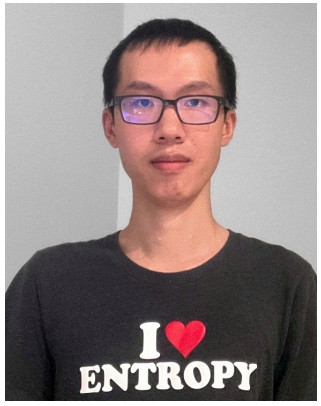
Regularization in the Face of Uncertainty

Yuejie Chi

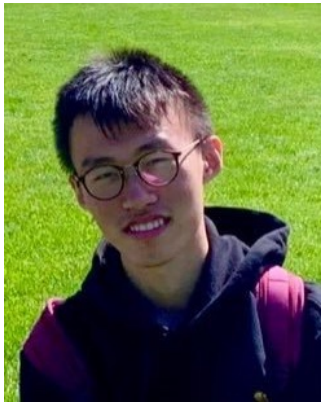
Yale & FAIR

Swiss CLOCK Summit
September 2025

Tong Yang
CMU



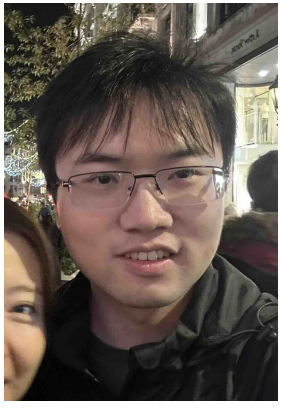
Shicong Cen
CMU→XTX



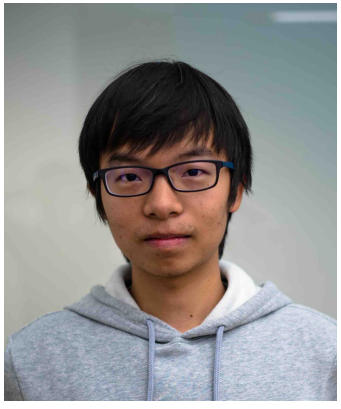
Zixin Wen
CMU

Lin Xiao
Meta

Bo Dai
GaTech/Google



Jincheng Mei
Google

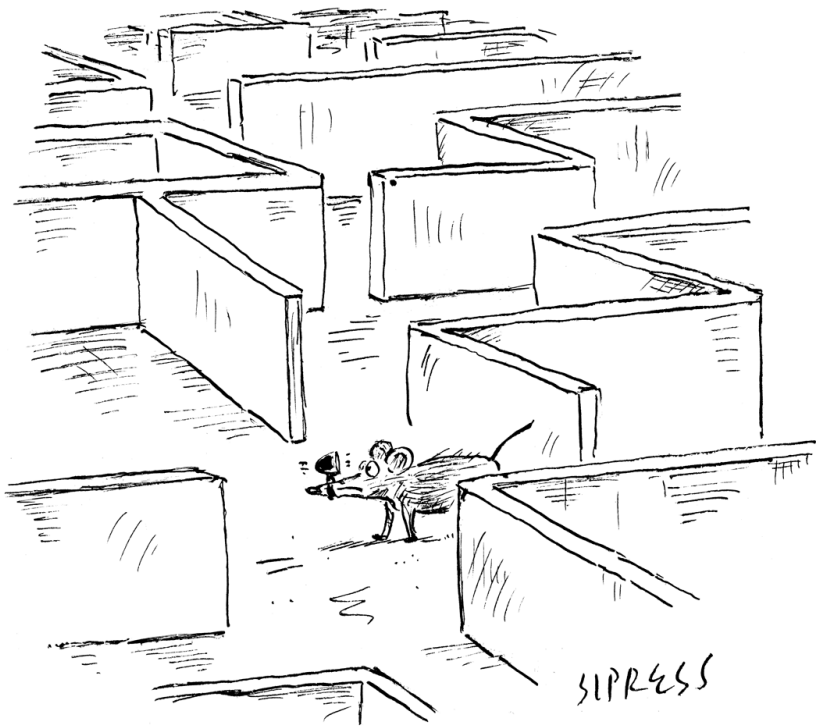


Hanjun Dai
Google

Dale Schuurmans
Alberta/Google

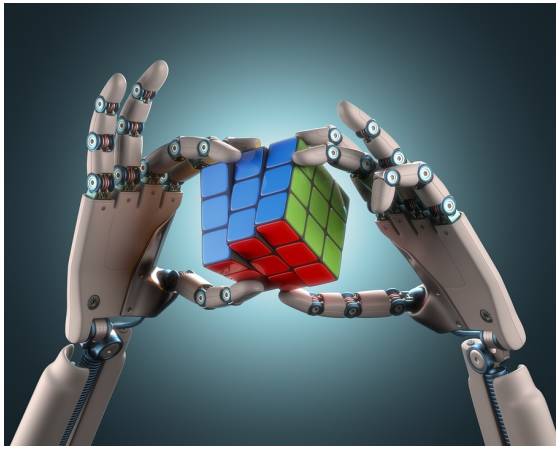
Reinforcement learning (RL)

In RL, an agent learns by interacting with an *unknown* environment through trial-and-error to maximize long-term total reward.

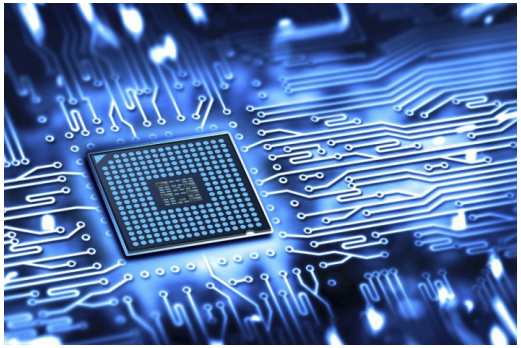


"Recalculating ... recalculating ..."

More successes of RL since AlphaGo



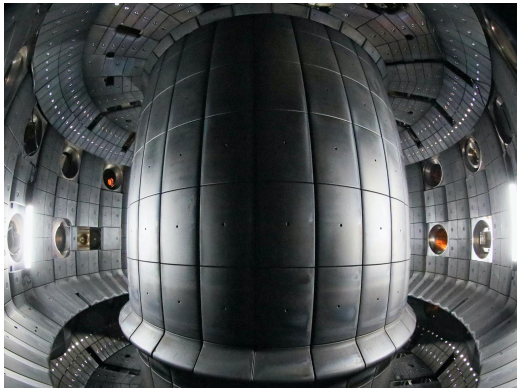
robotics



chip designs

resource management

strategic games



nuclear plant control

UAV and drones

One more: RL for foundation models

ChatGPT

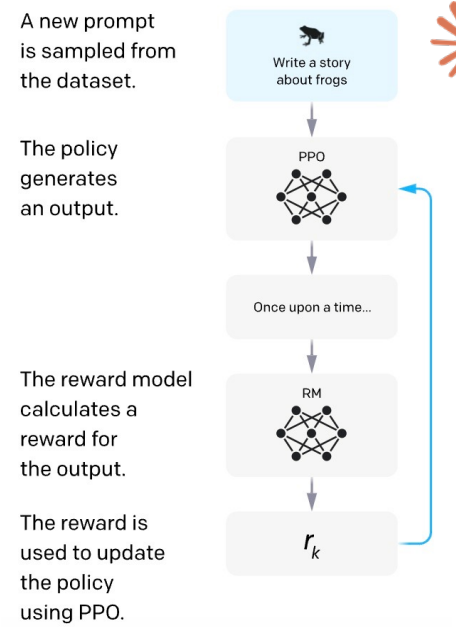
Gemini

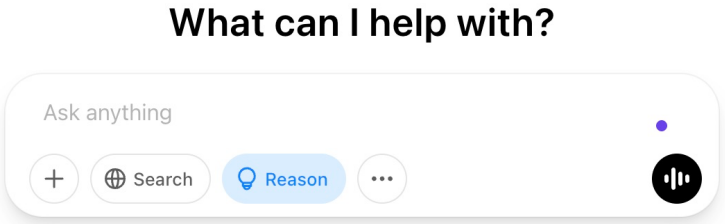
Claude

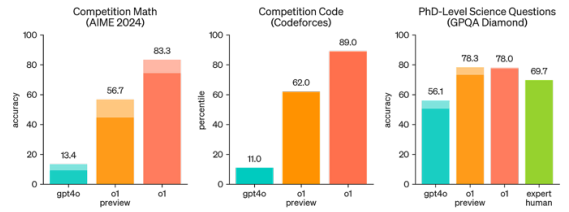
Meta AI

deepseek

Optimize a policy against the reward model using reinforcement learning.







Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a-\sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a-\sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a-\sqrt{a+x}}\right)^2 = x^2 \implies a-\sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a-x^2)^2 = a+x \implies a^2-2ax^2+(x^2)^2 = a+x \implies x^4-2ax^2-x+(a^2-a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a-\sqrt{a+x}} = x$

First, let's square both sides:

$a-\sqrt{a+x} = x^2 \implies \sqrt{a+x} = a-x^2$

Next, I could square both sides again, treating the equation: $\cdots$

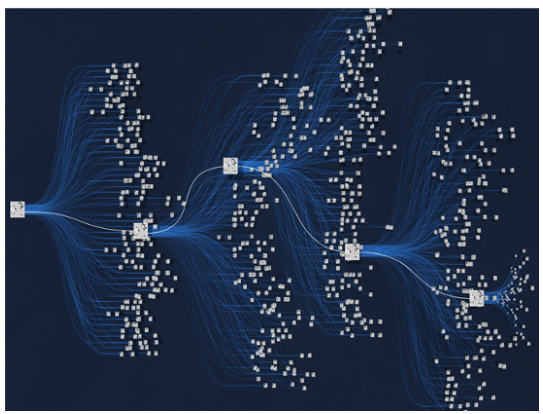
$\ldots$

Alignment: safety, human value..

