

# Controllable Generation

## Recap: KL-reg RL objective (traj-wise)

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

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

Stay close to  $\pi_{ref}$

Optimize reward

## Recap: DPO and REBEL

DPO:

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

REBEL:

$$\theta_{t+1} = \arg \min_{\theta} \mathbb{E}_{x, (\tau, \tau') \sim \pi_{\theta_t}(\cdot | x)} \left( \beta \left( \ln \frac{\pi_{\theta}(\tau | x)}{\pi_{\theta_t}(\tau | x)} - \ln \frac{\pi_{\theta}(\tau' | x)}{\pi_{\theta_t}(\tau' | x)} \right) - (r(x, \tau) - r(x, \tau')) \right)^2$$

# One more RL algorithm: GRPO (Deepseek-R1)

Basically some combination PPO clipping with RLoO (Reinforce w/ leave-one-out)

Given  $\pi_t$ , it updates policy to  $\pi_{t+1}$  as follows:

1. Sample a bunch of prompts, for each  $x$ , generate  $k$  i.i.d responses  $\tau^1, \tau^2, \dots, \tau^k$
2. Form the following clipping-based objective

$$\max_{\pi} \sum_{\{x, \tau^1, \tau^2, \dots, \tau^k\}} \sum_{i=1}^k \min \left\{ \frac{\pi(\tau^i | x)}{\pi_t(\tau^i | x)} A(x, \tau^i), \text{clip} \left( \frac{\pi(\tau^i | x)}{\pi_t(\tau^i | x)}, 1 - \epsilon, 1 + \epsilon \right) A(x, \tau^i) \right\}$$

where:

$$A(x, \tau^i) = \frac{r(\tau_i) - \bar{r}}{\text{std} (r(\tau^1), r(\tau^2), \dots, r(\tau^k))}$$

Normalize advantage use group responses  $\tau^1, \dots, \tau^k$ , per prompt;

## Today's question

So far, DPO, PPO, REBEL, and GRPO all optimize the entire LLM; when LLM is large (e.g., > 70B ), we cannot afford to do full parameter optimization...

Q: can we train small evaluation model (e.g., 3B) to **guide the generation** of a big large black-box model (e.g., 70B)?

# Outline

1. KL regularized RL again, but in token space (i.e.,  $s_h, a_h$ ) not traj space
2. Train value/Q functions
3. Controllable generation via guidance from Q/V functions

# Notation

Finite horizon MDP with deterministic transition, i.e.,  $s_{h+1} = f(s_h, a_h)$

$\pi_{ref}$  — a black-box large model, don't want to do full backpropagation on it...

We want to optimize KL-regularized RL objective:

$$\max_{\pi} \mathbb{E}_{\tau \sim \pi} \left[ \sum_{h=0}^{H-1} r(s_h, a_h) - \beta \text{KL} \left( \pi(\cdot | s_h) \parallel \pi_{ref}(\cdot | s_h) \right) \right]$$

Let's solve this via Dynamic Programming (backward in time)

# DP for solving the KL-regularized RL

$$\max_{\pi} \mathbb{E}_{\tau \sim \pi} \left[ \sum_{h=0}^{H-1} r(s_h, a_h) - \beta \text{KL} \left( \pi(\cdot | s_h) | \pi_{ref}(\cdot | s_h) \right) \right]$$

---

**Base case:**  $V^*(s_H) = 0$ , for the fictitious step  $H$

**Induction step:** given  $V^*(s_{h+1})$ , want to compute  $V^*(s_h)$

$$Q^*(s_h, a_h) = r(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim P(\cdot | s_h, a_h)} V^*(s_{h+1})$$

$$V^*(s_h) = \max_{\pi(\cdot | s_h) \in \Delta(A)} \mathbb{E}_{a \sim \pi(\cdot | s_h)} Q^*(s_h, a) - \beta \text{KL} \left( \pi(\cdot | s_h) | \pi_{ref}(\cdot | s_h) \right) \Rightarrow \pi^*(a | s_h) \propto \pi_{ref}(a | s_h) \exp(Q^*(s_h, a) / \beta)$$

$$V^*(s_h) = \beta \ln \mathbb{E}_{a_h \sim \pi_{ref}(\cdot | s_h)} \left[ \exp(Q^*(s_h, a_h) / \beta) \right]$$

(Exercise: show  $V^*(s) \rightarrow \max_a Q^*(s, a)$ , when  $\beta \rightarrow 0$ , assuming  $\pi_{ref}(a | s) > 0, \forall a$ )



## DP for solving the KL-regularized RL

$$V^*(s_h) = \beta \ln \mathbb{E}_{a_h \sim \pi_{ref}(\cdot | s_h)} [\exp(Q^*(s_h, a_h)/\beta)]$$

