Accelerating Nash Learning from Human Feedback via Mirror Prox

Daniil Tiapkin[1,2], Daniele Calandriello[3], Denis Belomestny[4,5], Éric Moulines[1,6], Alexey Naumov[5], Kashif Rasul[7], Michal Valko[8], Pierre Ménard[9]

[1]CMAP, CNRS, École Polytechnique, [2]LMO, Université Paris-Saclay, [3]Google DeepMind, [4]Duisburg-Essen University, [5]HSE University, [6]Mohamed Bin Zayed University of AI, [7]Hugging Face, [8]Stealth Startup / Inria / ENS, [9]ENS Lyon

- **Problem:** Traditional RLHF relies on reward models (e.g., Bradley-Terry) which fail to capture intransitive human preferences.
- **Alternative:** Nash Learning from Human Feedback (NLHF) frames the problem as finding a Nash Equilibrium (NE) of a preference game.
- **Our Contribution:** We introduce **Nash Mirror Prox (NashMP)**, a novel online NLHF algorithm.
- **Key Feature:** NashMP leverages the Mirror Prox optimization scheme to achieve faster convergence, which allows for **last-iterate linear convergence** to the regularized NE.
- **Practice:** Our method is compatible with existing methods, and shows competitive performance in fine-tuning Large Language Models (LLMs).

Setting: Regularized Nash Learning

- **Preference game:** Preferences $\mathcal{P}(y \succ y'|x)$ induces a bilinear form over preferences $\mathcal{P}(\pi \succ \pi')$ and thus we can define $\max_\pi \min_{\pi'} \mathcal{P}(\pi \succ \pi')$;
- **Goal:** Find a symmetric NE, or von Neumann Winner (VNW), a policy $\pi^\star$ that beats any other policy with probability at least $1/2 : \mathcal{P}(\pi^\star \succ \pi) \geq 1/2$.
- **Regularized Game:** For practical LLM fine-tuning, we must stay close to a reference policy π^{ref} (e.g., the SFT model). We solve a regularized game with the objective:

$$\max_\pi \min_{\pi'} \mathcal{P}_\beta(\pi \succ \pi') \triangleq \mathcal{P}(\pi \succ \pi') - \beta \mathrm{KL}(\pi \| \pi^{\mathrm{ref}}) + \beta \mathrm{KL}(\pi' \| \pi^{\mathrm{ref}})$$

- This regularized game has a unique NE, denoted $\pi^\star_\beta$. Finding it efficiently is the main objective.

Algorithm: Nash Mirror Prox (NashMP)

NashMP is an adaptation of the Mirror Prox method to the regularized preference game. It performs a two-step update at each iteration k:

❶ **Extrapolation Step:** Compute a best response against the *online* policy π_k, staying close to a *target* policy π_k and π^{ref}:

$$\pi_{k+1/2} = \arg\min_\pi \left\{ \mathcal{P}(\pi_k \succ \pi) + \beta \mathrm{KL}(\pi \| \pi^{\mathrm{ref}}) + \frac{\beta}{\eta} \mathrm{KL}(\pi \| \pi_k) \right\}.$$

❷ **Update Step:** Compute a best response against the *online* policy $\pi_{k+1/2}$, staying close to a *target* policy π_k and π^{ref}:

$$\pi_{k+1} = \arg\min_\pi \left\{ \mathcal{P}(\pi_{k+1/2} \succ \pi) + \beta \mathrm{KL}(\pi \| \pi^{\mathrm{ref}}) + \frac{\beta}{\eta} \mathrm{KL}(\pi \| \pi_k) \right\}.$$

Intuition: Two-step approximation of a more numerically stable discretization of the gradient flow ODE: *proximal point method*

$$\pi_{k+1} = \arg\min_\pi \left\{ \mathcal{P}(\pi_{k+1} \succ \pi) + \beta \mathrm{KL}(\pi \| \pi^{\mathrm{ref}}) + \frac{\beta}{\eta} \mathrm{KL}(\pi \| \pi_k) \right\}.$$

Theorem. For $\beta < 1/2$, for the last iterates $\pi_K, \pi_{K+1/2}$ of NashMP

- The KL-divergence decreases as: $\mathrm{KL}(\pi^\star_\beta \| \pi_K) = \mathcal{O}\left((1+2\beta)^{-K}/\beta\right)$;
- Exploitability gap satisfies $\mathrm{SubOpt}_\beta(\pi_{K+1/2}) = \mathcal{O}\left((1+2\beta)^{-K/2}/\beta\right)$;
- Span semi-norm in log-probs $\| \log \pi_K - \log \pi^\star_\beta \|_{\mathrm{span}} = \mathcal{O}\left((1+2\beta)^{-K/2}/\beta\right)$;

where K is the number of iterations ($N = 2K$ preference queries).

