Beyond Last-Iterate Convergence for Nash Learning from Human Feedback

Runlong Zhou

May 23 2025

IFDS Presentation

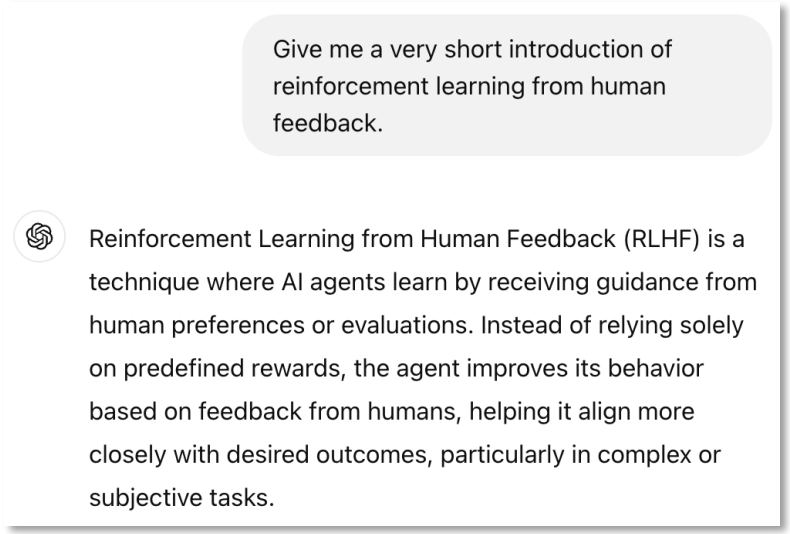
Acknowledgement

This is a joint work with

Maryam Fazel and **Simon Du**

Language Generation

- **Prompt (State)**: user queries
- **Response (Action)**: language model generation result



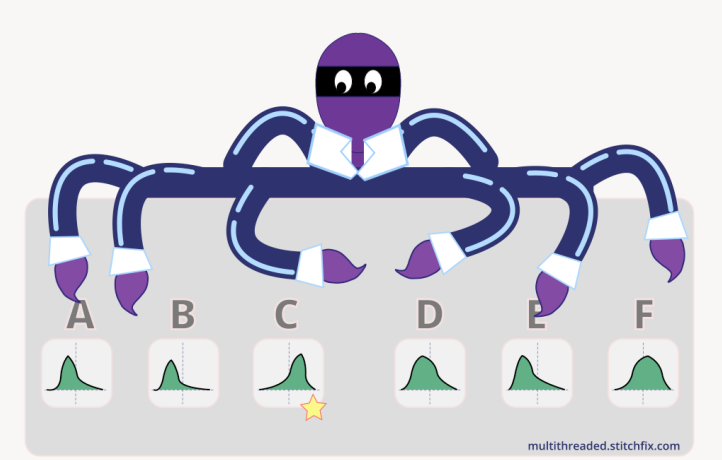
Bandits

Multi-armed bandits (MABs)

- **Arm** space $\mathcal{Y}$
- **Reward** function $r(y) \in [0,1]$

Contextual bandits (CBs)

- **Context (Prompt)** space $\mathcal{X}$
- **Arm (Response)** space $\mathcal{Y}$
- **Reward** function $r(x, y) \in [0,1]$



Picture from
https://multithreaded.stitchfix.com/blog/2020/08/05/bandits/

Results in this work can be easily adapted to multiple contexts, so focus on single-context for simplicity

Policy

- A **tabular softmax** policy π_θ satisfies

$$\pi_\theta(y) = \frac{e^{\theta_y}}{\sum_{y'} e^{\theta_{y'}}}$$

Reward-based v.s. Preference-based RL

Reward-based RL

After choosing an arm y, observe a sample $r \sim R(y)$ with mean $r(y)$

Preference-based RL

- A **preference** model $\mathcal{P}(y_1 \succ y_2)$ indicating the probability that y_1 is preferred over y_2
- After choosing a pair of arms (y_1, y_2), observe a sample $p \sim \text{Bernoulli}\big(\mathcal{P}(y_1 \succ y_2)\big)$

RL from Human Feedback (RLHF)

Christiano et al. (2017)

- Sigmoid function

$$\sigma(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

- Bradley-Terry (Bradley & Terry, 1952) preference model

$$\mathcal{P}(y_1 \succ y_2) = \sigma\big(r(y_1) - r(y_2)\big) = \frac{\mathrm{e}^{r(y_1)}}{\mathrm{e}^{r(y_1)} + \mathrm{e}^{r(y_2)}}$$

Core assumption of an underlying reward!

