Google DeepMind

Learning from Negative Feedback, or Positive Feedback or Both

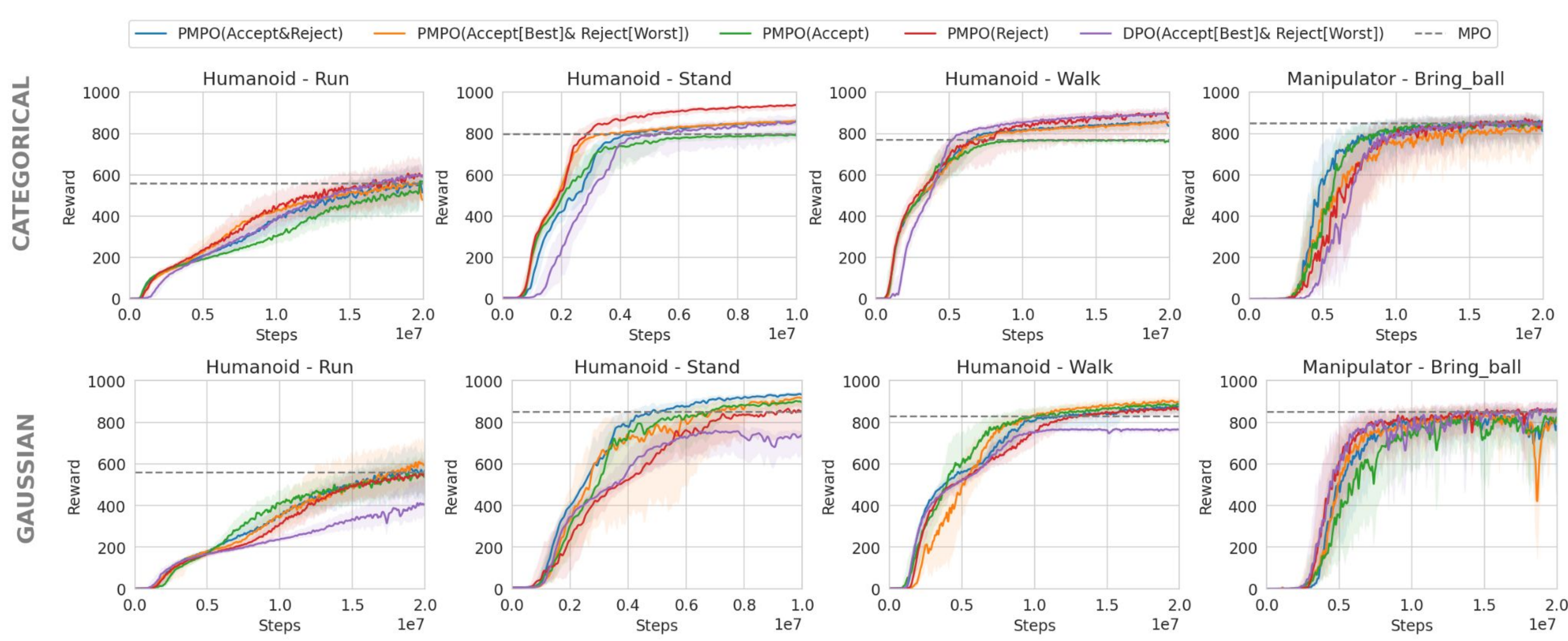
Abbas Abdolmaleki, Bilal Piot, Bobak Shahriari, Jost Tobias Springenberg, Tim Hertweck, Rishabh Joshi, Junhyuk Oh, Michael Bloesch, Thomas Lampe, Nicolas Heess, Jonas Buchli, Martin Riedmiller

