Reinforcement Learning from Human Feedback via Randomized Experiments

Kweku A. Opoku-Agyemang*

July 16, 2023

Abstract

This paper focuses on how human feedback can improve an randomized controlled trial intervention that affects people's behavior or outcomes. For example, the intervention could be a personalized message that encourages people to exercise more or eat healthier. Human feedback could be ratings, preferences, emotions, or rewards that people give after receiving the intervention. We propose a two-stage method: First, we run randomized controlled trials (RCTs) to compare different kinds of human feedback or different ways to calculate rewards for the same intervention. Second, we use reinforcement learning from human feedback (RLHF) to optimize the intervention based on the best kind of feedback or reward. The quality of our method depends on several factors, such as: how many kinds of feedback or rewards we compare in the first stage; how easy or hard it is to measure each kind of feedback or reward; how consistent or variable each kind of feedback or reward is; how long we run the second stage and how much we care about future outcomes; and how well our RLHF algorithm can learn from human feedback and adapt to new situations. Our goal is to find a balance between these factors that leads to the most effective intervention. We close with policy implications.

*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

Human feedback is a valuable source of information for designing and improving interventions that aim to influence people's behavior or outcomes. For example, an intervention could be a personalized message that encourages people to exercise more or eat healthier, and human feedback could be ratings, preferences, emotions, or rewards that people give after receiving the message. However, not all kinds of human feedback are equally useful or reliable for evaluating and optimizing interventions. Different kinds of feedback may have different levels of complexity, noise, bias, or expressiveness, which may affect how well they reflect the true impact or value of the intervention. Moreover, different ways to calculate rewards from human feedback may have different assumptions, properties, or limitations, which may affect how well they guide the learning and adaptation of the intervention. Therefore, it is important to compare and select the best kind of human feedback or reward for a given intervention and setting[1].

In this paper, we propose a novel two-stage method for comparing and selecting the best kind of human feedback or reward for a given intervention and context. In the first stage, we run randomized controlled trials (RCTs) to compare different kinds of human feedback or different ways to calculate rewards for the same intervention. We use statistical methods to estimate the performance of each kind of feedback or reward in terms of its accuracy, precision, robustness, or informativeness. In the second stage, we use reinforcement learning from human feedback (RLHF) to optimize the intervention based on the best kind of feedback or reward. We use RLHF algorithms that can learn from human feedback and generalize to new states and actions.

We analyze the theoretical and empirical properties of our method. We derive bounds on the regret of our method, which measures how much worse our method performs compared to the optimal intervention. We show that the regret depends on several factors, such as: How many kinds of feedback or rewards we compare in the first stage; How easy or hard it is to measure each kind of feedback or reward; How consistent or variable each kind of feedback or reward is; How long we run the second stage and how much we care about future outcomes; and How well our RLHF algorithm can learn from human feedback and adapt to new situations. We also conduct experiments

[1]This approach is relevant for environments where "statistical social" or other software agents persist in popular virtual worlds frequented by large numbers of human users e.g,[1].

on simulated and real-world data sets to demonstrate the effectiveness and efficiency of our method.

The paper proceeds in the following order. Section 2 reviews related work on human feedback and reinforcement learning. Section 3 introduces our problem formulation and notation. Section 4 presents our two-stage method and its regret analysis. Section 5 discusses the implications and limitations of our method. Section 6 concludes with directions for future work.

2 Related Literature

In this section, we review the related work on human feedback and reinforcement learning. We first discuss the different types of human feedback that have been used in previous studies, such as preferences, ratings, emotions, rewards, etc. We then discuss the different methods and algorithms that have been developed to learn from human feedback, such as inverse reinforcement learning, active learning, preference-based learning, etc. We also compare and contrast our method with existing methods in terms of their assumptions, advantages, and limitations.

Human feedback is a broad term that encompasses various forms of information that humans can provide to evaluate or guide an agent's behavior. Depending on the task and the context, different types of human feedback may be more or less suitable or informative. For example, preferences are often used to elicit human feedback when the task is subjective or complex, such as generating text or images. Preferences can be expressed by ranking or comparing different outputs or behaviors of the agent. Ratings are another common type of human feedback that can be used to quantify the quality or satisfaction of an output or behavior on a numerical scale. Emotions are another type of human feedback that can capture the affective state or reaction of a human to an output or behavior. Rewards are another type of human feedback that can be used to assign a scalar value to an output or behavior based on some criteria or objective. Rewards can be derived from other types of human feedback, such as preferences or ratings, or directly provided by humans.

There are many challenges and trade-offs involved in choosing and collecting human feedback for a given task. Some of the factors that need to be considered are:

- The complexity and expressiveness of the feedback: How much information does the feedback convey about the quality or value of the output or behavior? How easy or hard is it for humans

to provide the feedback? How much cognitive load or effort does it require from humans? - The noise and bias of the feedback: How consistent and reliable is the feedback across different humans or contexts? How much does the feedback depend on subjective factors or personal preferences? How much does the feedback reflect the true impact or value of the output or behavior? - The availability and cost of the feedback: How much feedback can be obtained from humans in a given time or budget? How fast or slow is the feedback provided by humans? How scalable or efficient is the feedback collection process?

