Alignment for Foundation Language Models: Mathematical Principle and Algorithmic Designs

Wei Xiong

Department of Computer Science
University of Illinois Urbana-Champaign

Why We Need RLHF?

- RLHF: A leading technique to adapt LLMs to being preferred by humans.

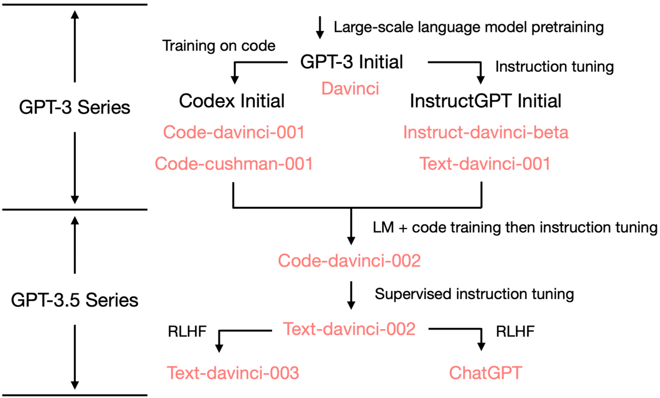


Figure 1: How does GPT Obtain its Ability [FK22]

Training Pipeline

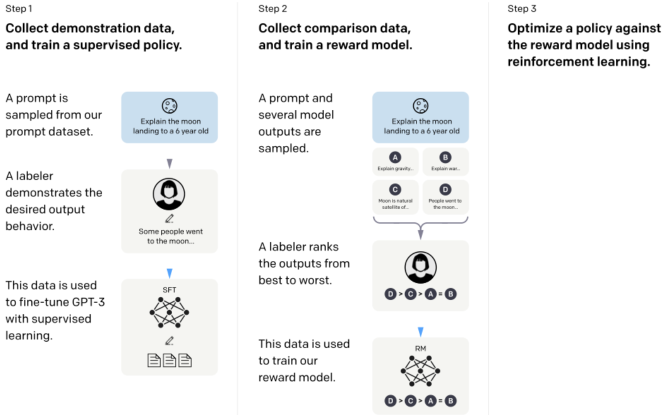


Figure 2: RLHF framework in Instruct-GPT [OWJ+22]

Mathematical Formulation

- Prompt space $\mathcal{X}$: "Human: Can you write a C++ program that prompts the user to enter the name of a country and checks if it borders the Mediterranean Sea? Assistant: ";
- Response space $\mathcal{A}$:
  - A high-quality and correct code ✓
  - A wrong code. ✗
  - "NO, I cannot."
- SFT-policy $\pi_0 : \mathcal{X} \to \Delta(\mathcal{A})$;
- Prompt distribution: $x \sim d_0$.

Mathematical Formulation Continued: Preference Oracle

- Examples: Human, Model from Inverse RL, GPT4 (RLAIF [BKK+22]);
- Goal: making LLM being preferred by $\mathcal{P}$ while stay close to π_0.
  - Fundamental issue: $\mathcal{P}$ is never perfect and can be hacked;
  - Training stability issue;

Definition 1 (General Preference Oracle)

There exists a preference oracle $\mathcal{P} : \mathcal{X} \times \mathcal{A} \times \mathcal{A} \to [0, 1]$, and we can query it to receive the preference signal:

$$y \sim \mathrm{Ber}(\mathcal{P}(a^1 \succ a^2 | x)),$$

where $y = 1$ means a^1 is preferred to a^2, and $y = 0$ means that a^2 is preferred.

RLHF v.s. SFT

SFT	RLHF
Dataset: Prompt+Response positive only	**Dataset:** Prompt+Response Pairs positive & negative
Loss: negative log-likelihood/cross entropy	**Loss:** - Learning reward - Maximize Reward + regularizations
Aim: Approximate the training data distribution	**Aim:** Learning beyond the training datasets, potentially generate better responses than positive data
Components: Generative Model Only	**Components:** Generative Model + Preference Model

Learning in Reward-based RLHF

Mathematical Formulation: Reward-based RLHF

- Implicitly assumes a *total order*: $A \succ B, B \succ C \rightarrow A \succ C$;
- MLE estimation from a preference dataset:

$$\ell_{\mathcal{D}}(\theta) = \sum_{(x,a^1,a^2,y)\in\mathcal{D}} \Big[y \log \Big(\sigma\big(r_\theta(x,a^1) - r_\theta(x,a^2)\big)\Big) + (1-y)\log\Big(\sigma\big(r_\theta(x,a^2) - r_\theta(x,a^1)\big)\Big)\Big]. \tag{1}$$

Definition 2 (Bradley-Terry (BT) model [BT52])

There exists a ground-truth reward function $r^* = r_{\theta^*}$ so that:

$$\mathcal{P}(a^1 \succ a^2 | x) = \frac{\exp(r^*(x,a^1))}{\exp(r^*(x,a^1)) + \exp(r^*(x,a^2))} = \sigma\big(r^*(x,a^1) - r^*(x,a^2)\big).$$

Additionally, we assume that the reward function is parameterized by $r_\theta(x,a) = \langle \theta, \phi(x,a)\rangle$ for feature extractor $\phi : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d$.

