CBAI Dylan Hadfield Menell Proposal

Ethan Chen

Intro

The failure mode I'll focus on is heterogeneous risk preferences, i.e. annotators with different risk tolerances systematically disagree about alternatives involving uncertainty. This is directly motivated by Example 1.1 in the DPL paper (students evaluating financial aid with different risk tolerances), but is never experimentally validated. The jailbreak experiment tests different labeling objectives, not different utility functions over outcomes.

When risk-seeking annotators are the majority, Borda count aggregation will favor risky alternatives, harming the risk-averse minority. DPL should detect this disagreement as high variance at risky alternatives, enabling risk-averse optimization to protect the minority.

Design

I construct a synthetic environment where hidden context z represents risk tolerance. Two annotator types:

- **Risk-seeking** ($z = 0$): convex utility $u_0(o) = o^2$
- **Risk-averse** ($z = 1$): concave utility $u_1(o) = \sqrt{o}$

with risk-seekers as majority. The alternative space contains options that induce preference reversal:

- **Safe**: deterministic outcome $o = 5$
- **Risky**: outcomes $\{10, 1\}$ with equal probability

Risk-seekers prefer risky ($U_0 = 50.5$ vs 25) while risk-averse prefer safe ($U_1 = 2.24$ vs 2.08). Pairwise comparisons are generated via BTL, sampling z per comparison and withholding it from the learner.

Compare standard RLHF (BTL reward model) against mean-variance DPL with $\beta \in \{0, 0.5, 1, 2\}$. Architecture, loss, and training regime follow Section 4 of the original paper, with the only modification being the data generation process.

Metrics

- **Risky selection rate**: how often the learned model prefers the risky alternative
- **Minority regret**: $\max_a U_1(a) - U_1(\hat{a})$, measuring harm to risk-averse users
- **Variance ratio**: $\hat{\sigma}(\text{risky})/\hat{\sigma}(\text{safe})$, validating that DPL detects disagreement

Expected Results

Standard RLHF should favor the risky alternative due to majority preference, with high minority regret. If DPL helps, it learns $\hat{\sigma}(\text{risky}) > \hat{\sigma}(\text{safe})$, and increasing β shifts selection toward safer alternatives, reducing minority regret. This would validate that DPL's variance estimation captures preference heterogeneity from different utility functions, which is the scenario motivating Example 1.1 in the original paper.