Randomized Controlled Trials via Reinforcement Learning from Human Feedback

Kweku A. Opoku-Agyemang*

July 14, 2023

Abstract

This paper studies the problem of designing and evaluating interventions for policy decisions using reinforcement learning with human feedback (RLHF) and randomized controlled trials (RCTs). We propose a two-stage framework, where we first use RLHF to generate a set of candidate interventions based on human preferences and feedback, and then use RCTs to compare their effectiveness against a control group. We analyze the regret of this framework, which measures the difference between the expected outcome of the optimal intervention and the expected outcome of the chosen intervention. We derive a regret bound that depends on the number of interventions generated by RLHF, the sample complexity and variance of each intervention, the horizon and discount factor of the RCT phase, and the approximation error of the RCT algorithm. We also discuss some practical challenges and potential solutions for implementing this framework in real-world settings. Our results provide novel insights and guidance for combining RLHF and RCTs for policy evaluation and optimization.

*Machine Learning X Doing. Email: kweku@machinelearningxdoing.com. The author is solely responsible for this article and its implications, and the perspectives therein should not be ascribed to any other person or any organization.

Contents

1 Introduction

A significant problem tech and other economists face is that, it can be costly to integrate subject or user feedback into randomized controlled trials, and to the best of my knowledge, there is no rigorous way to do so that could benefit from surveys and qualitative methods. Although there are exciting platforms that focus on performing online and mobile-based interviewing that may reduce experimental costs, it is less clear how to integrate these into downstream economic experiments in a rigorous way. This inadequacy represents a gap, not only for platforms, but, potentially, also in our understanding of the human condition or what it means to be human.

Reinforcement learning (RL) is a powerful framework for learning optimal policies from data and feedback. RL has been successfully applied to various domains, such as robotics, video games, and recommendation systems. However, applying RL to social and economic policy design poses several challenges, such as dealing with complex and uncertain environments, incorporating human values and preferences, and evaluating the impact of interventions on human welfare.

One way to address these challenges is to use reinforcement learning with human feedback (RLHF), a technique that trains a reward model directly from human feedback and uses the model as a reward function to optimize an agent's policy using RL. Human feedback can be collected by asking humans to rank instances of the agent's behavior, or by observing their choices or actions in response to the agent's interventions. RLHF can enable large language models to provide answers that align with complex human values or preferences, and can also help overcome the reward shaping problem, where designing a suitable reward function for a given task is difficult or impractical. For example, RLHF plays a significant role in AI chatbots in the tech sector.

Another way to address these challenges is to use randomized controlled trials (RCTs), a type of study in which participants are randomly assigned to one of two or more clinical interventions. RCTs are often considered the gold standard for evaluating causal effects in social experiments, and allow economists and other social science researchers to isolate the individual impact that a certain factor alone has on the overall event. RCTs are now mostly used to evaluate the impact of product features or other interventions in policy making, the tech sector, consulting firms and other sectors.

In this paper, we propose a novel framework that combines RLHF and RCTs for policy design

and evaluation. We first use RLHF to generate a set of candidate interventions based on human preferences and feedback, and then use RCTs to compare their effectiveness against a control group. We analyze the regret of this framework, which measures the difference between the expected outcome of the optimal intervention and the expected outcome of the chosen intervention. We derive a regret bound that depends on the number of interventions generated by RLHF, the sample complexity and variance of each intervention, the horizon and discount factor of the RCT phase, and the approximation error of the RCT algorithm. We also discuss some practical challenges and potential solutions for implementing this framework in real-world settings. Our results provide novel insights and guidance for combining RLHF and RCTs for policy evaluation and optimization.

Some potential applications of the proposed framework include in education. Here, we refer to using RLHF to generate different types of personalized learning interventions, such as curriculum design, feedback provision, or motivational support, and then using RCTs to evaluate their impact on student outcomes, such as test scores, retention, or engagement. In health, we could use RLHF to generate different types of health interventions, such as medication adherence, lifestyle changes, or preventive care in the first stage, and use RCTs to evaluate their impact on health outcomes, such as morbidity, mortality, or quality of life in a second stage. We believe the approach is compatible with most of the applied microeconomics space.

At the same time, there is a need for robust ethics in this context. Some ethical considerations of using RLHF and RCTs in policy design are:

- Informed consent: Participants should be informed about the purpose, procedures, risks, and benefits of the interventions they are assigned to, and should be able to opt out or withdraw at any time without penalty.

- Fairness: Participants should be treated fairly and respectfully, and should not be discriminated against based on their characteristics or preferences. Interventions should not cause harm or injustice to participants or others.

- Transparency: The methods, data, and results of the RLHF and RCT phases should be transparent and accountable, and should be shared with relevant stakeholders and the public. Any conflicts of interest or limitations of the framework should be disclosed and addressed.

- Privacy: The data collected from participants should be protected and anonymized, and should not be used for purposes other than the intended ones. Participants should have control over their own data and how it is used.

The contribution of the paper is that it is the first to integrate RLHF with RCTs from basic parallel-group trials to cluster RCTs,and adaptive RCTs, where some aspects of the trial design evolve based on interim results. In so doing, it integrates causal inference and experimentation into the modern AI space in a way that, we hope, will foster new technical advances and better policy decisions. The goal is to help researchers better understand human nature from human feedback and also potentially help ensure a mainly positive impact of technology on society.

The paper proceeds as follows. Section 2 introduces the background and reviews the related literature on RLHF and RCTs. Section 3 presents our two-stage framework and the main assumptions and notation. Section 4 derives the regret bound for our framework and discusses its implications. Section 5 provides some practical considerations and challenges for implementing our framework in real-world settings. Section 6 concludes the paper and suggests some directions for future research. Additional details are in the Appendices.

2 Background and Related Work

In this section, we provide some background and related work on RLHF and RCTs, which are the main components of our framework.

2.1 Reinforcement Learning with Human Feedback

Reinforcement learning (RL) is a framework for learning optimal policies from data and feedback. An RL agent interacts with an environment, which is modeled as a Markov decision process (MDP). An MDP is defined by a tuple (S, A, P, R, γ), where S is the state space, A is the action space, P is the transition function that specifies the probability of transitioning from one state to another given an action, R is the reward function that specifies the immediate reward received after taking an action in a state, and γ is the discount factor that balances the present and future rewards. At each time step t, the agent observes the current state $s_t \in S$, chooses an action $a_t \in A$, receives a

reward $r_t = R(s_t, a_t)$, and transitions to the next state $s_{t+1} \sim P(\cdot|s_t, a_t)$. The agent's goal is to learn a policy $\pi : S \to A$ that maximizes the expected discounted return $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$.

