (s): the context of the dialogue

Action space ($\mathcal{A}$): all tokens in the vocabulary

LLM (π): a stochastic policy

$\pi(a|s)$: the probability of generating a given state s

$\pi(\cdot\,|s)$: the probability distribution over all tokens

LLM as Policy – State and Action

$t = 0$:

$s_0 = what\ is\ the\ capital\ of\ France?$

$a_0 = the$

$t = 1$:

$s_1 = what\ is\ the\ capital\ of\ France?\ the$

$a_1 = capital$

…

$t = h$:

$s_h = what\ is\ the\ capital\ of\ France?\ the\ capital\ of\ France\ is\ Paris.$

$a_h =< EOS >$

LLM as Policy – State and Action

$t = 0$:

$s_0 = what\ is\ the\ capital\ of\ France?$

$a_0 = the$

$t = 1$:

$s_1 = what\ is\ the\ capital\ of\ France?\ the$

$a_1 = capital$

…

$t = h$:

$s_h = what\ is\ the\ capital\ of\ France?\ the\ capital\ of\ France\ is\ Paris.$

$a_h =< EOS >$

Final response: $\boldsymbol{the\ capital\ of\ France\ is\ Paris.}$

LLM as Policy – Reset

Prompt:

$$x = what\ is\ the\ capital\ of\ France?$$

Response:

$$y = the\ capital\ of\ France\ is\ Paris.$$

LLM as Policy – Reset

Prompt:

$$x = what\ is\ the\ capital\ of\ France?$$

Response:

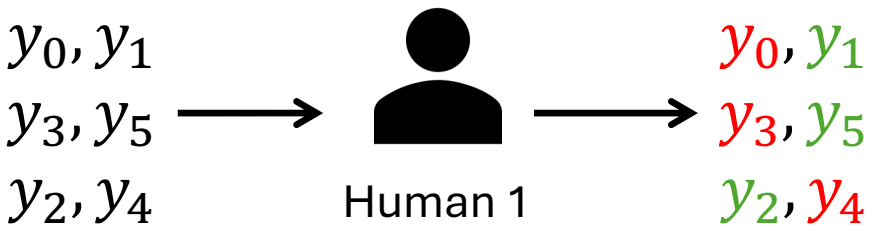
$$y_0 = the\ capital\ of\ France\ is\ Paris.$$

$$y_1 = Paris$$

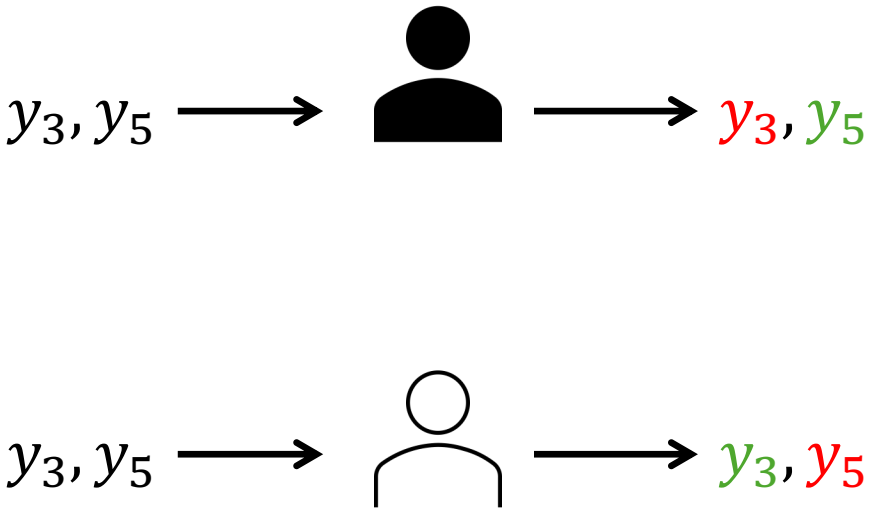
$$y_2 = It\ is\ Paris.$$

Bradley-Terry Model – Dataset

$$\mathcal{D} = \{x, y_{chosen}, y_{reject}\}$$



Bradley-Terry Model – Formulation



Bradley-Terry Model – Formulation

Given 2 responses:

$$y_i, y_j$$

And their rewards:

$$r(x, y_i), r(x, y_j)$$

Bradley-Terry model:

$$p\left(y_i \text{ is preferred over } y_j\right) = \frac{\exp(r(x, y_i))}{\exp(r(x, y_i)) + \exp(r(x, y_j))}$$

Bradley-Terry Model – MLE Optimization

Maximize the likelihood of human preferences (MLE):

$$\prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} p\left(y_{chosen} \text{ is preferred over } y_{reject}\right) =$$

Bradley-Terry Model – MLE Optimization

Maximize the likelihood of human preferences (MLE):

$$\prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} p\left(y_{chosen} \text{ is preferred over } y_{reject}\right) = \prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

Bradley-Terry Model – MLE Optimization

Maximize the likelihood of human preferences (MLE):

$$\prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} p\left(y_{chosen} \text{ is preferred over } y_{reject}\right) = \prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

$$r^* = \arg\max_r \prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

Reward Model Training

$$r^* = \arg\max_r \prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

Reward Model Training

$$r^* = \arg\max_r \prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

$$= \arg\max_r \sum_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \log \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

Reward Model Training

$$r^* = \arg\max_r \prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

$$= \arg\max_r \sum_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \log \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

$$= \arg\max_r \sum_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \log \frac{1}{1 + \exp\big(r(x, y_{reject}) - r(x, y_{chosen})\big)}$$

