A Markov Chain Approach to Preference Alignment

Takuya Koriyama Tengyuan Liang

July 4, 2026

The University of Chicago

Alignment is central to modern LLM fine-tuning

- Modern LLMs are first trained to predict the next token on large-scale text corpora.
- To make them useful assistants, we further fine-tune them to align with human preferences.
- A common source of supervision is *pairwise human preference*:

$$\text{response A} \prec \text{response B} \qquad \text{“B is preferred to A.”}$$

Preference alignment

- Let μ_{ref} be a reference distribution (e.g., pre-trained LLM)
- Given a dataset of pairwise human preference,

$$
\begin{aligned}
A &\prec B \\
B &\prec C \\
D &\prec B \\
E &\prec A \\
&\vdots
\end{aligned}
$$

we aim to fine-tune μ_{ref}.

Overview

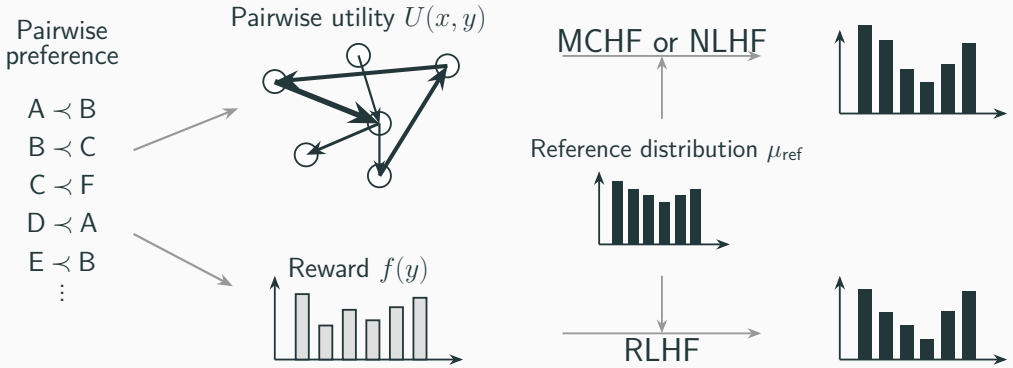
Existing approaches

- Reinforcement Learning from Human Feedback (RLHF) (e.g. Christiano et al. (2017))
- Nash Learning from Human Feedback (NLHF) (Munos et al. (2024))

Our proposal

- Markov Chain from Human Feedback (MCHF)

Overview


Pairwise preference
A ≺ B
B ≺ C
C ≺ F
D ≺ A
E ≺ B
⋮
Pairwise utility $U(x, y)$
Reward $f(y)$
MCHF or NLHF
Reference distribution μ_{ref}
RLHF

Reinforcement Learning from Human Feedback (RLHF)

Step 1: Reward modeling

- Let $\mathcal{P}(x \prec y)$ be the probability that a human prefers y to x.
- Assume BTL[1] model: $\exists$ reward function f such that

$$\mathcal{P}(x \prec y) = \text{sigmoid}(f(y) - f(x)).$$

- Given pairwise preference data $\{x_i \prec y_i\}_{i=1}^n$, we estimate f by the MLE $\hat{f}$:

[1] Bradley–Terry (1952) or Luce (1959) Model

Step 2: KL regularized distributional optimization

- Given a reward estimate $\hat{f}$, solve

$$\mu_{\mathsf{RL}} = \arg\max_{\mu} \int \mu(dy)\hat{f}(y) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}})$$

- The solution has an explicit formula:

$$\mu_{\mathsf{RL}}(dy) = \frac{\exp(\hat{f}(y))\mu_{\mathsf{ref}}(dy)}{\int \exp(\hat{f}(z))\mu_{\mathsf{ref}}(dz)} \propto \exp(\hat{f}(z))\mu_{\mathsf{ref}}(dz)$$

  but sampling from it directly is not feasible.

- In practice, μ is parameterized by a generative model μ_θ (e.g., diffusion model, transformer), and we optimize θ by gradient descent using RL algorithms.

Reward modeling enforces transitive preference

- Given a human preference $\mathcal{P}(\cdot \prec \cdot)$, define the average order $\prec_*$ as

$$x \prec_* y \quad \text{if and only if} \quad \mathcal{P}(x \prec y) > \frac{1}{2}$$

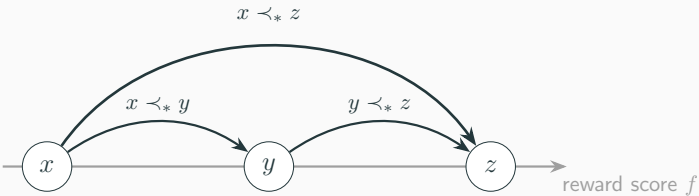
- Recall the BTL model assumes

$$\mathcal{P}(x \prec y) = \text{sigmoid}(f(y) - f(x)) = \frac{e^{f(y)}}{e^{f(y)} + e^{f(x)}},$$

so that

$$x \prec_* y \quad \Leftrightarrow \quad f(x) < f(y)$$

- Thus, the resulting average order $\prec_*$ must be transitive:



Non-transitive preference: Condorcet Paradox (1785)

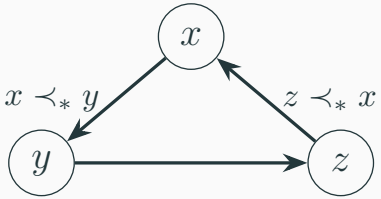
- Consider three outputs x, y, z and three annotators:

Annotator	Preference ranking
1	$x \prec y \prec z$
2	$y \prec z \prec x$
3	$z \prec x \prec y$

- Then,

$$\mathcal{P}(x \prec y) = \mathcal{P}(y \prec z) = \mathcal{P}(z \prec x) = \frac{2}{3} > \frac{1}{2}$$

- The resulting average order $\prec_*$ is non-transitive:



Step 1: Pairwise utility modeling

- Instead of the BTL model

$$\mathcal{P}(x \prec y) = \text{sigmoid}(-f(x) + f(y)),$$

we assume $\exists$ antisymmetric pairwise utility U such that

$$\mathcal{P}(x \prec y) = \text{sigmoid}(U(x, y))$$

and estimate U by the MLE

Nash Learning from Human Feedback (NLHF)

Step 2: Given an antisymmetric pairwise utility U, solve

$$\mu_{\mathsf{NL}} \in \arg\max_{\mu} \min_{\nu} F(\mu, \nu),$$

$$\text{where} \quad F(\mu, \nu) \equiv \iint U(x, y)\nu(dx)\mu(dy) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}}) + \mathrm{KL}(\nu \mid \mu_{\mathsf{ref}})$$

- Since U is antisymmetric, F is also antisymmetric: $F(\mu, \nu) = -F(\nu, \mu)$
- By standard minimax theorem, μ_{NL} is a unique solution satisfying

$$\mu_{\mathsf{NL}} = \arg\max_{\mu} F(\mu, \nu = \mu_{\mathsf{NL}})$$

- In practice, we solve μ_{NL} by iterations: $\mu^{t+1} = \arg\max_{\mu} F(\mu, \nu = \mu^t)$ with $\mu^0 = \mu_{\mathsf{ref}}$

Nash Learning from Human Feedback (NLHF)

$$\mu_{\mathsf{NL}} \in \arg\max_{\mu} \min_{\nu} \left\{ \iint U(x,y)\nu(dx)\mu(dy) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}}) + \mathrm{KL}(\nu \mid \mu_{\mathsf{ref}}) \right\}.$$

- NLHF formulates alignment as a simultaneous game; μ and ν compete at the same time
- However, alignment can be naturally defined as a non-simultaneous game:

$$\text{Given } x, \text{ choose } y = \arg\max_{y} U(x,y)$$

- However, we have to pick y such that the marginal distribution of y is close to μ_{ref}
- Motivated by these points, we propose a Markov chain approach

Our contribution 1: a Markov chain approach

- Given μ_{ref} and a pairwise utility $U(x, y)$, define the Markov kernel $\mathsf{P}(x, dy)$:

$$\mathsf{P}(x, dy) = \frac{1}{Z(x)} \exp(U(x, y))\mu_{\text{ref}}(dy)$$

  where $Z(x)$ is a normalizing constant s.t. $\int \mathsf{P}(x, dy) = 1$ for all x.

- Given a current distribution μ, we align it via conditional sampling

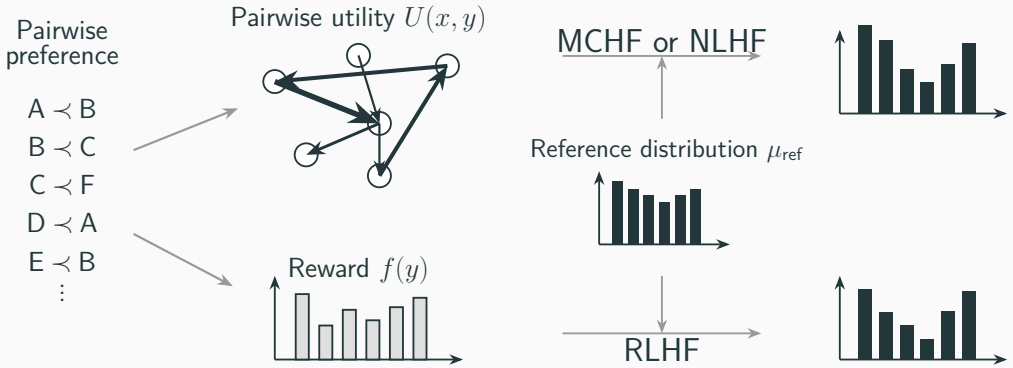
$$\mu_{\text{new}}(dy) = \mu\mathsf{P}(dy) = \int \mu(dx)\mathsf{P}(x, dy)$$