Reverse-KL Regularized Contextual Bandit [XDY+23]

For iteration t=1,2,...

- A context/prompt $x \sim d_0$;
- An agent selects $a^1, a^2 \in \mathcal{A}$;
- The preference signal y queried from $\mathcal{P}$ is revealed.

Learning objective:

$$\max_{\pi \in \Pi} J(\pi) = \max_{\pi \in \Pi} \mathbb{E}_{x \sim d_0}\Big[\underbrace{\mathbb{E}_{a \sim \pi(\cdot|x)}[r^*(x,a)]}_{\text{Optimize Reward}} - \underbrace{\eta D_{\mathrm{KL}}(\pi(\cdot|x) \| \pi_0(\cdot|x))}_{\text{Stay Close to } \pi_0}\Big].$$

Intractable closed-form Gibbs distribution:

$$\pi^*(a|x) \propto \pi_0(a|x) \exp\Big(\frac{1}{\eta} r^*(x,a)\Big).$$

We assume that we can compute the Gibbs distribution associated with any r by some information-theoretical Oracle $\mathcal{O}(r, \pi_0)$.

Offline Learning: Point-wise Pessimism

Given a pre-collected $\mathcal{D}_{\text{off}} = \{(x, a^1, a^2, y)\}$, we denote r_{MLE} as the MLE estimation on $\mathcal{D}_{\text{off}}$.

Intuition: being conservative at the point with high uncertainty.

- Point-wise confidence interval:

$$\begin{aligned}\underbrace{r_{\mathrm{MLE}}(x,a) - r^*(x,a)}_{\text{Out-of-sample error}} &= \langle \theta_{\mathrm{MLE}} - \theta^*, \phi(x,a)\rangle \\ &\leq \underbrace{\|\theta_{\mathrm{MLE}} - \theta^*\|_{\Sigma_{\mathcal{D}_{\mathrm{off}}}}}_{\text{In-sample error on } \mathcal{D}_{\mathrm{off}} \leq \beta = c\sqrt{d}} \cdot \underbrace{\|\phi(x,a)\|_{\Sigma^{-1}_{\mathcal{D}_{\mathrm{off}}}}}_{\text{Information Ratio}}\end{aligned}$$

- Point-wise pessimism:

$$\hat{\pi} = \mathcal{O}\Big(r_{\mathrm{MLE}}(x,a) - \beta\|\phi(x,a)\|_{\Sigma^{-1}_{\mathcal{D}_{\mathrm{off}}}}, \pi_0\Big).$$

- With high probability, we have

$$J(\pi) - J(\hat{\pi}) \leq 2\beta \cdot \mathbb{E}_{x\sim d_0, a\sim\pi(\cdot|x)}\|\phi(x,a)\|_{\Sigma^{-1}_{\mathcal{D}_{\mathrm{off}}}} - \eta \cdot \mathbb{E}_{x\sim d_0}\big[D_{\mathrm{KL}}(\pi(\cdot|x)\|\hat{\pi}(\cdot|x))\big].$$

Offline Learning: Version-space Pessimism

- Version space $\hat{\Theta}$ consists of θ^* with high probability:

$$\hat{\Theta} := \{\hat{\theta} : \|\theta_{\mathrm{MLE}} - \hat{\theta}\|_{\Sigma_{\mathcal{D}}} \leq \beta = c \cdot \sqrt{d}\}.$$

- Conservative value estimation:

$$\underline{J}(\pi) = \min_{\hat{\theta} \in \hat{\Theta}} \mathbb{E}_{x \sim d_0} \Big[\mathbb{E}_{a \sim \pi(\cdot|x)} [r_{\hat{\theta}}(x, a)] - \eta D_{\mathrm{KL}}(\pi(\cdot|x) \| \pi_0(\cdot|x)) \Big]$$

- Planning: $\hat{\pi} := \mathrm{argmax}_{\pi \in \Pi} \underline{J}(\pi)$, and with high probability,

$$J(\pi) - J(\hat{\pi}) \leq 2\beta \cdot \|\mathbb{E}_{x \sim d_0, a \sim \pi(\cdot|x)}[\phi(x, a)]\|_{\Sigma^{-1}_{\mathcal{D}_{\mathrm{off}}}}.$$

- Compared with point-wise pessimism:
    - Sharper bound due to Jensen's inequality;
    - Lacking general computational guidance.

