Provably Efficient Interaction-Grounded Learning with Personalized Reward

Mengxiao Zhang *
University of Iowa
mengxiao-zhang@uiowa.edu

Yuheng Zhang *
University of Illinois Urbana-Champaign
yuhengz2@illinois.edu

Haipeng Luo
University of Southern California
haipengl@usc.edu

Paul Mineiro
Microsoft Research
pmineiro@microsoft.com

Abstract

Interaction-Grounded Learning (IGL) [Xie et al., 2021] is a powerful framework in which a learner aims at maximizing unobservable rewards through interacting with an environment and observing reward-dependent feedback on the taken actions. To deal with personalized rewards that are ubiquitous in applications such as recommendation systems, Maghakian et al. [2022] study a version of IGL with context-dependent feedback, but their algorithm does not come with theoretical guarantees. In this work, we consider the same problem and provide the first provably efficient algorithms with sublinear regret under realizability. Our analysis reveals that the step-function estimator of prior work can deviate uncontrollably due to finite-sample effects. Our solution is a novel Lipschitz reward estimator which underestimates the true reward and enjoys favorable generalization performances. Building on this estimator, we propose two algorithms, one based on explore-then-exploit and the other based on inverse-gap weighting. We apply IGL to learning from image feedback and learning from text feedback, which are reward-free settings that arise in practice. Experimental results showcase the importance of using our Lipschitz reward estimator and the overall effectiveness of our algorithms.

1 Introduction

Traditional bandit problems [Auer et al., 2002] or reinforcement learning problems [Sutton and Barto, 2018] assume that the learner has access to the reward, which is her learning goal. However, such explicit reward information is usually difficult to obtain in many real-world scenarios, including human-computer interface applications [Pantic and Rothkrantz, 2003, Freeman et al., 2017] and recommender systems [Yi et al., 2014, Wu et al., 2017]. Recently, Xie et al. [2021] study a new setting called Interaction-Grounded Learning (IGL), where the learner interacts with the environment and receives some feedback on her actions instead of explicit reward signals. The learner needs to discover the implicit information about the reward in the feedback and maximizes the reward.

From an information-theoretic perspective, IGL is intractable unless further assumptions are made. Xie et al. [2021] introduce a conditional independence assumption which states that the feedback is conditionally independent of the action and context given the latent reward. However, this assumption is unrealistic in many scenarios. For example, different users interact with recommender systems in different ways even under the same latent reward [Maghakian et al., 2022]. This inspires us to study IGL with personalized rewards, a setting where the feedback depends on the context. Although

*Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3847

Maghakian et al. [2022] study this for recommender systems, they only provide empirical results of their approach, leaving the following question open:

Can we design provably efficient algorithms for interaction-grounded learning when the feedback depends on the context given the latent reward?

Contributions. In this paper, we answer the question in the positive and provide the first provably efficient algorithms with sublinear regret guarantee for IGL with personalized reward. To achieve this, in Section 3.1, we first propose a novel reward estimator via inverse kinematics. Specifically, using the samples collected by applying a uniform policy, we construct a *Lipschitz reward*, which underestimates the reward for all the policies but approximates the reward for the optimal policy well. With this reward estimator, we propose two algorithms, one based on Explore-then-Exploit (Section 3.2) and the other based on inverse-gap weighting (Section 3.3). Both algorithms achieve $\mathcal{O}(T^{\frac{2}{3}})$ regret. To the best of our knowledge, we are the first to propose provable regret guarantees for IGL with personalized reward. In Section 4, we implement both algorithms and apply them to both an image classification dataset and a conversational dataset. The empirical performance showcases the effectiveness of our algorithm and the importance of using the Lipschitz reward estimator.

1.1 Related Work

Interaction-Grounded Learning (IGL). Xie et al. [2021] is the first work studying IGL and assumes that the feedback is independent of the context and action conditioned on the latent reward. Xie et al. [2022] further relaxes the assumption to an action-inclusive version, where the feedback is independent of context conditioned on both the action and the reward.[2] Hu et al. [2024] follows the same setting as Xie et al. [2021] but proposes an information-theoretic approach to enforce the conditional independence assumption. However, as we mentioned before, context-dependent feedback is ubiquitous in real-word applications, so such conditional independence assumptions reduce the generality of the IGL framework. Maghakian et al. [2022] is the closest work to ours which also considers personalized rewards. Nonetheless, their work focuses on the empirical side and does not provide any theoretical guarantees.

Contextual online learning with partial feedback. Our work is closely related to the recent trend of designing efficient contextual learning algorithms with partial feedback, including contextual bandits [Langford and Zhang, 2007, Agarwal et al., 2012, 2014, Foster and Krishnamurthy, 2018, Foster et al., 2018, Foster and Rakhlin, 2020, Simchi-Levi and Xu, 2021], where the learner receives explicit reward signal of her taken action; contextual bandits with feedback graphs [Zhang et al., 2024a,b], where the learner's observation of the reward is determined based on a feedback graph; and contextual partial monitoring [Bartók and Szepesvári, 2012, Kirschner et al., 2020], where the learner's observation is defined by a signal matrix or a linear observation operator. We adopt and generalize the ideas for designing contextual bandits algorithms [Foster and Rakhlin, 2020] to design algorithms for IGL.

2 Preliminary

Problem setup. Throughout the paper, we use $[N]$ to denote the set $\{1, 2, \ldots, N\}$ for a positive integer N. The problem of online Interative-Grounded Learning (IGL) with personalized reward we consider is defined as follows. The interaction between the learner and the environment lasts for T rounds. At each round $t \in [T]$, the environment reveals a stochastic context $x_t \in \mathcal{X}$ i.i.d. drawn from an unknown distribution $\mathcal{D}$, and the learner decides an action $a_t \in [K]$ from a finite action set of size K based on this context. Then, different from the classic contextual bandits in which the learner observes the binary reward of her chosen action $r(x_t, a_t) \in \{0, 1\}$, she receives feedback $y_t \in \mathcal{Y}$.

[2] We point out that the idea behind Xie et al. [2022] cannot be used in our setting where feedback can be dependent on the context. Specifically, Xie et al. [2022] propose an algorithm which learns the value function and the reward decoder separately for each action. Generalizing their idea to our setting would then mean learning the value function and the reward decoder for each context, which is infeasible since there could be infinitely many contexts.

Feedback dependence assumption. We follow the assumption proposed in [Maghakian et al., 2022] that the feedback is conditionally independent of the action given the context and the realized reward.

Assumption 1. *For arbitrary (x, a, r, y) tuple where the reward r and the feedback y are generated based on the context x and action a, we assume y is conditionally independent of action a given context x and reward r. In other words, we assume that $y \perp a \mid r, x$.*

As mentioned, compared to prior work [Xie et al., 2021, 2022], this better captures many real-world applications such as recommender systems where different users interact with them in different ways [Beel et al., 2013, Shin, 2020, Maghakian et al., 2022].

Realizability. We assume that the learner has access to two function classes $\mathcal{F} \subseteq (\mathcal{X} \times [K] \mapsto [0, 1])$ and $\Phi \subseteq (\mathcal{X} \times \mathcal{Y} \mapsto \{0, 1\})$, where $\mathcal{F}$ characterizes the mean of the reward for a given context-action pair, and Φ characterizes the realized reward given the context and the received feedback. A policy $\pi : \mathcal{X} \to \Delta(K)$ specifies the action probability conditioned on the context. For each $f \in \mathcal{F}$, we use π_f to denote the induced policy which takes action greedily according to f, that is, $\pi_f(a|x) = \mathbb{1}\{a = \operatorname{argmax}_{a' \in [K]} f(x, a')\}, \forall a \in [K]$. We also use the shorthand $\pi^\star$ for the optimal policy $\pi_{f^\star}$. We then make the following realizability assumption following previous contextual bandits literature [Agarwal et al., 2012, Foster et al., 2018, Foster and Rakhlin, 2020, Simchi-Levi and Xu, 2021].

Assumption 2 (Realizability)**.** *There exists a regression function $f^\star \in \mathcal{F}$ such that $\mathbb{E}[r(x_t, a)|x_t] = f^\star(x_t, a)$ for all $a \in [K]$ and $t \in [T]$. Furthermore, there exists a feedback decoder $\phi^\star \in \Phi$ such that $\phi^\star(x_t, y_t) = r(x_t, a_t)$ for all $t \in [T]$.*

For simplicity, we also assume that $\mathcal{F}$ and Φ are finite with cardinality $|\mathcal{F}|$ and $|\Phi|$. Our results can be further generalized to broader function classes, which will be discussed in later sections.