RL from Human Feedback (RLHF)

- Human preference dataset $\mathcal{D} = \left\{\left(y_w^{(i)}, y_l^{(i)}\right)\right\}_{i=1}^{N}$
  - In the ith sample, $y_w^{(i)}$ is preferred over $y_l^{(i)}$
- **Step 1**: Learn reward function by minimizing negative log-likelihood

$$\mathcal{L}_r(\phi) = -\frac{1}{N}\sum_{i=1}^{N} \log \sigma\left(r_\phi\left(y_w^{(i)}\right) - r_\phi\left(y_l^{(i)}\right)\right)$$

- **Step 2**: Learn policy by maximizing regularized value using proximal policy optimization (PPO)

$$\theta_\phi^\star = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{y\sim\pi_\theta}[r_\phi(y)] - \beta \mathrm{KL}(\pi_\theta || \pi_{\mathrm{ref}})$$

Nash Learning from Human Feedback (NLHF)

Munos et al. (2023)

- Human preferences are **non-transitive**!
  - **Reward + BT model** implies

$$\left(\mathcal{P}(y_1 \succ y_2) > \frac{1}{2}\right) \wedge \left(\mathcal{P}(y_2 \succ y_3) > \frac{1}{2}\right) \Rightarrow \left(\mathcal{P}(y_1 \succ y_3) > \frac{1}{2}\right)$$

  - Think about rock-paper-scissor game

This step may not hold!

- But still has a property
  - $\mathcal{P}(y_1 \succ y_2) + \mathcal{P}(y_2 \succ y_1) = 1$

Nash Learning from Human Feedback (NLHF)

- Two-player (π, π') constant-sum matrix game
  - Value for player 1: $\mathbb{P}(\pi \succ \pi') = \sum_{y,y'} \pi(y)\pi'(y')\mathcal{P}(y \succ y') = \pi^\top \mathcal{P} \pi'$
  - Value for player 2: $\mathbb{P}(\pi' \succ \pi) = \sum_{y,y'} \pi(y)\pi'(y')\mathcal{P}(y' \succ y) = 1 - \pi^\top \mathcal{P} \pi'$
  - Focus on $V(\pi, \pi') = \pi^\top \mathcal{P} \pi'$
  - Minimax theorem: $\max_{\pi} \min_{\pi'} V(\pi, \pi') = \min_{\pi'} \max_{\pi} V(\pi, \pi')$
  - Nash equilibrium (NE): $\pi^\star = \text{argmax}_\pi \min_{\pi'} V(\pi, \pi')$
  - Duality gap: $\text{DualGap}(\pi) = \max_{\pi'} V(\pi', \pi) - \min_{\pi''} V(\pi, \pi'')$

$= 1 - \mathcal{P}(y \succ y')$

Nash Learning from Human Feedback (NLHF)

- Regularized game

$$V_\beta(\pi_1, \pi_2) := \pi_1^\top \mathcal{P} \pi_2 - \beta \mathsf{KL}(\pi_1 || \pi_{\mathsf{ref}}) + \beta \mathsf{KL}(\pi_2 || \pi_{\mathsf{ref}})$$

- Quantal response equilibrium (QRE)

$$\theta_1^\star = \arg\max_{\theta_1} \min_{\theta_2} V_\beta(\pi_1, \pi_2)$$

  - Denote as $\pi_\beta^\star$

- Duality gap

$$\mathsf{DualGap}_\beta(\pi) = \max_{\pi'} V_\beta(\pi', \pi) - \min_{\pi''} V_\beta(\pi, \pi'')$$

Nash Learning from Human Feedback (NLHF)

- Regularized game

$$V_\beta(\pi_1, \pi_2) := \pi_1^\top \mathcal{P} \pi_2 - \beta \mathsf{KL}(\pi_1 || \pi_{\text{ref}}) + \beta \mathsf{KL}(\pi_2 || \pi_{\text{ref}})$$

