ICML International Conference On Machine Learning

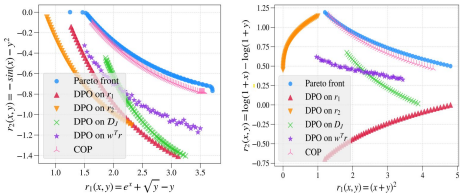
Google DeepMind

Multi-Objective Preference Optimization

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, Zheng Wen

Motivation

- Why multi-objective alignment? LLMs must satisfy multiple human objectives simultaneously: helpful, harmless, engaging, faithful, and so on.
- Existing preference optimization methods (DPO, IPO, RLHF) optimize **one** scalar objective or require **scalarizing** multiple rewards, often missing important trade-offs.
- **Goal**: Learn a policy that balances multiple objectives *without* requiring users to specify preference weights.

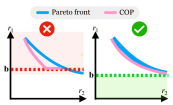


Problem Formulation

- What if we reformulate the problem as a constrained optimization problem?

$$\pi_{\text{COP}}(x) = \text{argmax}_y\, r_1(x,y) \;\; \text{s.t.} \;\; r_2(x,y) \geq b$$

- **Key insight**: Instead of scalarizing rewards, **MOPO** maximizes the *primary* objective preference while maintaining minimum performance on *secondary* objectives, where varying constraint thresholds traces the Pareto front.



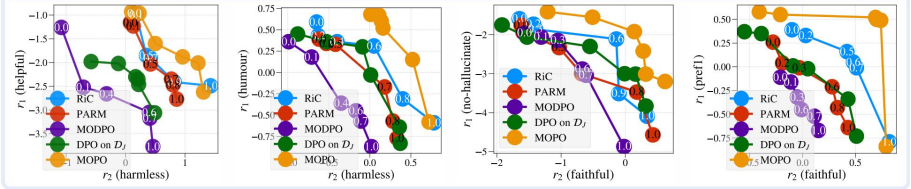
Experiments [Empirical results of multi-objective optimization of LLMs]

- LLMs considered: Phi-1.5, OpenChat-3.5, Llama-3.1-8B, Mistral-7B, Zephyr-7B.
- Evaluation across Helpful Assistant and Reddit Summarization tasks.
- **Goal**: Align LLMs to recover the empirical Pareto front in reward space. Vary the constraint thresholds b to find points on the Pareto front.

Takeaways

- **MOPO** consistently Pareto dominates baselines.
- **MOPO** Improves multiple objectives simultaneously, without probing users for preference weights.
- **MOPO** is stable across different model scales and scalable to more than two objectives as well.

RiC vs PARM vs MODPO vs DPO on $\mathcal{D}_J$ vs MOPO

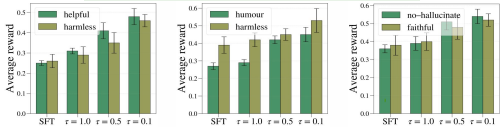


3 objective optimization

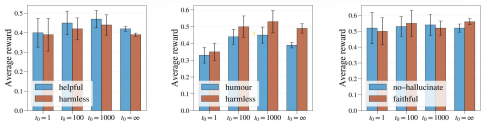
	helpful	humour	harmless
RLHF-r1	0.76	-0.42	-0.23
RLHF-r2	-0.81	0.53	-0.40
RLHF-r3	-0.79	-0.92	0.42
RiC	0.25	0.15	0.11
PARM	0.31	0.17	0.23
MODPO	0.04	-0.09	0.08
DPO on $\mathcal{D}_J$	0.18	0.09	0.11
MOPO-LB	0.30	0.19	0.18
MOPO-Lag	0.39	0.22	0.17

Ablations [Hyperparameter stability and regularization]

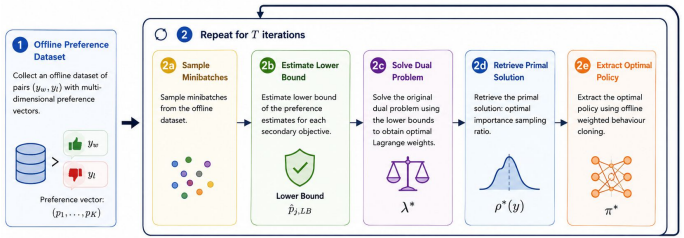
Effect of regularizer τ



Effect of lagged policy updates by t_0



Algorithm



Why Lower Bound Estimation?

Naively constraining preferences can violate constraints due to offline dataset estimation errors.

$$r_2(x,y) \geq b$$

$$\text{LowerBound}(r_2(x,y)) \geq b$$

Summary

- We introduced **MOPO**, a constrained optimization framework for multi-objective preference alignment.
- **MOPO** optimizes a *primary* objective and enforces lower-bound constraints on *secondary* objectives.
- The resulting algorithm is fully offline, preference-only, and scalable to 8b-parameter LLMs as well.
- Across synthetic and real-world benchmarks, **MOPO** consistently achieves better Pareto-optimal trade-offs than existing methods.