Offline Learning with Reference Policy

Theorem 3 (Offline Learning with Reference Policy [XDY+23])

If we add a reference in value estimation by:

$$\underline{J}(\pi) = \min_{\hat{\theta}\in\hat{\Theta}} \mathbb{E}_{x\sim d_0}\Big[\mathbb{E}_{a\sim\pi(\cdot|x)}[r_{\hat{\theta}}(x,a)] - \mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[r_{\hat{\theta}}(x,a)] - \eta D_{\mathrm{KL}}(\pi(\cdot|x)\|\pi_0(\cdot|x))\Big],$$

it holds that

$$J(\pi) - J(\hat{\pi}) \le c\cdot\sqrt{d}\cdot\|\mathbb{E}_{x\sim d_0}[\phi(x,\pi)] - \mathbb{E}_{x\sim d_0}[\phi(x,\pi_{\mathrm{ref}})]\|_{\Sigma_{\mathrm{off}}^{-1}}.$$

- Robust policy improvement: the resulting $\hat{\pi}$ is never worse than π_{ref} regardless of $\mathcal{D}_{\mathrm{off}}$;
- One common choice of π_{ref} is π_0 [OWJ+22, GSH23].

Online Learning

For iteration $t = 1, 2, \ldots$

- Compute r^t_{MLE} based on $\mathcal{D}^{1:t-1}$;
- The main agent computes $\pi^1_t = \mathcal{O}(r^t_{\mathrm{MLE}}, \pi_0)$;
- The enhancer computes the assistant policy by:

$$\pi^2_t = \mathop{\mathrm{argmax}}_{\pi \in \Pi} \|\phi(x, \pi^1_t) - \phi(x, \pi^2_t)\|_{\Sigma^{-1}_{t,m}}$$

$$\Sigma_{t,m} = \lambda I + \frac{1}{m} \sum_{i=1}^{t-1} \sum_{j=1}^{m} (\phi(x_{i,j}, a^1_{i,j}) - \phi(x_{i,j}, a^2_{i,j}))(\phi(x_{i,j}, a^1_{i,j}) - \phi(x_{i,j}, a^2_{i,j}))^\top.$$

- Collect m comparison pairs by (π^1_t, π^2_t) as $\mathcal{D}^t$.

Intuition: the two policies should be diverse to facilitate exploration.

Online Learning Continued

Theorem 4 (Batch Online Learning Guarantee [XDY+23])

With $m = c_1 \cdot d/\epsilon^2$ and $T = \tilde{\Theta}(d)$, we can find a $\pi^1_{t_0}$ such that

$$J(\pi^*) - J(\pi^1_{t_0}) \lesssim \epsilon - \eta \cdot \mathbb{E}_{x_{t_0} \sim d_0} \Big[D_{\mathrm{KL}}(\pi^*(\cdot|x_{t_0}) \| \pi^1_{t_0}(\cdot|x_{t_0})) \Big] \Big).$$

- Sample complexity scales with the complexity of reward space instance of the generator space;
- Sparse update for $\tilde{\Theta}(d)$ times;
- The techniques extend to regret $\sum_{t=1}^{T} \Big[\frac{2J(\pi^*) - J(\pi^1_t) - J(\pi^2_t)}{2} \Big]$, with enhancer selecting over

$$\Pi_t = \Big\{ \tilde{\pi} \in \Pi : \sqrt{\frac{cd \log(T/\delta)}{m}} \sum_{i=1}^{m} \|\phi(x_{t,i}, \tilde{\pi}) - \phi(x_{t,i}, \pi^1_t)\|_{\Sigma^{-1}_{t,m}} - \eta \sum_{i=1}^{m} D_{\mathrm{KL}}(\tilde{\pi}(\cdot|x_{t,i}) \| \pi^1_t(\cdot|x_{t,i})) \geq 0 \Big\}.$$

Hybrid Learning

For iteration $t = 1, 2, \ldots$

- Compute r_{MLE}^t based on $\mathcal{D}^{1:t-1}$ and $\mathcal{D}_{\mathrm{off}}$;
- The main agent computes $\pi_t^1 = \mathcal{O}(r_{\mathrm{MLE}}^t, \pi_0)$;
- The enhancer takes a fixed π_{ref};
- Collect m comparison pairs by $(\pi_t^1, \pi_{\mathrm{ref}})$ as $\mathcal{D}^t$.