Reward Model Training

$$r^* = \arg\max_r \prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

$$= \arg\max_r \sum_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \log \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

$$= \arg\max_r \sum_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \log \frac{1}{1 + \exp\big(r(x, y_{reject}) - r(x, y_{chosen})\big)}$$

$$= \arg\max_r \sum_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \log \sigma(r(x, y_{chosen}) - r(x, y_{reject}))$$

Reward Model Training

$$\mathcal{L}_{RM}(\theta) = -\sum_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \log \sigma(r_\theta(x, y_{chosen}) - r_\theta(x, y_{reject}))$$

Reward Model – Property

Invariance to Shifts

$$p\big(y_i \text{ is preferred over } y_j\big) = \frac{1}{1 + \exp\big(r(x, y_j) - r(x, y_i)\big)}$$

$$r(x, y_i)\,, r(x, y_j)$$

same under
the BT model

$$r(x, y_i) + \alpha\,, r(x, y_j) + \alpha$$

Outline

1. ✓ LLM as Policy
2. ✓ Bradley-Terry Model as reward
3. RLHF Setting
4. Algorithm: REBEL
5. Guarantee and Proof sketch

RLHF Setting – Deterministic Transition

$t = 0$:

$s_0 = what\ is\ the\ capital\ of\ France?$

$a_0 = the$

$t = 1$:

$s_1 = what\ is\ the\ capital\ of\ France?\ the$

$a_1 = capital$

…

$t = n$:

$s_n = what\ is\ the\ capital\ of\ France?\ the\ capital\ of\ France\ is\ Paris.$

$a_n =< EOS >$

RLHF Setting – Deterministic Transition

$t = 0$:

$s_0 = what\ is\ the\ capital\ of\ France?$

$a_0 = the$

$t = 1$:

$s_1 = what\ is\ the\ capital\ of\ France?\ the = s_0 + a_0$

$a_1 = capital$

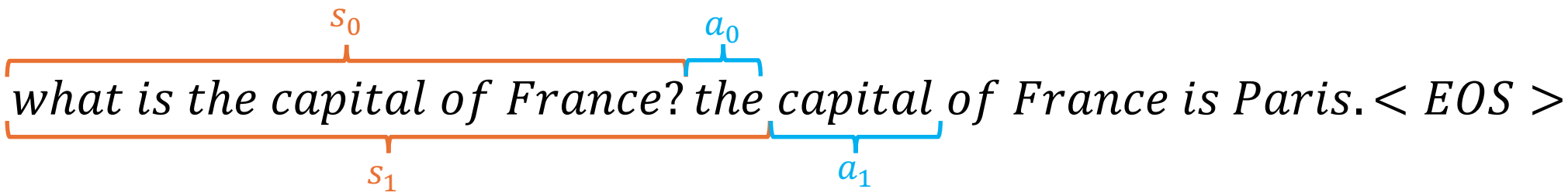
…

$t = n$:

$s_n = what\ is\ the\ capital\ of\ France?\ the\ capital\ of\ France\ is\ Paris. = s_{n-1} + a_{n-1}$

$a_n =< EOS >$

RLHF Setting – Deterministic Transition



Probability of generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \dots, s_h, a_h\}$:

$$\mathbb{P}^{\pi}(s_0, a_0, s_1, a_1, \dots, s_h, a_h)$$
$$= \mu(s_0)\pi(a_0|s_0)p(s_1|s_0, a_0) \dots p(s_h|s_{h-1}, a_{h-1})\pi(a_h|s_h)$$
$$= \mu(s_0)\pi(a_0|s_0) \dots \pi(a_h|s_h)$$

RLHF Setting – Bandit Setting

s_0 a_0

$what\ is\ the\ capital\ of\ France?\ the\ capital\ of\ France\ is\ Paris.{<}EOS{>}$

s_1 a_1

$$\mathbb{P}^{\pi}(s_0, a_0, s_1, a_1, \ldots, s_h, a_h) = \mu(s_0)\pi(a_0|s_0) \ldots \pi(a_h|s_h)$$

$what\ is\ the\ capital\ of\ France?\ the\ capital\ of\ France\ is\ Paris.{<}EOS{>}$

x y

$$x = s_0$$

$$\pi(y|x) = \pi(a_0|s_0) \ldots \pi(a_h|s_h)$$

RLHF Setting – Bandit Setting

s_0 a_0

$what\ is\ the\ capital\ of\ France?\ the\ capital\ of\ France\ is\ Paris.{<}EOS{>}$

s_1 a_1

$$\mathbb{P}^{\pi}(s_0, a_0, s_1, a_1, \ldots, s_h, a_h) = \mu(s_0)\pi(a_0|s_0) \ldots \pi(a_h|s_h)$$

$what\ is\ the\ capital\ of\ France?\ the\ capital\ of\ France\ is\ Paris.{<}EOS{>}$

x y

$$x = s_0$$

$$\pi(y|x) = \pi(a_0|s_0) \ldots \pi(a_h|s_h)$$

$\pi(y|x)$: the probability of generating y given state x

$\pi(\cdot\,|x)$: the probability distribution over all responses

RLHF Setting

$$\pi^* = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} r(x, y)$$

RLHF Setting – Prevent Reward Hacking

$$\pi^* = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} r(x, y)$$

$$\pi^* = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} [r(x, y) - \gamma(\ln \pi(y|x) - \ln \pi_0(y|x))]$$

Outline

1. ✓ LLM as Policy
2. ✓ Bradley-Terry Model as reward
3. ✓ RLHF Setting
4. Algorithm: REBEL
5. Guarantee and Proof sketch