Introduction

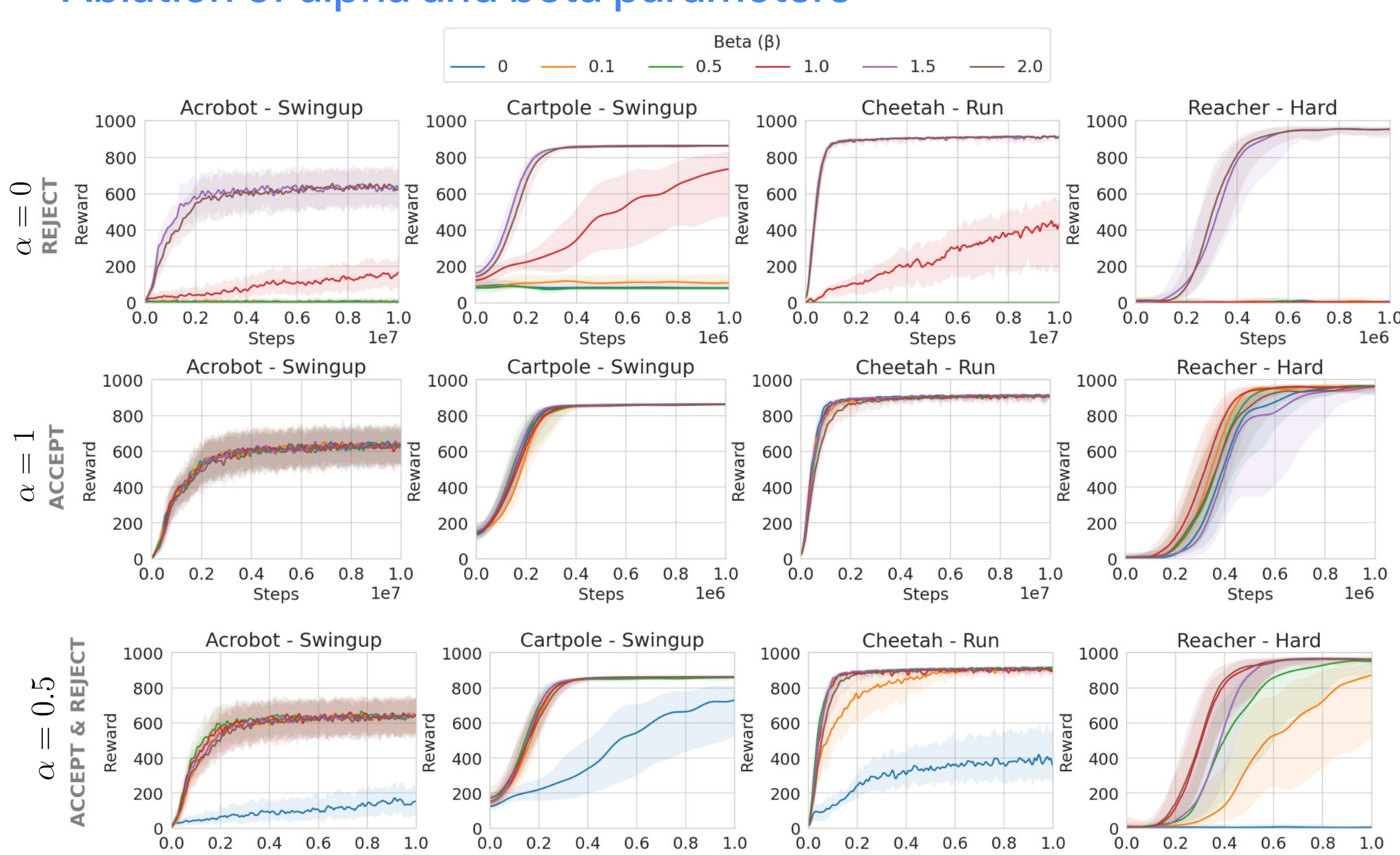
- Existing preference optimization methods often require paired feedback (positive vs. negative).
- This limits their use when only unpaired feedback (e.g., only positive or only negative) is available.
- We introduce **PMPO**, a novel approach decoupling learning from positive and negative feedback.
- This allows learning even when only one feedback type is present, including stable learning from *negative feedback alone*

Control tasks experiments

Sample 4 generations and rank them according to the Q function. Label the top 2 with positive feedback and the bottom 2 with negative feedback.