- Optimal response given θ_2
  - $\theta_1 = \theta_{\text{ref}} + \frac{\mathcal{P}\pi_2}{\beta} + C\mathbf{1}$
  - $\pi_1 \propto \pi_{\text{ref}} \exp\left(\frac{\mathcal{P}\pi_2}{\beta}\right)$
- Finding the QRE

$$\theta = \theta_{\text{ref}} + \frac{\mathcal{P}\pi_\theta}{\beta}$$

NLHF: Desirable Properties

- Convergence
  - Last-iterate convergence in $\mathrm{KL}(\pi^{(T)} || \pi_\beta^\star)$, $\mathrm{DualGap}(\pi^{(T)})$, and $\mathrm{DualGap}_\beta(\pi^{(T)})$
    - Averaging historical LLMs is computationally prohibitive
  - Convergence rate
    - Faster convergence = fewer rounds of costly human data collection
    - **Linear** for regularized games, **polynomial** for original games

NLHF: Desirable Properties

- Convergence
  - Last-iterate **linear/polynomial** convergence
- Robustness to **noise**
  - Human data is inherently noisy
  - Algorithm's convergence rate should ideally remain unchanged by noise, converging to a value influenced by noise variance

NLHF: Desirable Properties

- Convergence
  - Last-iterate **linear/polynomial** convergence
- Robustness to **noise**
- Implementation for **neural networks**
  - Theory often derived for **tabular** policies
  - Avoid constrained parametrization approach
    - Projection is intractable for NNs
  - Avoid nested optimization: $\theta^{(t+1)} = \arg\min_\theta \mathcal{L}_{\text{inner}}(\theta; \theta^{(t)})$

Our Contributions

- Extragradient Preference Optimization (EGPO)
  - Achieving all previous desirable properties
- A general approach to avoid nested optimization in NLHF
  - **Solution** for each inner optimization = online IPO (Azar et al., 2023) **gradient**

Extragradient Method

- Assume we use an **implicit** update rule

$$\theta^{(t+1)} = (1-\eta\beta)\theta^{(t)} + \eta\beta\left(\theta_{\text{ref}} + \frac{\mathcal{P}\pi^{(t+1)}}{\beta}\right)$$

- We have

$$\left\langle \log \pi^{(t+1)} - (1-\eta\beta)\log \pi^{(t)} - \eta\beta \log \pi_\beta^\star \, , \pi^{(t+1)} - \pi_\beta^\star \right\rangle = 0$$

- By definition of KL divergence

$$(1-\eta\beta)\mathrm{KL}\left(\pi^{(t+1)}||\pi^{(t)}\right) + \eta\beta\mathrm{KL}\left(\pi^{(t+1)}||\pi_\beta^\star\right) - (1-\eta\beta)\mathrm{KL}\left(\pi_\beta^\star||\pi^{(t)}\right) + \mathrm{KL}\left(\pi_\beta^\star||\pi^{(t+1)}\right) = 0$$

- So

$$\mathrm{KL}\left(\pi_\beta^\star||\pi^{(t+1)}\right) \leq (1-\eta\beta)\mathrm{KL}\left(\pi_\beta^\star||\pi^{(t)}\right)$$

Extragradient Method

- Use an extragradient step to estimate $\theta^{(t+1)}$!
- Generalizing the predictive update (PU) algo in Cen et al. (2021)

$$\theta^{(t+1/2)} = (1-\eta\beta)\theta^{(t)} + \eta\beta\left(\theta_{\text{ref}} + \frac{\mathcal{P}\pi^{(t)} + \epsilon^{(t)}}{\beta}\right), \tag{2}$$

$$\theta^{(t+1)} = (1-\eta\beta)\theta^{(t)} + \eta\beta\left(\theta_{\text{ref}} + \frac{\mathcal{P}\pi^{(t+1/2)} + \epsilon^{(t+1/2)}}{\beta}\right). \tag{3}$$

- $\mathbb{E}\left[\epsilon^{(i)}\middle|\pi^{(i)}\right] = \mathbf{0}, \epsilon_y^{(i)} \sim \text{subGaussian}(\sigma^2)$