Algorithm	KL to β-reg. VNW
NashMD (Munos et al., 2023)	$\mathcal{O}((\beta^2 N)^{-1})$
Online IPO (Calandriello et al., 2024)	Asymptotic
INPO (Zhang et al., 2025)	$\mathcal{O}((\beta^2 N)^{-1})$
MMD (Wang et al., 2025)	$\mathcal{O}((1+\beta^2)^{-N}/\beta)$
EGPO (Zhou et al., 2025)	$\mathcal{O}((1-\beta/(1+\beta+2Y))^N)$
NashMP (this paper)	$\mathcal{O}((1+2\beta)^{-N/2}/\beta)$

- **Original Game:** NashMP finds an ϵ-VNW of the *unregularized* game with $\tilde{\mathcal{O}}(1/\epsilon)$ queries, matching SOTA while providing stronger guarantees for the regularized setting.

Approximate NashMP

Problem: steps of NashMP are intractable under a functional approximation, thus we need an approximation for $p \in \{1, 2\}$

$$\hat{\pi}_{k+p/2} \approx \arg\min_{\pi \in \Pi} \left\{ \mathcal{P}(\hat{\pi}_{k+(p-1)/2} \succ \pi) + \beta \mathrm{KL}(\pi \| \pi^{\mathrm{ref}}) + (\beta/\eta) \mathrm{KL}(\pi \| \hat{\pi}_k) \right\},$$

Solution: approximate steps by policy gradients:

$$\theta_{k+\frac{p}{2}, t+1} = \theta_{k+\frac{p}{2}, t} - \gamma \hat{\nabla} J_{k+\frac{p}{2}}(\theta_{k+\frac{p}{2}, t}),$$

where

$$J_{k+p/2}(\theta) \triangleq \mathbb{E}_{y' \sim \pi_\theta}[\mathcal{P}(\hat{\pi}_{k+(p-1)/2} \succ y')] + \beta \mathrm{KL}(\pi_\theta \| \pi^{\mathrm{ref}}) + (\beta/\eta) \mathrm{KL}(\pi_\theta \| \hat{\pi}_k).$$

Lemma. Let $\bar{\varepsilon} < 1/3$ and assume that $\hat{\nabla} J_{k+\frac{p}{2}}$ is estimated using a batch size of size B, it holds $\log \hat{\pi}_{k+p/2} - \log \pi_{k+p/2} \leq \bar{\varepsilon}$ for all $k \in \{0, \ldots, K-1\}$ and $p \in \{1, 2\}$ with high probability after T steps, where

$$T = \mathcal{O}((c^\star_\beta)^{-1} \log(1/(\beta \bar{\varepsilon}))), \quad B = \mathcal{O}((c^\star_\beta \cdot \bar{\varepsilon})^{-2}).$$

Practical Implementation for LLMs

The exact updates are infeasible for LLMs. We propose a practical, approximate version.

- **Key Idea:** Instead of solving the inner minimization problems exactly, we take one (or few) gradient steps and use a slowly-updated target network.
- **Loss Function:** The online policy π_θ is updated using a loss that pits it against a target policy $\pi_{\theta^{\mathrm{target}}}$:

$$\mathcal{L}_{\mathrm{NashMP}}(\theta) = \mathbb{E}_{x \sim \rho, y, y' \sim \pi_\theta}\left[P(y \succ y'|x) + \beta \log \frac{\pi_\theta(y|x)}{\pi^{\mathrm{ref}}(y|x)} + \frac{\beta}{\eta} \log \frac{\pi_\theta(y|x)}{\pi_{\theta^{\mathrm{target}}}(y|x)} \right]$$