RLHF Setting – KL Constraint RL Problem

$$\pi_{t+1} = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} r(x, y)$$

RLHF Setting – KL Constraint RL Problem

$$\pi_{t+1} = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} r(x,y) - \frac{1}{\eta} \mathbb{E}_x \text{KL}(\pi(\cdot\,|x) || \pi_t(\cdot\,|x))$$

Deriving REBEL – Closed Form Solution

$$\pi_{t+1} = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} r(x,y) - \frac{1}{\eta} \mathbb{E}_x \text{KL}(\pi(\cdot\,|x) || \pi_t(\cdot\,|x))$$

$$= \text{argmax}_{\pi \in \Pi} \mathbb{E}_x [\mathbb{E}_{y \sim \pi(\cdot|x)} r(x,y) - \frac{1}{\eta} \text{KL}(\pi(\cdot\,|x) \big|\big| \pi_t(\cdot\,|x))]$$

Deriving REBEL – Closed Form Solution

$$\pi_{t+1} = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} r(x,y) - \frac{1}{\eta} \mathbb{E}_x \text{KL}(\pi(\cdot\,|x) || \pi_t(\cdot\,|x))$$

$$= \text{argmax}_{\pi \in \Pi} \mathbb{E}_x [\mathbb{E}_{y \sim \pi(\cdot|x)} r(x,y) - \frac{1}{\eta} \text{KL}(\pi(\cdot\,|x) \big|\big| \pi_t(\cdot\,|x))]$$

$$= \text{argmax}_{\pi \in \Pi} \mathbb{E}_x \left[\mathbb{E}_{y \sim \pi(\cdot|x)} r(x,y) - \frac{1}{\eta} \sum_y \pi(y|x) \log \frac{\pi(y|x)}{\pi_t(y|x)} \right]$$

Deriving REBEL – Closed Form Solution

$$\pi_{t+1} = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y\sim\pi(\cdot|x)} r(x,y) - \frac{1}{\eta}\mathbb{E}_x \text{KL}(\pi(\cdot\,|x)||\pi_t(\cdot\,|x))$$

$$= \text{argmax}_{\pi \in \Pi} \mathbb{E}_x [\mathbb{E}_{y\sim\pi(\cdot|x)} r(x,y) - \frac{1}{\eta}\text{KL}(\pi(\cdot\,|x)||\pi_t(\cdot\,|x))]$$

$$= \text{argmax}_{\pi \in \Pi} \mathbb{E}_x \left[\mathbb{E}_{y\sim\pi(\cdot|x)} r(x,y) - \frac{1}{\eta}\sum_y \pi(y|x)\log\frac{\pi(y|x)}{\pi_t(y|x)}\right]$$

$$= \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y\sim\pi(\cdot|x)} \left[r(x,y) - \frac{1}{\eta}\log\frac{\pi(y|x)}{\pi_t(y|x)}\right]$$

Deriving REBEL – Closed Form Solution

$$\pi_{t+1} = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} \left[r(x,y) - \frac{1}{\eta} \log \frac{\pi(y|x)}{\pi_t(y|x)} \right]$$

For each state x, we have an optimization problem with respect to $\pi(y|x)$:

$$\sum_y \pi(y|x) \left[r(x,y) - \frac{1}{\eta} \log \frac{\pi(y|x)}{\pi_t(y|x)} \right]$$

Deriving REBEL – Closed Form Solution

$$\pi_{t+1} = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} \left[r(x,y) - \frac{1}{\eta} \log \frac{\pi(y|x)}{\pi_t(y|x)} \right]$$

For each state x, we have an optimization problem with respect to $\pi(y|x)$:

$$\mathcal{L}(\pi(y|x)) = \pi(y|x) \left[r(x,y) - \frac{1}{\eta} \log \frac{\pi(y|x)}{\pi_t(y|x)} \right]$$

Deriving REBEL – Closed Form Solution

$$\pi_{t+1} = \text{argmax}_{\pi \in \Pi} \mathbb{E}_{x,y \sim \pi(\cdot|x)} \left[r(x,y) - \frac{1}{\eta} \log \frac{\pi(y|x)}{\pi_t(y|x)} \right]$$

For each state x, we have an optimization problem with respect to $\pi(y|x)$:

$$\mathcal{L}(\pi(y|x)) = \pi(y|x) \left[r(x,y) - \frac{1}{\eta} \log \frac{\pi(y|x)}{\pi_t(y|x)} \right]$$

$$\frac{d\mathcal{L}}{d\pi(y|x)} = r(x,y) - \frac{1}{\eta} \left(\log \frac{\pi(y|x)}{\pi_t(y|x)} + 1 \right) = 0$$

Deriving REBEL – Closed Form Solution

$$r(x,y) - \frac{1}{\eta}\left(\log\frac{\pi(y|x)}{\pi_t(y|x)} + 1\right) = 0$$

Deriving REBEL – Closed Form Solution

$$r(x,y) - \frac{1}{\eta}\left(\log\frac{\pi(y|x)}{\pi_t(y|x)} + 1\right) = 0$$

$$\log\pi(y|x) = \log\pi_t(y|x) + \eta r(x,y) - 1$$

$$\pi(y|x) = \pi_t(y|x)\exp(\eta r(x,y) - 1)$$

Deriving REBEL – Closed Form Solution

$$r(x,y) - \frac{1}{\eta}\left(\log\frac{\pi(y|x)}{\pi_t(y|x)} + 1\right) = 0$$

$$\log\pi(y|x) = \log\pi_t(y|x) + \eta r(x,y) - 1$$

$$\pi(y|x) = \pi_t(y|x)\exp(\eta r(x,y) - 1)$$

Normalize such that $\sum_y \pi(y|x) = 1$:

$$\pi(y|x) = \frac{\pi_t(y|x)\exp(\eta r(x,y) - 1)}{\sum_{y'}\pi_t(y'|x)\exp(\eta r(x,y') - 1)} = \frac{\pi_t(y|x)\exp(\eta r(x,y))}{\sum_{y'}\pi_t(y'|x)\exp(\eta r(x,y'))}$$