Extragradient Method

- Guarantees for the regularized game

Similar for $\pi^{\left(T+\frac{1}{2}\right)}$

Theorem 1. *For any initialization $\theta^{(0)}$, following the update rules defined by Equations (2) and (3) and setting $\eta \leqslant \frac{1}{\beta+2|\mathcal{Y}|+1}$, we have that for any $T \geqslant 1$,*

$$\mathbb{E}[\mathsf{KL}(\pi_\beta^\star||\pi^{(T)})] \leqslant \mathsf{KL}(\pi_\beta^\star||\pi^{(0)})(1-\eta\beta)^T + \frac{4\sigma^2\log(3\,|\mathcal{Y}|)}{\beta},$$

$$\mathbb{E}[\mathsf{KL}(\pi^{(T)}||\pi_\beta^\star)] \leqslant \frac{2\mathsf{KL}(\pi_\beta^\star||\pi^{(0)})}{\eta\beta}(1-\eta\beta)^T + \frac{8\sigma^2\log(3\,|\mathcal{Y}|)}{\eta\beta^2},$$

$$\mathbb{E}[\mathsf{DualGap}_\beta(\pi^{(T)})] \leqslant \left(\frac{2\,|\mathcal{Y}|^2}{\beta}+\frac{4}{\eta}\right)\mathsf{KL}(\pi_\beta^\star||\pi^{(0)})(1-\eta\beta)^T + \left(\frac{8\,|\mathcal{Y}|^2}{\beta^2}+\frac{16}{\eta\beta}\right)\sigma^2\log(3\,|\mathcal{Y}|).$$

*Under the **exact update** scheme where $\sigma^2 = 0$, all the expectations are removed.*

Extragradient Method

- Guarantee for the original game

Theorem 2. *Consider the* ***exact update*** *scheme, where* $\sigma^2 = 0$. *By setting* $\pi^{(0)} = \pi_{\text{ref}} = \text{Uniform}(\mathcal{Y})$, $\beta = \frac{\varepsilon}{4\log|\mathcal{Y}|}$, *and* $\eta \leqslant \frac{1}{\beta+2|\mathcal{Y}|+1}$ *we have that for any* $T \geqslant \widetilde{\Omega}(1/\eta\varepsilon)$,

$$\text{DualGap}(\pi^{(T)}), \text{DualGap}(\pi^{(T+1/2)}) \leqslant \varepsilon.$$

Remarks on EGPO

- Compared to results in Cen et al. (2021)
  - Added guarantees on $\mathrm{KL}\left(\pi^{(T)} || \pi_\beta^\star\right)$ and $\mathrm{DualGap}_\beta\left(\pi^{(T)}\right)$
  - Analyzed empirical updates

Remarks on EGPO

Algorithm	Convergence to Regularized QRE	Last-iterate Convergence	σ^2-noise Robustness	Convergence to Original ε-NE
Online Mirror Descent	$\tilde{O}(1/T)$	No	Not provided	$\tilde{O}(1/\varepsilon^2)$ iterations
Nash-MD (Munos et al., 2023)	$\tilde{O}((1-\eta\beta)^T + \eta/\beta)$	**Yes**	Not provided	Not provided
MTPO[1] (Shani et al., 2024)	$\tilde{O}(1/T)$			
SPO[2] (Swamy et al., 2024) SPPO (Wu et al., 2024)	Not provided	No	Not provided	$\tilde{O}(1/\varepsilon^2)$ iterations
INPO[3] Zhang et al. (2024)	$\tilde{O}(1/T)$	**Yes**	Not provided	Not provided
MPO Wang et al. (2024)	$\tilde{O}((\frac{1}{1+\eta\beta})^T)$	**Yes**	Not provided	Asymptotically
ONPO Zhang et al. (2025)	Not provided	No	Not provided	$\tilde{O}(1/\varepsilon)$ **iterations**
EGPO This work	$\tilde{O}((1-\eta\beta)^T)$	**Yes**	$+\tilde{O}(\sigma^2/(\eta\beta^2))$	$\tilde{O}(1/\varepsilon)$ **iterations**