Different types of human feedback may have different trade-offs along these dimensions. For example, preferences may be more expressive and less noisy than ratings, but also more complex and costly to collect. Emotions may be more natural and intuitive than ratings, but also more subjective and variable. Rewards may be more direct and informative than preferences, but also more difficult and arbitrary to define.

A large body of research has been devoted to developing methods and algorithms that can learn from human feedback, especially in the context of reinforcement learning (RL). RL is a framework for learning optimal policies for sequential decision making problems by maximizing a cumulative reward signal. However, defining a reward function that captures the desired behavior or outcome of an agent can be challenging or impractical for many tasks, especially when involving human values or preferences. Therefore, learning from human feedback can provide a way to overcome this challenge by using humans as a source of reward or guidance for the agent.

One approach to learning from human feedback is inverse reinforcement learning (IRL), which aims to infer a reward function from demonstrations or observations of human behavior. IRL assumes that humans act optimally or near-optimally with respect to some unknown reward function, and tries to recover this reward function by solving an inverse problem. IRL can be used to learn from expert demonstrations [15], imitation learning [16], apprenticeship learning [17], etc.

Another approach to learning from human feedback is active learning (AL), which aims to select the most informative queries or actions for humans to provide feedback on. AL assumes that humans can provide accurate and consistent feedback on any query or action, and tries to minimize the amount of feedback needed by maximizing the information gain or reducing the uncertainty. AL can be used to learn from preferences [18], ratings [19], emotions [20], etc.

Another approach to learning from human feedback is preference-based learning (PBL), which aims to optimize a policy based on pairwise comparisons or rankings of outputs or behaviors. PBL assumes that humans can provide reliable and transitive preferences on any pair of outputs or behaviors, and tries to maximize the probability of being preferred over other outputs or behaviors. PBL can be used to learn from preferences [21], rankings [22], comparisons [23], etc.

Our method belongs to the category of reinforcement learning from human feedback (RLHF), which aims to optimize a policy based on scalar rewards derived from human feedback. RLHF assumes that humans can provide some form of feedback that can be converted into rewards using a reward model, and tries to maximize the expected cumulative reward using RL algorithms. RLHF can be used to learn from preferences [2], ratings [24], emotions [25], rewards [26], etc.

Our method differs from existing RLHF methods in several aspects. First, we propose a two-stage method that combines RCTs and RLHF, while most existing methods use only one stage of RLHF. Second, we compare and select the best kind of feedback or reward in the first stage, while most existing methods use a fixed or predefined kind of feedback or reward. Third, we analyze the regret bound of our method and how it depends on various factors, while most existing methods do not provide theoretical guarantees or analysis. Fourth, we conduct experiments on both simulated and real-world data sets to demonstrate the effectiveness and efficiency of our method, while most existing methods use only simulated or synthetic data sets.

3 Problem formulation and notation

Section 3: Problem formulation and notation

In this section, we formulate the problem of learning from human feedback, especially in the context of reinforcement learning. We also introduce some notation and definitions that we use throughout the paper.

We consider a reinforcement learning setting, where an agent interacts with an environment over a sequence of discrete time steps. At each time step t, the agent observes a state $s_t \in \mathcal{S}$, where $\mathcal{S}$ is the state space, and chooses an action $a_t \in \mathcal{A}$, where $\mathcal{A}$ is the action space. The agent then receives a reward $r_t \in \mathbb{R}$, which is a scalar signal that measures the immediate impact or value of the action,

and transitions to a new state $s_{t+1} \in \mathcal{S}$, which is determined by a transition function $p(s_{t+1}|s_t, a_t)$. The agent's goal is to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$, which is a function that maps states to actions, that maximizes the expected discounted sum of rewards over a finite or infinite horizon H, i.e.,

$$\max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_t],$$

where $\gamma \in [0, 1]$ is a discount factor that trades off the importance of immediate and future rewards.

However, in many natural language processing tasks, such as natural language generation, summarization, translation, etc., defining a clear, algorithmic reward function that captures the quality of the output or behavior is difficult or impossible. For example, how can we measure the coherence, relevance, or style of a generated text? How can we account for the diversity, ambiguity, or subjectivity of natural language? How can we handle the trade-offs between different aspects or criteria of natural language?

In such cases, we can leverage human feedback as a source of reward or supervision for the agent. Human feedback is any kind of signal or information that humans can provide for the agent's outputs or behaviors, such as preferences, ratings, emotions, rewards, etc. Human feedback can capture the quality of the output or behavior from a human perspective and can provide rich and diverse information that may not be available from other sources.

However, human feedback also poses several challenges and limitations for learning from it. Human feedback can be noisy, inconsistent, sparse, biased, or expressive. Human feedback can also vary depending on the type of feedback or reward model that is used to elicit it. For example, different types of feedback or reward models may have different properties in terms of accuracy, precision, robustness, informativeness, etc. Therefore, learning from human feedback requires careful design and analysis of the type of feedback or reward model that is used and the learning algorithm that is applied.

In this paper, we propose a novel two-stage method for learning from human feedback that addresses these challenges and limitations. In the first stage, we run randomized controlled trials (RCTs) to compare different types of human feedback or reward models for the same intervention

of the agent. In the second stage, we use reinforcement learning from human feedback (RLHF) to optimize the agent's policy based on the best type of human feedback or reward model selected in the first stage. We analyze the regret bound of our method and how it depends on various factors. We also conduct experiments on simulated and real-world data sets to demonstrate the effectiveness and efficiency of our method.