Deriving REBEL – Closed Form Solution

$$\pi(y|x) = \frac{\pi_t(y|x)\exp(\eta r(x,y))}{\sum_{y'}\pi_t(y'|x)\exp(\eta r(x,y'))}$$

Closed-form solution:

$$\forall x, y : \pi_{t+1}(y|x) = \frac{\pi_t(y|x)\exp(\eta r(x,y))}{Z(x)}$$

$$Z(x) = \sum_{y'}\pi_t(y'|x)\exp(\eta r(x,y'))$$

Deriving REBEL – Closed Form Solution

$$\forall x, y : \pi_{t+1}(y|x) = \frac{\pi_t(y|x)\exp(\eta r(x,y))}{Z(x)}$$

Rewrite the reward in terms of policy:

$$\forall x, y : r(x,y) = \frac{1}{\eta}\left(\ln\big(Z(x)\big) + \ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right)\right)$$

Minimize the square difference objective:

$$\left(r(x,y) - \frac{1}{\eta}\left(\ln\big(Z(x)\big) + \ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right)\right)\right)^2$$

Deriving REBEL – Closed Form Solution

$$\forall x, y : \pi_{t+1}(y|x) = \frac{\pi_t(y|x)\exp(\eta r(x,y))}{Z(x)}$$

Rewrite the reward in terms of policy:

$$\forall x, y : r(x,y) = \frac{1}{\eta}\left(\ln\left(Z(x)\right) + \ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right)\right)$$

Minimize the square difference objective:

$$\left(r(x,y) - \frac{1}{\eta}\left(\ln\left(Z(x)\right) + \ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right)\right)\right)^2$$

Problem: partition function $Z(x)$ is hard to compute

Deriving REBEL – Closed Form Solution

$$\forall x, y : \pi_{t+1}(y|x) = \frac{\pi_t(y|x)\exp(\eta r(x,y))}{Z(x)}$$

Rewrite the reward in terms of policy:

$$\forall x, y : r(x,y) = \frac{1}{\eta}\left(\ln\big(Z(x)\big) + \ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right)\right)$$

Sample another response to cancel the partition function:

$$r(x,y) - r(x,y') = \frac{1}{\eta}\left(\ln\big(Z(x)\big) + \ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right)\right) - \frac{1}{\eta}\left(\ln\big(Z(x)\big) + \ln\left(\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)\right)$$

$$r(x,y) - r(x,y') = \frac{1}{\eta}\left(\ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right) - \ln\left(\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)\right)$$

Deriving REBEL – Closed Form Solution

$$\forall x, y : \pi_{t+1}(y|x) = \frac{\pi_t(y|x)\exp(\eta r(x,y))}{Z(x)}$$

Rewrite the reward in terms of policy:

$$\forall x, y : r(x,y) = \frac{1}{\eta}\left(\ln\big(Z(x)\big) + \ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right)\right)$$

Sample another response to cancel the partition function:

$$r(x,y) - r(x,y') = \frac{1}{\eta}\left(\ln\big(Z(x)\big) + \ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right)\right) - \frac{1}{\eta}\left(\ln\big(Z(x)\big) + \ln\left(\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)\right)$$

$$r(x,y) - r(x,y') = \frac{1}{\eta}\left(\ln\left(\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}\right) - \ln\left(\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)\right)$$

Regress the difference in rewards:

$$\left((r(x,y) - r(x,y')) - \frac{1}{\eta}\left(\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} - \ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)\right)^2$$

REgression to RElative REward Based RL (REBEL)

At iteration t with policy π_t

1. Sample data:

$$\mathcal{D}_t := \{x, y, y'\} \quad x \sim \rho, y \sim \pi_t(\cdot|x), y' \sim \pi_t(\cdot|x)$$

2. Regressing relative rewards (least square regression):

$$\pi_{t+1} = \arg\min_{\pi} \mathbb{E}_{\mathcal{D}_t} \left(\frac{1}{\eta} \underbrace{\left(\ln \frac{\pi(y|x)}{\pi_t(y|x)} - \ln \frac{\pi(y'|x)}{\pi_t(y'|x)} \right)}_{\text{Predictor}} - \underbrace{(r(x,y) - r(x,y'))}_{\text{Relative Reward}} \right)^2$$

Connection of REBEL to Mirror Descent

$$\pi_{t+1} = \arg\min_{\pi} \mathbb{E}_{\mathcal{D}_t} \left(\frac{1}{\eta} \left(\ln \frac{\pi(y|x)}{\pi_t(y|x)} - \ln \frac{\pi(y'|x)}{\pi_t(y'|x)} \right) - (r(x,y) - r(x,y')) \right)^2$$

Assume we solve the above regression “perfectly”:

A point-wise guarantee

$$\frac{1}{\eta} \left(\ln \frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} - \ln \frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)} \right) = (r(x,y) - r(x,y')), \forall x, y, y'$$