Now let's assume transition is deterministic, i.e.,  $s_{h+1} = f(s_h, a_h)$ , and see if we can further simplify  $V^*$

$$\exp(V^*(s_h)/\beta) = \mathbb{E}_{a_h \sim \pi_{ref}(\cdot | s_h)} [\exp(r_h/\beta + V^*(s_{h+1})/\beta)], \text{ where } s_{h+1} = f(s_h, a_h)$$

$$= \mathbb{E}_{a_h \sim \pi_{ref}(\cdot | s_h)} \exp(r_h/\beta) \exp(V^*(s_{h+1})/\beta) \quad \text{Recursion again}$$

$$= \mathbb{E}_{a_h \sim \pi_{ref}(\cdot | s_h)} \exp(r_h/\beta) \mathbb{E}_{a_{h+1} \sim \pi_{ref}(\cdot | s_{h+1})} \exp(r_{h+1}/\beta) \exp(V^*(s_{h+2}))$$

$$= \mathbb{E}_{\tau \sim \pi_{ref}(\cdot | s_h)} \exp \left( \sum_{\tau=h}^{H-1} r_\tau / \beta \right)$$

$\tau \sim \pi_{ref}(\cdot | s_h)$ : Denotes generating a future trajectory using  $\pi_{ref}$  from state  $s_h$

**In summary, when transition is deterministic, we have**

$$\forall h, s : \exp(V^*(s_h)/\beta) = \mathbb{E}_{\pi_{ref}} \left[ \exp \left( \sum_{\tau=h}^{H-1} r_{\tau}/\beta \right) \mid s_h \right]$$

$$\forall h, s, a : \exp(Q^*(s_h, a_h)/\beta) = \mathbb{E}_{\pi_{ref}} \left[ \exp \left( \sum_{\tau=h}^{H-1} r_{\tau}/\beta \right) \mid s_h, a_h \right]$$

Note the expectation is always wrt to the future generated from  $\pi_{ref}$

**In summary, when transition is deterministic, we have**

Also recall the optimal policy format:

$$\pi^*(a_h | s_h) \propto \pi_{ref}(a_h | s_h) \exp(Q^*(s_h, a_h) / \beta)$$

As long as we can learn  $\exp(Q^*(s, a) / \beta)$ , then we can use it to guide  $\pi_{ref}$

Q: why this format of  $\pi^*$  is tractable and implementable?

# Outline

1. KL regularized RL again, but in token space (i.e.,  $s_h, a_h$ ) not traj space
2. Train value/Q functions
3. Controllable generation via guidance from Q/V functions

**Recall the format of  $Q^*/V^*$**

$$\forall h, s : \exp(Q^*(s_h, a_h)/\beta) = \mathbb{E}_{\pi_{ref}} \left[ \exp \left( \sum_{\tau=h}^{H-1} r_{\tau}/\beta \right) \mid s_h, a_h \right]$$

## Learn $\exp(Q^*/\beta)$

1. Data collection: generate N i.i.d trajectories from  $\pi_{ref}$ ,  $\tau^1, \dots, \tau^N \sim \pi_{ref}$

2. For each  $s_h, a_h \in \tau^i$ , compute reward-to-go  $y = \sum_{\tau=h}^{H-1} r_\tau$  on  $\tau^i$

3 Given the data  $\{(s, a), y\}$ , train  $g$  via least square regression:

$$\hat{g} = \min_g \sum_{(s,a), y \in \mathcal{D}} (g(s, a) - \exp(y/\beta))^2$$

Q: what's the Bayes optimal of this regression problem?

**Learn  $\exp(V^*/\beta)$**

$$\hat{g} = \min_g \sum_{(s,a), y \in \mathcal{D}} (g(s, a) - \exp(y/\beta))^2$$

Bayes opt:  $\mathbb{E}_{\pi_{ref}} \left[ \exp\left( \sum_{\tau=h}^{H-1} r_{\tau}/\beta \right) \mid s_h = s, a_h = a \right] = \exp(Q^*(s, a)/\beta)$

Under reasonable conditions, let's assume  $\hat{g} \approx$  Bayes opt

# Outline

1. KL regularized RL again, but in token space (i.e.,  $s_h, a_h$ ) not traj space
2. Train value/Q functions
3. Controllable generation via guidance from Q/V functions



# Controlling generation via $Q^*$ guidance

Least square regression learns

$$\hat{g}(s_h, a_h) \approx \exp(Q^*(s_h, a_h)/\beta) = \mathbb{E}_{\pi_{ref}} \left[ \exp \left( \sum_{\tau=h}^{H-1} r_{\tau}/\beta \right) \mid s_h, a_h \right]$$

# Controlling generation via $Q^*$ guidance

During inference time, we can generate a trajectory use the following policy:

$$\pi(a | s_h) \propto \pi_{ref}(a | s_h) \hat{g}(s_h, a)$$

1. Note that we never do backpropagation on  $\pi_{ref}$
2.  $\hat{g}(s_h, a) \approx \exp(Q^*(s_h, a)/\beta)$  is an evaluation function, not a generator!