Theorem 5 (Batch Hybrid Learning Guarantee [XDY+23])

With $T = \tilde{\Theta}(d)$, we can find a $\pi_{t_0}^1$ such that

$$J(\pi^*) - J(\pi_{t_0}) \lesssim \underbrace{\sqrt{\frac{d}{\gamma^2 m}}}_{\text{Online exploration}} + \underbrace{\sqrt{d} \cdot \|\mathbb{E}_{x \sim d_0}[\phi(x, \pi^*) - \phi(x, \pi_{\mathrm{ref}})]\|_{\Sigma_{\mathrm{off}}^{-1}}}_{\text{Offline coverage}}$$

$$- \eta \mathbb{E}_{x_{t_0} \sim d_0}\left[D_{\mathrm{KL}}(\pi_{t_0}^1(\cdot|x_{t_0}) \| \pi^*(\cdot|x_{t_0}))\right],$$

Learning Paradigm

Offline	Online	Hybrid
Precondition: $\mathcal{D}_{\mathrm{off}}$	**Precondition:** -	**Precondition:** $\mathcal{D}_{\mathrm{off}}$
Condition: $\\|\mathbb{E}_{x\sim d_0}[\phi(x,\pi^*)-\phi(x,\pi_{\mathrm{ref}})]\\|_{\Sigma_{\mathrm{off}}^{-1}} \le \frac{C_{\mathrm{cov}}}{\sqrt{n_{\mathrm{off}}}}$	**Condition:** Low-rank reward	**Condition:** Both + $n_{\mathrm{off}} \approx mT$
Algorithmic idea: Pessimism	**Algorithmic idea:** π_{r^t} v.s. Optimism	**Algorithmic idea:** π_{r^t} v.s. π_{ref}
#Sample: $O\Big(\frac{dC_{\mathrm{cov}}}{\epsilon^2}\Big)$	**#Sample:** $O\Big(\frac{d^2}{\epsilon^2}\Big)$	**#Sample:** $O\Big(\frac{d^2+dC_{\mathrm{cov}}}{\epsilon^2}\Big)$

We omit the constant factors and some log factors like $\log(T/\delta)$.

Algorithmic Designs in Reward-based RLHF

RLHF v.s. Deep RL

Reward function is never perfect.

- Humans typically possess a set of intricate or even *contradictory* targets: helpful, harmless, honest, verbosity...
- Majority v.s. under-represented groups;
- Human expertise and opinions are diverse.

Essentially, all models are wrong, but some are useful. [Box76]

RL	**RLHF**
Reward:	**Reward:**
Ground truth (Gold reward)	Imperfect reward
	Learned reward / Human / GPT4
Model Selection:	**Model Selection:**
Model with highest reward	Evaluation Benchmarks + Human Raters

Toward a Practical Planning Oracle

Recall the planning oracle:

$$\operatorname*{argmax}_{\pi\in\Pi} \mathbb{E}_{x\sim d_0}\Big[\underbrace{\mathbb{E}_{a\sim\pi(\cdot|x)}[r^*(x,a)]}_{\text{Optimize Reward}} - \underbrace{\eta D_{\mathrm{KL}}(\pi(\cdot|x)\|\pi_0(\cdot|x))}_{\text{Stay Close to } \pi_0}\Big] \propto \pi_0(a|x)\exp\big(\frac{1}{\eta}r(x,a)\big)$$

- PPO [SWD+17] with $\tilde{r}(x,a) = r(x,a) - \eta\log\frac{\pi_\theta(a|x)}{\pi_0(a|x)}$.
  - Complicated hyper-parameter: learning rate, KL coefficient, update epoch, clip range... and code-level optimization;
  - Unstable convergence behavior;
  - Heavy burden on GPU memory: loading 4 models at the same time.
- Direct Preference Optimization (DPO) [RSM+23]:

$$\mathcal{L}(\theta,\pi_0) = -\sum_{(x,\tilde{a}^1,\tilde{a}^2)\in\mathcal{D}_{\mathrm{off}}} \log\sigma\big(\eta\log\frac{\pi_\theta(\tilde{a}^1|x)}{\pi_0(\tilde{a}^1|x)} - \eta\log\frac{\pi_\theta(\tilde{a}^2|x)}{\pi_0(\tilde{a}^2|x)}\big)$$

  - “Direct”: no reward modeling;
  - SFT-based learning: stable with less parameters;
  - DPO enjoys the same solution with PPO.

Iterative Hybrid DPO

- Dataset: Anthropic HH-RLHF, Helpful subset;
  - Human: What is the best way to apologize to someone? Assistant:
- Model: Open-LLaMA-3B V2;

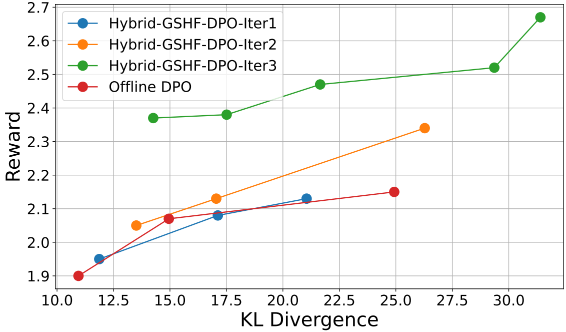


Figure 3: The figure of Reward-KL trade-off. Both the KL and reward are computed on a hand-out test set.