- The Markov kernel then induces an iterative alignment

$$\mu_{\text{ref}} \mapsto \mu_{\text{ref}}\mathsf{P} \mapsto \cdots \mu_{\text{ref}}\mathsf{P}^{\infty} \equiv \mu_{\text{MC}},$$

  where μ_{MC} is a stationary distribution of P (inspired by PageRank)

Meta comparison of MCHF, NLHF, RLHF



Meta comparison of NLHF and MCHF via implementation perspective

Consider the two algorithms:

$$\text{NLHF}: \quad \mu_{\mathsf{NL}}^{t+1} = \arg\max_{\mu} F(\mu, \nu = \mu_{\mathsf{NL}}^{t})$$
$$\text{MCHF}: \quad \mu_{\mathsf{MC}}^{t+1} = \mu_{\mathsf{MC}}^{t}\mathsf{P}$$

- NLHF repeatedly solves a RLHF solution with iterate-dependent reward

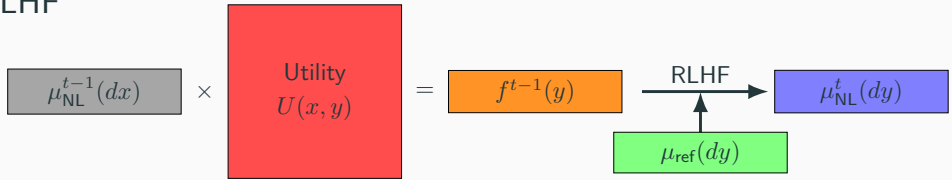
$$\mu_{\mathsf{NL}}^{t+1} = \arg\max_{\mu} \int f_{\mathsf{NL}}^{t}(y)\mu(dy) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}}), \quad f_{\mathsf{NL}}^{t}(y) = \int U(x,y)\mu_{\mathsf{NL}}^{t}(dx)$$

- MCHF repeatedly applies one fixed conditional sampler:

$$\mathsf{P}(x, dy) \propto \exp(U(x,y))\mu_{\mathsf{ref}}(dy)$$

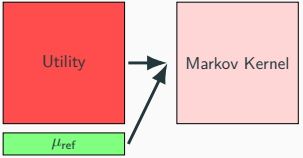
Meta comparison of NLHF and MCHF via implementation perspective

NLHF



MCHF

(Construct the Markov kernel at the beginning)



Our contributions 2: Unify MCHF, NLHF, and RLHF

Three methods are based on different premises:

- RLHF μ_{RL}: Reward-based approach
- NLHF μ_{NL}: Game-theoretic approach
- MCHF μ_{MC}: Markovian approach

Q: how can we compare these methods in a mathematically rigorous framework?

MCHF and RLHF reduce to RLHF when U is separable

Proposition

Let $\mu_{\mathsf{MC}}(U)$ and $\mu_{\mathsf{NL}}(U)$ be the MCHF/NLHF equilibria based on U. Then,

$$\Big(\exists f,\ U(x,y) = -f(x) + f(y)\Big) \quad \Rightarrow \quad \Big(\mu_{\mathsf{MC}}(U) = \mu_{\mathsf{NL}}(U) = \mu_{\mathsf{RL}}(f)\Big)$$

where μ_{RL} is the RLHF solution with reward f:

$$\mu_{\mathsf{RL}}(f)(dy) \propto \exp(f(y))\mu_{\mathsf{ref}}(dy)$$

- Recall

$$\mathcal{P}(x \prec y) = \mathsf{sigmoid}(U(x,y))$$

so $U(x,y) = -f(x) + f(y)$ means we impose BTL assumption.

Our contributions 2: Unify MCHF, NLHF, and RLHF

- Q. What if $U(x,y) \approx -f(x) + f(y)$ (but not necessarily in equality)?
- We address this question by introducing a seminorm

$$\|U\|_{\oplus} = \inf_{g,f \in L^{\infty}(\mu_{\mathsf{ref}})} \|U - g \oplus f\|_{\infty}, \quad \text{where} \quad (g \oplus f)(x,y) = g(x) + f(y)$$

- When $\|U\|_{\oplus}$ is small, there exists a common correction term such that

$$\text{for } * \in \{\mathsf{MC}, \mathsf{NL}\}, \quad \mu_*(U) \approx \mu_{\mathsf{RL}}(\hat{f}) + \text{Correction},$$

where

$$\hat{f}(y) = \int \mu_{\mathsf{ref}}(dx) U(x,y), \quad \text{Correction} = \text{a Linear functional of } U - (-\hat{f}) \oplus \hat{f}$$

Talk roadmap

1. Definition of seminorm $\|U\|_{\oplus}$ and some basic properties
2. Refined convergence analysis for MCHF and NLHF
   The smaller $\|U\|_{\oplus}$, the faster MCHF and NLHF iterations converge to their own equilibria
3. Asymptotic analysis of MCHF and NLHF when $\|U\|_{\oplus} \to 0$.