Identifiability. As mentioned in [Xie et al., 2021, 2022, Maghakian et al., 2022], it is impossible to learn if we do not break the symmetry between reward being 1 and being 0. Following [Maghakian et al., 2022], we make the following assumption:

Assumption 3 (Identifiability)**.** *For any $x \in \mathcal{X}$, $f^\star$ defined in Assumption 2 satisfies that: 1) $\sum_{a=1}^{K} f^\star(x, a) \leq \alpha$ for some $0 < \alpha < \frac{K}{2}$; 2) $\max_{a \in [K]} f^\star(x, a) \geq \theta$ where $\theta > \frac{\alpha}{K - \alpha}$.*

The first condition says that the sum of the expected reward over actions is less than $\frac{K}{2}$ given any context x. That is to say, the reward vector is sparse if $f^\star(x, a) \in \{0, 1\}$. The second condition says that for each context $x \in \mathcal{X}$, there exists an action that achieves a large enough expected reward. These two conditions are indeed satisfied by many real-world applications, including the s-multi-label classification problem (with $s < K/2$) where $\alpha = s$ and $\theta = 1$.

Regret. The learner's performance is measured via the notion of regret, which is defined as the expected difference between the learner's total reward and the one received by the optimal policy:

$$\mathbf{Reg}_{\mathsf{CB}} = \mathbb{E}\left[\sum_{t=1}^{T} f^\star(x_t, \pi^\star(x_t)) - \sum_{t=1}^{T} f^\star(x_t, a_t)\right],$$

Other notations. We denote the $(K-1)$-dimensional simplex by Δ_K. Let $\mathbf{1}$ be the all-one vector in an appropriate dimension and $\pi_{\text{Unif}} = \frac{1}{K} \cdot \mathbf{1} \in \Delta_K$. For a d-dimensional vector $v \in \mathbb{R}^d$, we denote its i-th entry by v_i. $\mathbb{1}\{\cdot\}$ is the indicator function and e_i is the i-th standard basis vector in an appropriate dimension.

3 Methodology

In this section, we discuss our methodology. In Section 3.1, we start from introducing how we construct a Lipschitz reward estimator based on uniform samples, which serves as the key for our algorithm construction. We prove that this estimator is an *underestimator* of the reward, and more importantly matches the reward for the optimal policy. Based on this estimator, we design two algorithms for this problem in Section 3.2 and Section 3.3, with the first one based on explore-then-exploit and the second one based on inverse-gap weighting (IGW).

https://doi.org/10.52202/079017-3847

3.1 Reward Estimator Construction via Uniform Exploration

We first show how we construct a reward estimator based on feedback collected from a uniform policy, which serves as the most important component in our two algorithms.

3.1.1 Inverse Kinematics

When the reward for each context-action pair is *deterministic* and binary, Maghakian et al. [2022] show that if the learner uniformly samples an action for any context and is able to accurately predict the posterior probability of her chosen action given the context and the feedback (inverse kinematics), then she is able to infer that the reward is 1 if that posterior probability is above certain threshold. Here, we generalize this thresholding reward estimator to the *randomized* binary reward case, and further prove that this estimator correctly models the reward for the optimal policy. To see this, we first prove the following lemma showing the exact posterior distribution over actions if the learner selects an action uniformly randomly. This is also proven in [Maghakian et al., 2022, Eq.(2)] as well, and we include it here for completeness.

Lemma 1. *For any context $x \in \mathcal{X}$, suppose that the learner picks a uniformly random action $a \in [K]$. Let r and y be its realized reward and the corresponding feedback. Then, under Assumption 1 and Assumption 2, the posterior distribution of a given the context x and feedback y equals to*

$$\mathsf{Pr}[a|x,y] = \frac{f^\star(x,a)\cdot\phi^\star(x,y)}{\sum_{a'=1}^K f^\star(x,a')} + \frac{(1-f^\star(x,a))(1-\phi^\star(x,y))}{K-\sum_{a'=1}^K f^\star(x,a')}, \tag{1}$$

where $f^\star$ and $\phi^\star$ are the true expected reward and feedback decoder defined in Assumption 2.

Now we show how to infer the true reward from this inverse kinematics. Specifically, we show:

Lemma 2. *For any context $x \in \mathcal{X}$ and action $a \in [K]$, let $r(x,a)$ be the realized reward and y be the feedback. Let $h^\star(x,y) \in \Delta_K$ be such that for each $a \in [K]$,*

$$h_a^\star(x,y) \triangleq \frac{f^\star(x,a)\cdot\phi^\star(x,y)}{\sum_{a'=1}^K f^\star(x,a')} + \frac{(1-f^\star(x,a))(1-\phi^\star(x,y))}{K-\sum_{a'=1}^K f^\star(x,a')}. \tag{2}$$

Then we have

- $r(x,a) = \phi^\star(x,y) \geq \mathbb{1}\{h_a^\star(x,y) \geq \frac{\theta}{\alpha}\}$ *for all* $a \in [K]$;
- $r(x,a) = \phi^\star(x,y) = \mathbb{1}\{h_a^\star(x,y) \geq \frac{\theta}{\alpha}\}$ *for* $a = \pi^\star(x)$.

Proof. Note that since $\phi^\star(x,y) \in \{0,1\}$, only one term can be non-zero in Eq. (2). If $h_a^\star(x,y) \geq \frac{\theta}{\alpha}$, where θ and α are defined in Assumption 3, then the reward $\phi^\star(x,y)$ has to be 1 since otherwise, we have $\phi^\star(x,y) = 0$ and

$$\frac{\theta}{\alpha} \leq h_a^\star(x,y) = \frac{1-f^\star(x,a)}{K-\sum_{a'=1}^K f^\star(x,a')} \leq \frac{1}{K-\alpha} < \frac{\theta}{\alpha},$$

where the second and the third inequality are due to Assumption 3. This leads to a contradiction. Therefore, we know that the realized reward is 1 if $h_a^\star(x,y)$ is no less than $\frac{\theta}{\alpha}$. Therefore, $\mathbb{1}\{h_a^\star(x,y) \geq \frac{\theta}{\alpha}\}$ can be viewed as an underestimator of $r(x,a)$. To prove the second property, consider the case in which $a = \pi^\star(x)$. Then, we know that when $\phi^\star(x,y) = 1$, we must also have $h_a^\star(x,y) \geq \frac{\theta}{\alpha}$ since

$$h_a^\star(x,y) = \frac{f^\star(x,\pi^\star(x))}{\sum_{i=1}^K f^\star(x,a)} \geq \frac{f^\star(x,\pi^\star(x))}{\alpha} \geq \frac{\theta}{\alpha},$$

where the first inequality uses the first property in Assumption 3 and the second inequality uses the second property in Assumption 3. □

Lemma 2 shows that the function $h^\star(x,y)$ defined in Eq. (2) satisfies two important properties. First, $\mathbb{1}\{h_a^\star(x,y) \geq \frac{\theta}{\alpha}\}$ serves as a reward *underestimator* for all the policies. Second, it matches the reward of the optimal policy. Note that this is different from [Maghakian et al., 2022, Eq.(3)], since

they only consider the *deterministic* binary reward for each context-action pair, and they do not show that the constructed reward estimator matches the reward of the optimal policy.

Based on these two properties, we know that if we have access to $h^\star$, then the policy π that maximizes the surrogate reward $\mathbb{E}\left[\mathbb{1}\{h^\star_{\pi(x)}(x,y)\} \geq \frac{\theta}{\alpha}\right]$ also maximizes the true expected reward.

3.1.2 Learning the Posterior Distribution via ERM

Next, we show how to learn this $h^\star$ via uniformly collected samples. Define the function class $\mathcal{H}$ as:

$$\mathcal{H} = \Bigg\{ h : \mathcal{X} \times \mathcal{Y} \mapsto \Delta_K, h_a(x,y) = \frac{f(x,a)\phi(x,y)}{\sum_{i=1}^K f(x,i)} + \frac{(1-f(x,a))(1-\phi(x,y))}{K - \sum_{i=1}^K f(x,i)} \Bigg| \; f \in \mathcal{F}, \phi \in \Phi, a \in [K] \Bigg\}.$$

Note that $|\mathcal{H}| = |\mathcal{F}| \cdot |\Phi|$. Since $h^\star$ models the posterior distribution over actions when the learner selects her action uniformly randomly, we collect N tuples of (x_t, a_t, y_t) by sampling a_t uniformly from $[K]$ and find the empirical risk minimizer (ERM) $\widehat{h} \in \mathcal{H}$ over these samples using squared loss. In the following lemma, we show that $\widehat{h}$ enjoys $\mathcal{O}\left(\frac{\log|\mathcal{H}|}{N}\right)$ excess risk with high probability.

Lemma 3. *Let $\{(x_i, a_i, y_i)\}_{i=1}^N$ be N i.i.d. samples where $x_i \in \mathcal{D}$, $a \in \pi_{\text{Unif}}$, and y_i is the corresponding feedback. Let $\widehat{h}$ be the ERM with respect to the squared loss defined as follows:*

$$\widehat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \left\{ \sum_{i=1}^N \|h(x_i, y_i) - e_{a_i}\|_2^2 \right\}. \tag{3}$$

Then, with probability at least $1-\delta$, we have

$$\mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_{\text{Unif}}, y|x,a}\left[\|\widehat{h}(x,y) - e_a\|_2^2 - \|h^\star(x,y) - e_a\|_2^2\right] \leq \mathcal{O}\left(\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}\right), \tag{4}$$

$$\mathbb{E}_{x\sim\mathcal{D}, a'\sim\pi_{\text{Unif}}, y|x,a'}\left[\|\widehat{h}(x,y) - h^\star(x,y)\|_2\right] \leq \mathcal{O}\left(\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right). \tag{5}$$

The full proof is deferred to Appendix A. Different from the classic one-dimensional squared loss, here we consider the ℓ_2-loss between two vectors in Δ_K. Directly applying the generalization bound for each entry leads to a K-factor worse bound. Instead, our proof is based on the observation that the loss function $\|h(x,y) - e_a\|_2^2$, when seen as a function of h, satisfies the so-called strong 1-central condition [Van Erven et al., 2015, Definition 7]. Moreover, these results can be extended to function classes with infinite size and bounded covering number based on Theorem 7.7 of [Van Erven et al., 2015].