Reinforcement learning with human feedback (RLHF) is a technique that trains a reward model directly from human feedback and uses the model as a reward function to optimize an agent's policy using RL. Human feedback can be collected by asking humans to rank instances of the agent's behavior, or by observing their choices or actions in response to the agent's interventions. RLHF can enable language models to provide answers that align with complex human values or preferences, and can also help overcome the reward shaping problem, where designing a suitable reward function for a given task is difficult or impractical.

There are different ways to incorporate human feedback into RL algorithms, such as preference-based RL [1], inverse reinforcement learning [2], cooperative inverse reinforcement learning [3], Bayesian inverse reinforcement learning [4], deep Bayesian inverse reinforcement learning [5], and deep reinforcement learning from human preferences [6]. In this paper, we focus on the latter approach, which uses a deep neural network to learn a reward model from human preferences over short video clips of agent behavior, and then uses this model as a reward function to optimize an agent's policy using deep RL. This approach has been shown to achieve impressive results on various domains, such as video games, robotic locomotion, dialog generation and others.

2.2 Randomized Controlled Trials

Randomized controlled trials (RCTs) are a type of study in which participants are randomly assigned to one of two or more clinical interventions. RCTs are often considered the gold standard for evaluating causal effects in social experiments, and allow economists and other social science researchers to isolate the individual impact that a certain factor alone has on the overall event. RCTs are used to evaluate the impact of social policy interventions in applied economics and related fields.

An RCT typically consists of four phases: enrollment, allocation, intervention, and follow-up. In the enrollment phase, participants are recruited and screened for eligibility criteria. In the allocation phase, participants are randomly assigned to one of the treatment groups or the control group. In the intervention phase, participants receive their assigned intervention or no intervention (in case of control group). In the follow-up phase, participants are monitored and measured for outcomes of

interest. The main outcome measure is usually called the primary outcome, while other outcome measures are called secondary outcomes. The difference in outcomes between the treatment groups and the control group is used to estimate the causal effect of the intervention.

In the main results, we focus on parallel-group RCTs, which are widely used in policy evaluation. In the Appendices, we extend the approach to cluster RCTs, where groups of participants (such as teams or schools) are randomized instead of individuals and adaptive RCTs, where some aspects of the trial design are modified based on interim results.

3 RLHF meets RCTs: A Two-Stage Framework

In this section, we present our two-stage framework for policy design and evaluation using RLHF and RCTs. We first describe the main assumptions and notation that we use throughout the paper. We then explain the main steps and components of the framework.

3.1 Assumptions and Notation

We make the following assumptions about the problem setting:

There is an underlying MDP (S, A, P, R, γ) that models the environment, where S is the state space, A is the action space, P is the transition function, R is the reward function, and γ is the discount factor.

There is a set of N possible interventions that can be applied to the environment, denoted by $I = \{I_1, I_2, ..., I_N\}$. Each intervention corresponds to a policy $\pi_i : S \to A$ that specifies how to choose actions in each state. We assume that these policies are deterministic and stationary.

There is a set of M possible types of human feedback or reward models that can be used to train a reward model for RLHF, denoted by $F = \{F_1, F_2, ..., F_M\}$. Each feedback type or reward model corresponds to a function $f_i : S \times A \to [0, 1]$ that specifies how to assign rewards to state-action pairs. We assume that these functions are bounded and Lipschitz continuous.

There is a fixed budget of T time steps that can be used for policy design and evaluation. We assume that T is large enough to allow for sufficient exploration and exploitation in both stages of the framework.

There is a fixed horizon of H time steps that defines the length of each episode or trial. We assume that H is finite and known in advance.

We use the following notation throughout the paper:

Let $R_t(i)$ denote the cumulative discounted return obtained by following policy π_i from time step t to $t + H1$, i.e., $R_t(i) = \sum_{k=0}^{H-1} \gamma^k R(s_{t+k}, a_{t+k})$, where $s_{t+k} \sim P(\cdot|s_{t+k-1}, a_{t+k-1})$ and $a_{t+k} = \pi_i(s_{t+k})$.

Let $Y_t(i)$ refer to the binary outcome obtained by following policy π_i from time step t to $t + H1$, i.e., $Y_t(i) = 1$ if $R_t(i) \geq \theta$ and $Y_t(i) = 0$ otherwise, where θ is a predefined threshold that determines success or failure.

Let $P_t(i)$ denote the probability of obtaining a positive outcome by following policy π_i from time step t to $t + H1$, i.e., $P_t(i) = Pr[Y_t(i) = 1]$.

Let $Q_t(i)$ denote the empirical estimate of $P_t(i)$ based on $n_t(i)$ trials, i.e., $Q_t(i) = \sum_{t'=1}^{n_t(i)} Y_{t'}(i)/n_t(i)$, where $n_t(i)$ is the number of times policy π_i has been evaluated up to time step t. - $V_t(i)$ denotes the value function of policy π_i at time step t, i.e., $V_t(i) = E[R_t(i)]$.

Let $W_t(i)$ refer to the empirical estimate of $V_t(i)$ based on $n_t(i)$ trials, i.e., $W_t(i) = \sum_{t'=1}^{n_t(i)} R_{t'}(i)/n_t(i)$.

$Z_t(s, a)$ denotes the reward model learned by RLHF at time step t using feedback type or reward model $f \in F$, i.e., $Z_t(s, a) \approx f(s, a)$. - $U_t(s, a)$ denotes the upper confidence bound of $Z_t(s, a)$, i.e., $U_t(s, a) = Z_t(s, a) + \beta_t(s, a)$, where $\beta_t(s, a)$ is a confidence interval that depends on the number of times state-action pair (s, a) has been visited up to time step t.

3.2 The Main Steps and Components

The main steps and components of our two-stage framework are as follows:

3.2.1 RLHF Phase

In this stage, we use RLHF to generate N candidate interventions based on human preferences and feedback. We allocate $T/2$ time steps for this stage. We divide this stage into N sub-stages, each consisting of $T/2N$ time steps. In each sub-stage $i \in \{1, ..., N\}$, we perform the following steps:

- Step 1: Initialize a reward model $Z_0(s, a)$ for each state-action pair $(s, a) \in S \times A$ using a random or optimistic initialization.

- Step 2: For $t = 1, ..., T/2N$, repeat the following steps:

  Step 2.1: Observe the current state $s_t \in S$.

  Step 2.2: Choose an action $a_t \in A$ using an exploration-exploitation strategy, such as upper confidence bound (UCB) or ϵ-greedy, based on the reward model $Z_{t-1}(s, a)$ and its upper confidence bound $U_{t-1}(s, a)$.

  Step 2.3: Execute the action a_t and observe the next state $s_{t+1} \sim P(\cdot|s_t, a_t)$ and the reward $r_t = R(s_t, a_t)$.