Neural Network Implementation

$$\theta^{(t+1/2)} = (1-\eta\beta)\theta^{(t)} + \eta\beta\left(\theta_{\text{ref}} + \frac{\mathcal{P}\pi^{(t)} + \cancel{\epsilon^{(t)}}}{\beta}\right)$$

- Minimizer for

$$\mathcal{L}_{\text{inner}}^{(t)}(\theta) = -\pi_\theta^\top\left(\mathcal{P}\pi^{(t)}\right) + \beta\text{KL}(\pi_\theta||\pi_{\text{ref}})$$

- EGPO update rule for NNs

$$\begin{cases}\theta^{(t+1/2)} = \theta^{(t)} - \eta\beta\text{argmin}_\theta\mathcal{L}_{\text{inner}}^{(t)}(\theta)\\ \theta^{(t+1)} = \theta^{(t)} - \eta\beta\text{argmin}_\theta\mathcal{L}_{\text{inner}}^{(t+1/2)}(\theta)\end{cases}$$

Avoiding Nested Optimization

- Identity preference optimization (IPO)

$$\mathcal{L}_{\text{IPO}}(\theta; \rho, \mu) = \mathbb{E}_{(y,y')\sim\rho}\left[\left(\log \frac{\pi_\theta(y)\pi_{\text{ref}}(y')}{\pi_\theta(y')\pi_{\text{ref}}(y)} - \frac{1}{\beta}\mathbb{E}_{y''\sim\mu}[\mathcal{P}(y \succ y'') - \mathcal{P}(y' \succ y'')]\right)^2\right]$$

- Online IPO (Calandriello et al., 2024)
  - When at least one of ρ and μ is current policy π_θ

Avoiding Nested Optimization

- IPO gradient

Define $\Sigma(\rho) := \mathbb{E}_{(y,y')\sim\rho}[(\mathbb{1}_y - \mathbb{1}_{y'})(\mathbb{1}_y - \mathbb{1}_{y'})^\top]$, then

$$\nabla_\theta \mathcal{L}_{\mathsf{IPO}}(\theta;\rho,\mu) = 2\mathbb{E}_{(y,y')\sim\rho}\left[\left(\theta - \theta_{\mathsf{ref}} - \frac{\mathcal{P}\mu}{\beta}\right)^\top (\mathbb{1}_y - \mathbb{1}_{y'}) \cdot (\mathbb{1}_y - \mathbb{1}_{y'})\right] = 2\Sigma(\rho)\left(\theta - \theta_{\mathsf{ref}} - \frac{\mathcal{P}\mu}{\beta}\right)$$

- Set $\pi^{\mathsf{s}} := \mathsf{Uniform}(\mathcal{Y}) \times \mathsf{Uniform}(\mathcal{Y})$, then $\Sigma(\pi^{\mathsf{s}}) = \frac{2}{|\mathcal{Y}|^2}(|\mathcal{Y}|\, I - \mathbb{1})$
- Finally,

$$\nabla_\theta \mathcal{L}_{\mathsf{IPO}}(\theta;\pi^{\mathsf{s}},\mu) = \frac{4}{|\mathcal{Y}|}\left(\theta - \theta_{\mathsf{ref}} - \frac{\mathcal{P}\mu}{\beta}\right) + C\mathbb{1}$$

Avoiding Nested Optimization

- EGPO update rule for NNs

$$\begin{cases} \theta^{(t+1/2)} = \theta^{(t)} - \eta\beta \operatorname{argmin}_\theta \mathcal{L}_{\text{inner}}^{(t)}(\theta) \\ \theta^{(t+1)} = \theta^{(t)} - \eta\beta \operatorname{argmin}_\theta \mathcal{L}_{\text{inner}}^{(t+1/2)}(\theta) \end{cases}$$

- Implement as

Stop gradient

$$\theta^{(t+1/2)} = \theta^{(t)} - \frac{\eta\beta\,|\mathcal{Y}|}{4} \nabla_\theta \mathcal{L}_{\text{IPO}}(\theta^{(t)}; \pi^{\text{s}}, \text{sg}[\pi^{(t)}]),$$