Ablation of alpha and beta parameters



PMPO Policy Improvement Objective

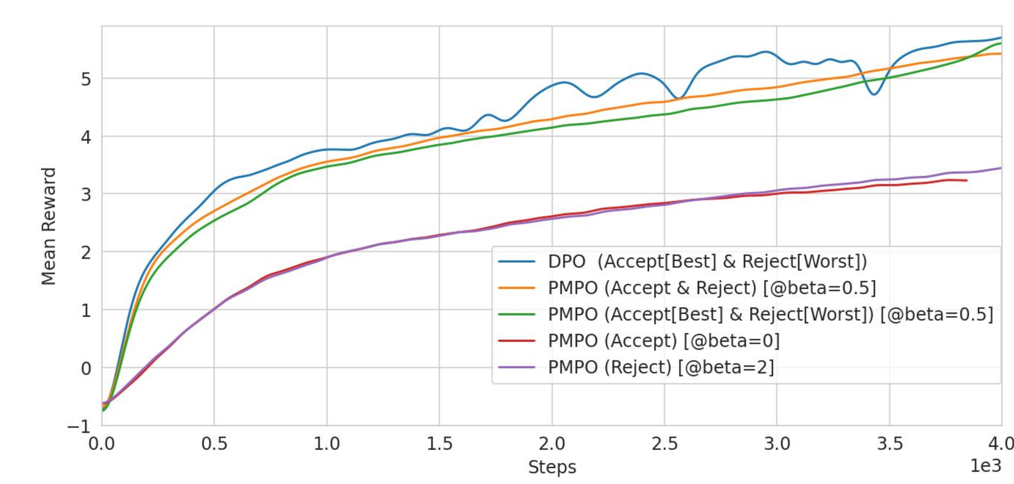
Policy improvement update rule for Improving over the current reference policy using the labeled positive/negative feedback samples:

$$\mathcal{J}(\pi_\theta;x) = \alpha \underbrace{\mathbb{E}_{y\sim\mathcal{D}_a}[\log \pi_\theta(y|x)]}_{\text{Learn from Positive Feedback}} -(1-\alpha)\underbrace{\mathbb{E}_{y\sim\mathcal{D}_r}[\log \pi_\theta(y|x)] - \beta \mathrm{KL}(\pi_{\mathrm{ref}}\|\pi_\theta;x)}_{\text{Learn from Negative Feedback}}$$

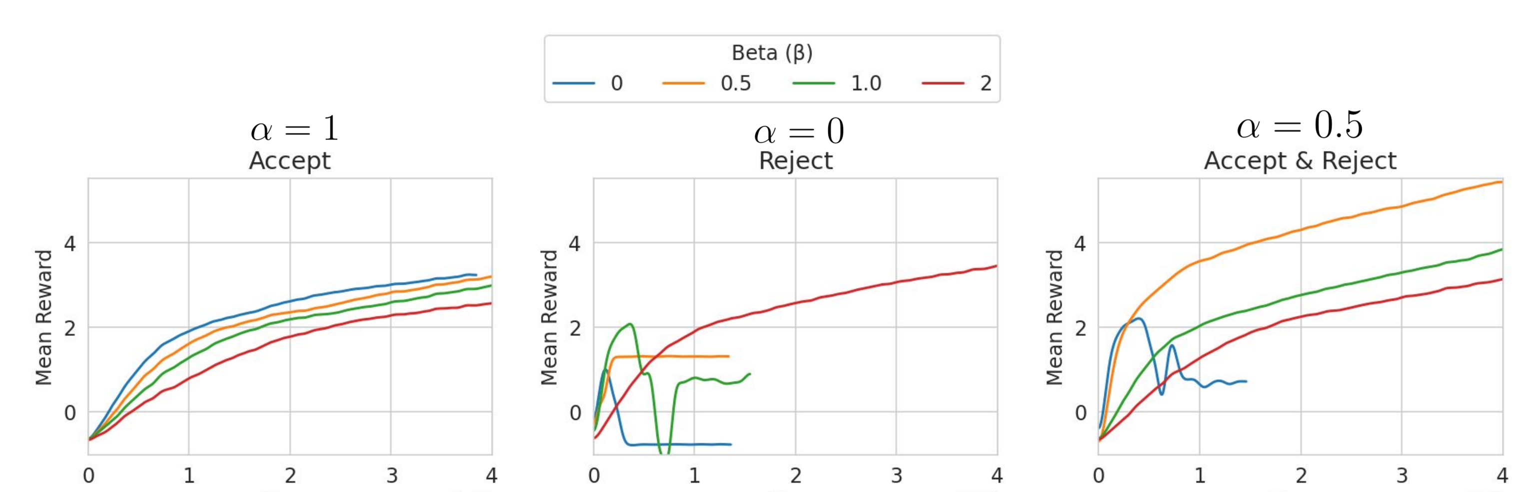
- x: Input context/conditioning variable (e.g., state, prompt).
- π_θ: Policy being optimized.
- π_{ref}: Reference policy (samples y are drawn from this).
- $\mathcal{D}_a$: Dataset providing positive ("accepted") feedback (Blue Term).
- $\mathcal{D}_r$: Dataset providing negative ("rejected") feedback (Red Term).
- α: Controls trade-off between positive ($\alpha \to 1$) and negative ($\alpha \to 0$) feedback objectives.
- β: Strength of KL regularization (essential when learning from negative feedback, i.e. when $\alpha < 1$).