Definition of seminorm: measuring non-transitive structure of pairwise preference

Notation

- Fix a reference measure μ_{ref}
- Assume that the pairwise utility U encodes human preference as

$$U(x, y) > 0 \Leftrightarrow \text{Humans prefer } y \text{ to } x$$

- Throughout this talk, U is given, and we assume $U \in L^\infty(\mu_{\text{ref}} \otimes \mu_{\text{ref}})$.

Definition of the seminorm $\|\cdot\|_{\oplus}$

Definition

Define the seminorm $\|\cdot\|_{\oplus}$ on $L^{\infty}(\mu_{\mathsf{ref}} \otimes \mu_{\mathsf{ref}})$ as

$$\|U\|_{\oplus} = \inf_{g,f \in L^{\infty}(\mu_{\mathsf{ref}})} \|U - g \oplus f\|_{\infty} \quad \text{where} \quad (g \oplus f)(x,y) = g(x) + f(y)$$

- $\|U\|_{\oplus} \leq \|U\|_{\infty}$ by the definition

Basic properties of the seminorm

Proposition

Suppose U is antisymmetric. Then,

$$\|U\|_{\oplus} \asymp \|\Delta U\|_{\infty},$$

where

$$\Delta U = U - (-\hat{f}) \oplus \hat{f}, \quad \hat{f}(y) = \int \mu_{\mathsf{ref}}(dx) U(x, y).$$

Furthermore,

$$\|U\|_{\oplus} \asymp \operatorname*{ess\,sup}_{x,y,z} |U(x, y) + U(y, z) + U(z, x)|$$

- $A \asymp B$ means $A \leq c_1 \cdot B$ and $c_2 \cdot A \leq B$ for absolute positive constants c_1, c_2.
- $\operatorname{ess\,sup}_{x,y,z}$ is the essential supremum with respect to $\mu_{\mathsf{ref}} \otimes \mu_{\mathsf{ref}} \otimes \mu_{\mathsf{ref}}$.

Refined convergence analysis for MCHF and NLHF

Refined convergence analysis for MCHF

Theorem

Let P be the Markov kernel based on U:

$$\mathsf{P}(x, dy) = \frac{1}{Z(x)} \exp(U(x, y)) \mu_{\mathsf{ref}}(dy)$$

Then, there exists a unique stationary distribution

$$\mu_{\mathsf{MC}} = \mu_{\mathsf{MC}} \mathsf{P}$$

and

$$\forall t \geq 1, \quad d_{\mathsf{TV}}(\mu_{\mathsf{ref}} \mathsf{P}^t, \mu_{\mathsf{MC}}) \leq c(\|U\|_{\oplus})^t \cdot d_{\mathsf{TV}}(\mu_{\mathsf{ref}}, \mu_{\mathsf{MC}})$$

where

$$c(r) = \min(r, 1 - e^{-2r}) \in [0, 1).$$

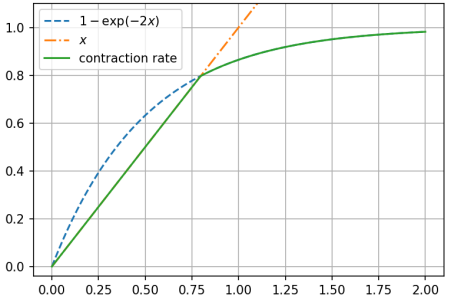
Refined convergence analysis for MCHF

$$\forall t \geq 1, \quad d_{\mathsf{TV}}(\mu_{\mathsf{ref}}\mathsf{P}^t, \mu_{\mathsf{MC}}) \leq c(\|U\|_{\oplus})^t \cdot d_{\mathsf{TV}}(\mu_{\mathsf{ref}}, \mu_{\mathsf{MC}})$$

where

$$c(r) = \min(r, 1 - e^{-2r}) \in [0, 1).$$

- $c(r)$ is increasing in $r \Rightarrow$ the smaller $\|U\|_{\oplus}$ is, the faster the Markov chain converges



Refined convergence analysis for NLHF

Theorem

Consider the mirror-descent iteration

$$\nu_t = \arg\min_{\nu} \left\{ \iint U(x,y)\nu(dx)\mu_t(dy) + \mathrm{KL}(\nu \mid \mu_{\mathsf{ref}}) \right\},$$
$$\mu_{t+1} = \arg\max_{\mu} \left\{ \iint U(x,y)\nu_t(dx)\mu(dy) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}}) - \eta^{-1}\mathrm{KL}(\mu \mid \mu_t) \right\}.$$

If $0 < \eta \leq \|U\|_{\oplus}^{-2}$, then

$$\mathrm{KL}(\mu_{\mathsf{NL}} \mid \mu_t) \leq \frac{1}{(1+\eta)^t}\mathrm{KL}(\mu_{\mathsf{NL}} \mid \mu_{\mathsf{ref}}).$$

- Previous works (e.g. Munos et al. (2024)) allow η at most $\eta = O(\|U\|_{\infty}^{-2})$,
- Since $\|U\|_{\oplus} \leq \|U\|_{\infty}$, our work provides refined convergence analysis.