Connection of REBEL to Mirror Descent

$$\pi_{t+1} = \arg\min_{\pi} \mathbb{E}_{\mathcal{D}_t} \left(\frac{1}{\eta} \left(\ln \frac{\pi(y|x)}{\pi_t(y|x)} - \ln \frac{\pi(y'|x)}{\pi_t(y'|x)} \right) - (r(x,y) - r(x,y')) \right)^2$$

Assume we solve the above regression "perfectly":

$$\frac{1}{\eta} \left(\ln \frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} - \ln \frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)} \right) = (r(x,y) - r(x,y')), \forall x, y, y'$$

A point-wise guarantee

Then, **REBEL *recovers the exact Mirror Descent update***:

$$\exists c(x), s.t. \forall x, y : \frac{1}{\eta} \ln \frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} = r(x,y) + c(x)$$

$$\Rightarrow \pi_{t+1}(y|x) \propto \pi_t(y|x) \exp(\eta r(y,x))$$

Connection of REBEL to Natural Policy Gradient

$$\pi_{t+1} = \arg\min_{\pi} \mathbb{E}_{\mathcal{D}_t} \left(\frac{1}{\eta} \left(\ln \frac{\pi(y|x)}{\pi_t(y|x)} - \ln \frac{\pi(y'|x)}{\pi_t(y'|x)} \right) - (r(x,y) - r(x,y')) \right)^2$$

Parameterized regressor:

$$\pi_{\theta_{t+1}} = \arg\min_{\pi_\theta} \mathbb{E}_{\mathcal{D}_t} \left(\frac{1}{\eta} \left(\ln \frac{\pi_\theta(y|x)}{\pi_{\theta_t}(y|x)} - \ln \frac{\pi_\theta(y'|x)}{\pi_{\theta_t}(y'|x)} \right) - (r(x,y) - r(x,y')) \right)^2$$

Claim: NPG is REBEL with Gauss-Newton Updates

Connection of REBEL to Natural Policy Gradient

NPG Update of the KL-constrained RL objective:

$$\theta_{t+1} - \theta_t = \eta F_t^{\dagger} \left(\mathbb{E}_{x,y \sim \pi_{\theta_t}(\cdot|x)} \nabla_\theta \ln \pi_{\theta_t}(y|x)\, r(x,y) \right)$$

Gauss-Newton Update of REBEL:

$$\theta_{t+1} - \theta_t = \frac{\eta}{2} F_t^{\dagger} \mathbb{E}_{x,y \sim \pi_{\theta_t}(\cdot|x), y' \sim \pi_{\theta_t}(\cdot|x)} \big(\nabla_\theta \ln \pi_{\theta_t}(y|x) - \nabla_\theta \ln \pi_{\theta_t}(y'|x) \big) (r(x,y) - r(x,y'))$$

$$= \frac{\eta}{2} F_t^{\dagger} \mathbb{E}_{x,y \sim \pi_{\theta_t}(\cdot|x), y' \sim \pi_{\theta_t}(\cdot|x)} \big(\nabla_\theta \ln \pi_{\theta_t}(y|x)\, r(x,y) + \nabla_\theta \ln \pi_{\theta_t}(y'|x)\, r(x,y') \big)$$

$$\theta_{t+1} - \theta_t = \eta F_t^{\dagger} \mathbb{E}_{x,y \sim \pi_{\theta_t}(\cdot|x)} \nabla_\theta \ln \pi_{\theta_t}(y|x)\, r(x,y)$$

Claim: NPG is REBEL with Gauss-Newton Updates

Outline

1. ✓ LLM as Policy
2. ✓ Bradley-Terry Model as reward
3. ✓ RLHF Setting
4. ✓ Algorithm: REBEL
5. Guarantee and Proof sketch

Assumption 1

Previously when we connect REBEL to mirror descent, we assume that the regression can be solved "perfectly":

$$\frac{1}{\eta}\left(\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} - \ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right) = (r(x,y) - r(x,y')), \forall x, y, y'$$

Here, we relax this assumption for the analysis:

Assumption 1: over T iterations, we have the following for some ϵ:

$$\mathbb{E}_{x,y\sim\pi_t(\cdot|x),y'\sim\pi_t(\cdot|x)}\left(\frac{1}{\eta}\left(ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} - ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right) - (r(x,y) - r(x,y'))\right)^2 \leq \epsilon$$

Theorem Proof – Data Coverage

Given a test policy π, we denote the concentrability coefficient as:

$$C_{\pi_t \to \pi} = \max_{x,y} \frac{\pi(y|x)}{\pi_t(y|x)}$$

π_t covers π if $C_{\pi_t \to \pi} < +\infty$.

Theorem

Under assumption 1, after T many iterations, among the learned policies $\pi_1, \dots, \pi_T$, there must exist a policy π_t such that:

$$\forall \pi^* : \mathbb{E}_{x,y\sim\pi^*(\cdot|x)} r(x,y) - \mathbb{E}_{x,y\sim\pi_t(\cdot|x)} r(x,y) \leq O\left(\sqrt{\frac{1}{T}} + \sqrt{C_{max}\epsilon}\right)$$

$$C_{max} = \max_{\pi\in\{\pi_1,\dots,\pi_T\}} C_{\pi\to\pi^*}$$

Theorem – Lemma 1

$$A_t(x,y) = \frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} - \frac{1}{\eta}\mathbb{E}_{y'\sim\pi_t(\cdot|x)}\ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}$$

Assume $\max_{x,y,t}|A_t(x,y)| \le A \in \mathbb{R}^+$, and $\pi_0(\cdot\,|x)$ is uniform over $\mathcal{Y}$ (response space). Then, with $\eta = \sqrt{\ln(|\mathcal{Y}|)/(A^2T)}$,

