Nash Learning From Human Feedback by Munos et al.

Clifford Lin, Yuzhou Liu, Travis McVoy

April 2026

Motivation: Why Move Beyond RLHF?

- Standard RLHF Pipeline:
  - Collect pairwise human preferences ("Which response is better?")
  - Train a reward model (Bradley-Terry) that assigns a single scalar score
  - Fine-tune LLM via RL to maximize that score
- Key Limitations of RLHF:
  - Single scores cannot capture non-transitive preferences
    - **Example:** Three restaruants A, B, C with $r(A) = 4.56$, $r(B) = 4$, and $r(C) = 3.8$
    - Rating system always picks A over C because $r(A) > r(B) > r(C)$
  - Reducing preferences to single number oversimplifies diverse human preferences
    - Some may prefer C over A even though $r(C) < r(A)$
  - Reward model is sensitive to data distribution; needs retraining when data shifts

NLHF Approach: Key Ideas

- Instead of a reward model (**single** score), learn a preference model:
  - Takes **two** responses as input and predicts which is preferred
  - Captures richer, more complex preference patterns
- Preference models can be constructed such that they implicitly define a **two-player constant sum game**
  - Maximizing preference prediction accuracy is achieved by finding a Nash Equilibrium
- Advantages over RLHF:
  - More expressive: can model non-transitive and diverse preferences
  - Better alignment with diverse human populations (mixture, not single winner)
  - Independent of data distribution — no need to retrain from scratch

Prior Work & Connections to Online Learning

- Preference-based RL: Learn directly from pairwise preferences, not scalar rewards
  - RLHF (Christiano et al., 2017; OpenAI, 2022): Learn reward model, then optimize via RL
  - Avoids reward hacking (gaming the score without real improvement)
- Optimization without reward function:
  - DPO (Rafailov et al., 2023): Optimizes policy via BT-based loss — still BT assumption
  - IPO (Azar et al., 2023), GPO (Tang et al., 2024): Directly optimize pairwise preference
  - Online IPO approximates Nash equilibrium of a preference model (Self-Play)
- Connection to Online Learning
  - NLHF frames alignment as a two-player constant-sum game
  - Uses online learning tools: mirror descent, regret minimization, fictitious play
  - Nash-MD is a novel variant of OMD for Nash equilibrium computation

Notation

- Context (prompt) space $\mathcal{X}$ and action (response) space $\mathcal{Y}$
- Given a prompt x and candidate responses y, y', the model assigns a preference $\mathcal{P}(y \succ y'|x)$ between 0 and 1
- Preferences are antisymmetric: $\mathcal{P}(y \succ y'|x) = 1 - \mathcal{P}(y' \succ y|x)$
  - **Interpretation:** $\mathcal{P}(y \succ y'|x)$ represents the probability a randomly sampled human prefers response y over y' to prompt x
- Responses are drawn from conditional distributions $\pi(\cdot|x)$
  - Policy preference given a state x:

$$\mathcal{P}(\pi \succ \pi'|x) = \mathbb{E}_{y\sim\pi(\cdot|x), y'\sim\pi'(\cdot|x)}[\mathcal{P}(y \succ y'|x)]$$

  - Preference of action over an entire policy

$$\mathcal{P}(y \succ \pi'|x) = \mathbb{E}_{y'\sim\pi'(\cdot|x)}[\mathcal{P}(y \succ y'|x)]$$

  - Preference between two polices:

$$\mathcal{P}(\pi \succ \pi') = \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x), y'\sim\pi'(\cdot|x)}[\mathcal{P}(y \succ y'|x)]$$

Learning Preference Models

- Learning preferences when comparing **actions** is done via regression with a cross entropy loss

$$\mathcal{P}^* = \arg\max_{\mathcal{P}(\cdot \succ \cdot | \cdot)} \mathbb{E}[\log \mathcal{P}(y_w^Z \succ y_l^Z | x)]$$

where $\mathbb{E}$ is taken with respect to $x \sim \rho, y \sim \pi, y' \sim \pi', Z \sim \nu$ and y_w^Z is the winner between y and y' for a randomly selected human Z