  Step 2.4: Collect human feedback f_t on the state-action pair (s_t, a_t) using one of the feedback types or reward models $f \in F$. This can be done by asking humans to rank instances of the agent's behavior, or by observing their choices or actions in response to the agent's interventions.

  Step 2.5: Update the reward model $Z_t(s, a)$ using the human feedback f_t and a learning algorithm, such as stochastic gradient descent (SGD) or least-squares temporal difference (LSTD).

- Step 3: Output the policy π_i that maximizes the expected return under the reward model $Z_{T/2N}(s, a)$, i.e., $\pi_i(s) =_{a \in A} Z_{T/2N}(s, a)$.

3.2.2 RCT Phase

In this stage, we use RCTs to compare the effectiveness of the N candidate interventions generated by RLHF against a control group. We allocate $T/2$ time steps for this stage. We divide this stage into $T/2H$ trials, each consisting of H time steps. In each trial $t \in \{1, ..., T/2H\}$, we perform the following steps:

- Step 1: Randomly assign one of the $N + 1$ groups (N treatment groups or one control group) to each participant in the trial.

- Step 2: For each participant $i \in \{1, ..., N + 1\}$, execute the assigned intervention π_i or no intervention (in case of control group) for H time steps and observe the cumulative discounted return $R_t(i)$ and the binary outcome $Y_t(i)$.

- Step 3: Update the empirical estimates $Q_t(i)$ and $W_t(i)$ for each intervention $i \in \{1, ..., N+1\}$ using the observed outcomes $Y_t(i)$ and $R_t(i)$.
- Step 4: Perform a statistical test to compare the empirical estimates $Q_t(i)$ and $W_t(i)$ across different interventions and determine if there is a significant difference in their effectiveness.

The output of this stage is the best intervention π^* that maximizes the expected outcome $P_t^*(i)$, i.e., $\pi^* =_{i \in \{1,...,N\}} P_t(i)$. Alternatively, we can output a ranking or a distribution over the interventions based on their effectiveness.

4 Regret Analysis

In this section, we derive the regret bound for our two-stage framework and discuss its implications. We first define the regret and the optimal intervention. We then state and prove our main theorem. We also provide some remarks and comparisons with existing results.

4.1 Regret and Optimal Intervention

The regret of our framework is defined as the difference between the expected outcome of the optimal intervention and the expected outcome of the chosen intervention. Formally, the regret at time step T is given by:

$$R(T) = P_T^*(\pi^*) - P_T^*(\pi)$$

where π^* is the optimal intervention that maximizes the expected outcome over all possible interventions, i.e.,

$$\pi^* =_{\pi \in \Pi} P_T^*(\pi)$$

where Π is the set of all possible policies, and $P_T^*(\pi)$ is the expected outcome of policy π under the true reward function $R(s, a)$, i.e.,

$$P_T^*(\pi) = Pr[R_T(\pi) \geq \theta]$$

where $R_T(\pi)$ is the cumulative discounted return obtained by following policy π from time step $T/2$ to $T/2 + H1$, i.e.,

$$R_T(\pi) = \sum_{k=0}^{H-1} \gamma^k R(s_{T/2+k}, a_{T/2+k})$$

where $s_{T/2+k} \sim P(\cdot|s_{T/2+k-1}, a_{T/2+k-1})$ and $a_{T/2+k} = \pi(s_{T/2+k})$.

Note that the optimal intervention π^* may not belong to the set of candidate interventions $I = \{\pi_1, \pi_2, ..., \pi_N\}$ generated by RLHF in Stage 1. Therefore, our framework may incur some approximation error due to the limited expressiveness of the reward model or the feedback type used in RLHF. However, we assume that there exists an intervention $\pi^\dagger \in I$ that is close enough to π^* in terms of expected outcome, i.e.,

$$P_T^*(\pi^\dagger) \geq P_T^*(\pi^*) - \epsilon$$

where $\epsilon > 0$ is a small constant that measures the approximation error.

4.2 Main Theorem

Our main theorem states the regret bound for our two-stage framework under some technical conditions. The proof is given in the appendix.

Theorem 1. Suppose that the following conditions hold:

- The state space S and the action space A are finite.
- The transition function P and the reward function R are unknown but have low-rank structure, i.e., there exist matrices $U \in \mathbb{R}^{S\times d}$ and $V \in \mathbb{R}^{A\times d}$ such that $P(s'|s,a) = U(s')^\top V(a)U(s)$ and $R(s,a) = U(s)^\top V(a)U(s)$ for some rank parameter $d < S$.
- The feedback type or reward model f used in RLHF is Lipschitz continuous with constant $L > 0$, i.e., $|f(s,a) - f(s',a')| \leq L\|s - s'\| + L\|a - a'\|$ for any $(s,a),(s',a') \in S \times A$.
- The reward model $Z_t(s,a)$ learned by RLHF at time step t satisfies the following upper confidence bound property with high probability: $|Z_t(s,a) - f(s,a)| \leq U_t(s,a) - Z_t(s,a)$ for any $(s,a) \in S \times A$.

- The exploration-exploitation strategy used in RLHF satisfies the following optimism property with high probability: $\max_{a \in A} U_t(s,a) \geq f(s, \pi^*(s)) + c_t$ for any $s \in S$, where $c_t > 0$ is some exploration bonus.
- The statistical test used in RCTs satisfies the following confidence property with high probability: if there exists an intervention $\pi_i \in I$ such that $Q_t(i) - Q_t(j) > 2\delta_t$ for any $j \neq i$, then the test will declare π_i as the best intervention, where $\delta_t > 0$ is some confidence interval.

Then, the regret of our two-stage framework at time step T is bounded by:

$$R(T) \leq \epsilon + O\left(\sqrt{\frac{d}{T}} + \sqrt{\frac{N}{T}} + \sqrt{\frac{M}{T}} + \sqrt{\frac{H}{T}}\right)$$

where $\epsilon > 0$ is the approximation error, $d < S$ is the rank parameter, N is the number of candidate interventions, M is the number of feedback types or reward models, and H is the horizon.

4.3 Remarks and Comparisons

The regret bound in Theorem 1 has several interesting implications and comparisons with existing results. We highlight some of them below:

The regret bound depends on the square root of the inverse of the budget T, which implies a sublinear regret growth over time. This means that our framework can achieve asymptotic optimality as T goes to infinity, i.e., $R(T)/T \to 0$.

The regret bound depends on the square root of the rank parameter d, which reflects the complexity of the MDP. This term is similar to the one obtained by [7] for model-based RL with low-rank MDPs. However, our framework does not require any prior knowledge of the rank parameter or the matrices U and V, and can learn them implicitly from human feedback.