Reasoning: math, coding…

Challenges of RL

- explore or exploit: unknown or changing environments
- credit assignment problem: delayed rewards or feedback
- enormous state and action space



This talk:
exploration with complex function approximation (like LLMs)

Classical wisdom

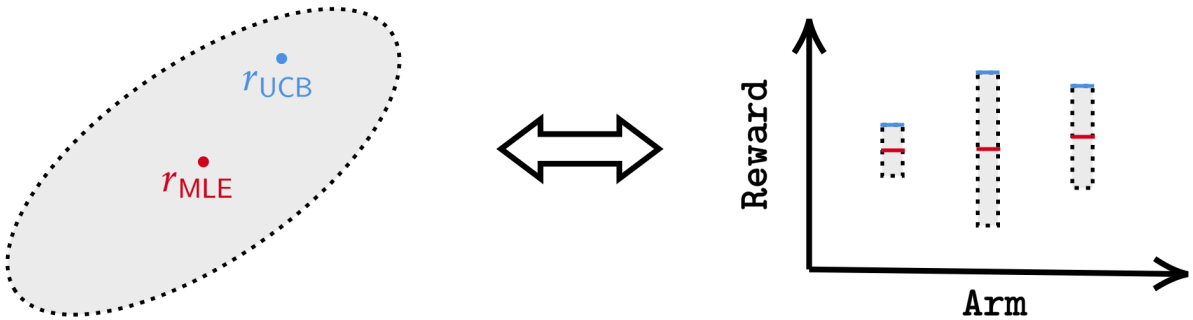
T. L. Lai

H. Robbins

Optimism via UCB in the face of uncertainty:

- explores the best optimistic estimates associated with the actions.
- a common framework: utilize upper confidence bounds (UCB)

accounts for estimates + uncertainty level



Issue: UCB performs poorly under complex function approximation.

Classical wisdom

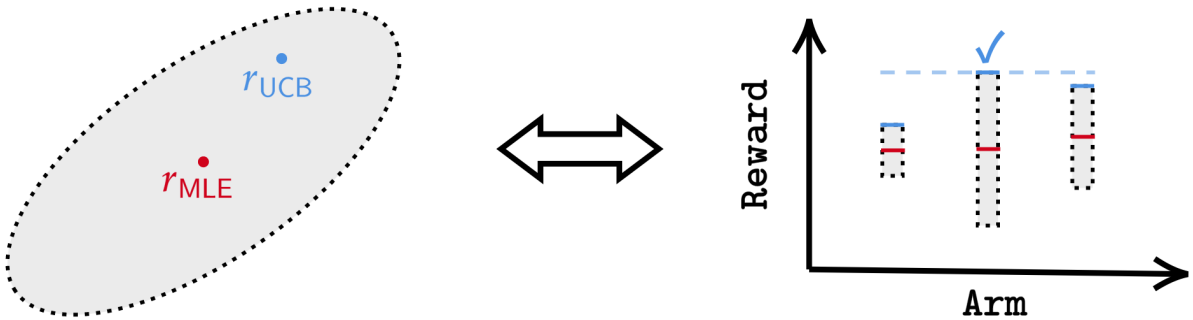
T. L. Lai

H. Robbins

Optimism via UCB in the face of uncertainty:

- explores the best optimistic estimates associated with the actions.
- a common framework: utilize upper confidence bounds (UCB)

accounts for estimates + uncertainty level



Issue: UCB performs poorly under complex function approximation.

This talk

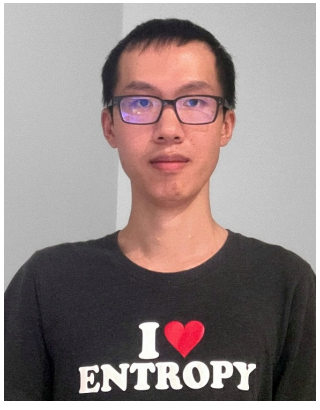
Goal: theoretically-grounded and optimization-friendly exploration scheme compatible with complex function approximation

Optimism via regularization in the face of uncertainty:

$$f_t = \arg\min_{f \in \mathcal{F}} \underbrace{\mathcal{L}_t(f; \mathcal{D}_t)}_{\text{data consistency}} - \alpha \underbrace{V^\star(f)}_{\text{optimal value}}$$

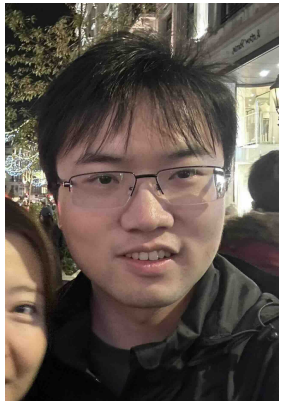
- f: can be either model-based or model-free.
- The key idea is inspired by the reward-biased estimation framework (Kumar and Lin, 1982 and follow-ups) for adaptive control, which is further developed recently for RL.
- This talk provides new computationally tractable and provably efficient vignettes for RLHF, RL and competitive games.

Value-incentivized exploration in RLHF



Shicong Cen
CMU→XTX

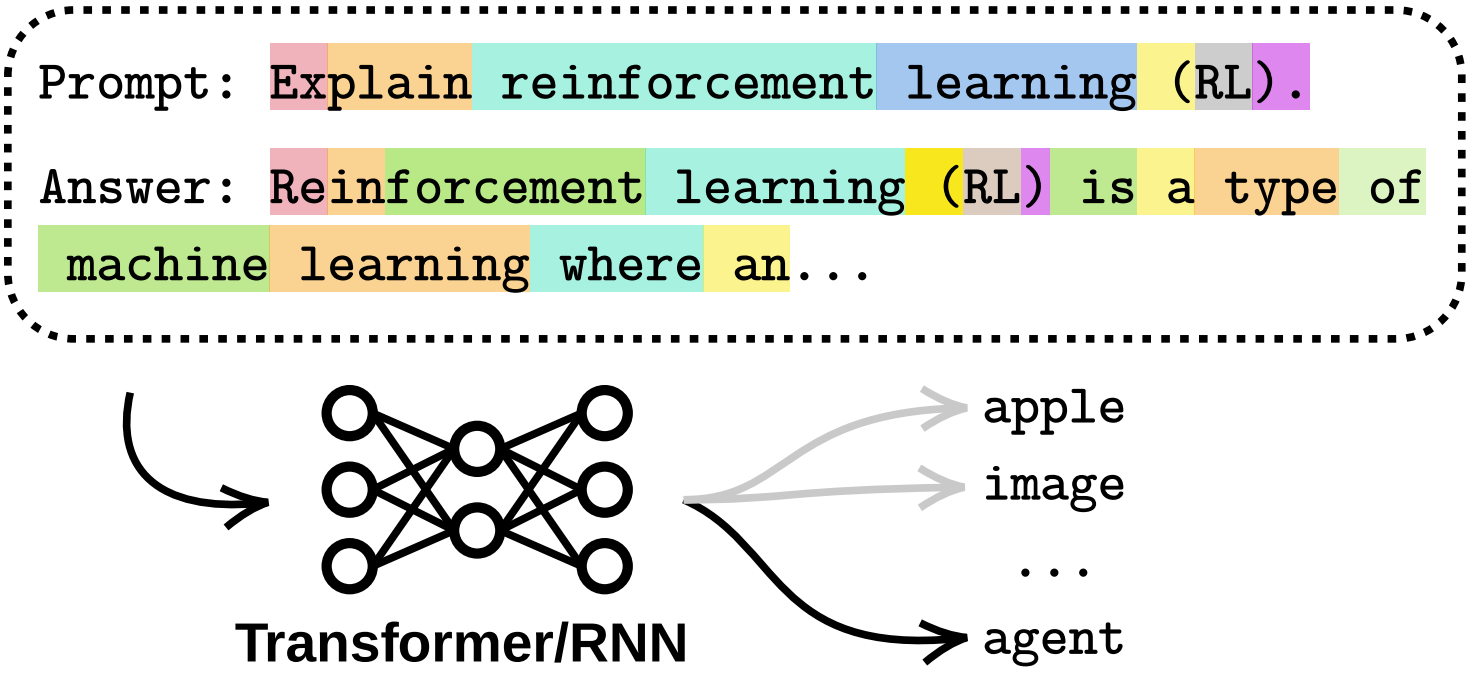
Bo Dai
GaTech/Google



Jincheng Mei
Google

Dale Schuurmans
Alberta/Google

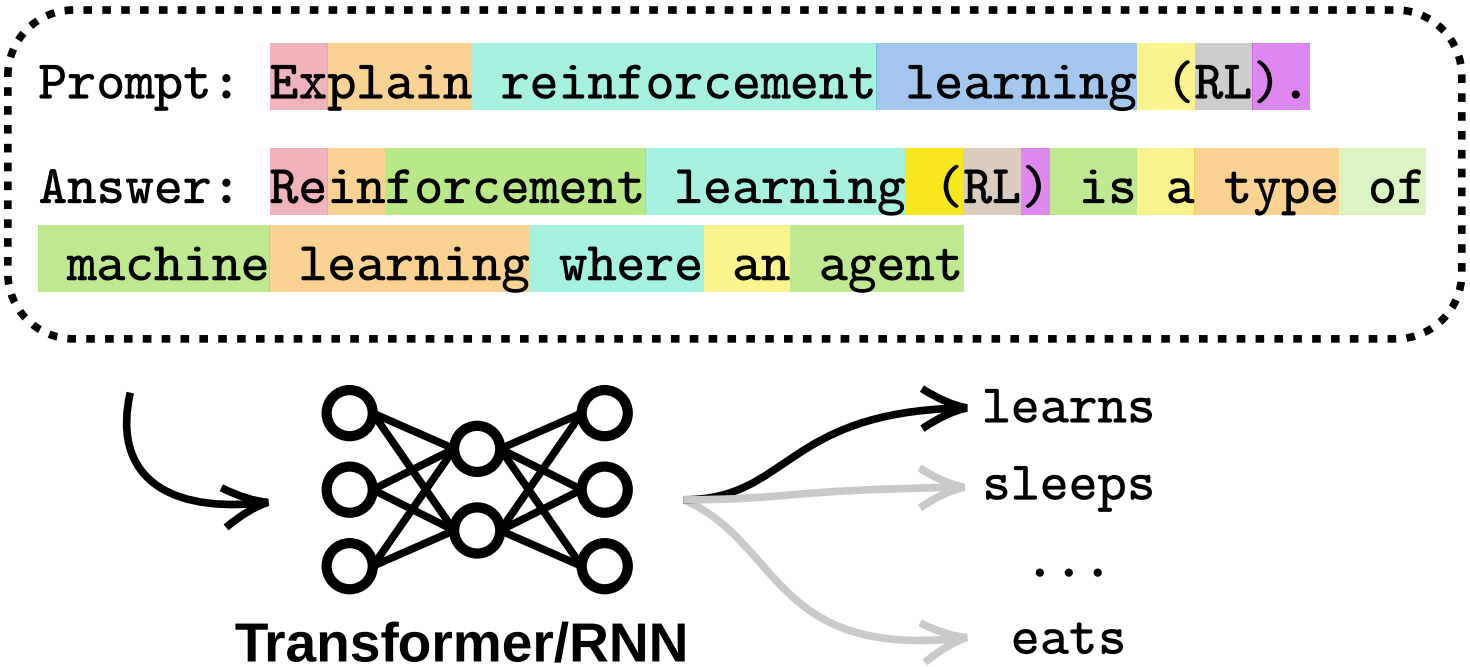
Language models as policies



Given prompt $x \in \mathcal{X}$, a language model generates an answer:

$$y \sim \underbrace{\pi(\cdot|x)}_{\text{parameterized by LLM}}$$

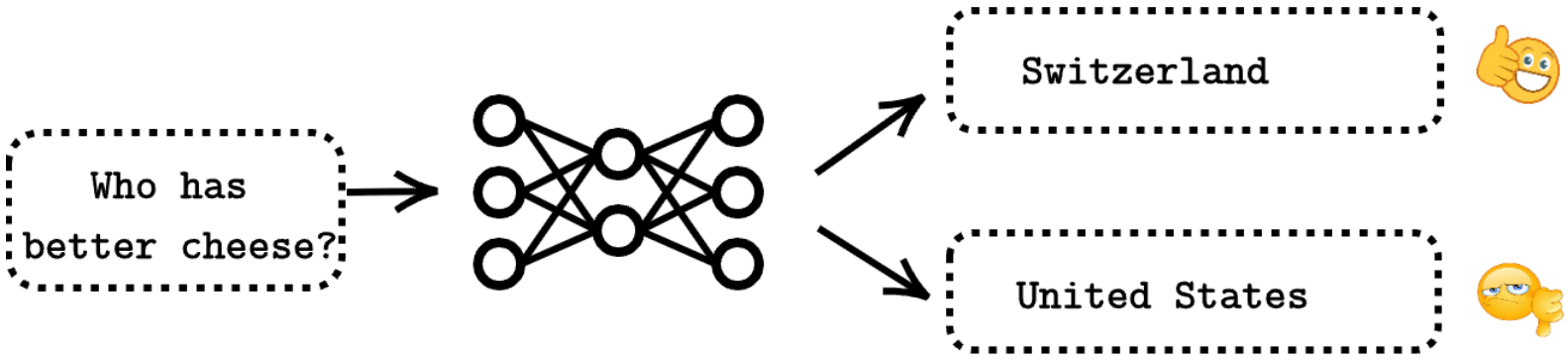
Language models as policies



Given prompt $x \in \mathcal{X}$, a language model generates an answer:

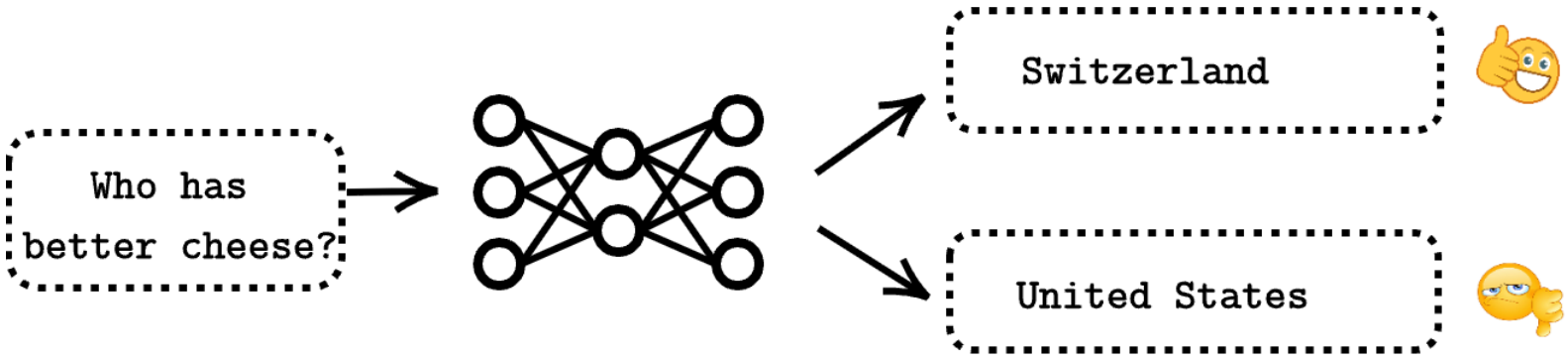
$$y \sim \underbrace{\pi(\cdot|x)}_{\text{parameterized by LLM}}$$

Reinforcement learning with human feedback (RLHF)



Goal: finetune the LLM to align with human preference

Reinforcement learning with human feedback (RLHF)



Goal: finetune the LLM to align with human preference

Prototypical pipeline:

- *Reward learning*: learn a reward model from preference data;
- *Policy optimization*: optimize the LLM to maximize the reward.

RLHF: reward learning

Bradly-Terry model

The probability of pairwise comparison $i \succ j$ is modeled by

$$\mathbb{P}(i \succ j) = \frac{\exp(r_i^\star)}{\exp(r_i^\star) + \exp(r_j^\star)} = \sigma(r_i^\star - r_j^\star),$$

where $r_i^\star \in \mathbb{R}$ is the score associated with item i.

RLHF: reward learning

Bradly-Terry model

The probability of pairwise comparison $i \succ j$ is modeled by

$$\mathbb{P}(i \succ j) = \frac{\exp(r_i^\star)}{\exp(r_i^\star) + \exp(r_j^\star)} = \sigma(r_i^\star - r_j^\star),$$

where $r_i^\star \in \mathbb{R}$ is the score associated with item i.

- **Reward model:** $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, evaluating the quality of a prompt-answer pair (x, y) that aligns with human preference;

RLHF: reward learning

Bradly-Terry model

The probability of pairwise comparison $i \succ j$ is modeled by

$$\mathbb{P}(i \succ j) = \frac{\exp(r_i^\star)}{\exp(r_i^\star) + \exp(r_j^\star)} = \sigma(r_i^\star - r_j^\star),$$

where $r_i^\star \in \mathbb{R}$ is the score associated with item i.

- **Reward model:** $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, evaluating the quality of a prompt-answer pair (x, y) that aligns with human preference;
- **Reward learning:** Given comparison data $\mathcal{D} = \{(x^i, y_+^i, y_-^i)\}_{i=1}^N$, the MLE of the reward function is given by

$$r_{\mathsf{MLE}} = \operatorname*{argmin}_r \ell(r, \mathcal{D}),$$

where $\ell(r, \mathcal{D}) = -\sum_{i=1}^N \log \sigma(r(x^i, y_+^i) - r(x^i, y_-^i))$.

RLHF: policy optimization

Policy optimization via reward maximization

Find π that (approximately) maximizes the objective w.r.t. r

$$J(r,\pi) = \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi(\cdot|x)}} \big[r(x,y)\big] - \beta \mathop{\mathbb{E}}_{x\sim\rho} \Big[\mathsf{KL}\big(\pi(\cdot|x) \,\|\, \pi_{\mathrm{ref}}(\cdot|x)\big)\Big]$$

- $\beta > 0$: KL regularization parameter;
- π_{ref}: a reference policy, typically the model after SFT;
- $\rho \in \Delta(\mathcal{X})$: prompt distribution.

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

1. Reward learning: $\widehat{r} \leftarrow \operatorname{argmin}_r \ell(r, \mathcal{D})$
2. Policy learning: $\widehat{\pi} \leftarrow \operatorname{argmax}_\pi J(\widehat{r}, \pi)$

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

Observation: the optimal π w.r.t. r admits a closed-form solution

$$\pi_r = \operatorname*{argmax}_{\pi} J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

Observation: the optimal π w.r.t. r admits a closed-form solution

$$\pi_r = \operatorname*{argmax}_{\pi} J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x,y)/\beta)}{Z(r,x)}.$$

- The reward function r in terms of its optimal π_r is

$$r(x,y) = \underbrace{\beta\big(\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x) + \log Z(r,x)\big)}_{=: r(\pi)}.$$

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

Observation: the optimal π w.r.t. r admits a closed-form solution

$$\pi_r = \operatorname*{argmax}_{\pi} J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x,y)/\beta)}{Z(r,x)}.$$

- The reward function r in terms of its optimal π_r is

$$r(x,y) = \underbrace{\beta\big(\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x) + \log Z(r,x)\big)}_{=: r(\pi)}.$$