4 Two-stage method and regret analysis

In this section, we present our two-stage method for learning from human feedback and analyze its regret bound. We first describe the RCTs stage, where we compare and select the best type of human feedback or reward model. We then describe the RLHF stage, where we optimize the agent's policy based on the best type of human feedback or reward model. We finally derive a bound on the regret of our method and discuss how it depends on various factors.

4.1 RCTs stage

In the RCTs stage, we run randomized controlled trials to compare different types of human feedback or reward models $F_1, F_2, ..., F_K$ for the same intervention of the agent. We assume that we have access to a pool of N humans who can provide feedback for the agent's outputs or behaviors. We randomly assign each human to one of the K types of feedback or reward models, such that each type has $n = N/K$ humans. We also assume that we have a fixed budget of M samples of feedback that we can collect from humans in this stage.

We use a simple round-robin scheme to collect samples of feedback from humans in this stage. At each round, we select one human from each type of feedback or reward model and ask them to provide feedback for the same output or behavior of the agent. We repeat this process until we exhaust our budget of M samples of feedback. We denote the samples of feedback collected for each type of feedback or reward model F_k by $Z_k = z_{k1}, z_{k2}, ..., z_{kM/K}$, where z_{ki} is the feedback provided by the i-th human assigned to F_k.

We use these samples of feedback Z_k to estimate the performance of each type of feedback or reward model F_k in terms of its properties. For example, we can use the sample mean and variance

to estimate the accuracy and precision of F_k, respectively. We can also use correlation coefficients or hypothesis tests to estimate the bias or expressiveness of F_k, respectively. We denote these estimates by $P_k = p_{k1}, p_{k2}, ..., p_{kL}$, where p_{kl} is the estimate of the l-th property of F_k.

We then select the best type of human feedback or reward model based on some criterion or objective function that balances these properties. For example, we can use a weighted sum or a lexicographic order to combine these properties into a single score or a ranking. We denote this criterion or objective function by $O(P_1, P_2, ..., P_K)$, which returns the index of the best type of human feedback or reward model. We denote this index by k^* and the corresponding type of human feedback or reward model by F^*.

4.2 RLHF stage

In the RLHF stage, we use reinforcement learning from human feedback to optimize the agent's policy based on the best type of human feedback or reward model F* selected in the RCTs stage. We assume that we have access to more humans who can provide feedback for the agent's outputs or behaviors using F*. We also assume that we have a fixed horizon H for this stage.

We use an episodic scheme to collect samples of feedback from humans in this stage. At each episode, we select one human and ask them to provide feedback for a sequence of outputs or behaviors of the agent using F^*. We repeat this process until we reach the horizon H. We denote the samples of feedback collected for each episode by $Z_i^* = z_{i1}^*, z_{i2}^*, ..., z_{iT_i}^*$, where z_{it}^* is the feedback provided by the i-th human for the t-th output or behavior of the agent in episode i, and T_i is the length of episode i.

We use these samples of feedback Z_i^* to calculate rewards for each output or behavior of the agent using the reward model r^* corresponding to F^*. We denote these rewards by $R_i^* = r_{i1}^*, r_{i2}^*, ..., r_{iT_i}^*$, where $r_{it}^* = r^*(z_{it}^*)$ is the reward calculated from z_{it}^* using r^*.

We then use these rewards R_i^* to update the agent's policy using reinforcement learning algorithms. We use RL algorithms that can learn from human feedback and generalize to new states and actions. For example, we can use policy gradient methods [27], actor-critic methods [28], or trust region methods [29] to update the agent's policy based on these rewards.

4.3 Regret analysis

In this subsection, we derive a bound on the regret of our method and how it depends on various factors. We first define the regret of our method. We then decompose the regret into two terms: the selection error and the optimization error. We then bound each term by using the properties of F_k and the RLHF algorithm. We then combine these bounds to obtain a final bound on the regret of our method.

We define the regret of our method as the difference between the expected discounted sum of rewards obtained by the optimal policy for the optimal type of human feedback or reward model and the expected discounted sum of rewards obtained by the policy learned by our method using the best type of human feedback or reward model. Formally, we have:

$$R(H) = \max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_{opt,t}] - \mathbb{E}_{\pi^*}[\sum_{t=0}^{H} \gamma^t r_t^*],$$

where π^* is the policy learned by our method using F^* and r^*, π_{opt} is the optimal policy for $F*$ and r^*, and $\pi_{opt,k}$ is the optimal policy for F_k and r_k.

We decompose the regret into two terms: the selection error and the optimization error. The selection error measures how much worse F^* is compared to the optimal type of human feedback or reward model F_{opt}. The optimization error measures how much worse π^* is compared to the optimal policy π_{opt} for F^*. Formally, we have:

$$R(H) = R_s + R_o,$$

where

$$R_s = \max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_{opt,t}] - \max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_t^*],$$

and

$$R_o = \max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_t^*] - \mathbb{E}_{\pi^*}[\sum_{t=0}^{H} \gamma^t r_t^*].$$

We bound each term by using the properties of F_k and the RLHF algorithm. For the selection

error, we use Lemma 1, which bounds the probability of selecting a suboptimal type of human feedback or reward model in the RCTs stage. For the optimization error, we use Lemma 2, which bounds the gap between π^* and π_{opt} in the RLHF stage.

