Statement of Purpose (SoP)

DSL501: Machine Learning Project

Team Name: He He!! We cook real stuff

August 23, 2025

1. Project Details

Project Title: LEARNING FROM NEGATIVE FEEDBACK, OR POSITIVE FEEDBACK OR BOTH

2. Problem Statement

Preference-based learning, where feedback can be positive, negative, or both. Generally, data sets contain only positive or negative samples, and paired preference data is rarely available. To address this problem, our project aims to design methods that can learn effectively from only positive data, only negative data, and mixed feedback signals. This enables broader applicability in settings such as reinforcement learning, offline RL, and LLM alignment, where structured paired data is not always feasible.

3. Methodology

- **Implement and compare preference optimization algorithms:** Preference-Based MPO (PMPO), Direct Preference Optimization (DPO), and Maximum a Posteriori Policy Optimization (MPO).
- **Explore key regularizers and divergences:** KL divergence, α-divergence, entropy regularization, and hybrid α–KL objectives.
- **Derive and test sample-based EM-style optimization** for both positive and negative feedback with the help of regularizers, following the RL-as-inference framework.
- **Investigate robustness** to unpaired preference feedback (positive-only, negative-only) and multi-sample preference settings.
- **Evaluate stability** under different KL weights (β) when learning from negative feedback alone.

4. Dataset Details

- **Primary dataset:** LMSYS-chat-1M, containing large-scale LLM conversations with prompts and responses.
- **Augmented with Bandit RL benchmarks** (Rosenbrock, Sphere, Schwefel functions) for synthetic preference testing.
- **Construct unpaired feedback datasets** using LLMs to simulate positive-only, negative-only, or mixed signals for fine-tuning experiments.

- **Preprocessing:** filter non-English prompts, generate positive/negative preference labels via human annotation or reward models. Human validation of the generated dataset.

5. Required Resources

- **Hardware:** 32GB GPU (NVIDIA A100 or equivalent) to support large-scale LLM fine-tuning.
- **Software:** PyTorch/TensorFlow, Hugging Face Transformers, LangChain, Qwen, SciPy.
- **APIs and libraries:** Reinforcement learning frameworks for Bandit, preference labeling pipelines, GPT-4 for A/B win-rate evaluations.
- **Benchmark tools:** Stable Baselines3, evaluation metrics for KL divergence and α-divergence.

6. Novelty of Approach

- **Extend beyond standard KL divergence** by incorporating α-divergence, mixtures of α and KL, and entropy-regularized objectives.
- **Explore the flexibility of PMPO** in handling positive-only, negative-only, and mixed datasets, unlike DPO/IPO which require paired feedback.
- **Evaluate robustness** against failure modes such as reward hacking in language alignment experiments.
- **Benchmark across diverse domains:** synthetic Bandit optimization, offline RL stacking tasks, and LLM alignment.
- **Empirically study the sensitivity of hyperparameters** (α, β) and provide tuning guidelines for preference-based optimization.

7. Team Composition and Individual Contributions

- **Member 1:** Ajay Chikate, 12340580 – Implementation of PMPO and divergence-based training experiments.
- **Member 2:** Farhan Alam, 12340740 – Dataset preprocessing, implementation of EM and Bandit RL
- **Member 3:** Pobitro Bhattacharya, 12341580 – Evaluation metrics, benchmarking against DPO/IPO, and analysis.
- **Member 4:** Sidhesh Kumar Patra, 12342060 – LLM fine-tuning, mathematical formulation, exploring different regularizers and divergences.

8. Expected Outcomes

- Replicating results from the PMPO paper across synthetic benchmarks, LMSYS dataset and LLM alignment tasks.
- Demonstrate stable learning from negative feedback alone using KL and other regularizations.
- Establish empirical comparisons between PMPO, DPO, and MPO across multiple domains.
- Provide insights into hyperparameter sensitivity (role of α and β).

- Deliverables: trained models, detailed evaluation report, reproducible code repository, and possible extension towards real-world RLHF datasets.

9. References

- Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. https://arxiv.org/pdf/1312.6114

- Son, K., Lee, J., Park, S. and Lee, S., 2007, August. Reinforcement Learning Using Negative Relevance Feedback. In Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007) (pp. 559-563). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4460701

- Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S. and Finn, C., 2023. Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems, 36, pp.53728-53741. https://arxiv.org/pdf/2305.18290

- Abdolmaleki, A., Springenberg, J.T., Tassa, Y., Munos, R., Heess, N. and Riedmiller, M., 2018. Maximum a posteriori policy optimisation. arXiv preprint arXiv:1806.06920. https://arxiv.org/abs/1806.06920