  - Equivalently,

$$\mathcal{P}^*(y \succ y' | x) = \Pr_{Z \sim \nu}[\text{Human } Z \text{ prefers } y \text{ given } x]$$

and therefore preferences between actions are independent of π, π', ρ

- The goal of the paper is to find a **policy** π^* that is preferred over all other polices:

$$\pi^* = \arg\max_{\pi} \min_{\pi'} \mathcal{P}(\pi \succ \pi')$$

which defines a game

  - Players correspond are the polices π and π'
  - Payoffs are preferences $\mathcal{P}(\pi \succ \pi')$ and $1 - \mathcal{P}(\pi \succ \pi')$

Regularization

- Regularized preference between **actions**:

$$\mathcal{P}_\tau^{\pi,\pi'} = \mathcal{P}(y \succ y'|x) - \tau \log \frac{\pi(y|x)}{\mu(y|x)} + \tau \log \frac{\pi'(y'|x)}{\mu(y'|x)}$$

  where μ is a reference policy and τ a temperature parameter
- KL-regularized preference between **polices**:

$$\begin{aligned}\mathcal{P}_\tau(\pi \succ \pi') &= \mathbb{E}_{x\sim\rho, y\sim\pi(\cdot|x), y'\sim\pi'(\cdot|x)}[\mathcal{P}_\tau^{\pi,\pi'}(y \succ y'|x)] \\ &= \mathcal{P}(\pi \succ \pi') - \tau \mathsf{KL}_\rho(\pi, \mu) + \tau \mathsf{KL}_\rho(\pi', \mu)\end{aligned}$$

  where $\mathsf{KL}_\rho(\pi, \mu) = \mathbb{E}_{x\sim\rho}[\mathsf{KL}\big(\pi(\cdot|x), \mu(\cdot|x)\big)]$
- Regularization simultaneously achieves two important feats
  - **Trustworthiness.** A policy should not violate the trained model's safety protocols just because user prefers the model to
  - **Unique Nash:** The regularized policy $\mathcal{P}_\tau$ has a unique Nash equilibrium, which will be important in main Theorem

Algorithms For Approximating Nash Equilibria

- Convergence on average:
  - Fictitious play: $\pi_{t+1} = \arg\max_{\pi} \mathcal{P}(\pi \succ \bar{\pi}_t)$
  - Drawback: need to store past polices
- Last Iterate Convergence:
  - For example: Optimistic OMD, extragradient
  - Drawback: More complicated structure, slower learning rate

Nash-MD

- Define a regularized policy π_t^μ as a **geometric mixture** between current policy π_t and reference policy μ:

$$\pi_t^\mu(y) = \frac{\pi_t(y)^{1-\eta_t\tau}\mu(y)^{\eta_t\tau}}{\sum_{y'}\pi_t(y')^{1-\eta_t\tau}\mu(y')^{\eta_t\tau}}$$

  with η_t the learning rate
- The Nash-MD algorithm is a step of mirror descent relative to the regularized policy π_t^μ:

$$\pi_{t+1} = \arg\max_\pi [\eta_t \mathcal{P}(\pi \succ \pi_t^\mu) - \mathsf{KL}(\pi, \pi_t^\mu)]$$

  which can be expressed explicitly with

$$\pi_{t+1}(y) \propto \pi_t^\mu(y)\exp(\eta_t \mathcal{P}(y \succ \pi_t^\mu))$$

  - **Intuition:**
    $\mathcal{P}(y \succ \pi_t^\mu) \to 1$ implies $\pi_{t+1}(y) \to \pi_t^\mu(y) + \epsilon$
    $\mathcal{P}(y \succ \pi_t^\mu) \to 0$ implies $\pi_{t+1}(y) \to \pi_t^\mu(y)$

Main Results

- **Proposition.** The regularized preference $\mathcal{P}_\tau$ has a unique Nash Equilibrium
  - Mappings $\pi \mapsto \mathcal{P}(\pi \succ \pi')$ and $\pi' \mapsto \mathcal{P}(\pi' \succ \pi)$ are linear with respect to π (and respectively π')
    - Implies $\pi \mapsto \mathcal{P}_\tau(\pi \succ \pi')$ is concave and $\pi' \mapsto \mathcal{P}_\tau(\pi' \succ \pi)$ convex
    - Implies existence of Nash via minimax Theorem from (Sion, 1958)
  - Uniqueness is argued via variational inequalities
- **Theorem.** Let π_τ^* be the Nash Equilibrium of the regularized preference model $\mathcal{P}_\tau(\pi \succ \pi') = \mathcal{P}(\pi \succ \pi') - \tau \mathsf{KL}_\rho(\pi, \mu) + \tau \mathsf{KL}_\rho(\pi', \mu)$. At every iteration t,