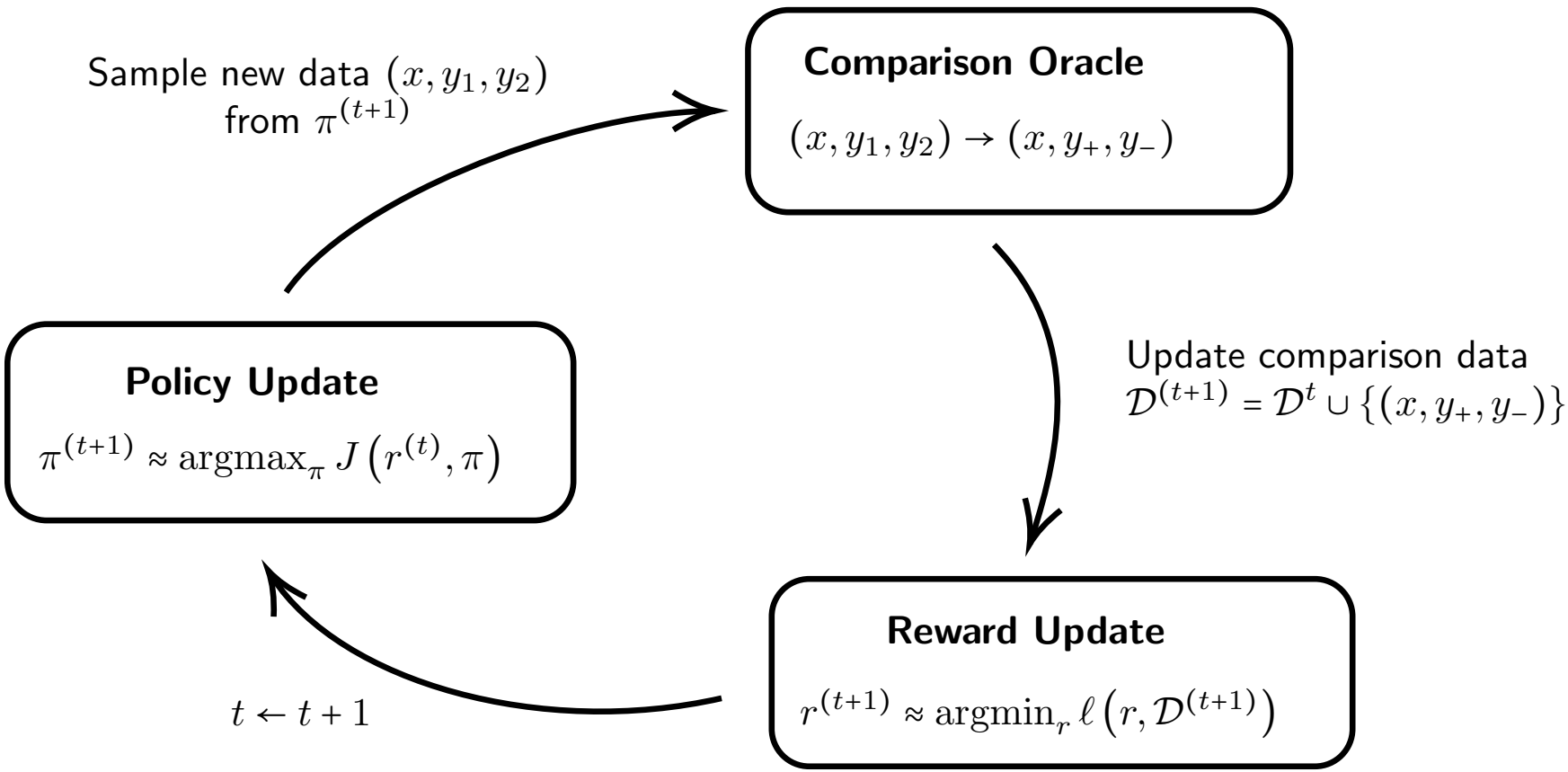
- The two-step procedure is equivalent to

$$\widehat{\pi} \leftarrow \operatorname*{argmin}_{\pi} \ell(r(\pi), \mathcal{D}) = -\sum_{i=1}^{N} \log \sigma \left(\beta \log \frac{\pi(y_+^i|x^i)}{\pi_{\text{ref}}(y_+^i|x^i)} - \beta \log \frac{\pi(y_-^i|x^i)}{\pi_{\text{ref}}(y_-^i|x^i)} \right).$$

Single-step and **policy-only**! Very popular in practice.

Online RLHF

Leverage online data collection to improve data coverage - how do we perform **exploration** in the policy space directly?



Exploration via optimistic MLE

- **Optimistic MLE:** Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \operatorname*{argmin}_r \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

  See (Kumar & Lin, 1982) and follow-ups.

Exploration via optimistic MLE

- **Optimistic MLE:** Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \operatorname*{argmin}_r \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

  See (Kumar & Lin, 1982) and follow-ups.

- The update is not well-defined: BT model cannot distinguish between r and $r + c \cdot \mathbf{1}$, while

$$J^\star(r + c \cdot \mathbf{1}) = J^\star(r) + c.$$

Exploration via optimistic MLE

- **Optimistic MLE:** Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \operatorname*{argmin}_{r \in \mathcal{R}} \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

  See (Kumar & Lin, 1982) and follow-ups.

- The update is not well-defined: BT model cannot distinguish between r and $r + c \cdot \mathbf{1}$, while

$$J^\star(r + c \cdot \mathbf{1}) = J^\star(r) + c.$$

- We can resolve the shift ambiguity by focusing on the following equivalent class of reward functions:

$$\mathcal{R} = \left\{ r : \mathop{\mathbb{E}}_{x \sim \rho, y \sim \pi_{\mathrm{cal}}(\cdot|x)} [r(x, y)] = 0 \right\}.$$

Exploration via optimistic MLE

- **Optimistic MLE:** Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \operatorname*{argmin}_{r \in \mathcal{R}} \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

  See (Kumar & Lin, 1982) and follow-ups.

- The update is not well-defined: BT model cannot distinguish between r and $r + c \cdot \mathbf{1}$, while

$$J^\star(r + c \cdot \mathbf{1}) = J^\star(r) + c.$$

- We can resolve the shift ambiguity by focusing on the following equivalent class of reward functions:

$$\mathcal{R} = \left\{ r : \mathop{\mathbb{E}}_{x \sim \rho, y \sim \pi_{\mathrm{cal}}(\cdot|x)} [r(x, y)] = 0 \right\}.$$

Can we avoid solving a bilevel optimization problem?

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \operatorname*{argmax}_{\pi} J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \operatorname*{argmax}_{\pi} J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[\log Z(r, x) \right] \end{aligned}$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \operatorname*{argmax}_{\pi} J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} [\log Z(r, x)] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log Z(r, x)] \end{aligned}$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \operatorname*{argmax}_{\pi} J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[\log Z(r, x) \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[\log Z(r, x) \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \end{aligned}$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \operatorname*{argmax}_{\pi} J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned}
J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\
&= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[\log Z(r, x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[\log Z(r, x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[\cancel{r(x, y)} - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right]
\end{aligned}$$

Value-incentivized preference optimization (VPO)

$$\pi^{(t+1)} \leftarrow \operatorname*{argmin}_{\pi} \{\ell(r(\pi), \mathcal{D}^{(t)}) - \alpha J^{\star}(r(\pi))\}.$$

- The negative log-likelihood term reformulates into DPO loss:

$$\begin{aligned}&\ell(r(\pi), \mathcal{D}^{(t)}) \\ &= -\sum_{(x,y_+,y_-)\in\mathcal{D}^{(t)}} \log\sigma\bigg(\beta\Big(\log\frac{\pi(y_+|x)}{\pi_{\mathrm{ref}}(y_+|x)} - \log\frac{\pi(y_-|x)}{\pi_{\mathrm{ref}}(y_-|x)}\Big)\bigg).\end{aligned}$$

- The reward bias term can be written as:

$$J^{\star}(r(\pi)) = -\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} \left[\log\pi(y|x) - \log\pi_{\mathrm{ref}}(y|x)\right],$$

  which is essentially becomes a reverse-KL regularization that maximizes $\mathrm{KL}\big(\pi_{\mathrm{cal}}(\cdot|x) \,\|\, \pi(\cdot|x)\big)$.

Main results - online VPO

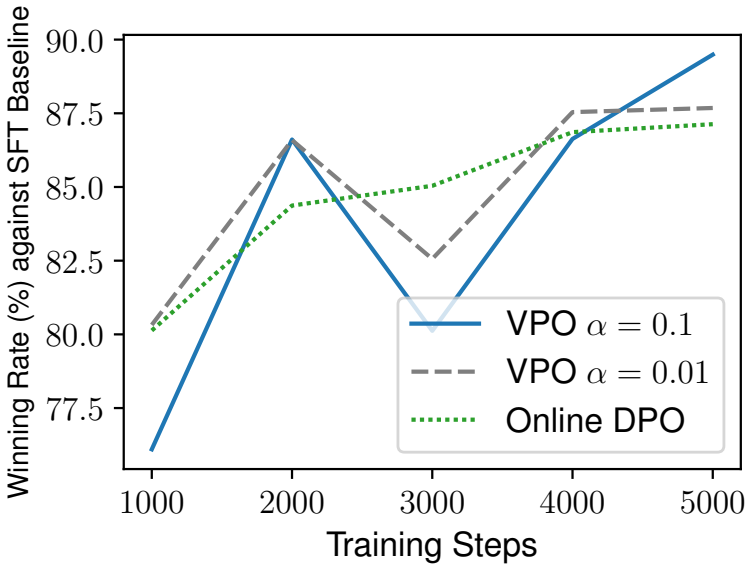
Theorem (Cen et al., ICLR 2025)

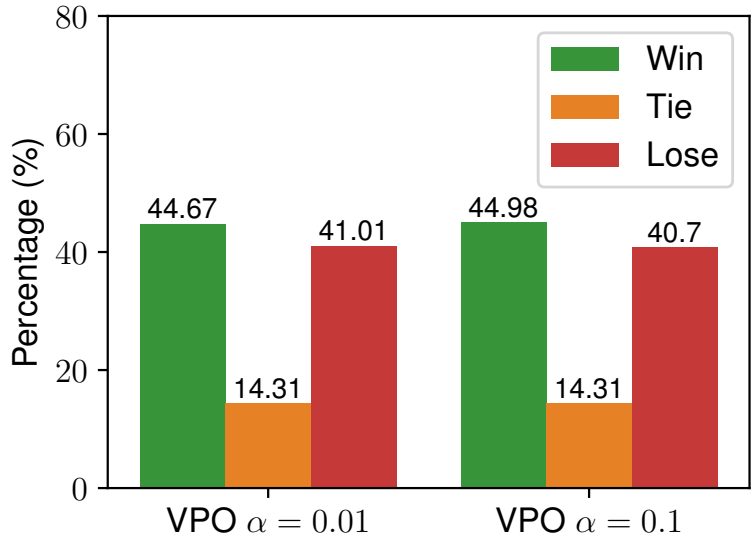
Assume that reward estimates $\|r^{(t)}\|_\infty \le B$ *and* $\|r^\star\|_\infty \le B$ *for some* $B > 0$. *With high probability we have*

$$\sum_{t=1}^{T} \left[J^\star(r^\star) - J(r^\star, \pi^{(t)}) \right] \le \widetilde{O}\big(\sqrt{T}\big).$$

- We can obtain similar regret bounds under general function approximation of the reward model.
- Consistent with the $\widetilde{O}(\sqrt{T})$ regret for online RL with UCB-type bonus.
- Offline RL: flipping the sign of α leads to a pessimistic algorithm.