Small-$\|U\|_{\oplus}$ asymptotics

Small-$\|U\|_\oplus$ perturbation: equilibrium level

Theorem

For each U, let p_{MC} and p_{NL} be the densities of MCHF and NLHF based on U,

$$p_*(\cdot) = \frac{d\mu_*(U)}{d\mu_{\mathsf{ref}}}(\cdot) \quad \text{for} \quad * \in \{\mathsf{MC}, \mathsf{NL}\},$$

and let p_{RL} be the density of RLHF based on the reward $\hat{f}$,

$$p_{\mathsf{RL}}(\cdot) = \frac{\exp(\hat{f}(\cdot))}{\int \mu_{\mathsf{ref}}(dz) \exp(\hat{f}(z))}, \quad \text{where} \quad \hat{f}(y) = \int U(x,y)\mu_{\mathsf{ref}}(dx)$$

and let $\Delta U = U - (-\hat{f}) \oplus \hat{f}$. Then,

$$\text{for } * \in \{\mathsf{MC}, \mathsf{NL}\}, \quad \|p_* - p_{\mathsf{RL}} - p_{\mathsf{RL}} \odot \Delta U^\star p_{\mathsf{RL}}\|_1 = o(\|U\|_\oplus)$$

where $\odot$ is the entrywise product and $\Delta U^\star p(\cdot) = \int \mu_{\mathsf{ref}}(dx)\Delta U(x,\cdot)p(x)$.

Small-$\|U\|_\oplus$ perturbation: equilibrium level

$$\text{for } * \in \{\mathsf{MC}, \mathsf{NL}\}, \quad \|p_* - p_{\mathsf{RL}} - p_{\mathsf{RL}} \odot \Delta U^\star p_{\mathsf{RL}}\|_1 = o(\|U\|_\oplus)$$

- MCHF and NLHF have the same correction term. Thus, by the triangle inequality,

$$\|p_{\mathsf{MC}} - p_{\mathsf{NL}}\|_1 = o(\|U\|_\oplus)$$

- MCHF and NLHF both incorporate the non-transitive component of U:

$$\Delta U = U - (-\hat{f}) \oplus \hat{f}$$

on top of the RLHF density p_{RL} based on the reward $\hat{f}(y) = \int U(x, y)\mu_{\mathsf{ref}}(dx)$

Theorem

Let p^t_{MC} and p^t_{NL} be the densities generated by the algorithms for MCHF and NLHF defined by:

$$\mu^t_{\mathsf{MC}} = \mu^{t-1}_{\mathsf{MC}}\mathsf{P}, \qquad \mu^t_{\mathsf{NL}} = \arg\max_{\mu} \left\{ \iint U(x,y)\mu^{t-1}_{\mathsf{NL}}(dx)\mu(dy) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}}) \right\}$$

initialized at $\mu^0_{\mathsf{MC}} = \mu^0_{\mathsf{NL}} = \mu_{\mathsf{ref}}$. Then, for any antisymmetric U such that $\|U\|_\oplus \to 0$,

$$\left\|p^1_* - p_{\mathsf{RL}}\right\|_1 = o(\|U\|_\oplus),$$
$$\sup_{t\geq 2} \left\|p^t_* - p_{\mathsf{RL}} - p_{\mathsf{RL}} \odot \Delta U^\star p_{\mathsf{RL}}\right\|_1 = o(\|U\|_\oplus).$$

Small-$\|U\|_{\oplus}$ perturbation: iteration level

for $* \in \{\mathsf{MC}, \mathsf{NL}\}$,

$$\left\|p_*^1 - p_{\mathsf{RL}}\right\|_1 = o(\|U\|_{\oplus}),$$
$$\sup_{t \geq 2} \left\|p_*^t - p_{\mathsf{RL}} - p_{\mathsf{RL}} \odot \Delta U^\star p_{\mathsf{RL}}\right\|_1 = o(\|U\|_{\oplus}).$$