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\eta\beta\,|\mathcal{Y}|}{4} \nabla_\theta \mathcal{L}_{\text{IPO}}(\theta^{(t)}; \pi^{\text{s}}, \pi^{(t+1/2)}),$$

Online IPO Population Loss

- Last step towards implementable training

Theorem 3. *Define*

$$\widehat{\mathcal{L}}(\theta; \pi^{\mathsf{s}}, \mu) := \mathbb{E}_{(y,y')\sim\pi^{\mathsf{s}}, y''\sim\mu}\left[\left(\log\frac{\pi_\theta(y)\pi_{\mathsf{ref}}(y')}{\pi_\theta(y')\pi_{\mathsf{ref}}(y)} - \frac{I(y,y'') - I(y',y'')}{\beta}\right)^2\right],$$

where $I(y,y''), I(y',y'')$ are unbiased estimators for $\mathcal{P}(y \succ y''), \mathcal{P}(y' \succ y'')$, respectively. Then

$$\mathbb{E}[\nabla_\theta \widehat{\mathcal{L}}(\theta; \pi^{\mathsf{s}}, \mu)] = \nabla_\theta \mathcal{L}_{\mathsf{IPO}}(\theta; \pi^{\mathsf{s}}, \mu).$$

Experiments

- Baselines
  - Online IPO 1 (Online Mirror Descent)

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \frac{\eta\beta\,|\mathcal{Y}|}{4}\nabla_\theta \mathcal{L}_{\text{IPO}}(\theta^{(t)}; \pi^{\text{s}}, \text{sg}[\pi^{(t)}])$$

  - Online IPO 2

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \frac{\eta\beta\,|\mathcal{Y}|}{4}\nabla_\theta \mathcal{L}_{\text{IPO}}(\theta^{(t)}; \text{sg}[\pi^{(t)}], \text{sg}[\pi^{(t)}])$$

  - Nash-MD

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \frac{\eta\beta\,|\mathcal{Y}|}{4}\nabla_\theta \mathcal{L}_{\text{IPO}}(\theta^{(t)}; \pi^{\text{s}}, \text{sg}[\widetilde{\pi}^{(t)}])$$

Experiments

- Baselines
  - Nash-MD

$$\pi^{(t+1)} = \arg\max_{\pi} \{\eta \mathcal{P}(\pi \succ \widetilde{\pi}^{(t)}) - \mathsf{KL}(\pi || \widetilde{\pi}^{(t)})\}$$

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \frac{\eta\beta\,|\mathcal{Y}|}{4} \nabla_\theta \mathcal{L}_{\mathsf{IPO}}(\theta^{(t)}; \pi^{\mathsf{s}}, \mathsf{sg}[\widetilde{\pi}^{(t)}])$$

  - Nash-MD-PG (nested optimization)

$$\theta^{(t+1)} = \theta^{(t)} + \eta \mathbb{E}_{y \sim \pi^{(t)}} \left[\left(P\widetilde{\pi}^{(t)} - \beta \log \frac{\pi_\theta(y)}{\pi_{\mathsf{ref}}(y)} \right) \nabla_\theta \log \pi^{(t)}(y) \right]$$