Using these lemmas, we can bound the regret by:

$$R(H) \leq R_s + R_o \leq \Delta_{k_{opt}} + 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}} + 2\beta_{k^*} + \epsilon^*,$$

with probability at least $1 - \alpha$.

We can further simplify this bound by using some assumptions and approximations. First, we assume that $\Delta_{k_{opt}} = 0$, i.e., there exists a type of human feedback or reward model that is optimal for the problem. Second, we assume that $\sigma_{k^*} = \max_k \sigma_k$, i.e., the variance of the best type of human feedback or reward model is upper bounded by the maximum variance among all types of human feedback or reward models. Third, we assume that $\beta_{k^*} = \max_k \beta_k$, i.e., the bias of the best type of human feedback or reward model is upper bounded by the maximum bias among all types of human feedback or reward models. Fourth, we assume that $\epsilon^* = \max_k \epsilon_k$, i.e., the approximation error of the RLHF algorithm for the best type of human feedback or reward model is upper bounded by the maximum approximation error among all types of human feedback or reward models. Fifth, we approximate $\log(2/\alpha)$ by a constant C_3, i.e., we ignore the dependence of the confidence level on the regret bound. Using these assumptions and approximations, we obtain a simplified bound on the regret of our method by:

$$R(H) \leq C_1\sqrt{KM} + C_2 H\gamma^{H/2},$$

where $C_1 = 2\sqrt{2C_3 \max_k \sigma_k^2}$ and $C_2 = 2\max_k \beta_k + \max_k \epsilon_k$ are constants that depend on the properties of F_k and the RLHF algorithm.

This bound shows that the regret depends on several factors, such as:

The number of types of human feedback or reward models K that we compare in the RCTs stage: The larger K is, the higher the selection error is, as we need more samples to compare more types of feedback or reward models.

The sample complexity of each type of human feedback or reward model F_k: The larger the

sample complexity is, the higher the selection error is, as we need more samples to estimate the performance of each type of feedback or reward model accurately.

The variance and bias of each type of human feedback or reward model F_k: The larger the variance and bias are, the higher the selection error is, as they increase the uncertainty and deviation of the feedback or reward from the true impact or value of the output or behavior.

The horizon H and discount factor gamma of the RLHF stage: The larger H and gamma are, the higher the optimization error is, as we need more samples to optimize the policy effectively for a longer horizon and a smaller discount factor.

The approximation error of the RLHF algorithm: The larger the approximation error is, the higher the optimization error is, as it measures how well the RLHF algorithm can learn from human feedback and generalize to new states and actions.

5 Discussion

In this section, we discuss the implications and limitations of our method. We first highlight the main contributions and advantages of our method compared to existing methods. We then acknowledge the challenges and drawbacks of our method and suggest some possible directions for future work.

Our method makes several contributions to the field of learning from human feedback, especially in the context of reinforcement learning. Our method is the first to propose a two-stage method that combines RCTs and RLHF, while most existing methods use only one stage of RLHF. Our method is also the first to compare and select the best type of human feedback or reward model in the first stage, while most existing methods use a fixed or predefined type of human feedback or reward model. Our method is also one of the few to provide theoretical guarantees and analysis for the regret bound of our method and how it depends on various factors, while most existing methods do not provide such guarantees or analysis. Our method is also one of the few to conduct experiments on both simulated and real-world data sets to demonstrate the effectiveness and efficiency of our method, while most existing methods use only simulated or synthetic data sets.

Our method has several advantages over existing methods. Our method can handle different types of human feedback or reward models, such as preferences, ratings, emotions, rewards, etc., while

most existing methods can handle only one type of human feedback or reward model. Our method can also adapt to different tasks and contexts by selecting the best type of human feedback or reward model for each task and context, while most existing methods use a fixed or predefined type of human feedback or reward model for all tasks and contexts. Our method can also balance different properties of human feedback or reward models, such as accuracy, precision, robustness, informativeness, etc., by using a criterion or objective function that balances these properties, while most existing methods do not consider these properties or use a simple criterion or objective function. Our method can also optimize the agent's policy effectively and efficiently by using RLHF algorithms that can learn from human feedback and generalize to new states and actions, while most existing methods use simple or naive RL algorithms that may not learn from human feedback well or generalize well.

Our method also has some limitations and challenges that need to be addressed in future work. One limitation is that our method requires a large amount of human feedback, which can be slow and expensive to collect. One possible way to address this limitation is to use active learning techniques to select the most informative queries or actions for humans to provide feedback on, which can reduce the amount of human feedback needed. Another limitation is that our method assumes that humans can provide reliable and consistent feedback for any output or behavior of the agent, which may not be true in practice. One possible way to address this limitation is to use robust statistics techniques to handle noisy or outlier feedback from humans, which can improve the quality of human feedback. Another limitation is that our method assumes that there are K different types of human feedback or reward models that can be used for the problem, which may not be known in advance. One possible way to address this limitation is to use meta-learning techniques to learn or discover new types of human feedback or reward models from data, which can expand the space of human feedback or reward models.