- **First step:** both MCHF and NLHF learn the RLHF solution with reward

$$\hat{f}(y) = \int U(x, y) \mu_{\mathsf{ref}}(dx).$$

- **Second step onward:** both incorporate the non-transitive residual

$$\Delta U = U - (-\hat{f}) \oplus \hat{f}.$$

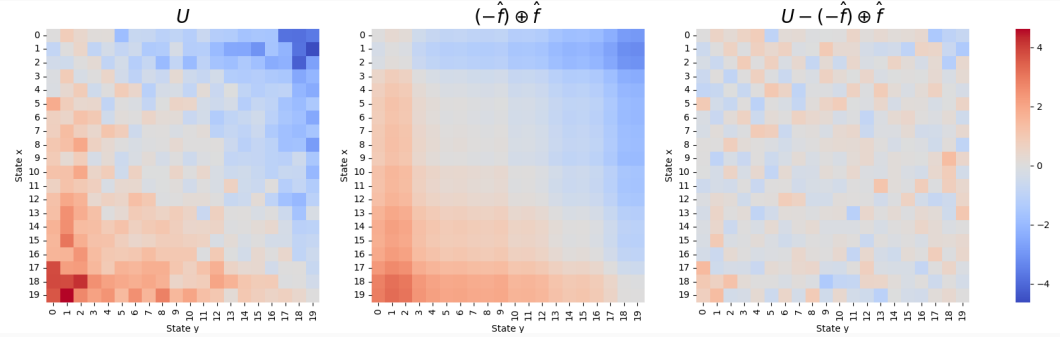
- **Practical implication:** when $\|U\|_{\oplus}$ is small, the main first-order benefit of iterative alignment appears within the first two steps.

Numerical illustration

Set $\mu_{\text{ref}} = \text{Uniform}[n]$ and take U as $U(x, y) = \text{sigmoid}^{-1}(\mathcal{P}(x \prec y))$, where

$$\mathcal{P}(x \prec y) = \Phi\Big(R(y) - R(x) + E(x, y) - E(y, x)\Big),$$

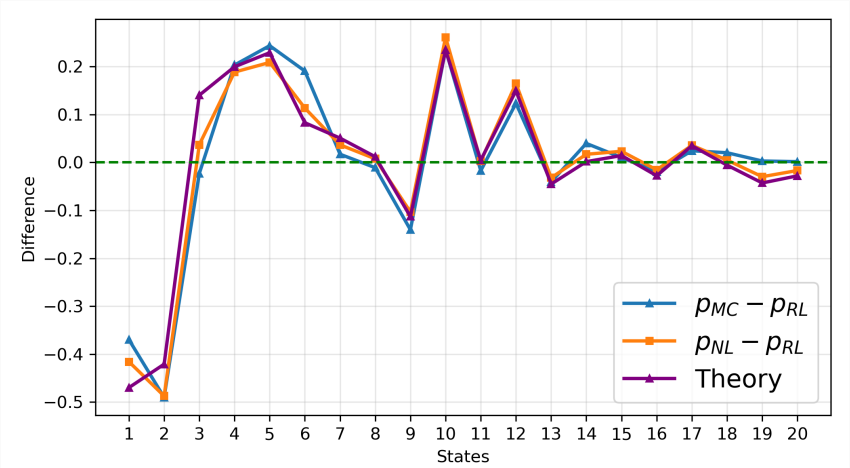
where Φ is the CDF of $N(0, 1)$, $R \sim N(0, I_n)$, and $E \sim N(0, I_{n\times n})$.



Numerical simulation

Based on the previous U, verify

$$\forall * \in \{\mathsf{MC}, \mathsf{NL}\}, \quad \|p_* - p_{\mathsf{RL}} - p_{\mathsf{RL}} \odot \Delta U^\star p_{\mathsf{RL}}\|_1 = o(\|U\|_\oplus)$$



Conclusion

Takeaways and future work

1. MCHF turns a pairwise utility into a preference-guided Markov kernel.
2. MCHF and NLHF iterations both converge geometrically to their own equilibria fast when $\|U\|_{\oplus}$ is small.
3. In the small-$\|U\|_{\oplus}$ asymptotics, MCHF and NLHF share the same first-order expansion around RLHF.

Future work

1. Practical implementation of the conditional sampler $\mathrm{P}(x, dy)$
2. Calculate $\|U\|_{\oplus}$ in real data

Paper: *A Markov Chain Approach to Preference Alignment*

arXiv:2606.22652

References

- Munos et al. (2024). Nash Learning from Human Feedback.
- Christiano et al. (2017). Deep reinforcement learning from human preferences.

Backup slides

Benefit of MCHF over NLHF: inference time refinement

Consider the Condorcet Paradox example

Annotator	Preference ranking
1	$x \prec y \prec z$
2	$y \prec z \prec x$
3	$z \prec x \prec y$

and construct U as the logit transform of average preference. Then, U can be written as

$$U = c \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

for a constant $c > 0$.

Q. Fix a distribution μ and suppose each annotator samples outputs from μ until they get their best output. Then, how should we choose μ from U so that the expected time to get the best output is small?