We have:

$$\forall\pi, x: \sum_{t=0}^{T-1}\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y) \le 2A\sqrt{\ln(|\mathcal{Y}|)T}$$

Theorem – Lemma 1 Proof

$$\pi_{t+1}(y|x) = \pi_t(y|x)\exp(\eta\frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)})$$

Theorem – Lemma 1 Proof

$$\pi_{t+1}(y|x) = \pi_t(y|x)\exp(\eta\frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)})$$

$$= \frac{\pi_t(y|x)\exp(\eta(\frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} - \frac{1}{\eta}\mathbb{E}_{y'\sim\pi_t(\cdot|x)}\ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}))}{\exp(\eta(-\frac{1}{\eta}\mathbb{E}_{y'\sim\pi_t(\cdot|x)}\ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}))}$$

Theorem – Lemma 1 Proof

$$\pi_{t+1}(y|x) = \pi_t(y|x)\exp(\eta\frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)})$$

$$= \frac{\pi_t(y|x)\exp(\eta(\frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} - \frac{1}{\eta}\mathbb{E}_{y'\sim\pi_t(\cdot|x)}\ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}))}{\exp(\eta(-\frac{1}{\eta}\mathbb{E}_{y'\sim\pi_t(\cdot|x)}\ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}))}$$

With $A_t(x,y) = \frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)} - \frac{1}{\eta}\mathbb{E}_{y'\sim\pi_t(\cdot|x)}\ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}$, we can rewrite $\pi_{t+1}(y|x)$ as:

$$\pi_{t+1}(y|x) = \frac{\pi_t(y|x)\exp(\eta A_t(x,y))}{Z_t(x)}$$

$$Z_t = \mathbb{E}_{y\sim\pi_t(\cdot|x)}\exp(\eta A_t(x,y))$$

Theorem – Lemma 1 Proof

$$Z_t = \mathbb{E}_{y\sim\pi_t(\cdot|x)} \exp(\eta A_t(x,y))$$

$$\ln Z_t = \ln \mathbb{E}_{y\sim\pi_t(\cdot|x)} \exp(\eta A_t(x,y))$$

Theorem – Lemma 1 Proof

$$Z_t = \mathbb{E}_{y\sim\pi_t(\cdot|x)} \exp(\eta A_t(x,y))$$

$$\ln Z_t = \ln \mathbb{E}_{y\sim\pi_t(\cdot|x)} \exp(\eta A_t(x,y))$$

With $\max_{x,y,t} |A_t(x,y)| \leq A \in \mathbb{R}^+$ and $\eta \leq 1/A$, we have $\eta A_t(x,y) \leq 1$. Using $\exp(x) \leq 1 + x + x^2$ for any $x \leq 1$:

$$\ln Z_t \leq \ln \mathbb{E}_{y\sim\pi_t(\cdot|x)}(1 + \eta A_t(x,y) + \eta^2 A_t(x,y)^2) \leq \ln(1 + 0 + \eta^2 A^2) \leq \eta^2 A^2$$

Theorem – Lemma 1 Proof

$$\pi_{t+1}(y|x) = \frac{\pi_t(y|x)\exp(\eta A_t(x,y))}{Z_t(x)}$$

$$\mathbb{E}_{y\sim\pi(\cdot|x)} \ln \pi_{t+1}(y|x) = \mathbb{E}_{y\sim\pi(\cdot|x)}(\ln \pi_t(y|x) + \eta A_t(x,y) - \ln Z_t(x))$$

Theorem – Lemma 1 Proof

$$\pi_{t+1}(y|x) = \frac{\pi_t(y|x)\exp(\eta A_t(x,y))}{Z_t(x)}$$

$$\mathbb{E}_{y\sim\pi(\cdot|x)} \ln \pi_{t+1}(y|x) = \mathbb{E}_{y\sim\pi(\cdot|x)}(\ln \pi_t(y|x) + \eta A_t(x,y) - \ln Z_t(x))$$

$$\begin{aligned}&\mathbb{E}_{y\sim\pi(\cdot|x)} \ln \pi_{t+1}(y|x) - \mathbb{E}_{y\sim\pi(\cdot|x)} \ln \pi(y|x)\\ &= \mathbb{E}_{y\sim\pi(\cdot|x)} \ln \pi_t(y|x) - \mathbb{E}_{y\sim\pi(\cdot|x)} \ln \pi(y|x) + \mathbb{E}_{y\sim\pi(\cdot|x)}\eta A_t(x,y) - \mathbb{E}_{y\sim\pi(\cdot|x)} \ln Z_t(x)\end{aligned}$$

$$-KL(\pi(\cdot\,|x)||\pi_{t+1}(\cdot\,|x)) = -KL(\pi(\cdot\,|x)||\pi_t(\cdot\,|x)) + \eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y) - \mathbb{E}_{y\sim\pi(\cdot|x)} \ln Z_t(x)$$

$$KL(\pi(\cdot\,|x)||\pi_{t+1}(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_t(\cdot\,|x)) = -\eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y) + \mathbb{E}_{y\sim\pi(\cdot|x)} \ln Z_t(x)$$

Theorem – Lemma 1 Proof

$$KL(\pi(\cdot\,|x)||\pi_{t+1}(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_t(\cdot\,|x)) = -\eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y) + \mathbb{E}_{y\sim\pi(\cdot|x)}\ln Z_t(x)$$