Multi-step Rejection Sampling

[LZJ+23] proposes Rejection Sampling Optimization (RSO):

- DPO: dataset from unknown data distribution, learning target: π^*;
- RSO with a trained reward r:
  - dataset: $a \sim \pi_r$ by ejection sampling with r;
  - label: r;
  - learning target: π_r by running DPO on newly generated dataset.

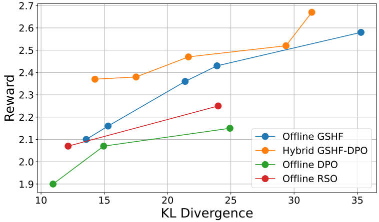
Approximating π_r by π_0 can be inefficient

- Multi-step RS [XDY+23] $\eta_1 > \eta_2 > ... > \eta$:

$$\pi_0 \to \pi_0 \exp(\frac{1}{\eta_1} r) \to \pi_0 \exp(\frac{1}{\eta_2} r) \cdots \to \pi_0 \exp(\frac{1}{\eta} r)$$

More Experimental Results

We remark that the boundary between online and offline is not strict. We can fix $\mathcal{P}$ as some learned preference oracle and use it for the online preference query.



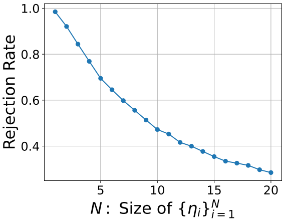


Figure 4: The figure of Reward-KL trade-off. Both the KL and reward are computed on a hand-out test set.

Online Iterative DPO: Uncertainty in LLMs

Essentially, we require

- Exploration: $\|\phi(x, \pi_t^1) - \phi(x, \pi_t^2)\|_{\Sigma_{t,m}^{-1}} \geq \|\phi(x, \pi_t^1) - \phi(x, \pi^*)\|_{\Sigma_{t,m}^{-1}}$;
- Exploitation: π_t^1, π_t^2 are "around" π_{r^t}.

Practical Implementation:

- Rejection sampling:
  - For prompt x, we independently sample (a^1, a^2, a^3, a^4) by π_{r^t};
  - Best-of-4 v.s. Worst-of-4, as ranked by $\mathcal{P}$.
- The concurrent work [HT24] implements rejection-sampling-based online iterative DPO
  - Mistral-7B + pairRM-0.4B;
  - Ranked 2nd in AlpacaEval.

Question: How to quantify the uncertainty in general LLMs?

RLHF Under General Preference

Disadvantage of Reward-based RLHF

- Implicitly assumes a *total order*: $A \succ B, B \succ C \to A \succ C$;
- The total order is not satisfied for the preference *averaged* over a diverse set of human groups;
- The reward function can be easily hacked;
- Win rate is a more robust metric that is close to real-world user experience.

Reverse-KL Regularized Two-player Game

- Relative preference: $R^*(x, a^1, a^2) = \log \mathcal{P}(a^1 \succ a^2|x) - \log \mathcal{P}(a^1 \prec a^2|x)$;
- With BT model, it holds that

$$R^*(x, a^1, a^2) = r^*(x, a^1) - r^*(x, a^2).$$

- Value function

$$\begin{aligned} J(\pi^1, \pi^2) &= \mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1\sim\pi^1, a^2\sim\pi^2}\Big[R^*(x, a^1, a^2) + \eta \log \frac{\pi_0(a^1|x)}{\pi^1(a^1|x)} - \eta \log \frac{\pi_0(a^2|x)}{\pi^2(a^2|x)}\Big] \\ &= \mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1\sim\pi^1, a^2\sim\pi^2}\Big[R^*(x, a^1, a^2) - \eta D_{\mathrm{KL}}(\pi^1(\cdot|x)\|\pi_0(\cdot|x)) + \eta D_{\mathrm{KL}}(\pi^2(\cdot|x)\|\pi_0(\cdot|x))\Big]. \end{aligned} \tag{2}$$

- Nash Equilibrium: $(\pi^1_*, \pi^2_*) = (\pi_*, \pi_*) = \mathrm{argmax}_{\pi^1\in\Pi}\, \mathrm{argmin}_{\pi^2\in\Pi} J(\pi^1, \pi^2)$, which implies $J(\pi_*, \pi_*) = 0$.
- Goal: find a $\hat{\pi}^1$ such that

$$J(\pi_*, \pi_*) - J(\hat{\pi}^1, \dagger) \le \epsilon,$$

where $J(\hat{\pi}^1, \dagger) = \min_{\pi'} J(\hat{\pi}^1, \pi')$.