Benefit of MCHF over NLHF: inference time refinement

Annotator	Preference ranking
1	$x \prec y \prec z$
2	$y \prec z \prec x$
3	$z \prec x \prec y$

$$\Longrightarrow U = c \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

- Suppose $\mu_{\text{ref}} = \text{Uniform}\{x, y, z\}$. Then $\mu_{\text{NL}} = \mu_{\text{ref}}$ and hence the hitting time to the best output for each annotator is 3.
- For a fixed $\gamma > 0$, consider the Markov chain

$$\mu_{\text{ref}} \to \mu_{\text{ref}}\mathsf{P}_\gamma \to \mu_{\text{ref}}\mathsf{P}_\gamma^2 \cdots \quad \text{where} \quad \mathsf{P}_\gamma(x, dy) \propto \exp(\gamma U(x, y))\mu_{\text{ref}}(dy).$$

  Then, one can show that the hitting time is a strictly decreasing function of $\gamma > 0$.
- We extend this phenomenon to general μ_{ref} and consider the hitting time to a general set A, showing that we can reduce the hitting time if and only if there is enough *energy* $U(x, y)^2$ from A^c to A.

MCHF reduces to RLHF when U takes an additive form

Proof

- Recall that μ_{MC} is the stationary distribution of the Markov kernel:

$$\mathsf{P}(x, dy) = \frac{\exp(U(x, y))\mu_{\mathsf{ref}}(dy)}{\int \exp(U(x, z))\mu_{\mathsf{ref}}(dz)}.$$

- If $U(x, y) = -f(x) + f(y)$,

$$\mathsf{P}(x, dy) = \frac{\exp(f(y))\mu_{\mathsf{ref}}(dy)}{\int \exp(f(z))\mu_{\mathsf{ref}}(dz)} = \mu_{\mathsf{RL}}(f)(dy)$$

- Thus, the stationary distribution μ_{MC} is $\mu_{\mathsf{RL}}(f)$.

NLHF reduces to RLHF when U takes an additive form

Proof

- Recall

$$\mu_{\mathsf{NL}} \in \arg\max_{\mu} \min_{\nu} F(\mu, \nu), \quad F(\mu, \nu) \equiv \iint U(x, y)\nu(dx)\mu(dy) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}}) + \mathrm{KL}(\nu \mid \mu_{\mathsf{ref}})$$

- If $U(x, y) = -f(x) + f(y)$,

$$\begin{aligned} F(\mu, \nu) &= -\int f(x)\nu(dx) + \int f(y)\mu(dy) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}}) + \mathrm{KL}(\nu \mid \mu_{\mathsf{ref}}) \\ &= \left(\int f(y)\mu(dy) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}})\right) - \left(\int f(x)\nu(dx) - \mathrm{KL}(\nu \mid \mu_{\mathsf{ref}})\right) \end{aligned}$$

- Thus,

$$\mu_{\mathsf{NL}} = \arg\max_{\mu} \int f(y)\mu(dy) - \mathrm{KL}(\mu|\mu_{\mathsf{ref}}) = \mu_{\mathsf{RL}}(f)$$

When local regularization is removed

Theorem

With $\eta = +\infty$, the NLHF updates satisfy

$$d_{\mathsf{TV}}(\nu_t, \nu_{\mathsf{NL}}) \le \|U\|_{\oplus} d_{\mathsf{TV}}(\mu_t, \mu_{\mathsf{NL}}), \qquad d_{\mathsf{TV}}(\mu_{t+1}, \mu_{\mathsf{NL}}) \le \|U\|_{\oplus} d_{\mathsf{TV}}(\nu_t, \nu_{\mathsf{NL}}).$$

Therefore, if $\|U\|_{\oplus} < 1$,

$$d_{\mathsf{TV}}(\mu_t, \mu_{\mathsf{NL}}) \le \|U\|_{\oplus}^{2t} d_{\mathsf{TV}}(\mu_{\mathsf{ref}}, \mu_{\mathsf{NL}}).$$

Proof idea for MCHF contraction

1. **Minorization / coupling:**

$$\exists \text{ prob. measure } \mu_*, \exists \delta > 0 \quad \mathsf{P}(x, dy) \geq \delta \cdot \mu_*(dy)$$
$$\Rightarrow \quad d_{\mathsf{TV}}(\mu\mathsf{P}, \nu\mathsf{P}) \leq (1-\delta) d_{\mathsf{TV}}(\mu, \nu).$$

2. **Dobrushin coefficient:**

$$\sup_{\mu \neq \nu} \frac{d_{\mathsf{TV}}(\mu\mathsf{P}, \nu\mathsf{P})}{d_{\mathsf{TV}}(\mu, \nu)} = \operatorname*{ess\,sup}_{x, x'} d_{\mathsf{TV}}(\mathsf{P}(x, \cdot), \mathsf{P}(x', \cdot)).$$

Backup: stability under utility perturbations

Let $B_{\oplus}^{\infty}(R) = \{U : \|U\|_{\oplus} \leq R\}$. For $* \in \{\mathsf{MC}, \mathsf{NL}\}$, there exists C_R such that

$$\sup_{U, \widehat{U} \in B_{\oplus}^{\infty}(R)} \frac{d_{\mathsf{TV}}(\mu_*(U), \mu_*(\widehat{U}))}{\|U - \widehat{U}\|_{\infty}} \leq C_R.$$

- Robustness matters because the utility is estimated from finite preference data.
- The map $U \mapsto \mu_*(U)$ is stable on bounded $\|\cdot\|_{\oplus}$ balls.