Toy experiments on LLM





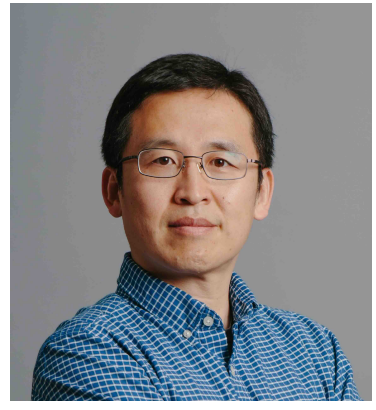
Left: Win rate of VPO and Online DPO against the SFT baseline on `TL;DR` task.

Right: Win/tie/loss rate of VPO with different exploration rate $\alpha = \{0.01, 0.1\}$, directly against Online DPO.

Value-incentivized exploration for online RL

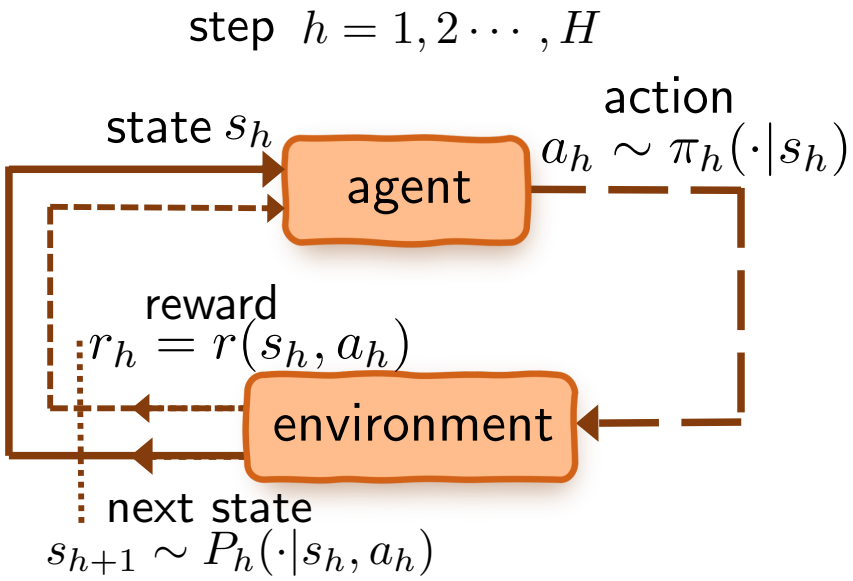
Tong Yang
CMU

Bo Dai
GaTech/Google



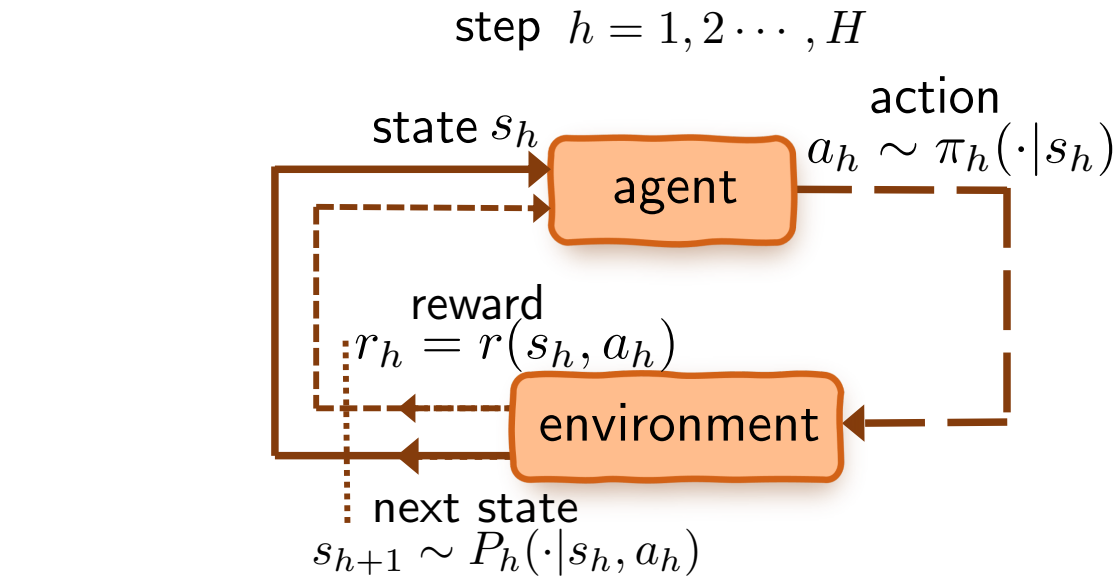
Lin Xiao
Meta

Finite-horizon Markov decision process (MDP)



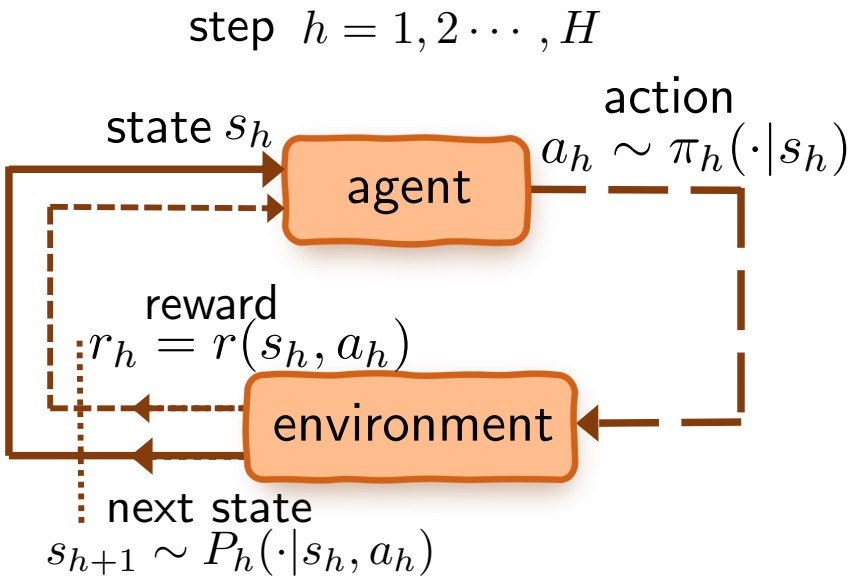
- H: horizon length; $\mathcal{S}$: state space; $\mathcal{A}$: action space

Finite-horizon Markov decision process (MDP)



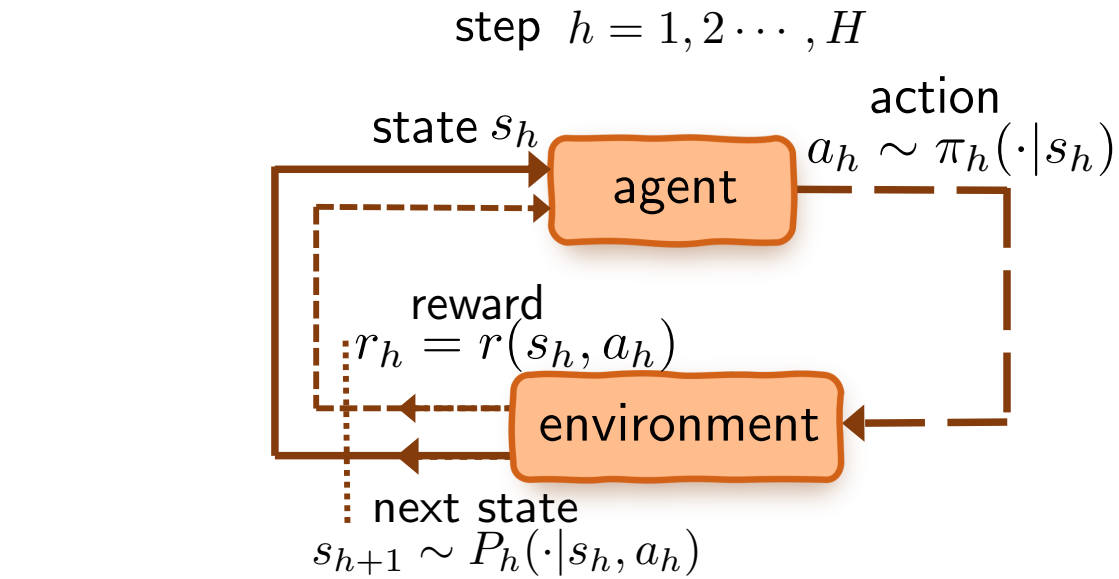
- H: horizon length; $\mathcal{S}$: state space; $\mathcal{A}$: action space
- $P_h(\cdot \,|\, s, a)$: transition probability in step h; $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h

Finite-horizon Markov decision process (MDP)



- H: horizon length; $\mathcal{S}$: state space; $\mathcal{A}$: action space
- $P_h(\cdot \,|\, s, a)$: transition probability in step h; $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h
- $\pi = \{\pi_h\}_{1 \le h \le H}$: policy

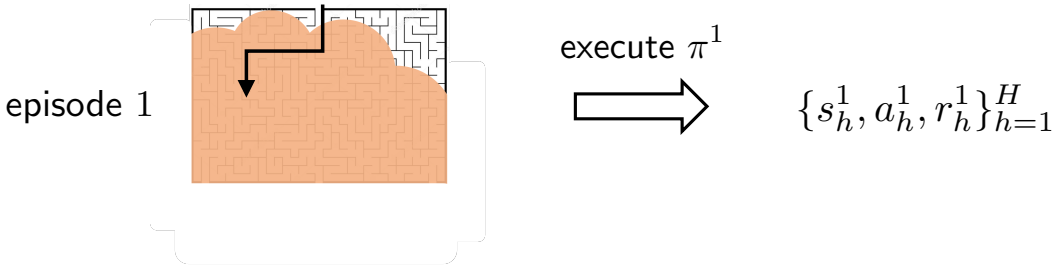
Finite-horizon Markov decision process (MDP)



- H: horizon length; $\mathcal{S}$: state space; $\mathcal{A}$: action space
- $P_h(\cdot \,|\, s, a)$: transition probability in step h; $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h
- $\pi = \{\pi_h\}_{1 \le h \le H}$: policy
- value and Q-functions: $V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^H r_t(s_t, a_t) \,\middle|\, s_h = s\right]$ and $Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{t=h}^H r_t(s_t, a_t) \,\middle|\, s_h = s, a_h = a\right]$.