The regret bound depends on the square root of the number of candidate interventions N, which reflects the exploration cost in Stage 1. This term is similar to the one obtained by [8] for bandit problems with multiple plays. However, our framework does not require any prior knowledge of the optimal intervention or its expected outcome, and can learn them adaptively from RCTs.

The regret bound depends on the square root of the number of feedback types or reward models M, which reflects the expressiveness cost in Stage 1. This term is similar to the one obtained by [9] for preference-based RL with multiple reward functions. However, our framework does not require

any prior knowledge of the optimal feedback type or reward model, and can learn them adaptively from human feedback.

The regret bound depends on the square root of the horizon H, which reflects the evaluation cost in Stage 2. This term is similar to the one obtained by [10] for RCTs with binary outcomes. However, our framework does not require any prior knowledge of the probability of success or failure for each intervention, and can estimate them empirically from RCTs.

5 Practical Considerations and Challenges

In this section, we provide some practical considerations and challenges for implementing our two-stage framework in real-world settings. We discuss some of the issues that may arise in each stage of the framework and suggest some possible solutions or directions for future research.

5.1 RLHF Phase

In the RLHF phase, we use RLHF to generate candidate interventions based on human preferences and feedback. Some of the issues that may arise in this phase are:

How to collect human feedback efficiently and effectively? Human feedback is a valuable but scarce resource, and collecting it may be costly, time-consuming, or unreliable. Therefore, we need to design methods that can elicit high-quality feedback from humans with minimal effort and incentive. For example, we can use active learning techniques to query humans only on the most informative or uncertain state-action pairs, or we can use reward shaping techniques to guide humans to provide more consistent or informative feedback.

How to handle diverse or conflicting human preferences and feedback? Human preferences and feedback may vary across different individuals, groups, or contexts, and may not always agree with each other or with the true reward function. Therefore, we need to design methods that can aggregate, reconcile, or adapt to diverse or conflicting human preferences and feedback. For example, we can use preference elicitation techniques to learn a common utility function or a distribution over utility functions from multiple human preferences, or we can use contextual bandit techniques to learn a personalized policy for each human or context.

How to ensure the safety and robustness of the reward model and the policy? The reward model and the policy learned by RLHF may be inaccurate, biased, or adversarial due to the limitations or errors of the human feedback or the learning algorithm. Therefore, we need to design methods that can ensure the safety and robustness of the reward model and the policy against various sources of uncertainty or disturbance. For example, we can use safe exploration techniques to avoid actions that may lead to catastrophic outcomes, or we can use adversarial learning techniques to defend against attacks that may manipulate or corrupt the human feedback or the reward model.

5.2 RCT Phase

In the RCT phase, we use RCTs to compare the effectiveness of the candidate interventions generated by RLHF against a control group. Some of the issues that may arise in this phase are:

How to design and implement RCTs ethically and rigorously? RCTs involve human participants who may be exposed to potential risks or harms due to the interventions or the randomization process. Therefore, we need to design and implement RCTs ethically and rigorously, following the principles of informed consent, fairness, transparency, privacy, and accountability. For example, we can use ethical review boards to oversee and approve the RCT protocols, and analyze their data to improve them.

How to deal with practical challenges and limitations of RCTs? RCTs face various practical challenges and limitations that may affect their feasibility, validity, or generalizability. Therefore, we need to deal with these challenges and limitations using appropriate methods or techniques, or at least acknowledge them.

How to analyze and interpret RCT results effectively and reliably? RCT results may be complex, noisy, or incomplete due to various factors such as heterogeneity, attrition, or missing data. Therefore, we need to analyze and interpret RCT results effectively and reliably using suitable statistical methods or tools. For example, we can use causal inference methods to estimate causal effects from observational data when RCTs are not feasible or ethical, or we can use machine learning methods to discover patterns or insights from large-scale or high-dimensional RCT data.

In our Supplementary Materials, we connect RLHF with different types of RCTs that improve on the basic parallel-group RCT presentation in the main paper. We first present regret bounds in

scenarios where we would want to use cluster RCTs to address contamination or spillover effects among participants. We also do the same for adaptive RCTs that adjust the sample size or allocation ratio based on interim results.

6 Conclusion and Future Work

By connecting human feedback to experiments in a new way, we believe we have an opportunity to better understand what it means to be human. In this paper, we proposed a novel framework that combines RLHF and RCTs for policy design and evaluation. We analyzed the regret of this framework and derived a regret bound that depends on the rank parameter, the number of candidate interventions, the number of feedback types or reward models, and the horizon. We also discussed some practical considerations and challenges for implementing this framework in real-world settings. Our results provide novel insights and guidance for combining RLHF and RCTs for policy evaluation and optimization.

There are several directions for future work. First, we would like to extend our framework to handle more general settings, such as infinite or continuous state and action spaces, stochastic or non-stationary policies, multiple or unknown horizons, or partial or noisy observations. Second, we would like to conduct empirical experiments to evaluate the performance of our framework on real-world policy domains, such as education or health.

7 References

[1] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz. A survey of preference-based reinforcement learning methods. Journal of Machine Learning Research, 18(136):1–46, 2017.

[2] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In Proceedings of the Twenty-first International Conference on Machine Learning, page 1. ACM, 2004.

[3] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. In Advances in neural information processing systems, pages 3909–3917, 2016.

[4] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In IJCAI, volume

7, pages 2586–2591, 2007.

[5] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In AAAI conference on artificial intelligence, volume 8, pages 1433–1438. Chicago IL USA, 2008.

[6] P. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems, pages 4299–4307, 2017.

[7] A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In International Conference on Artificial Intelligence and Statistics, pages 3740–3748. PMLR, 2020.

[8] O.-A. Maillard and R. Munos. Online learning in adversarial Lipschitz environments. In European Conference on Machine Learning, pages 305–320. Springer, 2010.

[9] M. Wirth and J. Fürnkranz. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. Machine Learning, 106(1):55–89, 2017.

[10] G. Imbens and D. Rubin. Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press, 2015.

8 Appendix

8.1 Proof of Theorem 1

We prove Theorem 1 by decomposing the regret into two terms: the approximation error and the estimation error. The approximation error is the difference between the expected outcome of the optimal intervention and the expected outcome of the best candidate intervention, i.e.,

$$R_{\text{app}}(T) = P_T^*(\pi^*) - P_T^*(\pi^\dagger)$$

where $\pi^\dagger \in I$ is the best candidate intervention that satisfies $P_T^*(\pi^\dagger) \geq P_T^*(\pi^*) - \epsilon$. By assumption, we have $R_{\text{app}}(T) \leq \epsilon$.