Theorem – Lemma 1 Proof

$$KL(\pi(\cdot\,|x)||\pi_{t+1}(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_t(\cdot\,|x)) = -\eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y) + \mathbb{E}_{y\sim\pi(\cdot|x)}\ln Z_t(x)$$

Given $\ln Z_t \le \eta^2 A^2$, we have:

$$KL(\pi(\cdot\,|x)||\pi_{t+1}(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_t(\cdot\,|x)) \le \eta^2 A^2 - \eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y)$$

Theorem – Lemma 1 Proof

$$KL(\pi(\cdot\,|x)||\pi_{t+1}(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_t(\cdot\,|x)) = -\eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y) + \mathbb{E}_{y\sim\pi(\cdot|x)}\ln Z_t(x)$$

Given $\ln Z_t \le \eta^2 A^2$, we have:

$$KL(\pi(\cdot\,|x)||\pi_{t+1}(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_t(\cdot\,|x)) \le \eta^2 A^2 - \eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y)$$

Sum and telescope:

$$KL(\pi(\cdot\,|x)||\pi_T(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_0(\cdot\,|x)) \le \sum_{t=0}^{T-1}(\eta^2 A^2 - \eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y))$$

Theorem – Lemma 1 Proof

$$KL(\pi(\cdot\,|x)||\pi_{t+1}(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_t(\cdot\,|x)) = -\eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y) + \mathbb{E}_{y\sim\pi(\cdot|x)}\ln Z_t(x)$$

Given $\ln Z_t \le \eta^2 A^2$, we have:

$$KL(\pi(\cdot\,|x)||\pi_{t+1}(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_t(\cdot\,|x)) \le \eta^2 A^2 - \eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y)$$

Sum and telescope:

$$KL(\pi(\cdot\,|x)||\pi_T(\cdot\,|x)) - KL(\pi(\cdot\,|x)||\pi_0(\cdot\,|x)) \le \sum_{t=0}^{T-1}(\eta^2 A^2 - \eta\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y))$$

$$\sum_{t=0}^{T-1}\mathbb{E}_{y\sim\pi(\cdot|x)}A_t(x,y) \le T\eta A^2 + KL(\pi(\cdot\,|x)||\pi_0(\cdot\,|x))/\eta \le T\eta A^2 + \ln(|\mathcal{Y}|)/\eta$$

Theorem – Lemma 1 Proof

$$\sum_{t=0}^{T-1} \mathbb{E}_{y\sim\pi(\cdot|x)} A_t(x,y) \leq T\eta A^2 + \ln(|\mathcal{Y}|)\,/\eta$$

With $\eta = \sqrt{\ln(|\mathcal{Y}|)/(A^2T)}$:

$$\sum_{t=0}^{T-1} \mathbb{E}_{y\sim\pi(\cdot|x)}\, A_t(x,y) \leq A\sqrt{\mathrm{T}\ln(|\mathcal{Y}|)} + A\sqrt{\mathrm{T}\ln(|\mathcal{Y}|)} = 2A\sqrt{\ln(|\mathcal{Y}|)T}$$

Theorem Proof

Assumption 1: over T iterations, we have the following for some ϵ:

$$\mathbb{E}_{x,y\sim\pi_t(\cdot|x),y\prime\sim\pi_t(\cdot|x)}\left(\frac{1}{\eta}\left(ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}-ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)-(r(x,y)-r(x,y'))\right)^2\leq\epsilon$$

Lemma 1: Assume $\max_{x,y,t}|A_t(x,y)|\leq A\in\mathbb{R}^+$, and $\pi_0(\cdot\,|x)$ is uniform over $\mathcal{Y}$ (response space). Then, with $\eta=\sqrt{\ln(|\mathcal{Y}|)/(A^2T)}$, We have:

$$\forall\pi,x:\sum_{t=0}^{T-1}\mathbb{E}_{y\sim\pi(\cdot|x)}\,A_t(x,y)\leq 2A\sqrt{ln(|\mathcal{Y}|)T}$$

$$A_t(x,y)=\frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}-\frac{1}{\eta}\mathbb{E}_{y\prime\sim\pi_t(\cdot|x)}\,ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}$$

Prove: after T many iterations, among the learned policies $\pi_1,\ldots,\pi_T$, there must exist a policy π_t such that:

$$\forall\pi^*:\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}r(x,y)-\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}r(x,y)\leq O\left(\sqrt{\frac{1}{T}}+\sqrt{C_{max}\epsilon}\right)$$

Theorem Proof

We start by bounding the performance difference between π^* and uniform mixture policy $\sum_{t=0}^{T-1} \pi_t / T$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)} r(x,y) - \mathbb{E}_{x,y\sim\pi_t(\cdot|x)} r(x,y)\right]$$

$$= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{x,y\sim\pi^*(\cdot|x)}\, A_t(x,y) + \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}(r(x,y) - A_t(x,y)) - \mathbb{E}_{x,y\sim\pi_t(\cdot|x)} r(x,y)\right]$$

Theorem Proof

We start by bounding the performance difference between π^* and uniform mixture policy $\sum_{t=0}^{T-1} \pi_t / T$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)} r(x,y) - \mathbb{E}_{x,y\sim\pi_t(\cdot|x)} r(x,y)\right]$$

$$= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{x,y\sim\pi^*(\cdot|x)}\, A_t(x,y) + \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}(r(x,y) - A_t(x,y)) - \mathbb{E}_{x,y\sim\pi_t(\cdot|x)} r(x,y)\right]$$

With Lemma 1:

$$\leq 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}(r(x,y) - A_t(x,y)) - \mathbb{E}_{x,y\sim\pi_t(\cdot|x)} r(x,y)\right]$$

$$= 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}(r(x,y) - \mathbb{E}_{y'\sim\pi_t(\cdot|x)} r(x,y') - A_t(x,y))\right]$$