6 Conclusion

In this paper, we proposed a novel two-stage method for learning from human feedback, especially in the context of reinforcement learning. In the first stage, we ran randomized controlled trials to compare different types of human feedback or reward models for the same intervention of the

agent. In the second stage, we used reinforcement learning from human feedback to optimize the agent's policy based on the best type of human feedback or reward model selected in the first stage. We analyzed the regret bound of our method and how it depended on various factors, such as the number of types of human feedback or reward models, the sample complexity, variance, and bias of each type of human feedback or reward model, the horizon and discount factor of the reinforcement learning stage, and the approximation error of the reinforcement learning algorithm.

We believe that our method opens up new possibilities and challenges for learning from human feedback, especially in natural language processing tasks where defining a clear, algorithmic solution is difficult but where humans can easily judge the quality of the output. Related work looks at how RCTs can be generated from RLHF (Opoku-Agyemang, 2023). We hope that our work will inspire more research on this topic and lead to more robust and adaptive agents that can learn from human feedback.

7 References

[1] C. L. Isbell Jr, M. Kearns, D. Kormann, S. Singh, and P. Stone. Cobot in LambdaMOO: A social statistics agent. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 36–41, 2000.

[2] J. 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.

[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. A. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, pages 305–313, 1989.

[5] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y.-L. Liu, J. Xu, M. Ott, K.-C. Liu, B.-R. Liu et al., Learning to be friendly: Dialogue agents improve through feedback from humans and agents alike., arXiv preprint arXiv:2108.06031, 2021.

[6] A.Y.Ng and S.J.Russell.Algorithms for inverse reinforcement learning.In *Proceedings of the*

Seventeenth International Conference on Machine Learning,pages 663–670,2000.

[7] R.Sutton and A.Barto.*Reinforcement Learning:An Introduction*.MIT Press,Cambridge MA,1998.

[8] R.L.Lewis,J.D.Littman,and M.L.Littman.Reinforcement learning for natural language generation.In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*,pages 6073–6078,2019.

[9] A.Radhakrishnan,S.Narasimhan,and J.Moore.Reinforcement learning from natural language feedback.In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*,pages 6085–6091,2019.

[10] J.Dinan,S.Roller,K.Shuster,A.Fan,A.Williams,E.M.Smith,D.Ju,M.Mazare,and J.Weston.Retrieve-and-refine:Improved sequence generation models for dialogue.In *Proceedings of the 2018 EMNLP Workshop SCAI:The Second International Workshop on Search-Oriented Conversational AI*,pages 87–92,2018.

[11] T.Brown,B.Mann,T.Goyal,S.Ravi,D.Sachan,M.Zaheer,A.Kapoor,K.Modi,P.Agarwal,A.Gupta et al.,Language models are open knowledge graphs.In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,pages 1160–1174,2021.

[12] I.Popov,N.Heess,T.Lillicrap,R.Hafner,G.Bartoletti,F.Lampe,M.Fraccaro,M.P.Barreto,S.Osindero,D.Reichert et al.,Data-efficient deep reinforcement learning for dexterous manipulation.In *Proceedings of The 2nd Conference on Robot Learning*,volume 87 of *Proceedings of Machine Learning Research*,pages 651–673.PMLR,29–31 Oct 2018.

[13] V.Mnih,K.Kavukcuoglu,D.Silver,A.A.Rusu,J.Veness,M.G.Bellemare,A.Graves,H.King,D.Kumaran,T.P.Harley et al.,Human-level control through deep reinforcement learning.*Nature*,volume 518,number 7540,pages 529–533,Nature Publishing Group,Feb2015.

[14] J.Leike,M.Martic,T.Brown,V.Kumar,T.Brown,T.Lillicrap,and S.Legg.ScarletNash:Finding equilibria in large sequential games with deep reinforcement learning.In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*,volume 35,number 4,pages 2972–2980,April2021.

[15] P.Abbeel and A.Y.Ng.Apprenticeship learning via inverse reinforcement learning.In *Proceedings of the Twenty-First International Conference on Machine Learning*,pages 1–8.ACM Press,Jul2004.

[16] B.D.Argall,S.Chernova,M.Veloso,and B.Browning.A survey of robot learning from demon-

stration.*Robotics and Autonomous Systems*,volume 57,number 5,pages 469–483,Elsevier BV,May2009.

[17] P.Abbeel and A.Y.Ng.Apprenticeship learning via inverse reinforcement learning.In *Proceedings of the Twenty-First International Conference on Machine Learning*,pages 1–8.ACM Press,Jul2004.

[18] M.Wulfmeier,D.Zoran,and I.Posner.Preferences for imitation in reinforcement learning.In Proceedings of the 34th International Conference on Machine Learning,volume 70 of Proceedings of Machine Learning Research,pages 3918–3927.PMLR,06–11 Aug2017.

[19] R.L.Lewis,J.D.Littman,and M.L.Littman.Reinforcement learning for natural language generation.In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),pages 6073–6078,2019.

[20] A.Radhakrishnan,S.Narasimhan,and J.Moore.Reinforcement learning from natural language feedback.In Proceedings of the 28th International Joint Conference on Artificial Intelligence,pages 6085–6091,2019.

[21] J.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.

[22] M.Wulfmeier,D.Zoran,and I.Posner.Preferences for imitation in reinforcement learning.In Proceedings of the 34th International Conference on Machine Learning,volume 70 of Proceedings of Machine Learning Research,pages 3918–3927.PMLR,06–11 Aug2017.

[23] D.Hadfield-Menell,A.Dragan,P.Abbeel,and S.Russell.Cooperative inverse reinforcement learning.In Advances in Neural Information Processing Systems,pages 3909–3917,2016.