# Experiment results

Guide black-box models to do math reasoning

**Reward:** 0/1 — 1 means model's final answer is correct solution, 0 otherwise

For reasoning tasks, we often prompt model to think **step-by-step before outputting a final answer**  
(aka chain-of-thought (CoT) reasoning)

**Problem from GSM8K:** According to its nutritional info, a bag of chips has 250 calories per serving. If a 300g bag has 5 servings, how many grams can you eat if your daily calorie target is 2000 and you have already consumed 1800 calories?

**Incorrect Solution from  $\pi^{\text{ref}}$ :** Let's break this down step by step!

1. The bag of chips has 5 servings, and each serving has 250 calories. To find the total calories in the bag, multiply the number of servings by the calories per serving:

$$5 \text{ servings} \times 250 \text{ calories/serving} = 1250 \text{ calories}$$

2. You have already consumed 1800 calories, and your daily calorie target is 2000. To find out how many more calories you can consume, subtract the calories you've already had from your daily target:

$$2000 - 1800 = 200 \text{ calories}$$

3. Since each serving of chips has 250 calories, you can consume:

$$200 \text{ calories} \div 250 \text{ calories/serving} = 0.8 \text{ servings}$$

4. Since there are 5 servings in the bag, you can eat:

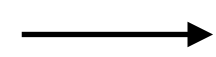
$$0.8 \text{ servings} \times 300\text{g}/\text{bag} = 240\text{g}$$

So, you can eat up to 240g of the bag of chips.

Here's the answer:

240 g

$\pi_{\text{ref}}$  made a  
mistake



# Experiment results

Train a small model (1B)  $\hat{g}(s, a)$  to approximate  $\exp(Q^*(s, a)/\beta)$

Define the new policy as  $\pi_{new} \propto \pi_{ref}(a | s) \cdot \hat{g}(s, a)$

Let's compare the generation from  $\pi_{ref}$  and  $\pi_{new}$



**Incorrect Solution from  $\pi^{\text{ref}}$ :** Let's break this down step by step!

1. The bag of chips has 5 servings, and each serving has 250 calories. To find the total calories in the bag, multiply the number of servings by the calories per serving:  
 $5 \text{ servings} \times 250 \text{ calories/serving} = 1250 \text{ calories}$
2. You have already consumed 1800 calories, and your daily calorie target is 2000. To find out how many more calories you can consume, subtract the calories you've already had from your daily target:  
 $2000 - 1800 = 200 \text{ calories}$
3. Since each serving of chips has 250 calories, you can consume:  
 $200 \text{ calories} \div 250 \text{ calories/serving} = 0.8 \text{ servings}$

4. **Since there are 5 servings in the bag, you can eat:**  
 $0.8 \text{ servings} \times 300\text{g/bag} = 240\text{g}$   
So, you can eat up to 240g of the bag of chips.  
Here's the answer:

240 g

Almost the same

**Correct Solution from  $Q\ddagger$ :** A delicious problem!  
Let's break it down step by step:

1. The bag of chips has 250 calories per serving, and there are 5 servings in the bag.
2. The total calories in the bag are:  $250 \text{ calories/serving} \times 5 \text{ servings} = 1250 \text{ calories}$ .
3. You have already consumed 1800 calories, so you have  $2000 - 1800 = 200 \text{ calories}$  left for the day.
4. Since each serving of chips has 250 calories, you can consume:  $200 \text{ calories} / 250 \text{ calories/serving} = 0.8 \text{ servings}$ .

5. **Since there are 5 servings in the bag, you can eat:**  $0.8 \text{ servings} \times 300\text{g}/5 \text{ servings} = 0.8 \times 60\text{g} = 48\text{g}$ .  
So, you can eat up to 48g of the bag of chips.  
Enjoy your snacking, but don't overdo it!

Corrected the mistake

# Summary

1. DP on the KL-regularized RL objective
2. With the KL-reg, now the hard  $\max_a$  is replaced by a soft-max operator  $\ln \mathbb{E} \exp(\dots)$
3. The  $V^*/Q^*$  has very simple form when transition is deterministic
4. Simple regression to learn a approximator of  $\exp(Q^*/\beta)$  directly, and use it to guide  $\pi_{ref}$  in generation