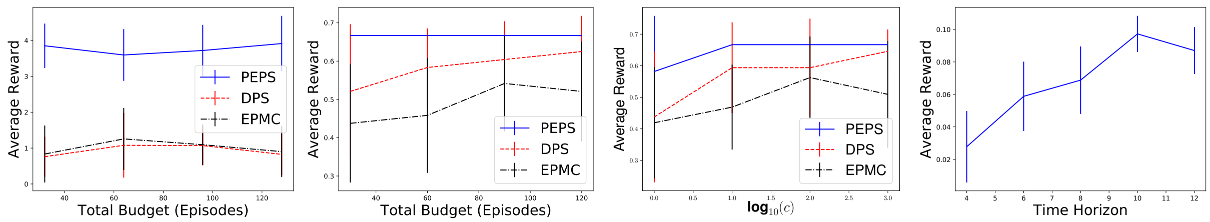

(a) Linear preference model, Random MDP (b) Deterministic preference, GridWorld (c) Varying c in BTL model, GridWorld (d) Varying H, GridWorld

We thank the reviewers for their thoughtful comments and we respond to the major questions below.

Reviewer 1 *Q1: Core contribution of the paper.* Our core contribution is to understand the theoretical rate of PbRL, and how they are different from traditional value-based RL. Our experiments do not aim to show strong performance in real applications; instead, we find that existing PbRL algorithms can suffer from long convergence time in estimating the value function. We hope our method can inspire better performing PbRL algorithms.

Q2: The assumption is very strong. If we replace the $C_0(v(\pi_1) - v(\pi_2))$ in Assumption 1 by $f(v(\pi_1) - v(\pi_2))$ for some function f (e.g., logistic for BTL model), then our result still holds as long as f is Lipschitz lower bounded; so our methods work for the BTL model. On the other hand, if we impose a BTL model on trajectory preferences, one cannot recover the correct optimal policy similar to the deterministic preference case. For example, suppose π_1 gets reward $0.5 + \varepsilon$ with probability 1 for some $\varepsilon > 0$, and π_2 gets reward 0.75 with probability $2/3$ and 0 otherwise. With some calculation one can show that π_1 loses to π_2 with probability larger than 0.5 for $\varepsilon = 0.001$. Actually, if we multiply all the rewards in Proposition 1 by a large constant, BTL will become close to deterministic and the proposition will hold for BTL. We conjecture that Proposition 1 will be true for *any* non-linear function f under the f(difference) comparisons model. Under deterministic transitions, our Assumption 1 holds for a large family of comparison models including BTL and deterministic. It is interesting future work to check the quality of the recovered policy when Assumption 1 holds with some error.

Q3: von Neumann winner. This is a very nice suggestion. However, von Neumann winner requires a distribution of policies whose support is the on the whole policy space. The policy space is exponentially large so it can be exponentially hard to recover the von Neumann policy. But we agree this would be an interesting avenue for future work.

Q4: Assumption on reward scaling and state space. We need an extra assumption that the total reward is between $[0, 1]$ so that c in Assumption 1 can be a constant. If we instead assume all rewards are in $[0, 1/H]$, the step and comparison complexities will be less by a factor of $O(H^2)$. Traditional value-based RL literature has considered this setting as well, see references 20 and 26 and Line 119-126 in our paper. The disjoint state space assumption is common in prior works, e.g., reference 14 and 23 in the paper. So our results can be compared fairly with previous work. We will make this point clear in our final version.

Questions on experiments. We have performed extra experiments, and we provide some examples above. We have tested the linear comparison model and deterministic (exact) comparisons (figure a,b), and tested the effect of c in BTL model (figure c). We focus on small-scale experiments as our goal is only to illustrate the ideas, similar to prior work (e.g., reference 14, 23 in paper). In Figure (d) we test the regret versus the time horizon H. It shows a close to linear relation, which fits our rate in Corollary 8 (the $O(H^2/\varepsilon^2)$ term translates to a linear dependence of ε on H). We will include plots verifying scaling with S, A, H in our final version.

Reviewer 2 *Q: Some assumptions might be too constraining.* Our assumptions are necessary to ensure that the true optimal policy can be recovered from the preferences (see Q2, Reviewer 1).

Reviewer 3 *Q1: Significance of the PbRL framework.* PbRL is widely applied in previous research to combat problems like reward hacking and help with reward engineering, and we refer the reviewer to reference 27 in our paper for an overview. By replacing numerical rewards with human preferences, PbRL not only reduces the effort in reward engineering but also in reward shaping, where the rewards help the agent to find the optimal policy. PbRL has a wide application in robot training [1] and game playing (reference 11,27 in the paper). As we stated in the introduction, there is NO existing work with a finite-time guarantee to the best of knowledge, and we propose the first PbRL algorithm with guaranteed performance. Our results (Proposition 1,2, Assumption 1) also establish the necessary conditions on preference probabilities to make sure that the optimal policy is recoverable.

Q2: Technical Details. Our technical contribution is mainly two-fold. Firstly, we characterize the conditions on the preference probabilities to recover the optimal policy. Different than dueling bandits, the deterministic or BTL model (see Q2 of Reviewer 1 above) does not work for PbRL. Secondly, we show a reward-free way to guide the exploration (our PEPS algorithm) when we do not have access to the reward values in each step. We cannot compute the value function in PbRL because reward values are hidden. We use a synthetic reward function (see Sec 4.1) to guide the exploration of PbRL. While our algorithm is based on existing results in dueling bandits, developing algorithms for PbRL is much harder and dueling bandits is just a building block.

Reviewer 4 *Q1: The assumptions are overly strong.* We believe that the reviewer has a misunderstanding of our assumption. Our definition of $\phi_s(\pi_1, \pi_2)$ (see first line of Proposition 1) is defined as the probability that a random *trajectory* from π_1 beats a random trajectory from π_2; it already **includes** the randomness in the transitions and preference probability. This does not mean that a good policy will never lose to a worse policy, and also it does not have to win under all trajectories; we only need to assume that it wins with a large probability under the distribution of trajectories. Our Assumption 1 states the exact point that a trajectory τ_1 from π_1 only beats a trajectory τ_2 from π_2 with a probability, and we assume that the overall probability of τ_1 beating τ_2 is at least $C_0(v(\pi_1) - v(\pi_2))$, over the random draws of the trajectories. We do not make assumptions on individual trajectories and our assumption is a relatively mild one. Moreover, we have shown that more traditional assumptions like deterministic and BTL (see Proposition 1 and Q2 in reviewer 1) cannot correctly recover the optimal policy.

The errors that the reviewer points out are typos that we will correct in our final version. On line 511, P_h^* is defined as the state distribution of π^*, so the equation should be exact "$=$".

References [1] Kupcsik, Andras, David Hsu, and Wee Sun Lee. "Learning dynamic robot-to-human object handover from human feedback." Robotics research. Springer, Cham, 2018. 161-176.