RLHF Experiments with Gemma 2B

Sample 4 generations and rank them according to the learned reward function. Label the top 2 with positive feedback and bottom 2 with negative feedback.



Ablation of alpha and beta parameters



Policy Optimization as Probabilistic Inference

Objective: Find a policy that maximizes the expected likelihood of generations with positive feedback, i.e,

$$\max_\theta \mathbb{E}_{y\sim\pi_\theta(y|x)}[p(\mathbf{S}=1|y,x)]$$

- $p(\mathbf{S}=1|y,x)$: Likelihood that output y receives positive feedback
- $p(\mathbf{S}=0|y,x)$: Likelihood that output y receives negative feedback

- Note: these likelihoods sum to one: $p(\mathbf{S}=1|y,x) + p(\mathbf{S}=0|y,x) = 1$

Learning from positive feedback

Using Expectation-Maximization (EM), we obtain a one-step policy improvement over the current estimate:

$$\pi_{\mathrm{new}}(y|x) = \arg\max_\theta \ \mathbb{E}_{y\sim q(y|x)}[\log \pi_\theta(y|x)]$$

Where

$$\overbrace{q(y|x)}^{\text{Posterior}} \propto \overbrace{\pi_{\mathrm{ref}}(y|x)}^{\text{Prior}} \ \overbrace{p(\mathbf{S}=1|y,x)}^{\text{Likelihood Function}}$$

If for y with positive feedback we choose $p(\mathbf{S}=1|y,x)=1$, we get the update rule in blue.

Learning from negative feedback

Applying reparameterization $p(\mathbf{S}=1|y,x) = 1 - p(\mathbf{S}=0|y,x)$ gives the update rule based on likelihood of negative feedback:

$$\pi_{\mathrm{new}}(y|x) = \arg\max_\theta \left[-\mathbb{E}_{y\sim t(y|x)}[\log \pi_\theta(y|x)] - \beta \mathrm{KL}(\pi_{\mathrm{ref}}\|\pi_\theta;x)\right]$$

Where

$$\overbrace{t(y|x)}^{\text{Posterior}} \propto \overbrace{\pi_{\mathrm{ref}}(y|x)}^{\text{Prior}} \ \overbrace{p(\mathbf{S}=0|y,x)}^{\text{Likelihood Function}}$$

If for y with negative feedback we choose $p(\mathbf{S}=0|y,x)=1$ we get the update rule in red.

Note: When only learning from negative feedback, derivations suggests **β >= 1** is required for effective learning:

$$\beta = \frac{1}{\int \pi_{\mathrm{ref}}(y|x)p(S=0|y,x)\,dy}$$

Conclusion

- PMPO provides a flexible, intuitive, and theoretically grounded algorithm for policy optimization.
- Successfully extends learning to utilize unpaired, unbalanced, positive-only, and crucially, **negative-only** feedback. Addresses limitations of existing methods requiring paired data.
- **Limitation:** Effective negative learning benefits from accurate KL estimation.