3.1.3 Constructing Lipschitz Reward Estimators Based on $\widehat{h}$

Now we show how to construct a reward estimator based on the ERM $\widehat{h}$. According to Section 3.1.1, an intuitive form of the reward (under)estimator is $\mathbb{1}\{\widehat{h}_a(x,y) \geq \frac{\theta}{\alpha}\}$. However, since the indicator function is not Lipschitz, $\mathbb{1}\{\widehat{h}_a(x,y) \geq \frac{\theta}{\alpha}\}$ can be very different from $\mathbb{1}\{h^\star_a(x,y) \geq \frac{\theta}{\alpha}\}$ even with the generalization bound proven in Lemma 3. To resolve this issue, we propose a Lipschitz variant of the indicator function (defined in Eq. (6)) and show that it also satisfies the two properties shown in Lemma 2.

Lemma 4. *Define $G(v,\beta,\sigma)$ as*

$$G(v,\beta,\sigma) = \frac{1}{\sigma}(v-\beta)\mathbb{1}\{\beta \leq v < \beta+\sigma\} + \mathbb{1}\{v \geq \beta+\sigma\}. \tag{6}$$

Then, for any context $x \in \mathcal{X}$, action $a \in [K]$, and feedback $y \in \mathcal{Y}$ generated via context x and the realized reward $r(x,a)$, we have the following two properties with $\sigma \triangleq \frac{1}{2}\left(\frac{\theta}{\alpha} - \frac{1}{K-\alpha}\right) > 0$.

https://doi.org/10.52202/079017-3847

Algorithm 1 Off-Policy IGL

Input: number of exploration samples N, parameters α, θ and $\sigma = \frac{1}{2}\left(\frac{\theta}{\alpha} - \frac{1}{K-\alpha}\right)$.
for $t = 1, 2, \cdots, 2N$ **do**
  Receive context x_t, sample $a_t \sim \pi_{\text{Unif}}$, and observe signal y_t.
Calculate the empirical risk minimizer $\widehat{h}$ as in Eq. (3).
Construct the reward decoder $\widehat{r}_\sigma(x, y, a) = G(\widehat{h}_a(x, y), \frac{\theta}{\alpha} - \sigma, \sigma)$ where G is defined in Eq. (6).
Calculate $\widehat{\pi} = \operatorname{argmax}_{\pi \in \Pi} \left\{ \sum_{i=N+1}^{2N} \pi(a_i|x_i) \cdot \widehat{r}_\sigma(x_i, y_i, a_i) \right\}$ where $\Pi = \{\pi_f : f \in \mathcal{F}\}$.
for $t = 2N + 1, \ldots, T$ **do**
  Execute policy $\widehat{\pi}$.

- $r(x, a) = \phi^\star(x, y) \geq G(h_a^\star(x, y), \frac{\theta}{\alpha} - \sigma, \sigma)$,
- $r(x, a) = \phi^\star(x, y) = G(h_a^\star(x, y), \frac{\theta}{\alpha} - \sigma, \sigma)$ *if* $a = \pi^\star(x)$,

The proof shares a similar spirit to Lemma 2 and is deferred to Appendix A. Notably, the function $G(v, \beta, \sigma)$ is $\frac{1}{\sigma}$-Lipschitz in v, which is important in order to show concentration between the reward estimator with respect to the true posterior distribution $G(h_a^\star(x, y), \frac{\theta}{\alpha} - \sigma, \sigma)$ and that constructed via the ERM function $\widehat{h}$: $G(\widehat{h}_a(x, y), \frac{\theta}{\alpha} - \sigma, \sigma)$. In the following, we will show how to use the reward estimator $G(\widehat{h}_a(x, y), \frac{\theta}{\alpha} - \sigma, \sigma)$ to design algorithms with provable guarantees.

3.2 Off-Policy Algorithm

Built on the reward estimator in Section 3.1, we present our off-policy algorithm, summarized in Algorithm 1. Our algorithm follows the explore-then-exploit idea and consists of two phases. In the exploration phase, we perform uniform exploration and collect the dataset $\{x_i, a_i, y_i\}_{i=1}^{2N}$. The first N samples are used to learn the reward estimator $\widehat{r}$ and the rest N samples are used to learn the policy $\widehat{\pi}$ with $\widehat{r}$. For the exploitation phase, we employ the learned policy $\widehat{\pi}$ in the remaining $T - 2N$ iterations. We present the regret bound of Algorithm 1 in the following theorem.

Theorem 1. *Under Assumptions 1-3, Algorithm 1 with* $N = T^{2/3} K^{2/3} \sigma^{-2/3} \log^{1/3}(|\mathcal{H}|T)$ *guarantees that* $\mathbf{Reg}_{\mathsf{CB}} \leq \mathcal{O}\left(T^{2/3} K^{2/3} \sigma^{-2/3} \log^{1/3}(|\mathcal{H}|T)\right)$.

We remark that this is the first provably efficient algorithm for IGL with personalized reward. The key of the proof is to show that the learned policy $\widehat{\pi}$ is near-optimal under the true reward. This is achieved by using the properties of function G in Lemma 4 and proving that the reward decoder $\widehat{r}_\sigma$ is close to the ground-truth. The full proof is deferred to Appendix B. Besides the dependence on T, K and $\log|\mathcal{H}|$, our regret bound also depends on σ^{-1}, which measures how sparse the reward vector is and characterizes the difficulty of the problem. For example, $\sigma^{-1} = \mathcal{O}(1)$ in the multi-class classification problem and $\sigma^{-1} \leq \mathcal{O}(s^2)$ in the s-multi-label classification problem.

3.3 On-Policy Algorithm

Since on-policy algorithms are more favorable in practice, we further introduce an on-policy algorithm based on the inverse-gap weighting strategy. Following [Foster and Rakhlin, 2020], we assume access to an online regression oracle AlgSq: at each round $t \in [T]$, the oracle AlgSq produces an estimator $\widehat{f}_t$ in the convex hull of $\mathcal{F}$, then receives a context-action-reward tuple (x_t, a_t, r_t). The squared loss of the oracle for this round is defined as $(\widehat{f}_t(x_t, a_t) - r_t)^2$, which is on average assumed to be close to that of the best predictor in $\mathcal{F}$.

Assumption 4 (Bounded squared loss regret)**.** *For any sequence* $\{(x_t, a_t, r_t)\}_{t=1}^T$*, the regression oracle* AlgSq *guarantees the following regret bound for some* $\mathbf{Reg}_{\mathsf{Sq}}$ *that depends on* T*,* K*, and* $\mathcal{F}$*:*

$$\sum_{t=1}^{T} \left(\widehat{f}_t(x_t, a_t) - r_t\right)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \left(f(x_t, a_t) - r_t\right)^2 \leq \mathbf{Reg}_{\mathsf{Sq}}.$$

Algorithm 2 On-policy IGL

Input: online regression oracle AlgSq, number of exploration samples N, parameters α, θ, γ and $\sigma = \frac{1}{2}\left(\frac{\theta}{\alpha} - \frac{1}{K-\alpha}\right)$.

for $t = 1, 2, \cdots, N$ **do**

  Receive context x_t, sample $a_t \sim \pi_{\text{Unif}}$, and observe signal y_t.

Calculate the empirical risk minimizer $\widehat{h}$ as in Eq. (3).

for $t = N+1, \ldots, T$ **do**

  Receive context x_t.

  Obtain an estimator $\widehat{f}_t$ from the oracle AlgSq.

  Calculate the action distribution p_t as

$$p_{t,a} = \begin{cases} \frac{1}{K+\gamma(\widehat{f}_t(x_t,\widehat{a}) - \widehat{f}_t(x_t,a))}, & a \neq \widehat{a}, \\ 1 - \sum_{a' \neq \widehat{a}} p_{a'}, & a = \widehat{a} \end{cases}, \tag{7}$$

  where $\widehat{a} = \operatorname{argmax}_{a \in [K]} \widehat{f}_t(x_t, a)$.

  Sample a_t from p_t and receive feedback y_t.

  Feed $(x_t, a_t, G(\widehat{h}_{a_t}(x_t, y_t), \frac{\theta}{\alpha} - \sigma, \sigma))$ to the oracle AlgSq where G is defined in Eq. (6).