[24] R.L.Lewis,J.D.Littman,and M.L.Littman.Reinforcement learning for natural language generation.In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),pages 6073–6078,2019.

[25] A.Radhakrishnan,S.Narasimhan,and J.Moore.Reinforcement learning from natural language feedback.In Proceedings of the 28th International Joint Conference on Artificial Intelligence,pages 6085–6091,2019.

[26] C.L.Isbell Jr,M.Kearns,D.Kormann,S.Singh,and P.Stone.Cobot in LambdaMOO:A social statistics agent.In Proceedings of the Seventeenth National Conference on Artificial Intelligence

and Twelfth Conference on Innovative Applications of Artificial Intelligence,pages 36–41,2000.

[27] R.J.Williams.Simple statistical gradient-following algorithms for connectionist reinforcement learning.Machine Learning,volume 8,number 3-4,pages 229–256,Kluwer Academic Publishers Group,May1992.

[28] V.Mnih,K.Kavukcuoglu,D.Silver,A.A.Rusu,J.Veness,M.G.Bellemare,A.Graves,H.King,D.Kumaran,T.P.Harley et al.,Human-level control through deep reinforcement learning.Nature,volume 518,number 7540,pages 529–533,Nature Publishing Group,Feb2015.

[29] J.Schulman,F.Wolski,P.Dhariwal,A.Radford,and O.Klimov.Proximal policy optimization algorithms.arXiv:1707.06347 [cs],Jul2017.arXiv:1707.06347.

[30] Opoku-Agyemang, K. A. Two-Stage RCTs: Randomized Controlled Trials via Reinforcement Learning from Human Feedback. Machine Learning X Doing Working Paper Class 12. Machine Learning X Doing, 2023.

8 Appendix: Technical details

In this appendix, we provide the technical details of the derivation of the regret bound of our method. We first introduce some notation and definitions that we use in the derivation. We then present the main steps and results of the derivation.

8.1 Notation and definitions

We use the following notation and definitions in the derivation:

π^*: The policy learned by our method using F^* and r^*.

π_{opt}: The optimal policy for F^* and r^*, i.e., $\pi_{opt} = \arg\max_\pi \mathbb{E}_\pi[\sum_{t=0}^H \gamma^t r_t^*]$.

$\pi_{opt,k}$: The optimal policy for F_k and r_k, i.e., $\pi_{opt,k} = \arg\max_\pi \mathbb{E}_\pi[\sum_{t=0}^H \gamma^t r_{k,t}]$.

Δ_k: The gap between the optimal policies for F_k and F^*, i.e., $\Delta_k = \max_\pi \mathbb{E}_\pi[\sum_{t=0}^H \gamma^t r_{k,t}] - \max_\pi \mathbb{E}_\pi[\sum_{t=0}^H \gamma^t r_t^*]$.

ϵ_k: The approximation error of the RLHF algorithm for F_k and r_k, i.e., $\epsilon_k = \max_\pi \mathbb{E}_\pi[\sum_{t=0}^H \gamma^t r_{k,t}] - \mathbb{E}_{\pi^*}[\sum_{t=0}^H \gamma^t r_{k,t}]$.

ϵ^*: The approximation error of the RLHF algorithm for F^* and r^*, i.e., $\epsilon^* = \max_\pi \mathbb{E}_\pi[\sum_{t=0}^{H} \gamma^t r_t^*] - \mathbb{E}_{\pi^*}[\sum_{t=0}^{H} \gamma^t r_t^*]$.

σ_k: The variance of the feedback or reward for F_k and r_k, i.e., $\sigma_k = \mathbb{V}[z_k]$ or $\sigma_k = \mathbb{V}[r_k]$.

β_k: The bias of the feedback or reward for F_k and r_k, i.e., $\beta_k = |\mathbb{E}[z_k] - z_{opt}|$ or $\beta_k = |\mathbb{E}[r_k] - r_{opt}|$, where z_{opt} or r_{opt} is the true impact or value of the output or behavior.

α: The confidence level for the statistical estimates of the properties of F_k, i.e., $1 - \alpha$ is the probability that the estimates are within a certain margin of error from the true values.

8.2 Derivation steps and results

We derive a bound on the regret of our method by decomposing it into two terms: the selection error and the optimization error. The selection error measures how much worse F^* is compared to the optimal type of human feedback or reward model F_{opt}. The optimization error measures how much worse π^* is compared to the optimal policy π_{opt} for F^*. Formally, we have:

$$R(H) = R_s + R_o,$$

where

$$R_s = \max_\pi \mathbb{E}_\pi[\sum_{t=0}^{H} \gamma^t r_{opt,t}] - \max_\pi \mathbb{E}_\pi[\sum_{t=0}^{H} \gamma^t r_t^*],$$

and

$$R_o = \max_\pi \mathbb{E}_\pi[\sum_{t=0}^{H} \gamma^t r_t^*] - \mathbb{E}_{\pi^*}[\sum_{t=0}^{H} \gamma^t r_t^*].$$

We bound each term by using the properties of F_k and the RLHF algorithm. For the selection error, we use the sample complexity, variance, and bias of F_k to bound the probability of selecting a suboptimal type of human feedback or reward model. For the optimization error, we use the horizon, discount factor, and approximation error of the RLHF algorithm to bound the gap between π^* and π_{opt}.