Numerical Simulations

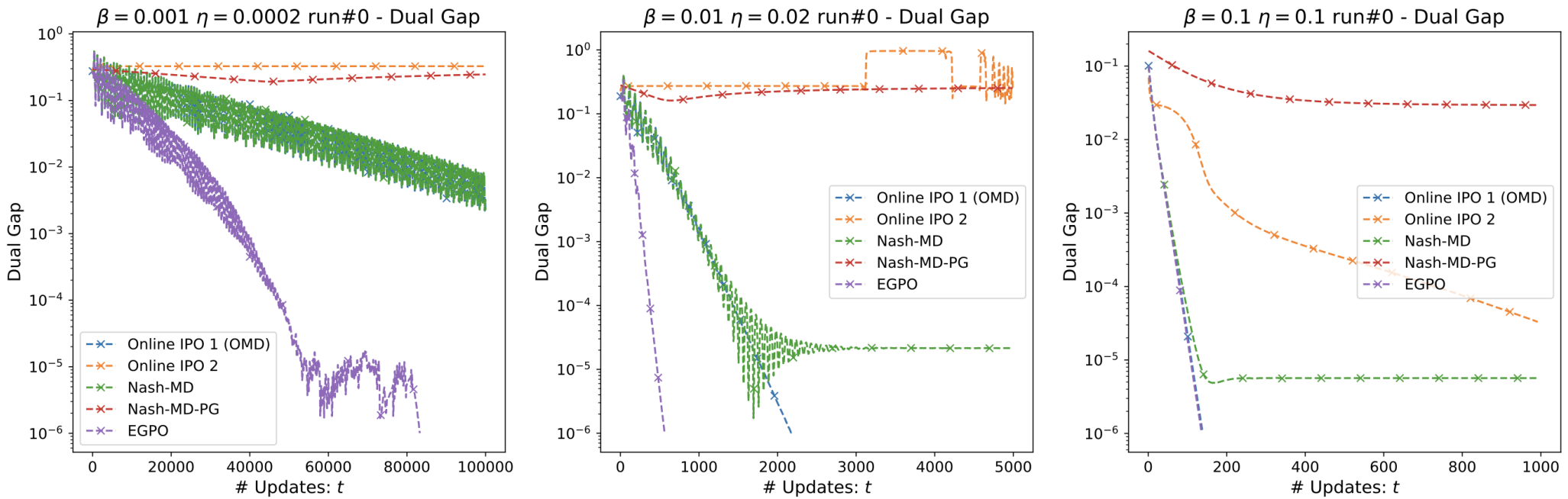
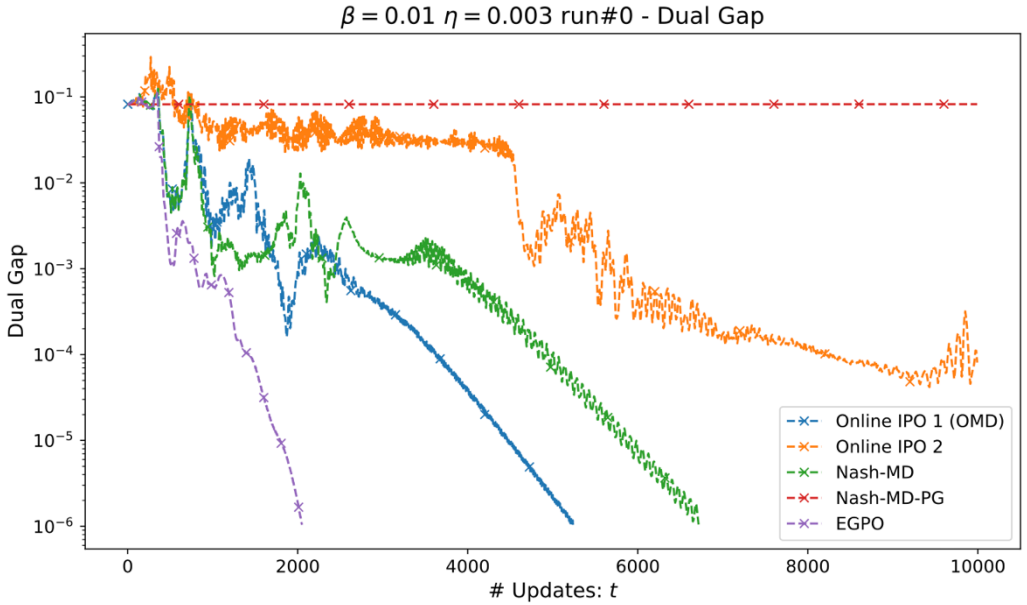
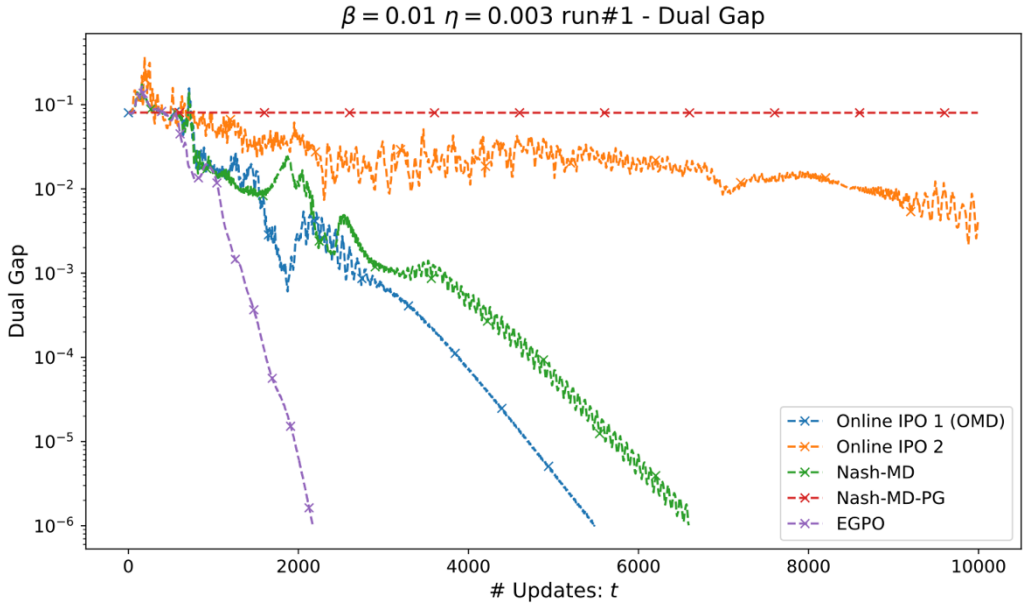