Based on the ground-truth inverse kinematics function $h^\star$, we define a function $\underline{f}^\star(x,a) := \mathbb{E}_{y|x,a}\left[G(h^\star_a(x,y), \frac{\theta}{\alpha} - \sigma, \sigma)\right]$, which is always a lower bound on $f^\star(x,a)$ according to Lemma 4. Since we only feed the surrogate reward to the oracle, we make another mild assumption that our regression function class $\mathcal{F}$ also realizes $\underline{f}^\star$.

Assumption 5 (Lower Bound Realizability)**.** *We also assume that $\underline{f}^\star \in \mathcal{F}$, where $\underline{f}^\star$ is defined as $\underline{f}^\star(x,a) = \mathbb{E}_{y|x,a}\left[G(h^\star_a(x,y), \frac{\theta}{\alpha} - \sigma, \sigma)\right]$*

We now summarize our algorithm in Algorithm 2. After obtaining the inverse kinematics predictor $\widehat{h}$ in the same manner as Algorithm 1, instead of uniform exploring, we use the estimated reward from the oracle and an inverse-gap weighting strategy [Abe and Long, 1999, Foster and Rakhlin, 2020] defined in Eq. (7). Different from the contextual bandit problem where the true reward is given and fed to the oracle, we feed the predicted reward $G(\widehat{h}_{a_t}(x_t, y_t), \frac{\theta}{\alpha} - \sigma, \sigma)$ to the oracle.

One might wonder how such misspecification in rewards would affect the regret bound. Our key observation is that since we use the uniform policy to collect data used to train $\widehat{h}$, the generalization error of $\widehat{h}$ is small for any a due to good coverage of the dataset on each action. Based on this observation, we prove the following theorem for Algorithm 2.

Theorem 2. *Under Assumptions 1-5, Algorithm 2 with certain choice of N and γ guarantees that* $\mathbf{Reg}_{\mathsf{CB}} = \mathcal{O}\left(\sqrt{KT\mathbf{Reg}_{\mathsf{Sq}}} + \sigma^{-2/3}(KT)^{2/3}\log^{1/3}(|\mathcal{H}|T)\right)$.

The proof mainly relies on the generalization bounds in Lemma 3 and the property of G in Lemma 4, and is deferred to Appendix C. We observe that Algorithm 2 enjoys the same dependence on T, K, σ^{-1} as Algorithm 1. For finite $\mathcal{F}$, we can use Vovk's aggregation algorithm [Vovk, 1995] as the regression oracle and achieve $\mathbf{Reg}_{\mathsf{Sq}} = \mathcal{O}(\log|\mathcal{F}|)$, making the second term in the regret bound negligible.[3] While in theory our on-policy algorithm does not seem to be more favorable than the off-policy algorithm (mostly because they both need to uniformly explore for a certain period in order to build $\widehat{h}$), it can still perform better in practice, as shown in our experiments.

4 Experiments

In this section, we apply IGL to learning from image feedback and learning from text feedback. Specifically, we conduct experiments on the MNIST dataset and a conversational dataset to verify the effectiveness of our algorithms and the Lipschitz reward estimator constructed by Eq. (6).

[3]We refer readers to Foster and Rakhlin [2020] for more examples of regression oracles when $\mathcal{F}$ is not finite.

https://doi.org/10.52202/079017-3847

Table 1: Performance of Algorithm 1 and Algorithm 2 on the MNIST dataset.

Algorithm	Reward Estimator	Average Progressive Reward	Test Accuracy
Off-policy Algorithm 1	Binary	0.614 (0.012)	72.6% (1.5%)
	Lipschitz	0.638 (0.042)	75.9% (4.9%)
On-policy Algorithm 2	Binary	0.718 (0.006)	89.4% (4.0%)
	Lipschitz	0.740 (0.022)	90.3% (3.7%)

4.1 Experiments on Learning from Image Feedback

Experiment Setup We first conduct experiments on MNIST dataset and implement both the off-policy algorithm Algorithm 1 and the on-policy algorithm Algorithm 2. The setup is as follows: at each step t, the learner receives an image x_t with ground-truth label l_t and picks action a_t from $\{0, \cdots, 9\}$ as the predicted label. If the prediction a_t is correct, the learner receives as feedback an image y_t with digit $(l_t + 1) \mod 10$; otherwise, the learner receives an image with digit $(l_t - 1) \mod 10$. Therefore, different from the experimental setup in [Xie et al., 2021], our feedback y_t does depend on both the context and the reward. Both function classes $\mathcal{H}$ and $\mathcal{F}$ are implemented as two-layer convolutional neural networks.

We use PyTorch framework [Paszke et al., 2019] and parameter-free optimizers [Orabona and Tommasi, 2017] to learn the reward estimator and the policy. For both algorithms, we set the number of exploration samples $N = 5000$ and pick the parameter $\sigma \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ and $\frac{\theta}{\alpha} \in \{\frac{1}{3} + \frac{\sigma}{2}, \frac{1}{2} + \frac{\sigma}{2}\}$. For the on-policy algorithm, we use a time-varying exploration parameter $\gamma_t = \sqrt{Kt}$ as suggested by Foster and Krishnamurthy [2021]. We run the experiments on one NVIDIA GeForce RTX 2080 Ti.

The interaction lasts for $T = 60000$ rounds. We use two metrics to evaluate the performance of the algorithm, including the average progressive reward during the interaction process and the test accuracy on a held-out test set containing 10000 samples. When evaluating the on-policy algorithm on the test set, we take actions greedily according to $\widehat{f}_T$. Since we use the uniform policy to collect data for learning the reward estimator, the progressive reward at that phase is counted as $\frac{1}{K}$.

Results We run the experiments with 4 different random seeds and report the mean value and standard deviation in Table 1. The running averaged reward over the entire T rounds are shown in Figure 1. We can see that both algorithms achieve good performance, despite never observing any true labels. While the theoretical regret guarantees for both algorithms are of the same order, empirically, the on-policy Algorithm 2 performs better than Algorithm 1 with over 90% test accuracy. This is because the on-policy algorithm uses the inverse-gap weighting strategy to achieve a better trade-off between exploration and exploitation, while the off-policy algorithm learns the policy from uniform exploration. On the other hand, to demonstrate the effectiveness of the Lipschitz reward estimator, we compare the performances of the Lipschitz estimator Eq. (6) with a binary reward estimator $\widehat{r}_{\text{binary}}(x, y, a) = \mathbb{1}\{\widehat{h}_a(x, y) \geq \frac{\theta}{\alpha}\}$, where the parameter $\frac{\theta}{\alpha}$ is searched over the same space. The results in Table 1 show that the Lipschitz reward estimator improves over the binary one by a clear margin for both algorithms. This matches our theoretical analysis that highlights the vital role of the Lipschitzness of the reward estimator in obtaining good regret guarantees. We also plot the performance of Algorithm 2 under both true reward and constructed reward in the left figure of Figure 2.[4] The figure shows that the constructed reward is indeed a lower bound of the true reward, and the policy can learn from the constructed reward effectively.

4.2 Experiments on Learning From Text Feedback

We further consider an application of learning with text feedback. The IGL framework fits this problem well since this is a natural reward-free setting involving learning from user's text feedback, which is idiosyncratically related to the quality of the actions taken and the question asked. Specifically, given the input of a question, the learner needs to select one of the K possible answers. Instead of

[4]Both the true reward and the constructed reward values are averaged over the rounds after the first N rounds of uniform exploration.

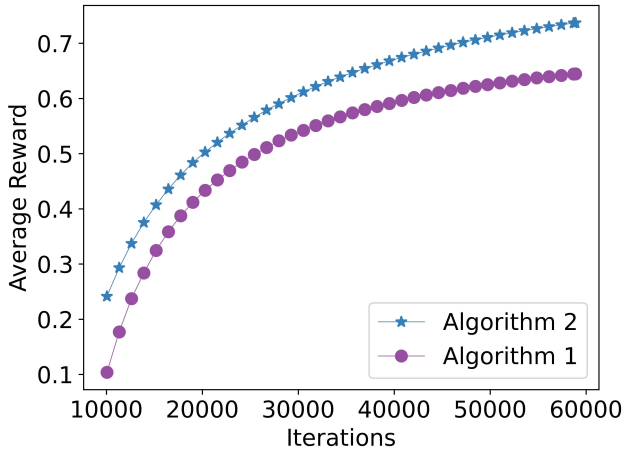

Figure 1: Running averaged reward of Algorithm 1 and Algorithm 2 on MNIST. Note that Algorithm 1 uniformly explores in the first $2N = 10000$ rounds, and thus its averaged reward at $t = 10000$ is about $1/K = 0.1$.

receiving whether the selected answer is correct, the learner only receives user's text feedback to the chosen answer. The goal of the learner is to choose the best answer only based on such text feedback.