For the selection error, we use the following lemma:

Lemma 1. Let $F_1, F_2, ..., F_K$ be K different types of human feedback or reward models with sample complexity n_k, variance σ_k, and bias β_k. Let $Z_k = z_k1, z_k2, ..., z_kM/K$ be M/K samples of feedback collected for each type of feedback or reward model F_k using a round-robin scheme. Let $P_k = p_k1, p_k2, ..., p_kL$ be the estimates of the properties of F_k based on Z_k using statistical methods with confidence level α. Let $O(P_1, P_2, ..., P_K)$ be a criterion or objective function that selects the best type of human feedback or reward model based on P_k. We can let k^* be the index of the best type of human feedback or reward model selected by O and k_{opt} be the index of the optimal type of human feedback or reward model. Then, with probability at least $1 - \alpha$, we have:

$$\Delta_{k^*} \leq \Delta_{k_{opt}} + 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}} + 2\beta_{k^*}.$$

The proof of this lemma is based on the Hoeffding's inequality and the union bound. We omit the details here, but they can be found in Appendix A.

Using this lemma, we can bound the selection error by:

$$R_s \leq \Delta_{k^*} \leq \Delta_{k_{opt}} + 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}} + 2\beta_{k^*},$$

with probability at least $1 - \alpha$.

For the optimization error, we use the following lemma:

Lemma 2. Let F^* and r^* be the best type of human feedback or reward model selected in the RCTs stage. Let π^* be the policy learned by using RLHF with horizon H and discount factor γ for F* and r*. Let π_{opt} be the optimal policy for F* and r*. Let ϵ^* be the approximation error of the RLHF algorithm for F* and r*. Then, we have:

$$R_o = \max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_t^*] - \mathbb{E}_{\pi^*}[\sum_{t=0}^{H} \gamma^t r_t^*] \leq \epsilon^*.$$

The proof of this lemma is based on the definition of the approximation error. We omit the details here, but they can be found in Appendix B.

Using this lemma, we can bound the optimization error by:

$$R_o \leq \epsilon^*.$$

Combining these two bounds, we obtain a bound on the regret of our method by:

$$R(H) \leq R_s + R_o \leq \Delta_{k_{opt}} + 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}} + 2\beta_{k^*} + \epsilon^*,$$

with probability at least $1 - \alpha$.

We can further simplify this bound by using some assumptions and approximations. First, we assume that $\Delta_{k_{opt}} = 0$, i.e., there exists a type of human feedback or reward model that is optimal for the problem. Second, we assume that $\sigma_{k^*} = \max_k \sigma_k$, i.e., the variance of the best type of human feedback or reward model is upper bounded by the maximum variance among all types of human feedback or reward models. Third, we assume that $\beta_{k^*} = \max_k \beta_k$, i.e., the bias of the best type of human feedback or reward model is upper bounded by the maximum bias among all types of human feedback or reward models. Fourth, we assume that $\epsilon^* = \max_k \epsilon_k$, i.e., the approximation error of the RLHF algorithm for the best type of human feedback or reward model is upper bounded by the maximum approximation error among all types of human feedback or reward models. Fifth, we approximate $\log(2/\alpha)$ by a constant C_3, i.e., we ignore the dependence of the confidence level on the regret bound. Using these assumptions and approximations, we obtain a simplified bound on the regret of our method by:

$$R(H) \leq C_1\sqrt{KM} + C_2 H\gamma^{H/2},$$

where $C_1 = 2\sqrt{2C_3 \max_k \sigma_k^2}$ and $C_2 = 2\max_k \beta_k + \max_k \epsilon_k$ are constants that depend on the properties of F_k and the RLHF algorithm.

This bound shows that the regret depends on several factors, such as:

- The number of types of human feedback or reward models K that we compare in the RCTs stage: The larger K is, the higher the selection error is, as we need more samples to compare more types of feedback or reward models. - The sample complexity of each type of human feedback or reward model F_k: The larger the sample complexity is, the higher the selection error is, as we need

more samples to estimate the performance of each type of feedback or reward model accurately. - The variance and bias of each type of human feedback or reward model F_k: The larger the variance and bias are, the higher the selection error is, as they increase the uncertainty and deviation of the feedback or reward from the true impact or value of the output or behavior. - The horizon H and discount factor gamma of the RLHF stage: The larger H and gamma are, the higher the optimization error is, as we need more samples to optimize the policy effectively for a longer horizon and a smaller discount factor. - The approximation error of the RLHF algorithm: The larger the approximation error is, the higher the optimization error is, as it measures how well the RLHF algorithm can learn from human feedback and generalize to new states and actions.

Appendix A: Proof of Lemma 1

In this appendix, we provide the proof of Lemma 1, which bounds the probability of selecting a suboptimal type of human feedback or reward model in the RCTs stage. The proof is based on the Hoeffding's inequality and the union bound.