MLE and Information Ratio

We consider function approximation by $\mathcal{R}$ with $R^* \in \mathcal{R}$.

- Maximum Likelihood Estimation:
$$\ell_{\mathcal{D}_{\mathrm{off}}}(R) = \sum_{(x,a^1,a^2,y)\in\mathcal{D}_{\mathrm{off}}} y \log \sigma(R(x,a^1,a^2)) + (1-y)\log \sigma(-R(x,a^1,a^2)).$$

- Information ratio between prediction error and in-sample error:
$$\Gamma(x,\pi^1,\pi^2) = \sup_{R\in\mathcal{R}} \frac{|R(x,\pi^1,\pi^2) - \hat{R}(x,\pi^1,\pi^2)|}{\sqrt{\lambda + \sum_{i=1}^n (R(x_i,a_i^1,a_i^2) - \hat{R}(x_i,a_i^1,a_i^2))^2}},$$

- Linear reward + BT model:
$$\Gamma(x,\pi^1,\pi^2) \le \|\phi(x,\pi^1) - \phi(x,\pi^2)\|_{\Sigma^{-1}_{\mathcal{D}_{\mathrm{off}}}}.$$

Offline Learning

- Version space consists of R^* with high probability:

$$\widehat{\mathcal{R}} = \Big\{ R \in \mathcal{R} : \sum_{i=1}^{n} (R(x_i, a_i^1, a_i^2) - \hat{R}(x_i, a_i^1, a_i^2))^2 \le O\big(\log(|\mathcal{R}|/\delta)\big) \Big\}.$$

- Conservative value estimations:

$$\underline{J}(\pi^1, \pi^2) = \min_{R \in \widehat{\mathcal{R}}} \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1 \sim \pi^1, a^2 \sim \pi^2} \Big[R(x, a^1, a^2) + \eta \ln \frac{\pi_0(a^1|x)}{\pi^1(a^1|x)} - \eta \ln \frac{\pi_0(a^2|x)}{\pi^2(a^2|x)} \Big],$$
$$\overline{J}(\pi^1, \pi^2) = \max_{R \in \widehat{\mathcal{R}}} \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1 \sim \pi^1, a^2 \sim \pi^2} \Big[R(x, a^1, a^2) + \eta \ln \frac{\pi_0(a^1|x)}{\pi^1(a^1|x)} - \eta \ln \frac{\pi_0(a^2|x)}{\pi^2(a^2|x)} \Big].$$

- Compute pessimistic Nash equilibrium

$$\begin{aligned} (\hat{\pi}^1, \tilde{\pi}^2) &= \operatorname*{argmax}_{\pi^1 \in \Pi} \min_{\pi^2 \in \Pi} \underline{J}(\pi^1, \pi^2), \\ (\tilde{\pi}^1, \hat{\pi}^2) &= \operatorname*{argmax}_{\pi^1 \in \Pi} \min_{\pi^2 \in \Pi} \overline{J}(\pi^1, \pi^2). \end{aligned} \tag{3}$$

- Return $(\hat{\pi}^1, \hat{\pi}^2)$.

Theoretical Results of Offline Learning

Theorem 6 (Offline learning guarantee [YXZ+24])

With $\beta^2 = O(\log|\mathcal{R}|/\delta)$, with high probability, we have

$$\mathrm{DuaGP}(\hat{\pi}^1, \hat{\pi}^2) \le 4\beta\sqrt{\frac{\mathcal{C}((\tilde{\pi}^1, \pi_*^2), \pi_D, \mathcal{R})}{n}} + 4\beta\sqrt{\frac{\mathcal{C}((\pi_*^1, \tilde{\pi}^2), \pi_D, \mathcal{R})}{n}}$$

where the coverage coefficient of some policy pair of $\mathcal{D}$ is

$$\mathcal{C}((\pi^1, \pi^2), \pi_D, \mathcal{R}) = \sup_{R \in \widehat{\mathcal{R}}} \frac{(\mathbb{E}_{x \sim d_0}[R(x, \pi^1, \pi^2) - \hat{R}(x, \pi^1, \pi^2)])^2}{\mathbb{E}_{x \sim d_0, a^1 \sim \pi_D^1, a^2 \sim \pi_D^2}(R(x, a^1, a^2) - \hat{R}(x, a^1, a^2))^2}.$$