Dataset Construction Our dataset is constructed as follows. Specifically, we construct our question set $S = \{q_i\}_{i\in[20000]}$ from Chatbot Arena datasets [Zheng et al., 2024]. Then, for each question $q_i \in S$ we ask a larger language model $\mathcal{G}$ with a high ELO score on the chatsys leaderboard [Chiang et al., 2024] to generate a "good" answer $g_{i,0}$ with $r_{i,0} = 1$; and ask a (much) smaller language model $\mathcal{B}$ with a (much) lower ELO score to generate 4 "bad" answers $b_{i,j}$ with reward $r_{i,j} = 0$, $j \in \{1, 2, 3, 4\}$. Specifically, we pick $\mathcal{G}$ to be "Qwen1.5-32B-Chat" with ELO score 1134 and $\mathcal{B}$ to be "Qwen1.5-0.5B-Chat" with ELO score[5] less than 804 [Bai et al., 2023].[6] For each question-answer tuple $(q_i, g_{i,0}, b_{i,1}, b_{i,2}, b_{i,3}, b_{i,4})$, we ask another large language model $\mathcal{R}$ to simulate a user response $f_{i,j}$, $j \in \{0, 1, 2, 3, 4\}$ to the good (bad) answers under the instruction that the user is satisfied (unsatisfied) with the answer. We pick $\mathcal{R}$ to be Qwen1.5-32B-Chat as well. This forms our final dataset $S_{\text{Conv}} = \{(q_i, (g_{i,0}, f_{i,0}, r_{i,0}), \{(b_{i,j}, f_{i,j}, r_{i,j})\}_{j\in[4]})\}_{i\in[20000]}$. Again, the true reward is never revealed to the learner, and we only use this reward to measure the performance of our algorithm. The prompt we use is deferred to Appendix D.1. We generate our dataset using one A100 GPU for two weeks.

4.2.1 Algorithm Configurations and Results

Given the superior performance of Algorithm 2 over Algorithm 1 on the MNIST dataset shown in Section 4.1, we only test Algorithm 2 on the conversational dataset. We use the first $N = 10000$ data points to learn $\widehat{h}$ and the remaining $|S| - N = 10000$ data points to learn the optimal policy based on the reward function constructed via $\widehat{h}$. We use the same parameter-free optimizer as the one in Section 4.1. Next, we introduce the construction of $\widehat{h}$ and the regression oracle:

Inverse kinematic model. The model class $\mathcal{H}$ we consider is the pretrained language model Llama-3-8B-Instruct [AI@Meta, 2024] with an additional multi-class classification head.[7] The language model is prompted with a question-answer-feedback tuple (x, a, y); see Appendix D.2 for the prompt. To learn the inverse kinematic model, we use parameter efficient fine-tuning [Mangrulkar et al., 2024] with a rank-1 LORA adapter [Hu et al., 2021] and binary cross entropy loss, which is with respect to the indicator of whether-or-not the predicted action corresponds to the action selected in the tuple.

[5]The bad model $\mathcal{B}$ is not displayed on the leaderboard; 804 is the lowest displayed score.

[6]"Qwen1.5-32B-Chat" is available at https://huggingface.co/Qwen/Qwen1.5-32B-Chat and "Qwen1.5-0.5B-Chat" is available at https://huggingface.co/Qwen/Qwen1.5-0.5B-Chat.

[7]"Llama-3-8B-Instruct": https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct.

https://doi.org/10.52202/079017-3847

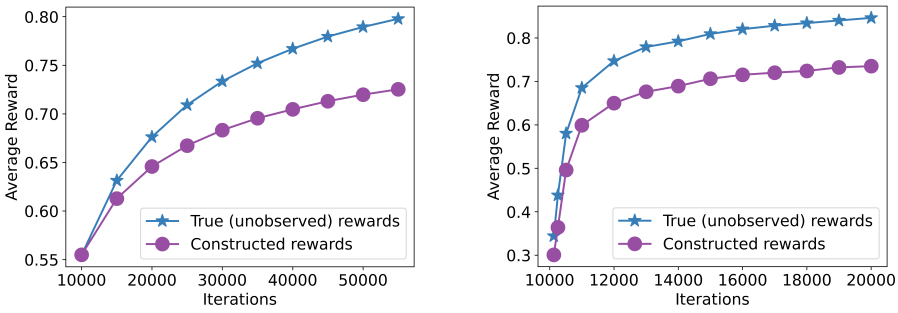

Figure 2: Performance of Algorithm 2 under true (unobserved) rewards and constructed rewards. **Left figure**: Results on MNIST dataset after the first N uniform exploration rounds. **Right figure**: Results on our conversational dataset.

We successfully learn $\widehat{h}$ using one A100 GPU within 6 hours. After obtaining $\widehat{h}$, we construct the reward estimator $G(\widehat{h}_a(x,y), \frac{\theta}{\alpha} - \sigma, \sigma)$ with $\sigma = 0.1$, $\alpha = \frac{K}{2} = \frac{5}{2}$, and $\theta = 1$.

Regression oracle. Similar to the inverse kinematic model, the reward prediction function class $\mathcal{F}$ is again based on the pretrained Llama-3-8B-Instruct model but with an additional regression head. Specifically, the language model is prompted only with a question-answer pair (x, a) using the prompt deferred to Appendix D.3 and predicts a score in $[0, 1]$ for this question-answer pair. The regression oracle again applies parameter efficient fine-tuning with a different rank-1 LORA adapter on the regression loss, which measures the error of predicting the output of the reward predictor $G(\widehat{h}_a(x,y), \frac{\theta}{\alpha} - \sigma, \sigma)$. This process is done on one A100 GPU within 3 hours.

Results. The empirical results on the conversational dataset are shown in Figure 2. We show the running averaged true reward and the running average constructed reward received by Algorithm 2 after the first $N = 10000$ rounds of learning $\widehat{h}$. The x-axis is the time horizon and y-axis is the value of average reward. Similar to our experiment results in Section 4.1, the right figure in Figure 2 shows that our constructed reward estimator is a lower bound on the true reward, matching our theoretical results in Lemma 4, and Algorithm 2 is able to learn the reward effectively through the text feedback with the constructed reward estimator.

Acknowledgments and Disclosure of Funding

HL and MZ were supported by NSF Award IIS-1943607.

References

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer, 1999.

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Gábor Bartók and Csaba Szepesvári. Partial monitoring with side information. In *International Conference on Algorithmic Learning Theory*, pages 305–319. Springer, 2012.

Joeran Beel, Stefan Langer, Andreas Nürnberger, and Marcel Genzmehr. The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems. In *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings 3*, pages 396–400. Springer, 2013.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan Foster, Alekh Agarwal, Miroslav Dudik, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

Dylan J Foster and Akshay Krishnamurthy. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34: 18907–18919, 2021.

Euan Freeman, Graham Wilson, Dong-Bach Vo, Alex Ng, Ioannis Politis, and Stephen Brewster. Multimodal feedback in hci: haptics, non-speech audio, and their applications. In *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations-Volume 1*, pages 277–317. 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xiaoyan Hu, Farzan Farnia, and Ho-fung Leung. An information theoretic approach to interaction-grounded learning. *arXiv preprint arXiv:2401.05015*, 2024.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 2328–2369, 2020.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

Jessica Maghakian, Paul Mineiro, Kishan Panaganti, Mark Rucker, Akanksha Saran, and Cheng Tan. Personalized reward learning with interaction-grounded learning (igl). *arXiv preprint arXiv:2211.15823*, 2022.

S Mangrulkar, S. Gugger, L. Debut, Y. Belkada, and Paul S. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2024. Accessed: 2024-05-14.

 https://doi.org/10.52202/079017-3847

Francesco Orabona and Tatiana Tommasi. Training deep networks without learning rates through coin betting. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://github.com/bremen79/parameterfree.

Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Donghee Shin. How do users interact with algorithm recommender systems? the interaction of users, algorithms, and performance. *Computers in human behavior*, 109:106344, 2020.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 2021.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Tim Van Erven, Peter Grunwald, Nishant A Mehta, Mark Reid, Robert Williamson, et al. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 54(6), 2015.

Vladimir G Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 51–60, 1995.

Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1927–1936, 2017.

Tengyang Xie, John Langford, Paul Mineiro, and Ida Momennejad. Interaction-grounded learning. In *International Conference on Machine Learning*, pages 11414–11423. PMLR, 2021.

Tengyang Xie, Akanksha Saran, Dylan J Foster, Lekan Molu, Ida Momennejad, Nan Jiang, Paul Mineiro, and John Langford. Interaction-grounded learning with action-inclusive feedback. *Advances in Neural Information Processing Systems*, 35:12529–12541, 2022.

Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. Beyond clicks: dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 113–120, 2014.

Mengxiao Zhang, Yuheng Zhang, Haipeng Luo, and Paul Mineiro. Efficient contextual bandits with uninformed feedback graphs. *International Conference on Machine Learning*, 2024a.

Mengxiao Zhang, Yuheng Zhang, Olga Vrousgou, Haipeng Luo, and Paul Mineiro. Practical contextual bandits with feedback graphs. *Advances in Neural Information Processing Systems*, 36, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

A Proofs in Section 3.1

Lemma 1. *For any context $x \in \mathcal{X}$, suppose that the learner picks a uniformly random action $a \in [K]$. Let r and y be its realized reward and the corresponding feedback. Then, under Assumption 1 and Assumption 2, the posterior distribution of a given the context x and feedback y equals to*

$$\mathsf{Pr}[a|x,y] = \frac{f^\star(x,a)\cdot\phi^\star(x,y)}{\sum_{a'=1}^K f^\star(x,a')} + \frac{(1-f^\star(x,a))(1-\phi^\star(x,y))}{K-\sum_{a'=1}^K f^\star(x,a')}, \tag{1}$$

where $f^\star$ and $\phi^\star$ are the true expected reward and feedback decoder defined in Assumption 2.