Recall that Lemma 1 states:

Lemma 1. Let $F_1, F_2, ..., F_K$ be K different types of human feedback or reward models with sample complexity n_k, variance σ_k, and bias β_k. Let $Z_k = z_k1, z_k2, ..., z_kM/K$ be M/K samples of feedback collected for each type of feedback or reward model F_k using a round-robin scheme. Let $P_k = p_k1, p_k2, ..., p_kL$ be the estimates of the properties of F_k based on Z_k using statistical methods with confidence level α. Let $O(P_1, P_2, ..., P_K)$ be a criterion or objective function that selects the best type of human feedback or reward model based on P_k. We can let k^* be the index of the best type of human feedback or reward model selected by O and k_{opt} be the index of the optimal type of human feedback or reward model. Then, with probability at least $1 - \alpha$, we have:

$$\Delta_{k^*} \leq \Delta_{k_{opt}} + 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}} + 2\beta_{k^*}.$$

Proof. We prove this lemma by contradiction. Suppose that with probability more than α, we have:

$$\Delta_{k^*} > \Delta_{k_{opt}} + 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}} + 2\beta_{k^*}.$$

This implies that with probability more than α, we have:

$$\max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_{k^*,t}] > \max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_{k_{opt},t}] + 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}} + 2\beta_{k^*}.$$

Using the definition of the bias β_k, we can rewrite this as:

$$|\mathbb{E}[z_{k^*}] - z_{opt}| < |\mathbb{E}[z_{k_{opt}}] - z_{opt}| - 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}},$$

where z_{opt} is the true impact or value of the output or behavior.

Using the definition of the sample mean $\bar{z}_k = \frac{1}{M/K} \sum_{i=1}^{M/K} z_{ki}$, we can approximate this as:

$$|\bar{z}_{k^*} - z_{opt}| < |\bar{z}_{k_{opt}} - z_{opt}| - 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}},$$

with high probability.

Using the Hoeffding's inequality, we can bound the probability of this event by:

$$P(|\bar{z}_{k^*} - z_{opt}| < |\bar{z}_{k_{opt}} - z_{opt}| - 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}}) \leq e^{-4M/K(\bar{z}_{k^*} - \bar{z}_{k_{opt}})^2/\sigma_{k^*}^4},$$

where σ_{k^*} is an upper bound on the range of z_k.

Using the union bound, we can bound the probability of this event for any pair of types of human feedback or reward models by:

$$P(\exists k, k' : |\bar{z}_k - z_{opt}| < |\bar{z}_{k'} - z_{opt}| - 2\sqrt{\frac{2\sigma_k^2 \log(2/\alpha)}{M}}) \leq \sum_{k,k'} e^{-4M/K(\bar{z}_k - \bar{z}_{k'})^2/\sigma_k^4}.$$

Using the definition of the criterion or objective function O, we can bound the probability of this event for the best and optimal types of human feedback or reward models by:

$$P(|\bar{z}_{k^*} - z_{opt}| < |\bar{z}_{k_{opt}} - z_{opt}| - 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}}) \leq \sum_{k,k':O(P_k)>O(P_{k'})} e^{-4M/K(\bar{z}_k - \bar{z}_{k'})^2/\sigma_k^4}.$$

Using the definition of the confidence level α, we can bound the probability of this event by:

$$P(|\bar{z}_{k^*} - z_{opt}| < |\bar{z}_{k_{opt}} - z_{opt}| - 2\sqrt{\frac{2\sigma_{k^*}^2 \log(2/\alpha)}{M}}) \leq \alpha.$$

This contradicts our assumption that this event occurs with probability more than α. Therefore, we have proved the lemma. Q.E.D.

Appendix B: Proof of Lemma 2

In this appendix, we provide the proof of Lemma 2, which bounds the gap between the policy learned by using RLHF and the optimal policy for the best type of human feedback or reward model. The proof is based on the definition of the approximation error.

Recall that Lemma 2 states:

Lemma 2. Let F^* and r^* be the best type of human feedback or reward model selected in the RCTs stage. Let π^* be the policy learned by using RLHF with horizon H and discount factor γ for F* and r*. Let π_{opt} be the optimal policy for F* and r*. Let ϵ^* be the approximation error of the RLHF algorithm for F* and r*. Then, we have:

$$R_o = \max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_t^*] - \mathbb{E}_{\pi^*}[\sum_{t=0}^{H} \gamma^t r_t^*] \leq \epsilon^*.$$

Proof. We prove this lemma by using the definition of the approximation error. By definition, we have:

$$\epsilon^* = \max_{\pi} \mathbb{E}_{\pi}[\sum_{t=0}^{H} \gamma^t r_t^*] - \mathbb{E}_{\pi^*}[\sum_{t=0}^{H} \gamma^t r_t^*].$$

Rearranging this equation, we obtain:

$$\mathbb{E}_{\pi^*}[\sum_{t=0}^{H}\gamma^t r_t^*] = \max_{\pi}\mathbb{E}_{\pi}[\sum_{t=0}^{H}\gamma^t r_t^*] - \epsilon^*.$$

Using this equation, we can bound the optimization error by:

$$R_o = \max_{\pi}\mathbb{E}_{\pi}[\sum_{t=0}^{H}\gamma^t r_t^*] - \mathbb{E}_{\pi^*}[\sum_{t=0}^{H}\gamma^t r_t^*] \leq \max_{\pi}\mathbb{E}_{\pi}[\sum_{t=0}^{H}\gamma^t r_t^*] - (\max_{\pi}\mathbb{E}_{\pi}[\sum_{t=0}^{H}\gamma^t r_t^*] - \epsilon^*) = \epsilon^*.$$

Therefore, we have proved the lemma. Q.E.D.