Theorem Proof

We start by bounding the performance difference between π^* and uniform mixture policy $\sum_{t=0}^{T-1} \pi_t / T$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)} r(x,y) - \mathbb{E}_{x,y\sim\pi_t(\cdot|x)} r(x,y)\right]$$

$$= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{x,y\sim\pi^*(\cdot|x)}\, A_t(x,y) + \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}(r(x,y) - A_t(x,y)) - \mathbb{E}_{x,y\sim\pi_t(\cdot|x)} r(x,y)\right]$$

With Lemma 1:

$$\leq 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}(r(x,y) - A_t(x,y)) - \mathbb{E}_{x,y\sim\pi_t(\cdot|x)} r(x,y)\right]$$

$$= 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}(r(x,y) - \mathbb{E}_{y\prime\sim\pi_t(\cdot|x)} r(x,y') - A_t(x,y))\right]$$

$$\leq 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}\left(r(x,y) - \mathbb{E}_{y'\sim\pi_t(\cdot|x)} r(x,y') - A_t(x,y)\right)^2\right]^{1/2}$$

$$= 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}\frac{\pi^*(y|x)}{\pi_t(y|x)}\left(r(x,y) - \mathbb{E}_{y'\sim\pi_t(\cdot|x)} r(x,y') - A_t(x,y)\right)^2\right]^{1/2}$$

$$\leq 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\left[C_{\pi_t\to\pi^*}\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}\left(r(x,y) - \mathbb{E}_{y'\sim\pi_t(\cdot|x)} r(x,y') - A_t(x,y)\right)^2\right]^{1/2}$$

Theorem Proof

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}r(x,y)-\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}r(x,y)\right]$$

$$\leq\ 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\frac{1}{T}\sum_{t=0}^{T-1}\left[C_{\pi_t\to\pi^*}\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}\left(r(x,y)-\mathbb{E}_{y'\sim\pi_t(\cdot|x)}r(x,y')-A_t(x,y)\right)^2\right]^{1/2}$$

$$=\ 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\frac{1}{T}\sum_{t=0}^{T-1}\left[C_{\pi_t\to\pi^*}\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}\left(r(x,y)-\mathbb{E}_{y'\sim\pi_t(\cdot|x)}r(x,y')-\frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}+\frac{1}{\eta}\mathbb{E}_{y'\sim\pi_t(\cdot|x)}\,ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)^2\right]^{1/2}$$

Theorem Proof

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}r(x,y)-\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}r(x,y)\right]$$

$$\leq\ 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\frac{1}{T}\sum_{t=0}^{T-1}\left[C_{\pi_t\to\pi^*}\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}\left(r(x,y)-\mathbb{E}_{y'\sim\pi_t(\cdot|x)}r(x,y')-A_t(x,y)\right)^2\right]^{1/2}$$

$$=\ 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\frac{1}{T}\sum_{t=0}^{T-1}\left[C_{\pi_t\to\pi^*}\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}\left(r(x,y)-\mathbb{E}_{y'\sim\pi_t(\cdot|x)}r(x,y')-\frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}+\frac{1}{\eta}\mathbb{E}_{y'\sim\pi_t(\cdot|x)}\,ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)^2\right]^{1/2}$$

$$\leq\ 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\frac{1}{T}\sum_{t=0}^{T-1}\left[C_{\pi_t\to\pi^*}\mathbb{E}_{x,y\sim\pi_t(\cdot|x),y'\sim\pi_t(\cdot|x)}\left(\frac{1}{\eta}\left(ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}-ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)-(r(x,y)-r(x,y'))\right)^2\right]^{1/2}$$

Theorem Proof

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}r(x,y)-\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}r(x,y)\right]$$

$$\leq\ 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\frac{1}{T}\sum_{t=0}^{T-1}\left[C_{\pi_t\to\pi^*}\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}\left(r(x,y)-\mathbb{E}_{y'\sim\pi_t(\cdot|x)}r(x,y')-A_t(x,y)\right)^2\right]^{1/2}$$

$$=\ 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\frac{1}{T}\sum_{t=0}^{T-1}\left[C_{\pi_t\to\pi^*}\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}\left(r(x,y)-\mathbb{E}_{y'\sim\pi_t(\cdot|x)}r(x,y')-\frac{1}{\eta}\ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}+\frac{1}{\eta}\mathbb{E}_{y'\sim\pi_t(\cdot|x)}\,ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)^2\right]^{1/2}$$

$$\leq\ 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\frac{1}{T}\sum_{t=0}^{T-1}\left[C_{\pi_t\to\pi^*}\mathbb{E}_{x,y\sim\pi_t(\cdot|x),y'\sim\pi_t(\cdot|x)}\left(\frac{1}{\eta}\left(ln\frac{\pi_{t+1}(y|x)}{\pi_t(y|x)}-ln\frac{\pi_{t+1}(y'|x)}{\pi_t(y'|x)}\right)-(r(x,y)-r(x,y'))\right)^2\right]^{1/2}$$

With Assumption 1:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\mathbb{E}_{x,y\sim\pi^*(\cdot|x)}r(x,y)-\mathbb{E}_{x,y\sim\pi_t(\cdot|x)}r(x,y)\right]\leq 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{C_{\pi_t\to\pi^*}\epsilon}\leq 2A\sqrt{\frac{ln(|\mathcal{Y}|)}{T}}+\sqrt{C_{max}\epsilon}$$