Proof.

$$\begin{aligned}
\mathsf{Pr}[a|x,y] &= \frac{\mathsf{Pr}[a|x]\cdot\mathsf{Pr}[y|x,a]}{\mathsf{Pr}[y|x]} \\
&= \mathsf{Pr}[a|x]\frac{\mathsf{Pr}[r=1|x,a]\cdot\mathsf{Pr}[y|x,a,r=1]+\mathsf{Pr}[r=0|x,a]\cdot\mathsf{Pr}[y|x,a,r=0]}{\mathsf{Pr}[y|x]} \\
&= \mathsf{Pr}[a|x]\frac{f^\star(x,a)\cdot\mathsf{Pr}[y|x,r=1]+(1-f^\star(x,a))\cdot\mathsf{Pr}[y|x,r=0]}{\mathsf{Pr}[y|x]} \\
&= \mathsf{Pr}[a|x]\frac{f^\star(x,a)\cdot\mathsf{Pr}[r=1|x,y]}{\mathsf{Pr}[r=1|x]} + \mathsf{Pr}[a|x]\frac{(1-f^\star(x,a))\cdot\mathsf{Pr}[r=0|x,y]}{\mathsf{Pr}[r=0|x]} \\
&= \mathsf{Pr}[a|x]\frac{f^\star(x,a)\cdot\mathsf{Pr}[r=1|x,y]}{\sum_{a'=1}^K\mathsf{Pr}[a'|x]\mathsf{Pr}[r=1|x,a']} + \mathsf{Pr}[a|x]\frac{(1-f^\star(x,a))\cdot\mathsf{Pr}[r=0|x,y]}{\sum_{a'=1}^K\mathsf{Pr}[a'|x]\mathsf{Pr}[r=0|x,a']} \\
&= \mathsf{Pr}[a|x]\frac{f^\star(x,a)\cdot\mathsf{Pr}[r=1|x,y]}{\sum_{a'=1}^K\mathsf{Pr}[a'|x]f^\star(x,a')} + \mathsf{Pr}[a|x]\frac{(1-f^\star(x,a))\cdot\mathsf{Pr}[r=0|x,y]}{\sum_{a'=1}^K\mathsf{Pr}[a'|x](1-f^\star(x,a'))}.
\end{aligned}$$

Recall that $\pi_{\text{Unif}} = \frac{1}{K}\cdot\mathbf{1}$ is the policy which uniformly samples an action regardless of the context. Under π_{Unif}, we know that

$$\mathsf{Pr}[a|x,y] = \frac{f^\star(x,a)\cdot\phi^\star(x,y)}{\sum_{a'=1}^K f^\star(x,a')} + \frac{(1-f^\star(x,a))(1-\phi^\star(x,y))}{K-\sum_{a'=1}^K f^\star(x,a')}.$$

□

Lemma 3. *Let $\{(x_i,a_i,y_i)\}_{i=1}^N$ be N i.i.d. samples where $x_i \in \mathcal{D}$, $a \in \pi_{\text{Unif}}$, and y_i is the corresponding feedback. Let $\widehat{h}$ be the ERM with respect to the squared loss defined as follows:*

$$\widehat{h} = \operatorname*{argmin}_{h\in\mathcal{H}}\left\{\sum_{i=1}^N \|h(x_i,y_i)-e_{a_i}\|_2^2\right\}. \tag{3}$$

Then, with probability at least $1-\delta$, we have

$$\mathbb{E}_{x\sim\mathcal{D},a\sim\pi_{\text{Unif}},y|x,a}\left[\|\widehat{h}(x,y)-e_a\|_2^2 - \|h^\star(x,y)-e_a\|_2^2\right] \le \mathcal{O}\left(\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}\right), \tag{4}$$

$$\mathbb{E}_{x\sim\mathcal{D},a'\sim\pi_{\text{Unif}},y|x,a'}\left[\|\widehat{h}(x,y)-h^\star(x,y)\|_2\right] \le \mathcal{O}\left(\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right). \tag{5}$$

Proof. For notational convenience, we denote (x,y,a) by Z and define $\ell_h(Z) = \|h(x,y)-e_a\|_2^2$. Now we aim to show that $\ell_h(Z)$ satisfies the strong η-central condition for some $\eta > 0$ [Van Erven et al., 2015]. Specifically, we aim to show that

$$\mathbb{E}_Z\left[\exp(-\eta(\ell_h(Z)-\ell_{h^\star}(Z)))\right] \le 1. \tag{8}$$

To show this, direct calculation shows that

$\mathbb{E}_Z\left[\exp(-\eta(\ell_h(Z)-\ell_{h^\star}(Z)))\right]$

https://doi.org/10.52202/079017-3847

$$= \mathbb{E}_{x,y}\left[\exp(-\eta\|h(x,y)-h^\star(x,y)\|_2^2)\cdot \mathbb{E}_{a|x,y}\left[\exp\left(-2\eta(h(x,y)-h^\star(x,y))^\top(h^\star(x,y)-e_a)\right)\right]\right].$$

Since $\mathbb{E}_{a|x,y}[e_a] = h^\star(x,y)$, the random variable $(h(x,y)-h^\star(x,y))^\top(h^\star(x,y)-e_a)$ given x and y is zero-mean and is within the range $[-2\|h(x,y)-h^\star(x,y)\|_2, 2\|h(x,y)-h^\star(x,y)\|_2]$. Therefore, we know that $(h(x,y)-h^\star(x,y))^\top(h^\star(x,y)-e_a)$ is $\|h(x,y)-h^\star(x,y)\|_2^2$-sub-Gaussian and we have

$$\begin{aligned}
&\mathbb{E}_Z\left[\exp(-\eta(\ell_h(Z)-\ell_{h^\star}(Z)))\right]\\
&\le \mathbb{E}_{x,y}\left[\exp(-\eta\|h(x,y)-h^\star(x,y)\|_2^2)\exp\left(\frac{4\eta^2\|h(x,y)-h^\star(x,y)\|_2^2}{2}\right)\right]\\
&= \mathbb{E}_{x,y}\left[\exp\left(\left(2\eta^2-\eta\right)\|h(x,y)-h^\star(x,y)\|_2^2\right)\right].
\end{aligned}$$

Picking $\eta = \frac{1}{2}$ proves Eq. (8). Noticing the fact that $|\ell_h(Z)| \le 4$ for all $h \in \mathcal{H}$ and applying Theorem 7.6 in [Van Erven et al., 2015] proves Eq. (4).

Now we prove Eq. (5). Noticing the fact that $\mathbb{E}_{a|x,y}[e_a] = h^\star(x,y)$ and applying this to Eq. (4), we know that with probability at least $1-\delta$,

$$\mathbb{E}_{x,y}\left[\|\widehat{h}(x,y)-h^\star(x,y)\|_2^2\right] \le \mathcal{O}\left(\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}\right).$$

This means that

$$\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{a'\sim\pi_{\text{Unif}}}\mathbb{E}_{y|x,a'}\left[\|\widehat{h}(x,y)-h^\star(x,y)\|_2^2\right] \le \mathcal{O}\left(\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}\right). \tag{9}$$

Further applying Jensen's inequality shows that

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{a'\sim\pi_{\text{Unif}}}\mathbb{E}_{y|x,a'}\left[\|\widehat{h}(x,y)-h^\star(x,y)\|_2\right]\\
&= \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{a'\sim\pi_{\text{Unif}}}\mathbb{E}_{y|x,a'}\left[\sqrt{\|\widehat{h}(x,y)-h^\star(x,y)\|_2^2}\right]\\
&\le \sqrt{\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{a'\sim\pi_{\text{Unif}}}\mathbb{E}_{y|x,a'}\left[\|\widehat{h}(x,y)-h^\star(x,y)\|_2^2\right]} && \text{(Jensen's inequality)}\\
&\le \mathcal{O}\left(\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right).
\end{aligned}$$

□

Lemma 4. *Define* $G(v,\beta,\sigma)$ *as*

$$G(v,\beta,\sigma) = \frac{1}{\sigma}(v-\beta)\mathbb{1}\{\beta \le v < \beta+\sigma\} + \mathbb{1}\{v \ge \beta+\sigma\}. \tag{6}$$

Then, for any context $x \in \mathcal{X}$*, action* $a \in [K]$*, and feedback* $y \in \mathcal{Y}$ *generated via context* x *and the realized reward* $r(x,a)$*, we have the following two properties with* $\sigma \triangleq \frac{1}{2}\left(\frac{\theta}{\alpha} - \frac{1}{K-\alpha}\right) > 0$*.*

- $r(x,a) = \phi^\star(x,y) \ge G(h_a^\star(x,y), \frac{\theta}{\alpha}-\sigma, \sigma)$*,*
- $r(x,a) = \phi^\star(x,y) = G(h_a^\star(x,y), \frac{\theta}{\alpha}-\sigma, \sigma)$ *if* $a = \pi^\star(x)$*,*

Proof. To prove the first property, we show $\phi^\star(x,y) = 1$ if $h_a^\star(x,y) \ge \frac{\theta}{\alpha}-\sigma$. Specifically, if $h_a^\star(x,y) \ge \frac{\theta}{\alpha}-\sigma$ and $\phi^\star(x,y) = 0$, then we know that

$$\frac{\theta}{\alpha} - \sigma \le h_a^\star(x,y) = \frac{1-f^\star(x,a)}{K-\sum_{i=1}^K f^\star(x,i)} \le \frac{1}{K-\alpha},$$

leading to a contradiction according to the definition of σ. Therefore, when $h_a^\star(x,y) \geq \frac{\theta}{\alpha} - \sigma$, we have $\phi^\star(x,y) = 1$, which is surely an upper bound of $G(h_a^\star(x,y), \frac{\theta}{\alpha} - \sigma, \sigma)$ since $G(h_a^\star(x,y), \frac{\theta}{\alpha} - \sigma, \sigma) \leq 1$. When $h_a^\star(x,y) < \frac{\theta}{\alpha} - \sigma$, by definition, we know that $G(h_a^\star(x,y), \frac{\theta}{\alpha} - \sigma, \sigma) = 0 \leq \phi^\star(x,y)$. Combining the two cases finishes the proof for the first property.