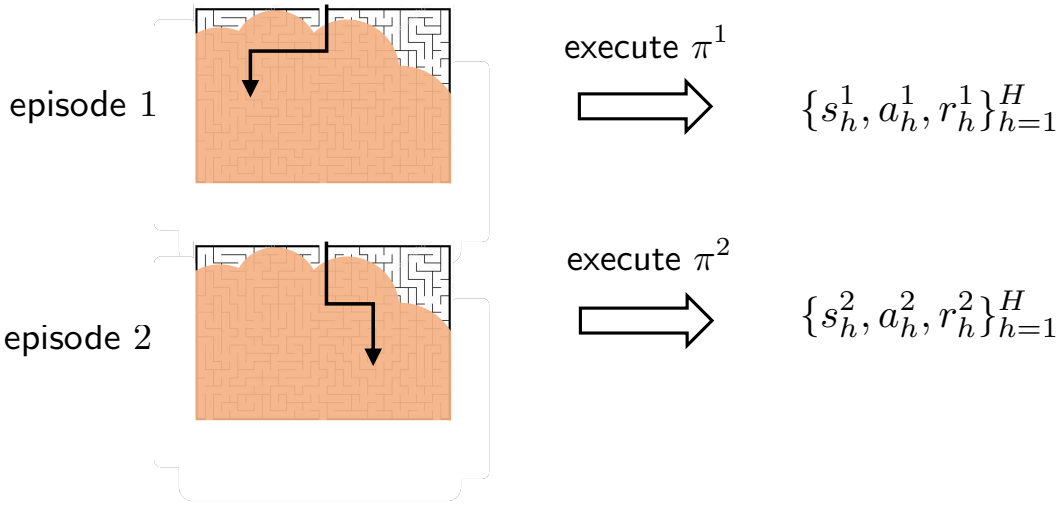
Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps



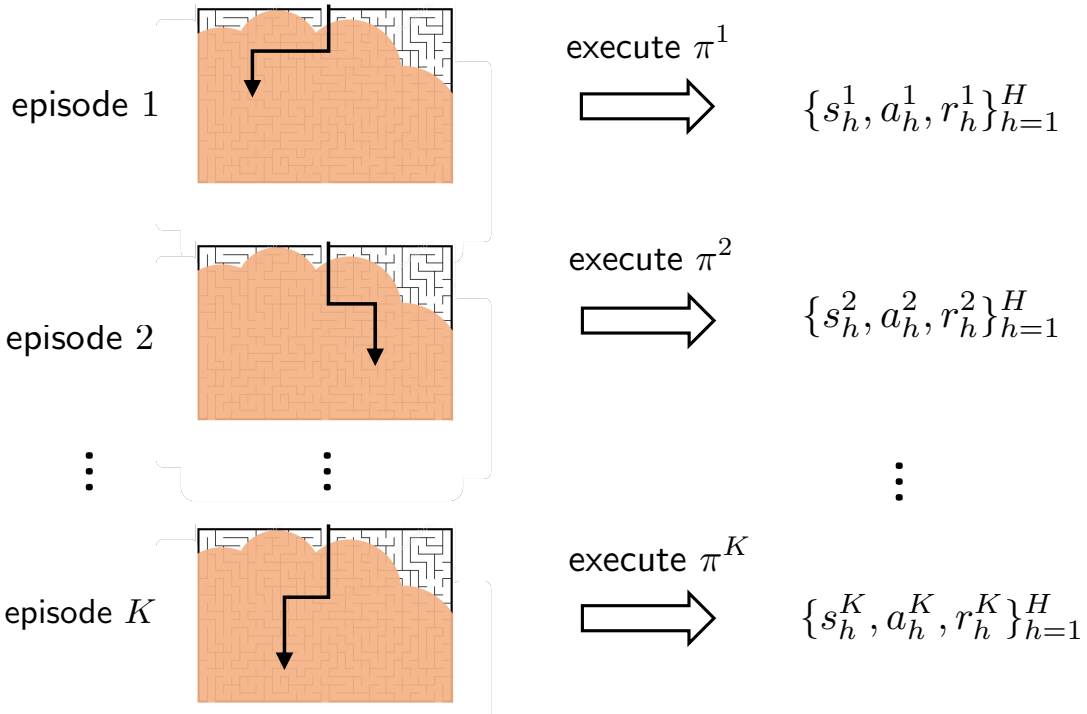
Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps



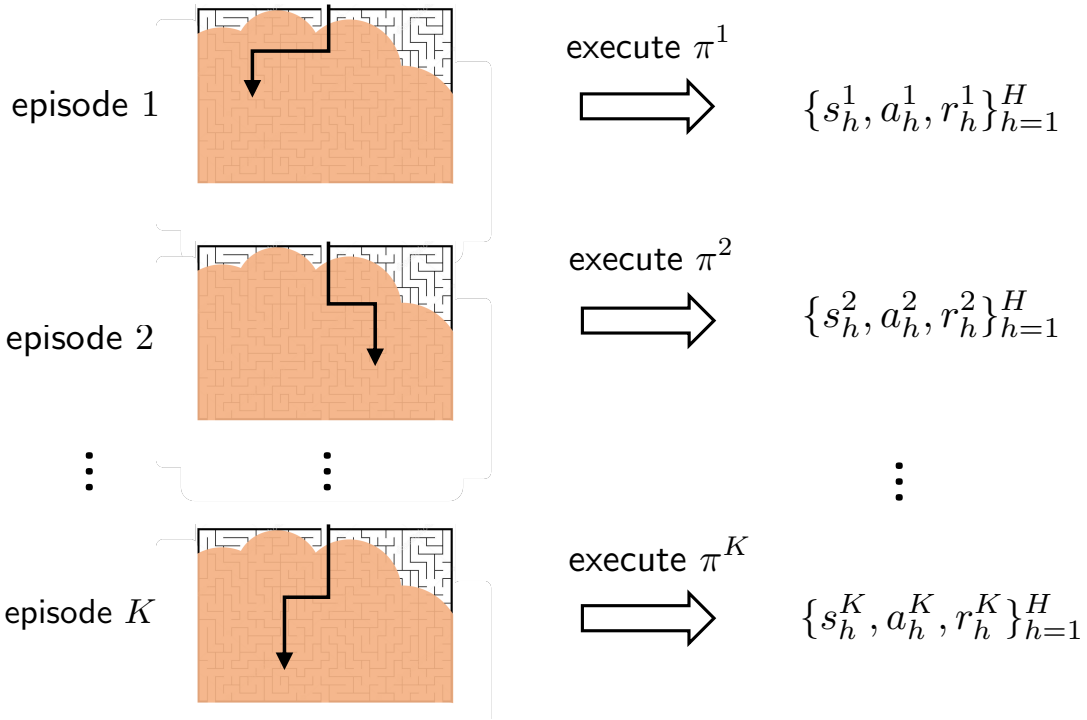
Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps



Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps



Goal: given initial states $s_1^k \sim \rho$, minimize

$$\text{Regret}(T) \;:=\; \sum_{k=1}^{K} \left(V_1^\star(\rho) - V_1^{\pi^k}(\rho) \right).$$

Optimistic regularization via MEX

MEX (maximize to explore) (Liu et al, 2024) optimizes $f := Q_f$ via

$$f_t \;=\; \arg\sup_{f\in\mathcal{Q}} \underbrace{\mathbb{E}_{s_1\sim\rho}\Big[\max_{a\in\mathcal{A}} Q_{f,1}(s_1,a)\Big]}_{\text{optimal value}} - \alpha \underbrace{\mathcal{L}_t(f)}_{\text{data consistency}} .$$

- $\mathcal{L}_t(f)$ is the data consistency term that minimizes the Bellman error using the collected data $\{\mathcal{D}_{t-1,h}\}_{h=1}^{H}$:

$$\mathcal{L}_t(f) = \sum_{h=1}^{H}\Bigg[\sum_{\xi_h\in\mathcal{D}_{t-1,h}} \big(r_h(s_h,a_h) + \max_{a\in\mathcal{A}} Q_{f,h+1}(s_{h+1},a) - Q_{f,h}(s_h,a_h)\big)^2 - \inf_{g_h\in\mathcal{Q}_h}\sum_{\xi_h\in\mathcal{D}_{t-1,h}} \big(r_h(s_h,a_h) + \max_{a\in\mathcal{A}} Q_{f,h+1}(s_{h+1},a) - g_h(s_h,a_h)\big)^2\Bigg],$$

where $\xi_h = (s_h, a_h, s_{h+1})$ is the transition tuple.

Optimistic regularization via MEX

MEX (maximize to explore) (Liu et al, 2024) optimizes $f := Q_f$ via

$$f_t = \arg\sup_{f \in \mathcal{Q}} \underbrace{\mathbb{E}_{s_1 \sim \rho}\Big[\max_{a \in \mathcal{A}} Q_{f,1}(s_1, a)\Big]}_{\text{optimal value}} - \alpha \underbrace{\mathcal{L}_t(f)}_{\text{data consistency}} .$$

- $\mathcal{L}_t(f)$ is the data consistency term that minimizes the Bellman error using the collected data $\{\mathcal{D}_{t-1,h}\}_{h=1}^H$:

$$\mathcal{L}_t(f) = \sum_{h=1}^{H} \Bigg[\sum_{\xi_h \in \mathcal{D}_{t-1,h}} \Big(r_h(s_h, a_h) + \max_{a \in \mathcal{A}} Q_{f,h+1}(s_{h+1}, a) - Q_{f,h}(s_h, a_h)\Big)^2 - \inf_{g_h \in \mathcal{Q}_h} \sum_{\xi_h \in \mathcal{D}_{t-1,h}} \Big(r_h(s_h, a_h) + \max_{a \in \mathcal{A}} Q_{f,h+1}(s_{h+1}, a) - g_h(s_h, a_h)\Big)^2 \Bigg],$$

  where $\xi_h = (s_h, a_h, s_{h+1})$ is the transition tuple.