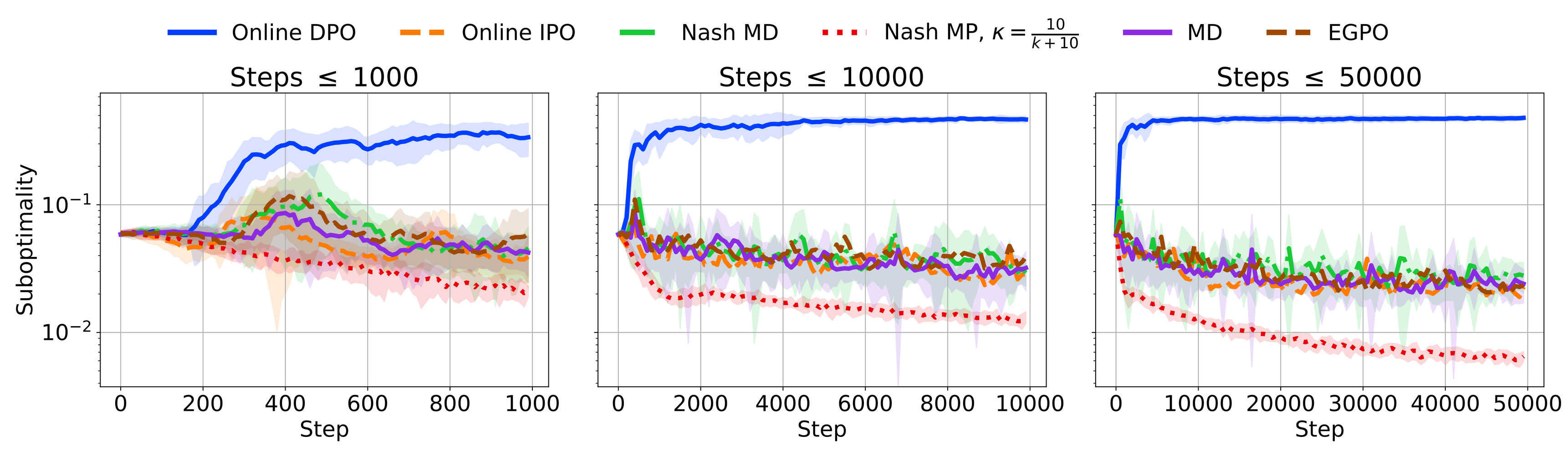
- **Target Update:** The target network parameters θ^{target} are updated via an exponential moving average (EMA) of the online parameters θ:

$$\theta^{\mathrm{target}}_{t+1} = \kappa \theta_t + (1-\kappa) \theta^{\mathrm{target}}_t$$

- The EMA parameter κ controls the trade-off, with $1/\kappa$ acting as the effective number of inner optimization steps. This approach is common in deep RL and stabilizes training.

Experiments: Matrix Games

- **Setup:** A contextual dueling bandit game designed to lack a Bradley-Terry reward model (i.e., has intransitivity).



Experiments: LLM Alignment

- **Setup:** Fine-tuning a Gemma-2B model on the RLHFlow dataset. We compare against Online DPO, Online IPO, NashMD, and "Regularized Self-Play" (NashMP without the target network).
- **Results:** Pairwise win rates judged by a more powerful Gemma-9B model.

Win rate	SFT	Online DPO	Online IPO	NashMD	Reg. Self-Play	NashMP, $\kappa = 0.1$
SFT	–	0.1623±0.0087	0.1554±0.0091	0.1974±0.0098	0.1536±0.0087	0.1283±0.0081
Online DPO	**0.8377**±0.0087	–	0.4743±0.0115	**0.5788**±0.0116	0.4730±0.0113	0.4392±0.0116
Online IPO	**0.8446**±0.0091	**0.5257**±0.0115	–	**0.6115**±0.0121	0.5036±0.0118	0.4706±0.0117
NashMD	**0.8026**±0.0098	0.4212±0.0116	0.3885±0.0121	–	0.4031±0.0119	0.3605±0.0115
Reg. Self-Play	**0.8464**±0.0087	**0.5270**±0.0113	0.4964±0.0118	**0.5969**±0.0119	–	0.4620±0.0118
NashMP, $\kappa = 0.1$	**0.8717**±0.0081	**0.5608**±0.0116	**0.5294**±0.0117	**0.6395**±0.0115	**0.5380**±0.0118	–