To prove the second property, note that when $\phi^\star(x,y) = 1$ and $a = \pi^\star(x)$, we have

$$h^\star_{\pi^\star(x)}(x,y) = \frac{f^\star(x,\pi^\star(x))}{\sum_{i=1}^K f^\star(x,i)} \geq \frac{\theta}{\alpha},$$

where the last inequality is by Assumption 3. This means that $G(h_a^\star(x,y), \frac{\theta}{\alpha} - \sigma, \sigma) = 1 = \phi^\star(x,y)$. When $\phi^\star(x,y) = 0$, we have

$$h^\star_{\pi^\star(x)}(x,y) = \frac{1 - f^\star(x,\pi^\star(x))}{K - \sum_{i=1}^K f^\star(x,i)} \leq \frac{1}{K-\alpha} \leq \frac{\theta}{\alpha} - \sigma,$$

meaning that $G(h_a^\star(x,y), \frac{\theta}{\alpha} - \sigma, \sigma) = 0$. This finishes the proof. □

B Proofs in Section 3.2

For any policy π and reward estimator r_σ, we define the value function $V(\pi, r_\sigma)$ and the empirical value function $\widehat{V}(\pi, r_\sigma)$ as follows:

$$\widehat{V}(\pi, r_\sigma) = \frac{1}{N} \sum_{i=N+1}^{2N} K \cdot \pi(a_i|x_i) \cdot r_\sigma(x_i, y_i, a_i), \tag{10}$$

$$V(\pi, r_\sigma) = \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_{\text{Unif}}, y|x,a} \left[K \cdot \pi(a|x) \cdot r_\sigma(x,y,a)\right], \tag{11}$$

where $\{(x_i, a_i, y_i)\}_{i=N+1}^{2N}$ is collected with the uniform policy π_{Unif}. The true value function $V(\pi)$ is defined as:

$$V(\pi) = \mathbb{E}_{x\sim\mathcal{D}} \left[\sum_{a=1}^K \pi(a|x) \cdot f^\star(x,a)\right].$$

Using Lemma 4, we can establish the equivalence between $V(\pi^\star)$ and $V(\pi^\star, r_\sigma^\star)$ where $r_\sigma^\star(x,y,a) = G(h_a^\star(x,y), \frac{\theta}{\alpha} - \sigma, \sigma)$.

$$\begin{aligned}
V(\pi^\star) &= \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_{\text{Unif}}} \left[K \cdot \pi^\star(a|x) \cdot f^\star(x,a)\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_{\text{Unif}}, y|x,a} \left[K \cdot \pi^\star(a|x) \cdot \phi^\star(x,y)\right] && \text{(Assumption 2)} \\
&= \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_{\text{Unif}}, y|x,a} \left[K \cdot \pi^\star(a|x) \cdot G\left(h_a^\star(x,y), \frac{\theta}{\alpha} - \sigma, \sigma\right)\right] && \text{(Lemma 4)} \\
&= V(\pi^\star, r_\sigma^\star). && (12)
\end{aligned}$$

In the following lemma, we prove that $\widehat{\pi}$ is near optimal under the true reward.

Lemma 5. *Under Assumption 2 and Assumption 3, the following holds with probability at least $1 - 3\delta$:*

$$V(\pi^\star) - V(\widehat{\pi}) \leq \mathcal{O}\left(\frac{K}{\sigma}\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right),$$

where $\widehat{\pi}$ is defined in Algorithm 1.

Proof. We first introduce some high probability events that the following analysis is based on. First, according to Hoeffding's inequality together with a union bound over $\pi \in \Pi$ (note that $|\Pi| = |\mathcal{F}| \leq |\mathcal{H}|$), with probability at least $1 - \delta$, we have for any $\pi \in \Pi$

$$\left|V(\pi, r_\sigma^\star) - \widehat{V}(\pi, r_\sigma^\star)\right| \leq \mathcal{O}\left(K \cdot \sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right). \tag{13}$$

https://doi.org/10.52202/079017-3847

In addition, according to Hoeffding's inequality with a union bound over $h \in \mathcal{H}$, with probability at least $1-\delta$, we have for any $h \in \mathcal{H}$, the following inequality holds:

$$\left| \frac{1}{N} \sum_{i=N+1}^{2N} \|h(x_i, y_i) - h^\star(x_i, y_i)\|_2 - \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_{\text{Unif}}, y|x,a} \left[\|h(x, y) - h^\star(x, y)\|_2\right] \right| \leq \mathcal{O}\left(\sqrt{\frac{\log \frac{|\mathcal{H}|}{\delta}}{N}} \right). \tag{14}$$

Combining Eq. (14) with Eq. (5) in Lemma 3, we know that with probability at least $1-2\delta$,

$$\begin{aligned} \frac{1}{N} \sum_{i=N+1}^{2N} \left| \widehat{h}_{a_i}(x_i, y_i) - h^\star_{a_i}(x_i, y_i) \right| &\leq \frac{1}{N} \sum_{i=N+1}^{2N} \|\widehat{h}(x_i, y_i) - h^\star(x_i, y_i)\|_2 \\ &\leq \mathcal{O}\left(\sqrt{\frac{\log \frac{|\mathcal{H}|}{\delta}}{N}} \right). \end{aligned} \tag{15}$$

The following analysis is based on the condition that Eq. (13) and Eq. (15) hold.

Bounding $\left|\widehat{V}(\pi, \widehat{r}_\sigma) - \widehat{V}(\pi, r^\star_\sigma)\right|$. We first show that for any policy $\pi \in \Pi$, the prediction from $\widehat{r}_\sigma$ is close to $r^\star_\sigma$ in terms of the empirical value function defined in Eq. (10), where $\widehat{r}_\sigma(x, y, a) \triangleq G(\widehat{h}_a(x, y), \frac{\theta}{\alpha} - \sigma, \sigma)$.

$$\begin{aligned} \left|\widehat{V}(\pi, \widehat{r}_\sigma) - \widehat{V}(\pi, r^\star_\sigma)\right| &\leq \frac{1}{N} \sum_{i=N+1}^{2N} K \cdot \pi(a_i|x_i) \left|\widehat{r}_\sigma(x_i, y_i, a_i) - r^\star_\sigma(x_i, y_i, a_i)\right| \\ &\leq \frac{K}{N} \sum_{i=N+1}^{2N} \left|\widehat{r}_\sigma(x_i, y_i, a_i) - r^\star_\sigma(x_i, y_i, a_i)\right| \\ &= \frac{K}{N} \sum_{i=N+1}^{2N} \left| G(\widehat{h}_{a_i}(x_i, y_i), \frac{\theta}{\alpha} - \sigma, \sigma) - G(h^\star_{a_i}(x_i, y_i), \frac{\theta}{\alpha} - \sigma, \sigma) \right| \\ &\leq \frac{K}{N\sigma} \sum_{i=N+1}^{2N} \left| \widehat{h}_{a_i}(x_i, y_i) - h^\star_{a_i}(x_i, y_i) \right| \\ &\leq \frac{K}{N\sigma} \sum_{i=N+1}^{2N} \|\widehat{h}(x_i, y_i) - h^\star(x_i, y_i)\|_2 \\ &\leq \mathcal{O}\left(\frac{K}{\sigma} \sqrt{\frac{\log \frac{|\mathcal{H}|}{\delta}}{N}} \right). \end{aligned} \tag{16}$$

The third inequality is because $G(v, \beta, \sigma)$ is $\frac{1}{\sigma}$-Lipschitz in v and the last inequality is from Eq. (15).