✓ Near-optimal regret without explicit uncertainty estimation
✗ The optimization is intractable.

A primal-dual perspective of MEX

How do we understand MEX? Introducing the primal problem:

$$\sup_{f \in \mathcal{Q}} \ \mathbb{E}_{s_1 \sim \rho}\Big[\max_{a \in \mathcal{A}} Q_{f,1}(s_1, a) \Big]$$

$$\text{s.t.} \ \underbrace{Q_{f,h}(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\Big[\max_{a \in \mathcal{A}} Q_{f,h+1}(s', a) \Big]}_{\text{Bellman's optimality equation}} \quad \forall (s, a, h).$$

A primal-dual perspective of MEX

How do we understand MEX? Introducing the primal problem:

$$\sup_{f\in\mathcal{Q}} \ \mathbb{E}_{s_1\sim\rho}\Big[\max_{a\in\mathcal{A}} Q_{f,1}(s_1,a)\Big]$$

$$\text{s.t.} \ \underbrace{Q_{f,h}(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\Big[\max_{a\in\mathcal{A}} Q_{f,h+1}(s',a)\Big]}_{\text{Bellman's optimality equation}} \quad \forall (s,a,h).$$

- With the dual variables $\{\lambda_h\}_{h\in[H]}$, its regularized Lagrangian can be written as

$$\sup_{f\in\mathcal{Q}} \inf_{\{\lambda_h\}_{h\in[H]}} \mathbb{E}_{s_1\sim\rho}\Big[\max_{a\in\mathcal{A}} Q_{f,1}(s_1,a)\Big]$$

$$+ \sum_{h=1}^{H} \mathbb{E}_{(s,a,s')\sim\mathcal{D}_h}\Big\{\lambda_h(s,a)\Big(r_h(s,a) + \max_{a\in\mathcal{A}} Q_{f,h+1}(s',a) - Q_{f,h}(s,a)\Big) + \frac{\beta}{2}\lambda_h(s,a)^2\Big\},$$

  where $\beta > 0$ is the regularization parameter of the dual variable.

- Reparameterizing $\lambda_h := (Q_{f,h} - g_h)/\beta$ leads to (population) of MEX.

An exploratory actor-critic framework

Actor-critic framework: introduce an equivalent primal problem that jointly optimizes over both the Q-function and the policy π:

$$\sup_{f \in \mathcal{Q},\, \pi \in \mathcal{P}} \mathbb{E}_{s_1 \sim \rho,\, a_1 \sim \pi_1(\cdot|s_1)} \big[Q_{f,1}(s_1, a_1) \big]$$

$$\text{s.t.} \quad \underbrace{Q_{f,h}(s,a) = r_h(s,a) + \mathbb{E}_{\substack{s' \sim P_h(\cdot|s,a) \\ a' \sim \pi_{h+1}(\cdot|s')}} \big[Q_{f,h+1}(s', a') \big]}_{\text{Bellman's consistency equation}}, \ \forall\ (s,a,h).$$

An exploratory actor-critic framework

Actor-critic framework: introduce an equivalent primal problem that jointly optimizes over both the Q-function and the policy π:

$$\sup_{f\in\mathcal{Q},\ \pi\in\mathcal{P}} \mathbb{E}_{s_1\sim\rho,\ a_1\sim\pi_1(\cdot|s_1)}\big[Q_{f,1}(s_1,a_1)\big]$$

$$\text{s.t.} \quad \underbrace{Q_{f,h}(s,a) = r_h(s,a) + \mathbb{E}_{\substack{s'\sim P_h(\cdot|s,a)\\ a'\sim\pi_{h+1}(\cdot|s')}}\big[Q_{f,h+1}(s',a')\big],\ \forall\ (s,a,h).}_{\text{Bellman's consistency equation}}$$

Following similar arguments gives rise to a (population-level) actor-critic method that optimizes jointly over the Q-function Q_f and policy π:

$$\sup_{f,\pi\in\mathcal{P}} \Big\{\mathbb{E}_{s\sim\rho,a\sim\pi_1(\cdot|s)}\left[Q_{f,1}(s,a)\right]$$

$$-\sum_{h=1}^{H}\frac{1}{2\beta}\sup_{g_h\in\mathcal{Q}_h}\mathbb{E}_{(s,a,s')\sim\mathcal{D}_h}\mathbb{E}_{a'\sim\pi_{h+1}(\cdot|s')}\Big[\big(r_h(s,a)+Q_{f,h+1}(s',a')-Q_{f,h}(s,a)\big)^2$$

$$-\big(r_h(s,a)+Q_{f,h+1}(s',a')-g_h(s,a)\big)^2\Big]\Big\}.$$

Value-incentivized actor-critic method

Value-incentivized actor-critic (VAC) method

For $t = 1, \cdots, T$,

1. *Update Q-function estimation and policy:*

$$(f_t, \pi_t) \leftarrow \arg \sup_{f \in \mathcal{Q}, \pi \in \mathcal{P}} \Big\{ \underbrace{V_f^{\pi}(\rho)}_{\text{value incentive}} - \alpha \underbrace{\mathcal{L}_t(f, \pi)}_{\text{data consistency}} \Big\}.$$

2. *Data collection: run* π_t *to obtain a trajectory* $\{s_{t,h}, a_{t,h}\}_{h=1}^{H}$, *and update the dataset* $\mathcal{D}_{t,h} = \mathcal{D}_{t-1,h} \cup \{(s_{t,h}, a_{t,h}, s_{t,h+1})\}$, $\forall h \in [H]$.

- $V_f^{\pi}(\rho) = \mathbb{E}_{s \sim \rho, a \sim \pi_1(\cdot|s)} [Q_{f,1}(s, a)]$ is the optimistic regularization;
- $\mathcal{L}_t(f, \pi)$ is the data consistency term

$$\mathcal{L}_t(f, \pi) = \sum_{h=1}^{H} \Bigg\{ \sum_{\xi_h \in \mathcal{D}_{t-1,h}} \mathbb{E}_{a' \sim \pi_{h+1}(\cdot|s_{h+1})} \big(r_h(s_h, a_h) + Q_{f,h+1}(s_{h+1}, a') - Q_{f,h}(s_h, a_h) \big)^2$$

$$- \inf_{g_h \in \mathcal{Q}_h} \sum_{\xi_h \in \mathcal{D}_{t-1,h}} \mathbb{E}_{a' \sim \pi_{h+1}(\cdot|s_{h+1})} \big(r_h(s_h, a_h) + Q_{f,h+1}(s_{h+1}, a') - g_h(s_h, a_h) \big)^2 \Bigg\},$$

where $\xi_h = (s_h, a_h, s_{h+1})$ is the transition tuple.

Theoretical guarantees

Theorem (Yang et al, 2025)

Under the linear MDP model, with linear function approximation on the Q function and the policy class, with high probability, the regret of VAC is bounded by

$$\widetilde{\mathcal{O}}\left(dH^2\sqrt{T}\right),$$

where d is the dimension of the linear MDP.

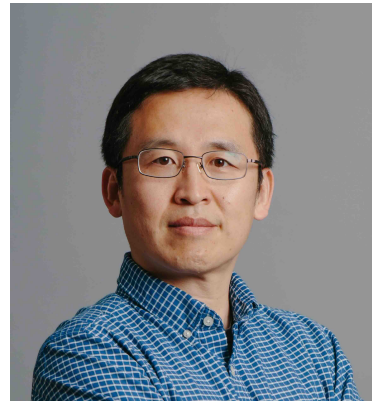
- The regret bound is near-optimal for linear MDP up to a factor of $\sqrt{H}$, matching UCB-type guarantees.
- We also achieve comparable statistical guarantees as MEX in the more general function approximation setting.

An optimization-friendly actor-critic framework with provably-efficient exploration without explicit uncertainty estimation!

Value-incentivized exploration for competitive games

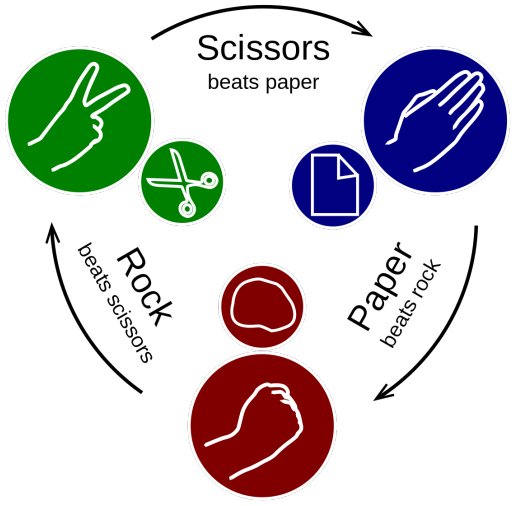
Tong Yang
CMU

Bo Dai
GaTech/Google



Lin Xiao
Meta

Zero-sum two-player matrix game



	0	-1	1
	1	0	-1
	-1	1	0

Zero-sum two-player matrix game

$$\max_{\mu \in \Delta(\mathrm{A})} \min_{\nu \in \Delta(\mathrm{B})} \mu^\top A \nu - \beta \mathrm{KL}\big(\mu \,\|\, \mu_{\mathrm{ref}}\big) + \beta \mathrm{KL}\big(\nu \,\|\, \nu_{\mathrm{ref}}\big)$$

- $\mathcal{A}$, $\mathcal{B}$: action space of the two players;
- $\mu \in \Delta(\mathcal{A})$, $\nu \in \Delta(\mathcal{B})$: policies of the two players;
- $\mu_{\mathrm{ref}} \in \Delta(\mathcal{A})$, $\nu_{\mathrm{ref}} \in \Delta(\mathcal{B})$: reference policies of the two players;
- $A \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{B}|}$: payoff matrix.

Motivation: game-theoretic view of RLHF

RLHF suffers from reward hacking. **Call for a game-theoretic view!**(Swamy et al., 2023, Munos et al., 2023, Gui et al., 2024)

- Given a policy pair π, π', the *win rate* of π over π' is given as

$$P(\pi \succ \pi') := \mathop{\mathbb{E}}_{\substack{x\sim\rho, y\sim\pi(\cdot|x),\\ y'\sim\pi'(\cdot|x)}} P_x(y, y') = \mathbb{E}_{x\sim\rho} \underbrace{\pi^\top(\cdot|x) P_x(\cdot,\cdot) \pi'(\cdot|x)}_{\text{bilinear}}.$$

  where $P_x : \mathcal{Y} \times \mathcal{Y}$ is a *symmetric* payoff matrix between (y, y') for prompt x.