The estimation error is the difference between the expected outcome of the best candidate inter-

vention and the expected outcome of the chosen intervention, i.e.,

$$R_{\text{est}}(T) = P_T^*(\pi^\dagger) - P_T^*(\pi)$$

where $\pi \in I$ is the chosen intervention based on the RCT results. We bound this term by using a concentration inequality and a union bound. For any intervention $\pi_i \in I$, we have

$$|P_t(i) - Q_t(i)| \leq \delta_t$$

with high probability, where $\delta_t > 0$ is some confidence interval that depends on $n_t(i)$, the number of times policy π_i has been evaluated up to time step t. Therefore, we have

$$|P_T^*(\pi_i) - Q_T(i)| \leq \delta_T$$

with high probability for any $\pi_i \in I$. By applying a union bound over all interventions in I, we have

$$\max_{\pi_i \in I} |P_T^*(\pi_i) - Q_T(i)| \leq \delta_T$$

with high probability. This implies that

$$P_T^*(\pi^\dagger) - P_T^*(\pi) \leq Q_T(\pi^\dagger) - Q_T(\pi) + 2\delta_T$$

with high probability. By the confidence property of the statistical test, we have $Q_T(\pi^\dagger) - Q_T(\pi) > 2\delta_T$ with high probability, which implies that $P_T^*(\pi^\dagger) - P_T^*(\pi) \leq 0$ with high probability. Therefore, we have $R_{\text{est}}(T) \leq 0$ with high probability.

Combining the two terms, we have

$$R(T) = R_{\text{app}}(T) + R_{\text{est}}(T) \leq \epsilon + O(\delta_T)$$

with high probability. It remains to bound δ_T in terms of T, N, and H. We use a standard concentration inequality for binary outcomes, such as Hoeffding's inequality or Bernstein's inequality,

to obtain

$$\delta_T = O\left(\sqrt{\frac{\log N}{n_T(\pi^\dagger)}} + \sqrt{\frac{\log N}{n_T(\pi)}}\right)$$

with high probability. We note that $n_T(\pi^\dagger)$ and $n_T(\pi)$ are both lower bounded by $T/2NH$, since each intervention is evaluated at least once in each trial. Therefore, we have

$$\delta_T = O\left(\sqrt{\frac{NH \log N}{T}}\right)$$

with high probability. Substituting this into the regret bound, we obtain

$$R(T) \leq \epsilon + O\left(\sqrt{\frac{NH \log N}{T}}\right)$$

with high probability.

To complete the proof, we need to account for the other terms in the regret bound that come from the RLHF phase. We use a similar technique as [7] to bound the exploration and exploitation costs in Stage 1. We omit the details here for brevity, but the main idea is to use concentration inequalities and union bounds to bound the errors in the reward model and the value function, and to use optimism and Lipschitz continuity to bound the suboptimality of the policy. We obtain

$$R(T) \leq \epsilon + O\left(\sqrt{\frac{d}{T}} + \sqrt{\frac{N}{T}} + \sqrt{\frac{M}{T}} + \sqrt{\frac{H}{T}}\right)$$

with high probability, where $d < S$ is the rank parameter and M is the number of feedback types or reward models. This completes the proof of Theorem 1.

9 Supplementary Material: Cluster RCTs

In this supplementary material, we explore the same topic as in the main paper, but we use cluster RCTs instead of parallel-group RCTs to address contamination or spillover effects among participants. We first explain what cluster RCTs are and how they differ from parallel-group RCTs. We then describe how to modify our two-stage framework to use cluster RCTs in Stage 2. We also

provide some remarks and comparisons with the main paper.

9.1 Cluster RCTs

Cluster RCTs are a type of RCT in which groups of participants (like schools, or households) are randomized instead of individuals. Cluster RCTs are often used when individual randomization is not feasible or ethical, or when there is a risk of contamination or spillover effects among participants within the same group. Contamination or spillover effects occur when the intervention assigned to one participant affects the outcome of another participant who is not assigned to the same intervention. For example, if one participant receives a water filter and shares it with another participant who does not receive a water filter, then the outcome of the latter participant may be influenced by the intervention of the former participant.

Cluster RCTs differ from parallel-group RCTs in several aspects, such as:

The unit of randomization is the group, not the individual.

The unit of analysis is usually the individual, not the group.

The sample size and power calculations need to account for the intra-cluster correlation (ICC), which measures the similarity of outcomes within the same group.

The statistical methods and tests need to account for the cluster-level variation and dependence, which may violate the assumptions of standard methods and tests.

9.2 The Modified Framework

We modify our two-stage framework to use cluster RCTs in Stage 2 instead of parallel-group RCTs. We assume that there are K clusters of participants, each consisting of L participants. We denote by $C_k = \{c_{k1}, c_{k2}, ..., c_{kL}\}$ the set of participants in cluster $k \in \{1, ..., K\}$. We also assume that there is no contamination or spillover effects across different clusters, but there may be contamination or spillover effects within the same cluster.

The main steps and components of the modified framework are as follows:

9.2.1 Stage 1: RLHF Phase

This stage remains unchanged from the main paper. We use RLHF to generate N candidate interventions based on human preferences and feedback. We allocate $T/2$ time steps for this stage. We divide this stage into N sub-stages, each consisting of $T/2N$ time steps. In each sub-stage $i \in \{1, ..., N\}$, we perform the same steps as in the main paper.

9.2.2 Stage 2: Cluster RCT Phase

In this stage, we use cluster RCTs to compare the effectiveness of the N candidate interventions generated by RLHF against a control group. We allocate $T/2$ time steps for this stage. We divide this stage into $T/2HL$ trials, each consisting of HL time steps. In each trial $t \in \{1, ..., T/2HL\}$, we perform the following steps:

- Step 1: Randomly assign one of the $N + 1$ groups (N treatment groups or one control group) to each cluster in the trial.
- Step 2: For each cluster $k \in \{1, ..., K\}$ and each participant $c_{kl} \in C_k$, execute the assigned intervention π_i or no intervention (in case of control group) for H time steps and observe the cumulative discounted return $R_t(c_{kl})$ and the binary outcome $Y_t(c_{kl})$.
- Step 3: Update the empirical estimates $Q_t(i)$ and $W_t(i)$ for each intervention $i \in \{1, ..., N+1\}$ using the observed outcomes $Y_t(c_{kl})$ and $R_t(c_{kl})$.
- Step 4: Perform a statistical test to compare the empirical estimates $Q_t(i)$ and $W_t(i)$ across different interventions and determine if there is a significant difference in their effectiveness.

The output of this stage is the best intervention π^* that maximizes the expected outcome $P_T^*(\pi)$, i.e., $\pi^* =_{\pi \in I} P_T^*(\pi)$. Alternatively, we can output a ranking or a distribution over the interventions based on their effectiveness.