Backup: uniform Fréchet differentiability

Let $p_*(U) = d\mu_*(U)/d\mu_{\mathsf{ref}}$. For $* \in \{\mathsf{MC}, \mathsf{NL}\}$,

$$\lim_{\|E\|_\infty \to 0} \sup_{\|U\|_\oplus \leq R} \frac{\|p_*(U+E) - p_*(U) - Dp_*(U)[E]\|_1}{\|E\|_\infty} = 0.$$

- This is the technical engine behind the first-order comparison.
- At additive antisymmetric utilities, the MCHF and NLHF derivatives coincide.

Backup: derivative at additive antisymmetric utilities

For $U = -f \oplus f$ and antisymmetric perturbation E,

$$Dp_*(-f \oplus f)[E] = p_{\mathsf{RL}}(f) \odot E^\star p_{\mathsf{RL}}(f), \qquad * \in \{\mathsf{MC}, \mathsf{NL}\}.$$

Here

$$p_{\mathsf{RL}}(f)(y) = \frac{\exp(f(y))}{\int \exp(f(y'))\mu_{\mathsf{ref}}(dy')}, \qquad E^\star h(\cdot) = \int E(x, \cdot)h(x)\mu_{\mathsf{ref}}(dx).$$

Coupling view: MCHF optimizes over a larger class

Suppose μ^0 is a current distribution, and we update it via the following two different rules:

$$\mu^1_{\mathsf{MC}} = \mu^0 \mathsf{P}, \qquad \mu^1_{\mathsf{NL}} \in \arg\max_{\mu} \int \mu^0(dx)\mu(dy)U(x,y) - \mathrm{KL}(\mu \mid \mu_{\mathsf{ref}})$$

Fix a current distribution μ^0 and consider the couplings induced by MCHF and NLHF

$$\pi_{\mathsf{MC}}(dx,dy) = \mu^0(dx)\mathsf{P}(x,dy), \qquad \pi_{\mathsf{NL}}(dx,dy) = \mu^0(dx)\mu^1_{\mathsf{NL}}(dy).$$

Coupling view: MCHF optimizes over a larger class

$$\pi_{\mathsf{MC}}(dx, dy) = \mu^0(dx)\mathsf{P}(x, dy), \qquad \pi_{\mathsf{NL}}(dx, dy) = \mu^0(dx)\mu^1_{\mathsf{NL}}(dy).$$

The two couplings π_{MC} and π_{NL} maximize the same objective function:

$$F(\pi) = \iint U(x, y)\pi(dx, dy) - \mathrm{KL}(\pi||\mu^0 \otimes \mu_{\mathsf{ref}}),$$

but over different feasible spaces of couplings:

$$\pi_{\mathsf{MC}} = \underset{\int \pi(\cdot, dy) = \mu^0(\cdot)}{\arg\max} F(\pi), \quad \pi_{\mathsf{NL}} = \underset{\exists \mu, \pi = \mu^0 \otimes \mu}{\arg\max} F(\pi).$$

General hitting-time analysis

For $\gamma \geq 0$, define

$$\mathsf{P}_\gamma(x, dy) = \frac{\exp(\gamma U(x,y))\mu_{\mathsf{ref}}(dy)}{\int \exp(\gamma U(x,z))\mu_{\mathsf{ref}}(dz)}.$$

For a target set A, let $T_\gamma(A)$ be the expected time to hit A starting from μ_{ref}. At $\gamma = 0$,

$$T_0(A) = \frac{1}{\mu_{\mathsf{ref}}(A)}.$$

Theorem: First derivative

For antisymmetric U,

$$\dot{T}_0(A) = -\frac{\mu_{\mathsf{ref}}(A^c)}{\mu_{\mathsf{ref}}(A)} \int s(x,A)\mu_{\mathsf{ref}}(dx), \qquad s(x,A) = \frac{1}{\mu_{\mathsf{ref}}(A)} \int_A U(x,y)\mu_{\mathsf{ref}}(dy).$$

Second-order refinement when the first-order effect vanishes

If $\int U(x,y)\mu_{\mathsf{ref}}(dy) = 0$ for all x, then $\dot{T}_0(A) = 0$ and

$$\ddot{T}_0(A) = \frac{1}{\mu_{\mathsf{ref}}(A)} \int_{A^c} \left\{ e(x,\mathcal{X}) - e(x,A) - 2\mu_{\mathsf{ref}}(A) s(x,A)^2 \right\} \mu_{\mathsf{ref}}(dx),$$

where

$$e(x,B) = \frac{1}{\mu_{\mathsf{ref}}(B)} \int_B U(x,y)^2 \mu_{\mathsf{ref}}(dy).$$

Consequence
If enough preference energy points from A^c to A, then $\ddot{T}_0(A) < 0$, so small positive γ reduces hitting time.