$$\mathsf{KL}(\pi_\tau^*, \pi_{t+1}) \leq (1 - \eta_t \tau)\mathsf{KL}(\pi_\tau^*, \pi_t) + 2\eta_t^2,$$

and consequently, a choice of $\eta_t = 2/(\tau(t+2))$ yields

$$\mathsf{KL}(\pi_\tau^*, \pi_T) \leq \frac{8}{\tau^2(T+1)}$$

Analysis of Nash-MD

- Using a Lemma from a previous Munos et al. paper, they obtain the bound

$$\mathsf{KL}(\pi, \pi_{t+1}) \leq \mathsf{KL}(\pi, \pi_t^\mu) + \eta_t \sum_y (\pi_t^\mu(y) - \pi(y)) \mathcal{P}(y \succ \pi_t^\mu) + 2\eta_t^2 \tag{1}$$

  for any policy π
- **Lemma.** For any π and $0 \leq \eta_t \tau \leq 1$,

$$\mathsf{KL}(\pi, \pi_t^\mu) \leq \eta_t \tau \mathsf{KL}(\pi, \mu) + (1 - \eta_t \tau) \mathsf{KL}(\pi, \pi_t) + \eta_t \tau \mathsf{KL}(\pi_t^\mu, \mu)$$

- Apply lemma to (1) with the choice $\pi = \pi_\tau^*$ and after some rewriting and simplifying via definitions, they achieve the bound

$$\mathsf{KL}(\pi_\tau^*, \pi_{t+1}) \leq (1 - \eta_t \tau) \mathsf{KL}(\pi_\tau^*, \pi_t) + 2\eta_t^2 \tag{2}$$

- Setting $\eta_t = 2/(\tau(t+2))$ and applying (2) at $t = 0$ gives

$$\mathsf{KL}(\pi_\tau^*, \pi_1) \leq \frac{2}{\tau^2}$$

Analysis of Nash-MD

- Observe that $\frac{2}{\tau^2} \leq \frac{8}{\tau^2(t+1)}$ for $t = 0$ so the bound we achieved previously becomes a base case

$$\mathsf{KL}(\pi_\tau^*, \pi_1) \leq \frac{2}{\tau^2} \leq \frac{8}{\tau^2(0+1)}$$

- Assuming $\mathsf{KL}(\pi_\tau^*, \pi_t) \leq \frac{8}{\tau^2(t+1)}$ and recalling we set $\eta_t = 2/(\tau(t+2))$, we apply the bound $\mathsf{KL}(\pi_\tau^*, \pi_{t+1}) \leq (1-\eta_t\tau)\mathsf{KL}(\pi_\tau^*, \pi_t) + 2\eta_t^2$ to get

$$\begin{aligned}
\mathsf{KL}(\pi_\tau^*, \pi_{t+1}) &\leq \left(1 - \frac{2}{t+2}\right)\frac{8}{\tau^2(t+1)} + \frac{8}{\tau^2(t+2)^2} \\
&\leq \left(1 - \frac{2}{t+2}\right)\frac{8}{\tau^2(t+1)} + \frac{8}{\tau^2(t+2)(t+1)} \\
&= \left(1 - \frac{2}{t+2} + \frac{1}{t+2}\right)\frac{8}{\tau^2(t+1)} \\
&= \frac{8}{\tau^2(t+2)}
\end{aligned}$$

which completes the inductive step

Deep Learning Implementation of NLHF

- Two practical gradient-based algorithms derived from Nash-MD:
  - **Nash-MD-PG:** Plays against geometric mixture of current policy and reference policy
    - Reference policy acts as regularizer
    - $\beta \in [0, 1]$ controls blend between current policy ($\beta = 0$ = Self-Play) and reference policy ($\beta = 1$ = Best-Response)
  - **Nash-EMA-PG:** Uses exponential moving average of past policy checkpoints as opponent
    - More adaptive, opponent evolves as policy improves
    - More challenging and informative
- Both use policy gradient updates
  - $\nabla_\theta \log \pi_\theta(y|x) \left(\mathcal{P}(y \succ y'|x) - 1/2 - \tau \log \frac{\pi_\theta(y|x)}{\mu(y|x)} \right)$
- Key advantage: No need to store past policies; only most recent checkpoint needed