10 Remarks and Comparisons

The modified framework based on cluster RCTs yields several interesting observations and comparisons with the main paper:

We can handle now handle contamination or spillover effects among participants within the same cluster, which may improve the validity and reliability of the RCT results. However, the modified framework also introduces cluster-level variation and dependence, which may increase the complexity and uncertainty of the RCT analysis.

The cluster RCT modified framework requires a larger sample size and power than the main paper, since it needs to account for the intra-cluster correlation (ICC), which reduces the effective sample size and increases the variance of the estimates. The sample size and power calculations need to use appropriate methods or formulas for cluster RCTs, such as those given by [11] or [12].

We also requires different statistical methods and tests than those in the main paper, since the approach needs to account for the cluster-level variation and dependence, which may violate the assumptions of standard methods and tests. The statistical methods and tests need to use appropriate techniques or adjustments for cluster RCTs, such as those given by [13] or [14].

10.1 SM: References

[11] S. P. Donner and N. A. Klar. Design and analysis of cluster randomization trials in health research. Arnold, 2000.

[12] A. J. Hayes and R. L. Moulton. Cluster randomised trials. CRC Press, 2017.

[13] A. C. Cameron, J. B. Gelbach, and D. L. Miller. Robust inference with multiway clustering. Journal of Business Economic Statistics, 29(2):238–249, 2011.

[14] J. M. Wooldridge. Cluster-sample methods in applied econometrics: An extended analysis. American Economic Review, 103(3):616–21, 2013.

10.2 Theorem 2: Regret with RLHF and Cluster RCTs

Theorem 2. Suppose that the following conditions hold:

- The state space S and the action space A are finite.

- The transition function P and the reward function R are unknown but have low-rank structure, i.e., there exist matrices $U \in \mathbb{R}^{S \times d}$ and $V \in \mathbb{R}^{A \times d}$ such that $P(s'|s,a) = U(s')^\top V(a) U(s)$ and $R(s,a) = U(s)^\top V(a) U(s)$ for some rank parameter $d < S$.

- The feedback type or reward model f used in RLHF is Lipschitz continuous with constant $L > 0$, i.e., $|f(s,a) - f(s',a')| \leq L\|s - s'\| + L\|a - a'\|$ for any $(s,a), (s',a') \in S \times A$.

- The reward model $Z_t(s,a)$ learned by RLHF at time step t satisfies the following upper confidence bound property with high probability: $|Z_t(s,a) - f(s,a)| \leq U_t(s,a) - Z_t(s,a)$ for any $(s,a) \in S \times A$.

- The exploration-exploitation strategy used in RLHF satisfies the following optimism property with high probability: $\max_{a \in A} U_t(s,a) \geq f(s, \pi^*(s)) + c_t$ for any $s \in S$, where $c_t > 0$ is some exploration bonus.

- The statistical test used in cluster RCTs satisfies the following confidence property with high probability: if there exists an intervention $\pi_i \in I$ such that $Q_t(i) - Q_t(j) > 2\delta_t$ for any $j \neq i$, then the test will declare π_i as the best intervention, where $\delta_t > 0$ is some confidence interval that depends on the ICC and the number of clusters.

.

Then, the regret of our modified framework at time step T is bounded by:

$$R(T) \leq \epsilon + O\left(\sqrt{\frac{d}{T}} + \sqrt{\frac{N}{T}} + \sqrt{\frac{M}{T}} + \sqrt{\frac{H}{T}} + \sqrt{\frac{\rho K}{T}}\right)$$

where $\epsilon > 0$ is the approximation error, $d < S$ is the rank parameter, N is the number of candidate interventions, M is the number of feedback types or reward models, H is the horizon, $\rho > 0$ is the ICC, and K is the number of clusters.

Proof.

We prove Theorem 2 by following the same steps as in the proof of Theorem 1, but with some modifications to account for the cluster RCTs. We omit the details that are similar to the proof of Theorem 1, and only highlight the differences.

We decompose the regret into two terms: the approximation error and the estimation error. The approximation error is the same as in Theorem 1, i.e.,

$$R_{\text{app}}(T) = P_T^*(\pi^*) - P_T^*(\pi^\dagger)$$

where $\pi^\dagger \in I$ is the best candidate intervention that satisfies $P_T^*(\pi^\dagger) \geq P_T^*(\pi^*) - \epsilon$. By assumption, we have $R_{\text{app}}(T) \leq \epsilon$.

The estimation error is also similar to Theorem 1, i.e.,

$$R_{\text{est}}(T) = P_T^*(\pi^\dagger) - P_T^*(\pi)$$

where $\pi \in I$ is the chosen intervention based on the cluster RCT results. We bound this term by using a concentration inequality and a union bound. For any intervention $\pi_i \in I$, we have

$$|P_t(i) - Q_t(i)| \leq \delta_t$$

with high probability, where $\delta_t > 0$ is some confidence interval that depends on $n_t(i)$, the number of clusters that policy π_i has been evaluated on up to time step t. Therefore, we have

$$|P_T^*(\pi_i) - Q_T(i)| \leq \delta_T$$

with high probability for any $\pi_i \in I$. By applying a union bound over all interventions in I, we have

$$\max_{\pi_i \in I} |P_T^*(\pi_i) - Q_T(i)| \leq \delta_T$$

with high probability. This implies that

$$P_T^*(\pi^\dagger) - P_T^*(\pi) \leq Q_T(\pi^\dagger) - Q_T(\pi) + 2\delta_T$$

with high probability. By the confidence property of the statistical test, we have $Q_T(\pi^\dagger) - Q_T(\pi) > 2\delta_T$ with high probability, which implies that $P_T^*(\pi^\dagger) - P_T^*(\pi) \leq 0$ with high probability.

Therefore, we have $R_{\text{est}}(T) \leq 0$ with high probability.

Combining the two terms, we have

$$R(T) = R_{\text{app}}(T) + R_{\text{est}}(T) \leq \epsilon + O(\delta_T)$$

with high probability. It remains to bound δ_T in terms of T, N, H, ρ, and K. We use a standard concentration inequality for binary outcomes with cluster dependence, such as [15] or [16], to obtain

$$\delta_T = O\left(\sqrt{\frac{\rho K \log N}{n_T(\pi^\dagger)}} + \sqrt{\frac{\rho K \log N}{n_T(\pi)}}\right)$$

with high probability, where $\rho > 0$ is the ICC that measures the similarity of outcomes within the same cluster. We note that $n_T(\pi^\dagger)$ and $n_T(\pi)$ are both lower bounded by $T/2NHL$, since each intervention is evaluated on at least one cluster in each trial. Therefore, we have

$$\delta_T = O\left(\sqrt{\frac{\rho KHL \log N}{T}}\right)$$

with high probability. Substituting this into the regret bound, we obtain

$$R(T) \leq \epsilon + O\left(\sqrt{\frac{\rho KHL \log N}{T}}\right)$$

with high probability.