- Unilateral Coverage [ZXT+22] by $\pi^1 = \tilde{\pi}^1, \pi^2 = \tilde{\pi}^2$:

$$4\beta\sqrt{\tilde{\mathcal{C}}((\tilde{\pi}^1, \pi_*^2), \pi_D, \mathcal{R})/n} + 4\beta\sqrt{\tilde{\mathcal{C}}((\pi_*^1, \tilde{\pi}^2), \pi_D, \mathcal{R})/n} \le \sup_{\pi \in \Pi} \beta\sqrt{\tilde{\mathcal{C}}((\pi, \pi_*), \pi_D, \mathcal{R})/n}.$$

- For reward-based RLHF, the suboptimality is

$$\inf_{\pi \in \Pi} \beta\sqrt{\tilde{\mathcal{C}}((\pi, \pi_*), \pi_D, \mathcal{R})/n}.$$

Refined Coverage Condition

Corollary 7 (Refined Guarantee for Offline Learning)

Under the same condition, we have

$$\mathrm{DuaGP}(\hat{\pi}^1, \hat{\pi}^2) \le \min_{\pi^1, \pi^2} \Big\{ 4\beta \sqrt{\frac{\mathcal{C}((\pi^1, \pi^2_*), \pi_D, \mathcal{R})}{n}} + 4\beta \sqrt{\frac{\mathcal{C}((\pi^1_*, \pi^2), \pi_D, \mathcal{R})}{n}} + \mathrm{subopt}^{\tilde{\pi}^1, \pi^2_*}(\pi^1) + \mathrm{subopt}^{\pi^1_*, \tilde{\pi}^2}(\pi^2) \Big\},$$

where $\mathrm{subopt}^{\tilde{\pi}^1, \pi^2_*}(\pi^1) = \overline{J}(\tilde{\pi}^1, \pi^2_*) - \overline{J}(\pi^1, \pi^2_*)$, $\mathrm{subopt}^{\pi^1_*, \tilde{\pi}^2}(\pi^2) = \underline{J}(\pi^1_*, \pi^2) - \underline{J}(\pi^1_*, \tilde{\pi}^2)$.

- If we take $(\pi^1, \pi^2) = (\tilde{\pi}^1, \tilde{\pi}^2)$, such an upper bound reduces to unilateral coverage case;
- If $\tilde{\mathcal{C}}((\tilde{\pi}^1, \pi^2), \pi_D, \mathcal{R})$ is large, the refined bound adapts to an alternate π^1 in the coverage term;
- Note $\tilde{\pi}^1$ is the best response to $\hat{\pi}^2$ but not necessarily to π^2_* so the subopt can be negative.

Batch Online Learning

For $t = 1, 2, \ldots$

- Compute the MLE $\hat{R}_t$ based on $\mathcal{D}_{1:t-1}$ and construct the bonus:

$$\Gamma_t^m(x, \pi^1, \pi^2) := \sup_{R \in \mathcal{R}} \frac{|R(x, \pi^1, \pi^2) - \hat{R}(x, \pi^1, \pi^2)|}{\sqrt{\lambda + \frac{1}{m}\sum_{s=1}^{t-1}\sum_{j=1}^{m}(R(x_{s,j}, a^1_{s,j}, a^2_{s,j}) - \hat{R}(x_{s,j}, a^1_{s,j}, a^2_{s,j}))^2}},$$

- Compute optimistic Nash policy for the max-player $(\hat{\pi}_t^1, \tilde{\pi}_t^2)$:

$$\operatorname*{argmax}_{\pi^1 \in \Pi} \operatorname*{argmin}_{\pi^2 \in \Pi} \mathbb{E}_{a^1 \sim \pi^1, a^2 \sim \pi^2}\Big[\hat{R}_t(x, a^1, a^2) + \beta \Gamma_t^m(x, \pi^1, \pi^2) + \eta^{-1} \log \frac{\pi_0(a^1|x)}{\pi^1(a^1|x)} - \eta^{-1} \log \frac{\pi_0(a^2|x)}{\pi^2(a^2|x)}\Big],$$

- The min-player aims to approximate the best response for the max-player:

$$\hat{\pi}_t^2 = \operatorname*{argmin}_{\pi^2 \in \Pi} \mathbb{E}_{a^1 \sim \hat{\pi}_t^1, a^2 \sim \pi^2}\Big[\hat{R}_t(x, a^1, a^2) - \beta \Gamma_t^m(x, \hat{\pi}_t^1, \pi^2) - \eta^{-1} \log \frac{\pi_0(a^2|x)}{\pi^2(a^2|x)}\Big].$$

- Collect m comparison pairs by $(\hat{\pi}_t^1, \hat{\pi}_t^2)$ as $\mathcal{D}^t$.