Experiments: Text Summarization Task

- Evaluates NLHF on task of text summarization using TL;DR dataset
  - Goal is to summarize reddit posts
- Setup: fine-tuned TX5-L language model starting from supervised checkpoint
- Preference model trained on human preference data
  - Model generates summaries during training, preference model scores them pairwise
- Evaluation: Pairwise comparisons judged by TX5-XL
  - Head to head win rate
  - Tested multiple β values
- Models compared:
  - SFT (Supervised Fine-Tuned, baseline), RLHF, Self-Play (SP), Nash-MD-PG (MD1–MD6, $\beta \in \{0.125 \dots 0.75\}$), Best-Response (BR), Nash-EMA-PG (EMA1, EMA2)

Results: Table 1 (Pairwise Win Rates)

Nash Learning from Human Feedback

Table 1. PaLM 2 preference $\mathcal{P}^*(\pi_c \succ \pi_r)$ model between column policy π_c against row policy π_r.

$\mathcal{P}^*$	SFT	RLHF	SP	**MD1**	MD2	MD3	MD4	MD5	MD6	BR	EMA1	EMA2	EMA1*	EMA2*
SFT	0.500	0.990	0.983	**0.982**	0.989	0.987	0.985	0.982	0.965	0.943	0.970	0.961	0.977	0.980
RLHF	0.010	0.500	**0.489**	**0.598**	**0.519**	**0.561**	**0.501**	**0.436**	**0.284**	**0.148**	0.468	0.320	0.477	0.510
SP	0.017	0.511	0.500	**0.592**	0.504	0.545	0.499	0.451	0.310	0.211	0.445	0.362	0.464	0.488
MD1	0.018	0.402	0.408	**0.500**	0.425	0.470	0.369	0.362	0.238	0.163	0.391	0.270	0.400	0.447
MD2	0.011	0.481	0.496	**0.575**	0.500	0.513	0.491	0.434	0.298	0.196	0.460	0.351	0.430	0.496
MD3	0.013	0.439	0.455	**0.530**	0.487	0.500	0.484	0.408	0.273	0.187	0.429	0.323	0.413	0.472
MD4	0.015	0.499	0.501	**0.631**	0.509	0.516	0.500	0.428	0.265	0.161	0.468	0.358	0.437	0.503
MD5	0.018	0.564	0.549	**0.638**	0.566	0.592	0.572	0.500	0.329	0.210	0.532	0.389	0.518	0.539
MD6	0.035	0.716	0.690	**0.762**	0.702	0.727	0.735	0.671	0.500	0.342	0.652	0.548	0.651	0.691
BR	0.057	0.852	0.789	**0.837**	0.804	0.813	0.839	0.790	0.658	0.500	0.743	0.640	0.752	0.774
EMA1	0.030	0.532	0.555	**0.609**	0.540	0.571	0.532	0.468	0.348	0.257	0.500	0.381	0.480	0.556
EMA2	0.039	0.680	0.638	**0.730**	0.649	0.677	0.642	0.611	0.452	0.360	0.619	0.500	0.585	0.659
EMA1*	0.023	0.523	0.536	**0.600**	0.570	0.587	0.563	0.482	0.349	0.248	0.520	0.415	0.500	0.555
EMA2*	0.020	0.490	0.512	**0.553**	0.504	0.528	0.497	0.461	0.309	0.226	0.444	0.341	0.445	0.500

- Key findings:
  - Nash-MD-PG with $\beta \in \{0.125, 0.25, 0.375\}$ (MD1, MD2, MD3) outperforms **all** other methods
  - Extreme values $\beta = 0$ (Self-Play) and $\beta = 1$ (Best-Response) under-perform intermediate β
  - Both NLHF methods consistently outperform their counterparts

Conclusion

- NLHF is a compelling alternative to RLHF
  - Learns preference model instead of reward model
  - Finds Nash equilibrium of resulting two player game
  - Avoids Bradley-Terry assumptions
- Nash-MD optimizes through self improvement
  - β controls exploitation vs exploration tradeoff
  - Intermediate β values work best
- Possible Future Works
  - Explore different mixture strategies
  - Incorporate decaying coefficient for β

Thank You