Figure 1: Duality gap (DualGap_β) of **exact tabular** algorithms with different βs. Values are cut off below 10^{-6} due to floating point precision. These figures are the 0th experiments shown in Figures 2 to 4.

Numerical Simulations

- Exact Neural





Benchmarks

- Pairwise win-rates on PKU-SafeRLHF
- Checkpoints are top-2 against π_{ref}

ALG	Ep	π_{ref}	OIPO1		OIPO2		NMD		NMDPG		MPO		EGPO	
			6	8	6	9	8	10	4	8	7	8	5	8
OIPO1	6	72.8%			58.6%	57.6%	47.7%	46.4%	68.4%	69.4%	45.2%	47.0%	42.6%	42.8%
	8	71.8%			58.9%	58.7%	48.1%	47.0%	68.2%	68.0%	45.7%	47.2%	42.1%	43.6%
OIPO2	6	66.8%	41.4%	41.1%			39.8%	38.5%	62.3%	61.3%	41.3%	42.8%	33.8%	35.2%
	9	66.3%	42.4%	41.3%			38.5%	38.7%	61.2%	61.3%	40.8%	42.7%	34.2%	33.8%
NMD	8	72.8%	52.3%	51.9%	60.2%	61.5%			70.0%	71.1%	46.4%	48.3%	44.0%	46.7%
	10	72.9%	53.6%	53.0%	61.5%	61.3%			70.6%	71.2%	47.3%	49.2%	44.6%	45.8%
NMDPG	4	55.2%	31.6%	31.8%	37.7%	38.8%	30.0%	29.4%			31.5%	33.2%	26.2%	26.4%
	8	55.1%	30.6%	32.0%	38.7%	38.7%	28.9%	28.8%			31.1%	32.2%	26.2%	25.8%
MPO	7	71.9%	54.8%	54.3%	58.7%	59.2%	53.6%	52.7%	68.5%	68.9%			49.4%	47.9%
	8	70.2%	53.0%	52.8%	57.2%	57.3%	51.7%	50.8%	66.8%	67.8%			47.2%	46.9%
EGPO	5	76.9%	57.4%	57.9%	66.2%	65.8%	56.0%	55.4%	73.8%	73.8%	50.6%	52.8%		
	8	77.4%	57.2%	56.4%	64.8%	66.2%	53.3%	54.2%	73.6%	74.2%	52.1%	53.1%		

References

- Paul Francis Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. ArXiv, abs/1706.03741, 2017.
- 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. ISSN 00063444.
- R´ emi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. arXiv preprint arXiv:2312.00886, 2023.
- Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. Advances in Neural Information Processing Systems, 34: 27952–27964, 2021.
- Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human alignment of large language models through online preference optimisation, 2024. URL https://arxiv.org/abs/2403.08635.

Thank You