Eluder Coefficient

Definition 8 (Eluder Coefficient [GWZ22, YXGZ23])

We define the information ratio as

$$\tilde{\Gamma}_t(\lambda, \pi_t^1, \pi_t^2) = \sup_{R \in \mathcal{R}} \frac{\mathbb{E}_{x \sim d_0, a^1 \sim \pi_t^1, a^2 \sim \pi_t^2} |R(x, a^1, a^2) - \hat{R}(x, a^1, a^2)|}{\sqrt{\lambda + \sum_{s=1}^{t-1} \mathbb{E}_{x_s \sim d_0, a_s^1 \sim \hat{\pi}_s^1, a_s^2 \sim \hat{\pi}_s^2} (R(x_s, a_s^1, a_s^2) - \hat{R}(x_s, a_s^1, a_s^2))^2}}.$$

Then, the eluder coefficient is given by

$$d(\mathcal{R}, \lambda, T) := \sup_{\pi_{1:T}^1, \pi_{1:T}^2} \sum_{t=1}^{T} \min(1, \tilde{\Gamma}_t^2(\lambda, \pi_t^1, \pi_t^2)).$$

- Information ratio: *Prediction error* (target) v.s. *In-sample error* (guarantee);
- Eluder coefficient: limits the extent to which we can be "surprised" by the new out-of-sample distributions, given the historical data collected so far.

Online Learning Guarantee

Theorem 9 (Online learning guarantee [YXZ+24])

For any $\epsilon > 0$, *with* $T = \min\{n \in \mathbb{N}^+ : n \geq 2d(\mathcal{R}, \lambda, n)\}$, *batch size as* $m = O\big(T \log(T|\mathcal{R}|/\delta)/\epsilon^2\big)$, $\beta = \sqrt{T \log(T|\mathcal{R}|/\delta)/m}$,*then, with high probability, we can find a* $t_0 \in [T]$,

$$J(\pi_1^*, \pi_2^*) - J(\pi_{t_0}^1, \dagger) \leq \epsilon.$$

- $\hat{\pi}^1$ almost beats any competing policy:

$$\min_{\pi^2 \in \Pi} \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1 \sim \hat{\pi}^1, a^2 \sim \pi^2} \Big[R^*(x, a^1, a^2) - \eta D_{\mathrm{KL}}(\hat{\pi}^1(\cdot|x) \| \pi_0(\cdot|x)) + \eta D_{\mathrm{KL}}(\pi^2(\cdot|x) \| \pi_0(\cdot|x)) \Big] \geq -\epsilon.$$

- $\hat{\pi}^1$ is consistently preferred by $\mathcal{P}$ with small η:

$$\min_{\pi^2 \in \Pi} \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1 \sim \hat{\pi}^1, a^2 \sim \pi^2} \mathcal{P}(x, a^1, a^2) > \frac{1}{1 + \exp(\epsilon)} \approx 0.5.$$

- $\hat{\pi}^1$ automatically maximizes the regularized reward under BT model:

$$\mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1 \sim \hat{\pi}^1} \Big[r^*(x, a^1) - \eta D_{\mathrm{KL}}(\hat{\pi}^1(\cdot|x) \| \pi_0(\cdot|x)) \Big]$$
$$\geq \max_{\pi^2} \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^2 \sim \pi^2} \Big[r^*(x, a^2) - \eta D_{\mathrm{KL}}(\pi^2(\cdot|x) \| \pi_0(\cdot|x)) \Big] - \epsilon.$$

What we know? & What is next?

What we know:

- Iterative DPO is provably efficient and demonstrates impressive empirical performance;
- Reward hacking: Snorkel-Mistral-PairRM-DPO with output length (2616) beats GPT4 (1365);
- The complexity scales with the complexity of reward space: good if discriminator is simpler than generator (weak to strong?).

What is next?

- More efficient exploration strategy beyond rejection sampling;
- Efficient and effective reward modeling;
- Alignment tax: performance degeneration after RLHF [LLX+24];
- Practical implementation of NLHF.

Thanks for listening!

[BKK+22] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[Box76] George EP Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[BT52] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[FK22] Hao Fu, Yao; Peng and Tushar Khot. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*, Dec 2022.

[GSH23] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[GWZ22] Claudio Gentile, Zhilei Wang, and Tong Zhang. Fast rates in pool-based batch active learning. *arXiv preprint arXiv:2202.05448*, 2022.

[HT24] Braden Hancock Hoang Tran, Chris Glaze. Snorkel-mistral-pairrm-dpo. 2024.

[LLX+24] Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of rlhf, 2024.

[LZJ+23] Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

[OWJ+22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[RSM+23] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[SWD+17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[XDY+23] Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*, 2023.

[YXGZ23] Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*, pages 39834–39863. PMLR, 2023.

[YXZ+24] Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. A theoretical analysis of nash learning from human feedback under general kl-regularized preference, 2024.

[ZXT+22] Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pages 27117–27142. PMLR, 2022.