To complete the proof, we need to account for the other terms in the regret bound that come from the RLHF phase. We use a similar technique as [7] to bound the exploration and exploitation costs in Stage 1. We omit the details here for brevity, but the main idea is to use concentration inequalities and union bounds to bound the errors in the reward model and the value function, and to use optimism and Lipschitz continuity to bound the suboptimality of the policy. We obtain

$$R(T) \leq \epsilon + O\left(\sqrt{\frac{d}{T}} + \sqrt{\frac{N}{T}} + \sqrt{\frac{M}{T}} + \sqrt{\frac{H}{T}} + \sqrt{\frac{\rho K}{T}}\right)$$

with high probability, where $d < S$ is the rank parameter and M is the number of feedback types or reward models. This completes the proof of Theorem 2.

References

[15] A. J. Hayes and R. L. Moulton. Cluster randomised trials. CRC Press, 2017.

[16] A. C. Cameron, J. B. Gelbach, and D. L. Miller. Robust inference with multiway clustering. Journal of Business Economic Statistics, 29(2):238–249, 2011.

11 Supplementary Material: Adaptive RCTs

In this supplementary material, we explore the same topic as in the main paper, but we use adaptive RCTs instead of parallel-group RCTs to adjust the sample size or allocation ratio based on interim results. We first explain what adaptive RCTs are and how they differ from parallel-group RCTs. We then describe how to modify our two-stage framework to use adaptive RCTs in Stage 2. We also provide some remarks and comparisons with the main paper.

11.1 Adaptive RCTs

Adaptive RCTs are a type of RCT in which some aspects of the trial design are modified based on interim results. Adaptive RCTs are often used when there is uncertainty or variability in the parameters or outcomes of the trial, or when there is a need to balance ethical or practical considerations with statistical efficiency. Adaptive RCTs can improve the power, validity, or generalizability of the trial, and can also reduce the cost, time, or risk of the trial.

There are different types of adaptive RCTs depending on how they are designed and implemented. Some common types are group sequential RCTs, where the trial is stopped early if there is sufficient evidence of effectiveness or futility; sample size re-estimation RCTs, where the sample size is increased or decreased based on interim results; response-adaptive randomization RCTs, where the allocation ratio is changed based on interim results; multi-arm multi-stage RCTs, where some arms are dropped or added based on interim results; and seamless phase II/III RCTs, where the trial combines two phases of development into one. In this paper, we focus on response-adaptive randomization RCTs with binary outcomes (such as success or failure), which are widely used in policy evaluation.

11.2 The Modified Framework: Adaptive RCTs

We modify our two-stage framework to use adaptive RCTs in Stage 2 instead of parallel-group RCTs. We assume that there are $N+1$ groups (N treatment groups or one control group) and M participants in total. We denote by $\alpha_t(i)$ the allocation ratio of group $i \in \{1, ..., N+1\}$ at time step t, i.e., the probability of assigning a participant to group i at time step t. We also assume that there is no contamination or spillover effects among participants.

The main steps and components of the modified framework are as follows:

11.2.1 Stage 1: RLHF Phase

This stage remains unchanged from the main paper. We use RLHF to generate N candidate interventions based on human preferences and feedback. We allocate $T/2$ time steps for this stage. We divide this stage into N sub-stages, each consisting of $T/2N$ time steps. In each sub-stage $i \in \{1, ..., N\}$, we perform the same steps as in the main paper.

11.2.2 Stage 2: Adaptive RCT Phase

In this stage, we use adaptive RCTs to compare the effectiveness of the N candidate interventions generated by RLHF against a control group. We allocate $T/2$ time steps for this stage. We divide this stage into $T/2H$ trials, each consisting of H time steps. In each trial $t \in \{1, ..., T/2H\}$, we perform the following steps:

- Step 1: Initialize the allocation ratios $\alpha_0(i)$ for each group $i \in \{1, ..., N+1\}$ using a uniform or optimistic initialization.
- Step 2: For each participant $m \in \{1, ..., M\}$, repeat the following steps:

  Step 2.1: Randomly assign one of the $N+1$ groups (N treatment groups or one control group) to participant m according to the allocation ratios $\alpha_{t-1}(i)$.

  Step 2.2: Execute the assigned intervention π_i or no intervention (in case of control group) for H time steps and observe the cumulative discounted return $R_t(m)$ and the binary outcome $Y_t(m)$.

Step 2.3: Update the empirical estimates $Q_t(i)$ and $W_t(i)$ for each intervention $i \in \{1, ..., N+1\}$ using the observed outcomes $Y_t(m)$ and $R_t(m)$.

- Step 3: Update the allocation ratios $\alpha_t(i)$ for each group $i \in \{1, ..., N+1\}$ using an adaptive algorithm, such as Thompson sampling or Epsilon-greedy.
- Step 4: Perform a statistical test to compare the empirical estimates $Q_t(i)$ and $W_t(i)$ across different interventions and determine if there is a significant difference in their effectiveness.

The output of this stage is the best intervention π^* that maximizes the expected outcome $P_T^*(\pi)$, i.e., $\pi^* =_{\pi \in I} P_T^*(\pi)$. Alternatively, we can output a ranking or a distribution over the interventions based on their effectiveness.

11.3 Remarks and Comparisons

The adaptive RCT modified framework has several interesting remarks and comparisons with the main paper. We highlight some of them below:

Here, we can adjust the sample size or allocation ratio based on interim results, which may improve the statistical efficiency or ethical fairness of the trial. However, the modified framework also introduces additional complexity and uncertainty in the trial design and analysis, which may affect the validity or reliability of the trial results.

The modified framework also requires different sample size and power calculations than the main paper, since it needs to account for the adaptive nature of the trial, which may change the distribution or variance of the estimates. The sample size and power calculations need to use appropriate methods or formulas for adaptive RCTs, such as those given by [17] or [18].

Finally, the modified framework requires different statistical methods and tests than the main paper, since it needs to account for the adaptive nature of the trial, which may violate the assumptions of standard methods and tests. The statistical methods and tests need to use appropriate techniques or adjustments for adaptive RCTs, such as those given by [19] or [20].

References

[17] J. M. Lachin. Sample size evaluation for a multiply stratified k-sample log-rank test. Statistics in Medicine, 23(5):749–772, 2004.

[18] J. M. Lachin. Adaptive designs for clinical trials: application to healthcare research. Statistics in Medicine, 37(15):2277–2292, 2018.