Motivation: game-theoretic view of RLHF

RLHF suffers from reward hacking. **Call for a game-theoretic view!**(Swamy et al., 2023, Munos et al., 2023, Gui et al., 2024)

- Given a policy pair π, π', the *win rate* of π over π' is given as

$$P(\pi \succ \pi') := \mathop{\mathbb{E}}_{\substack{x\sim\rho, y\sim\pi(\cdot|x),\\ y'\sim\pi'(\cdot|x)}} P_x(y, y') = \mathbb{E}_{x\sim\rho} \underbrace{\pi^\top(\cdot|x) P_x(\cdot,\cdot) \pi'(\cdot|x)}_{\text{bilinear}}.$$

  where $P_x : \mathcal{Y} \times \mathcal{Y}$ is a *symmetric* payoff matrix between (y, y') for prompt x.

- Consider the two-player *win-rate* game:

$$\max_{\pi} \min_{\pi'} P(\pi \succ \pi') - \beta \mathsf{KL}\big(\pi \,\|\, \pi_{\mathrm{ref}}\big) + \beta \mathsf{KL}\big(\pi' \,\|\, \pi_{\mathrm{ref}}\big)$$

  which is an KL-regularized matrix game.

Self-play and win rate dominance

Theorem (Yang et al., AISTATS 2025)

The Nash equilibrium of the win-rate game $(\pi^\star_\beta, \pi^\star_\beta)$ *exists. Moreover, when* $\beta > 0$, $(\pi^\star_\beta, \pi^\star_\beta)$ *is the unique Nash equilibrium. We have*

$$\pi^\star_\beta \in \arg\max_\pi P(\pi \succ \pi^\star_\beta) - \beta \mathsf{KL}\big(\pi \,\|\, \pi_{\mathrm{ref}}\big)$$

Self-play and win rate dominance

Theorem (Yang et al., AISTATS 2025)

The Nash equilibrium of the win-rate game $(\pi^\star_\beta, \pi^\star_\beta)$ *exists. Moreover, when* $\beta > 0$, $(\pi^\star_\beta, \pi^\star_\beta)$ *is the unique Nash equilibrium. We have*

$$\pi^\star_\beta \in \arg\max_\pi P(\pi \succ \pi^\star_\beta) - \beta \mathsf{KL}\big(\pi \,\|\, \pi_{\mathrm{ref}}\big)$$

Win rate dominance: the fixed-point equation identifies a policy with a higher winning probability against any other policy. When $\beta = 0$,

$$\pi^\star_0 \in \arg\max_\pi P(\pi \succ \pi^\star_0) \qquad \Longrightarrow \qquad P(\pi \succ \pi^\star_0) \le 1/2 \quad \forall \pi.$$

NE of win rate matrix game = win rate dominance policy

Value-incentivized matrix game with bandit feedback

How do we optimistically optimize the payoff matrix with bandit feedback?

Value-incentivized matrix game with bandit feedback

How do we optimistically optimize the payoff matrix with bandit feedback?

Value-incentivized model update:

$$\omega_t = \operatorname*{argmin}_{\omega\in\Omega} \sum_{(i,j,\widehat{A}(i,j))\in\mathcal{D}_{t-1}} \big(A_\omega(i,j) - \widehat{A}(i,j)\big)^2 \underbrace{-\alpha f^{\star,\nu_t}(A_\omega) + \alpha f^{\mu_t,\star}(A_\omega)}_{\text{duality gap}},$$

where (μ_t, ν_t) is the NE of the matrix game with the param. ω_{t-1}.

Value-incentivized matrix game with bandit feedback

How do we optimistically optimize the payoff matrix with bandit feedback?

Value-incentivized model update:

$$\omega_t = \operatorname*{argmin}_{\omega\in\Omega} \sum_{(i,j,\widehat{A}(i,j))\in\mathcal{D}_{t-1}} \left(A_\omega(i,j) - \widehat{A}(i,j)\right)^2 \underbrace{-\alpha f^{\star,\nu_t}(A_\omega) + \alpha f^{\mu_t,\star}(A_\omega)}_{\text{duality gap}},$$

where (μ_t, ν_t) is the NE of the matrix game with the param. ω_{t-1}.

Computational tractability: the regularization term can be computed in closed form:

$$-\beta\left[\log\left(\sum_{i=1}^{n}\mu_{\text{ref},i}\exp\left(\frac{A_\omega(i,:)\nu_t}{\beta}\right)\right) + \log\left(\sum_{j=1}^{m}\nu_{\text{ref},j}\exp\left(-\frac{\mu_t^\top A_\omega(:,j)}{\beta}\right)\right)\right] + C,$$

allowing single-loop gradient-based updates on the model parameter.

Regret for value-incentivized matrix game

$$\begin{aligned}\text{regret}(T) &:= \sum_{t=1}^{T} \text{Dualgap}(\mu_t, \nu_t) \\ &= \underbrace{\sum_{t=1}^{T} \left(f^{\star,\nu_t}(A) - f^{\star}(A)\right)}_{\text{regret for min-player}} + \underbrace{\sum_{t=1}^{T} \left(f^{\star}(A) - f^{\mu_t,\star}(A)\right)}_{\text{regret for max-player}}\end{aligned}$$

Theorem (Yang et al., ICML 2025)

Under the linear function approximation of dimension d and realizability assumption, with high probability, the regret is on the order of

$$\widetilde{\mathcal{O}}(d\sqrt{T}).$$

- near-optimal regret as it matches with the $\Omega(d\sqrt{T})$ lower bound.

Value-incentivized exploration for Markov games

- The algorithm can be generalized to the Markov game setting for finding both NE and CCEs:

$$f_t = \arg\min_{f \in \mathcal{F}} \mathcal{L}_t(f) - \alpha \sum_{n=1}^{N} V_{f,n}^{\star, \boldsymbol{\pi}_t^{-n}}(\rho)$$

  Here, n is the number of agents,

Value-incentivized exploration for Markov games

- The algorithm can be generalized to the Markov game setting for finding both NE and CCEs:

$$f_t = \arg\min_{f \in \mathcal{F}} \mathcal{L}_t(f) - \alpha \sum_{n=1}^{N} V_{f,n}^{\star, \boldsymbol{\pi}_t^{-n}}(\rho)$$

Here, n is the number of agents,

- $\mathcal{L}_t(f)$, is the negative log-likehood of sample transitions,

Value-incentivized exploration for Markov games

- The algorithm can be generalized to the Markov game setting for finding both NE and CCEs:

$$f_t = \arg\min_{f \in \mathcal{F}} \mathcal{L}_t(f) - \alpha \sum_{n=1}^{N} V_{f,n}^{\star, \boldsymbol{\pi}_t^{-n}}(\rho)$$

Here, n is the number of agents,

- $\mathcal{L}_t(f)$, is the negative log-likehood of sample transitions,
- $V_{f,n}^{\star, \boldsymbol{\pi}_t^{-n}}(\rho)$ is the **best-response** values of each agent when other agents' policies are fixed.

Value-incentivized exploration for Markov games

- The algorithm can be generalized to the Markov game setting for finding both NE and CCEs:

$$f_t = \arg\min_{f \in \mathcal{F}} \mathcal{L}_t(f) - \alpha \sum_{n=1}^{N} V_{f,n}^{\star, \boldsymbol{\pi}_t^{-n}}(\rho)$$

Here, n is the number of agents,

- $\mathcal{L}_t(f)$, is the negative log-likehood of sample transitions,
- $V_{f,n}^{\star, \boldsymbol{\pi}_t^{-n}}(\rho)$ is the **best-response** values of each agent when other agents' policies are fixed.

Value-incentivized exploration for Markov games

- The algorithm can be generalized to the Markov game setting for finding both NE and CCEs:

$$f_t = \arg\min_{f \in \mathcal{F}} \mathcal{L}_t(f) - \alpha \sum_{n=1}^{N} V_{f,n}^{\star, \boldsymbol{\pi}_t^{-n}}(\rho)$$

  Here, n is the number of agents,
  - $\mathcal{L}_t(f)$, is the negative log-likehood of sample transitions,
  - $V_{f,n}^{\star, \boldsymbol{\pi}_t^{-n}}(\rho)$ is the **best-response** values of each agent when other agents' policies are fixed.

- Achieves near-optimal regret, and much easier to implement than using the optimal game value as in previous work (Liu et al., 2024).

Summary

Optimism via regularization in the face of uncertainty:

$$f_t = \arg\min_{f \in \mathcal{F}} \underbrace{\mathcal{L}_t(f; \mathcal{D}_t)}_{\text{data consistency}} - \alpha \underbrace{V^\star(f)}_{\text{optimal value}}$$

What makes it tractable: smoothing and actor-critic frameworks.

Theoretically-principled and practically-performant exploration via *optimistic regularization* under complex function approximation

Summary

Optimism via regularization in the face of uncertainty:

$$f_t = \arg\min_{f \in \mathcal{F}} \underbrace{\mathcal{L}_t(f; \mathcal{D}_t)}_{\text{data consistency}} - \alpha \underbrace{V^\star(f)}_{\text{optimal value}}$$

What makes it tractable: smoothing and actor-critic frameworks.

Theoretically-principled and practically-performant exploration via *optimistic regularization* under complex function approximation

Future work:

- Break the curse of multi-agency in the game setting.
- Applications to finetuning LLMs.

Thanks!

- Value-Incentivized Preference Optimization: A Unified Approach to Online and Offline RLHF, *ICLR*, 2025.
- Faster WIND: Accelerating Iterative Best-of-N Distillation for LLM Alignment, *AISTATS*, 2025.
- Incentivize without Bonus: Provably Efficient Model-based Online Multi-agent RL for Markov Games, *ICML*, 2025.
- Exploration from a Primal-Dual Lens: Value-Incentivized Actor-Critic Methods for Sample-Efficient Online RL, arXiv:2506.22401, 2025.

https://yuejiechi.github.io/