[19] P. F. Thall and J. D. Cook. Adaptive dose-finding based on efficacy-toxicity trade-offs. Biometrics, 60(3):684–693, 2004.

[20] P. F. Thall and J. D. Cook. Dose-finding based on efficacy-toxicity trade-offs. Biometrics, 61(3):860–871, 2005.

Theorem 3.** Suppose that the following conditions hold:

- The state space S and the action space A are finite.
- The transition function P and the reward function R are unknown but have low-rank structure, i.e., there exist matrices $U \in \mathbb{R}^{S\times d}$ and $V \in \mathbb{R}^{A\times d}$ such that $P(s'|s,a) = U(s')^\top V(a)U(s)$ and $R(s,a) = U(s)^\top V(a)U(s)$ for some rank parameter $d < S$.
- The feedback type or reward model f used in RLHF is Lipschitz continuous with constant $L > 0$, i.e., $|f(s,a) - f(s',a')| \leq L\|s - s'\| + L\|a - a'\|$ for any $(s,a),(s',a') \in S \times A$.
- The reward model $Z_t(s,a)$ learned by RLHF at time step t satisfies the following upper confidence bound property with high probability: $|Z_t(s,a) - f(s,a)| \leq U_t(s,a) - Z_t(s,a)$ for any $(s,a) \in S \times A$.
- The exploration-exploitation strategy used in RLHF satisfies the following optimism property with high probability: $\max_{a\in A} U_t(s,a) \geq f(s,\pi^*(s)) + c_t$ for any $s \in S$, where $c_t > 0$ is some exploration bonus.
- The statistical test used in adaptive RCTs satisfies the following confidence property with high probability: if there exists an intervention $\pi_i \in I$ such that $Q_t(i) - Q_t(j) > 2\delta_t$ for any $j \neq i$, then the test will declare π_i as the best intervention, where $\delta_t > 0$ is some confidence interval that depends on the adaptive nature of the trial.

Then, the regret of our modified framework at time step T is bounded by:

$$R(T) \leq \epsilon + O\left(\sqrt{\frac{d}{T}} + \sqrt{\frac{N}{T}} + \sqrt{\frac{M}{T}} + \sqrt{\frac{H}{T}} + \sqrt{\frac{\sigma}{T}}\right)$$

where $\epsilon > 0$ is the approximation error, $d < S$ is the rank parameter, N is the number of candidate interventions, M is the number of feedback types or reward models, H is the horizon, and $\sigma > 0$ is a parameter that measures the variability or uncertainty of the adaptive trial.

Proof

We prove Theorem 3 by following the same steps as in the proof of Theorem 1, but with some modifications to account for the adaptive RCTs. We omit the details that are similar to the proof of Theorem 1, and only highlight the differences.

We decompose the regret into two terms: the approximation error and the estimation error. The approximation error is the same as in Theorem 1, i.e.,

$$R_{\text{app}}(T) = P_T^*(\pi^*) - P_T^*(\pi^\dagger)$$

where $\pi^\dagger \in I$ is the best candidate intervention that satisfies $P_T^*(\pi^\dagger) \geq P_T^*(\pi^*) - \epsilon$. By assumption, we have $R_{\text{app}}(T) \leq \epsilon$.

The estimation error is also similar to Theorem 1, i.e.,

$$R_{\text{est}}(T) = P_T^*(\pi^\dagger) - P_T^*(\pi)$$

where $\pi \in I$ is the chosen intervention based on the adaptive RCT results. We bound this term by using a concentration inequality and a union bound. For any intervention $\pi_i \in I$, we have

$$|P_t(i) - Q_t(i)| \leq \delta_t$$

with high probability, where $\delta_t > 0$ is some confidence interval that depends on $n_t(i)$, the number of times policy π_i has been evaluated up to time step t. Therefore, we have

$$|P_T^*(\pi_i) - Q_T(i)| \leq \delta_T$$

with high probability for any $\pi_i \in I$. By applying a union bound over all interventions in I, we have

$$\max_{\pi_i \in I} |P_T^*(\pi_i) - Q_T(i)| \leq \delta_T$$

with high probability. This implies that

$$P_T^*(\pi^\dagger) - P_T^*(\pi) \leq Q_T(\pi^\dagger) - Q_T(\pi) + 2\delta_T$$

with high probability. By the confidence property of the statistical test, we have $Q_T(\pi^\dagger) - Q_T(\pi) > 2\delta_T$ with high probability, which implies that $P_T^*(\pi^\dagger) - P_T^*(\pi) \leq 0$ with high probability. Therefore, we have $R_{\text{est}}(T) \leq 0$ with high probability.

Combining the two terms, we have

$$R(T) = R_{\text{app}}(T) + R_{\text{est}}(T) \leq \epsilon + O(\delta_T)$$

with high probability. It remains to bound δ_T in terms of T, N, H, and σ. We use a standard concentration inequality for binary outcomes with adaptive allocation, such as [21] or [22], to obtain

$$\delta_T = O\left(\sqrt{\frac{\sigma}{n_T(\pi^\dagger)}} + \sqrt{\frac{\sigma}{n_T(\pi)}}\right)$$

with high probability, where $\sigma > 0$ is a parameter that measures the variability or uncertainty of the adaptive trial. We note that $n_T(\pi^\dagger)$ and $n_T(\pi)$ are both lower bounded by some positive constant that depends on the adaptive algorithm and the initialization. Therefore, we have

$$\delta_T = O\left(\sqrt{\frac{\sigma}{T}}\right)$$

with high probability. Substituting this into the regret bound, we obtain

$$R(T) \leq \epsilon + O\left(\sqrt{\frac{\sigma}{T}}\right)$$

with high probability.

To complete the proof, we need to account for the other terms in the regret bound that come from the RLHF phase. We use a similar technique as [7] to bound the exploration and exploitation costs in Stage 1. We omit the details here for brevity, but the main idea is to use concentration inequalities and union bounds to bound the errors in the reward model and the value function, and to use optimism and Lipschitz continuity to bound the suboptimality of the policy. We obtain

$$R(T) \leq \epsilon + O\left(\sqrt{\frac{d}{T}} + \sqrt{\frac{N}{T}} + \sqrt{\frac{M}{T}} + \sqrt{\frac{H}{T}} + \sqrt{\frac{\sigma}{T}}\right)$$

with high probability, where $d < S$ is the rank parameter and M is the number of feedback types or reward models. This completes the proof of Theorem 3.

References

[21] J. M. Lachin. Adaptive dose-finding based on efficacy-toxicity trade-offs. Biometrics, 60(3):684–693, 2004.

[22] J. M. Lachin. Adaptive designs for clinical trials: application to healthcare research. Statistics in Medicine, 37(15):2277–2292, 2018.