Lower bound $V(\pi)$ by $\widehat{V}(\pi, \widehat{r}_\sigma)$. Then, we show that for any policy $\pi \in \Pi$, $V(\pi)$ is lower bounded by $\widehat{V}(\pi, \widehat{r}_\sigma)$:

$$\begin{aligned} V(\pi) &= \mathbb{E}_{x\sim\mathcal{D}} \left[\sum_{a=1}^{K} \pi(a|x) f^\star(x, a) \right] \\ &= \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_{\text{Unif}}} \left[K\pi(a|x) f^\star(x, a) \right] \\ &= \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_{\text{Unif}}} \mathbb{E}_{y|x,a} \left[K\pi(a|x) \phi^\star(x, y) \right] && \text{(Assumption 2)} \\ &\geq \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi_{\text{Unif}}} \mathbb{E}_{y|x,a} \left[K\pi(a|x) G(h^\star_a(x, y), \frac{\theta}{\alpha} - \sigma, \sigma) \right] \end{aligned}$$

$$
\begin{aligned}
&= V(\pi, r^\star_\sigma) && \text{(Eq. (11))}\\
&= V(\pi, r^\star_\sigma) - \widehat{V}(\pi, r^\star_\sigma) + \widehat{V}(\pi, r^\star_\sigma) - \widehat{V}(\pi, \widehat{r}_\sigma) + \widehat{V}(\pi, \widehat{r}_\sigma)\\
&\geq \widehat{V}(\pi, \widehat{r}_\sigma) - \mathcal{O}\left(\frac{K}{\sigma}\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right). && (17)
\end{aligned}
$$

The first inequality is from Lemma 4 and the second inequality is from Eq. (13) and Eq. (16). Recall that $\widehat{\pi} = \max_{\pi\in\Pi} \widehat{V}(\pi, \widehat{r}_\sigma)$, we then know that with probability at least $1-3\delta$,

$$
\begin{aligned}
V(\pi^\star) - V(\widehat{\pi}) &\leq V(\pi^\star) - \widehat{V}(\widehat{\pi}, \widehat{r}_\sigma) + \mathcal{O}\left(\frac{K}{\sigma}\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right) && \text{(Eq. (17))}\\
&\leq V(\pi^\star) - \widehat{V}(\pi^\star, \widehat{r}_\sigma) + \mathcal{O}\left(\frac{K}{\sigma}\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right) && \text{(optimality of } \widehat{\pi} \text{ on } \widehat{V}(\pi, \widehat{r}_\sigma))\\
&\leq V(\pi^\star) - \widehat{V}(\pi^\star, r^\star_\sigma) + \mathcal{O}\left(\frac{K}{\sigma}\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right) && \text{(Eq. (16))}\\
&\leq V(\pi^\star) - V(\pi^\star, r^\star_\sigma) + \mathcal{O}\left(\frac{K}{\sigma}\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right) && \text{(Eq. (13))}\\
&= \mathcal{O}\left(\frac{K}{\sigma}\sqrt{\frac{\log\frac{|\mathcal{H}|}{\delta}}{N}}\right), && \text{(Eq. (12))}
\end{aligned}
$$

which finishes the proof. □

Next, we prove Theorem 1.

Theorem 1. *Under Assumptions 1-3, Algorithm 1 with* $N = T^{2/3}K^{2/3}\sigma^{-2/3}\log^{1/3}(|\mathcal{H}|T)$ *guarantees that* $\mathbf{Reg}_{\mathsf{CB}} \leq \mathcal{O}\left(T^{2/3}K^{2/3}\sigma^{-2/3}\log^{1/3}(|\mathcal{H}|T)\right)$.

Proof. We bound $\mathbf{Reg}_{\mathsf{CB}}$ as follows:

$$
\begin{aligned}
\mathbf{Reg}_{\mathsf{CB}} &= \mathbb{E}\left[\sum_{t=1}^{T} f^\star(x_t, \pi^\star(x_t)) - \sum_{t=1}^{T} f^\star(x_t, a_t)\right]\\
&\leq 2N + \mathbb{E}\left[\sum_{t=2N+1}^{T} f^\star(x_t, \pi^\star(x_t)) - \sum_{t=2N+1}^{T} f^\star(x_t, a_t)\right]\\
&= 2N + \mathbb{E}\left[\sum_{t=2N+1}^{T} f^\star(x_t, \pi^\star(x_t)) - \sum_{t=2N+1}^{T} f^\star(x_t, \widehat{\pi}(x_t))\right]\\
&\leq 2N + T\cdot\mathbb{E}\left[V(\pi^\star) - V(\widehat{\pi})\right]\\
&\leq \mathcal{O}\left(N + \frac{TK}{\sigma}\sqrt{\frac{\log(|\mathcal{H}|T)}{N}}\right).
\end{aligned}
$$

The last step is from Lemma 5 with $\delta = \frac{1}{T}$. Picking $N = T^{2/3}K^{2/3}\sigma^{-2/3}\log^{1/3}(|\mathcal{H}|T)$ finishes the proof. □

C Proofs in Section 3.3

Theorem 2. *Under Assumptions 1-5, Algorithm 2 with certain choice of* N *and* γ *guarantees that* $\mathbf{Reg}_{\mathsf{CB}} = \mathcal{O}\left(\sqrt{KT\mathbf{Reg}_{\mathsf{Sq}}} + \sigma^{-2/3}(KT)^{2/3}\log^{1/3}(|\mathcal{H}|T)\right)$.

https://doi.org/10.52202/079017-3847

Proof. According to Eq. (9), we know that with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{a=1}^{K} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y|x,a}\left(\widehat{h}_a(x,y) - h_a^\star(x,y)\right)^2 \\
&\le K\cdot \mathbb{E}_{x\sim\mathcal{D},a\sim\pi_{\text{Unif}},y|x,a}\left[\left(\widehat{h}_a(x,y) - h_a^\star(x,y)\right)^2\right] \\
&\le K\cdot \mathbb{E}_{x\sim\mathcal{D},a\sim\pi_{\text{Unif}},y|x,a}\left[\|\widehat{h}(x,y) - h^\star(x,y)\|_2^2\right] \\
&\le \mathcal{O}\left(\frac{K\log\frac{|\mathcal{H}|}{\delta}}{N}\right), \qquad (18)
\end{aligned}$$

where the last inequality uses Eq. (9). Let $a_t^\star = \operatorname{argmax}_a f^\star(x_t,a)$ and recall that $\underline{f}^\star(x,a) := \mathbb{E}_{y|x,a}\left[G(h_a^\star(x,y), \frac{\theta}{\alpha}-\sigma,\sigma)\right] \in \mathcal{F}$. We have

$$\begin{aligned}
\mathbf{Reg}_{\mathsf{CB}} &= \mathbb{E}\left[\sum_{t=1}^{T} f^\star(x_t,a_t^\star) - \sum_{t=1}^{T} f^\star(x_t,a_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{y|x_t,a_t^\star}[\phi^\star(x_t,y)] - \sum_{t=1}^{T}\mathbb{E}_{y'|x_t,a_t}[\phi^\star(x_t,y')]\right] && \text{(From Assumption 2)} \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\underline{f}^\star(x_t,a_t^\star) - \sum_{t=1}^{T}\underline{f}^\star(x_t,a_t)\right] && \text{(From Lemma 4)} \\
&\le \frac{\gamma}{4}\mathbb{E}\left[\sum_{t=1}^{T}(\widehat{f}_t(x_t,a_t) - \underline{f}^\star(x_t,a_t))^2\right] + \mathcal{O}\left(\frac{TK}{\gamma}\right) && (19)
\end{aligned}$$

The last step is from Foster and Rakhlin [2020, Lemma 3]. Recall that $\widehat{r}_\sigma(x_t,y_t,a_t) = G(\widehat{h}_{a_t}(x_t,y_t), \frac{\theta}{\alpha}-\sigma,\sigma)$. Direct calculation shows that

$$\begin{aligned}
\mathbf{Reg}_{\mathsf{Sq}} &\ge \mathbb{E}\left[\sum_{t=1}^{T}(\widehat{f}_t(x_t,a_t) - \widehat{r}_\sigma(x_t,y_t,a_t))^2 - \sum_{t=1}^{T}(\underline{f}^\star(x_t,a_t) - \widehat{r}_\sigma(x_t,y_t,a_t))^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}(\widehat{f}_t(a_t) - \underline{f}^\star(x_t,a_t))(\widehat{f}_t(x_t,a_t) + \underline{f}^\star(x_t,a_t) - 2\widehat{r}_\sigma(x_t,y_t,a_t))\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}(\widehat{f}_t(x_t,a_t) - \underline{f}^\star(x_t,a_t))^2\right] \\
&\quad + 2\mathbb{E}\left[\sum_{t=1}^{T}(\widehat{f}_t(x_t,a_t) - \underline{f}^\star(x_t,a_t))(\underline{f}^\star(x_t,a_t) - \mathbb{E}_{y_t}[\widehat{r}_\sigma(x_t,y_t,a_t)])\right] \\
&\ge \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}(\widehat{f}_t(x_t,a_t) - \underline{f}^\star(x_t,a_t))^2\right] - 2\mathbb{E}\left[\sum_{t=1}^{T}(\underline{f}^\star(x_t,a_t) - \mathbb{E}_{y_t}[\widehat{r}_\sigma(x_t,y_t,a_t)])^2\right], \qquad (20)
\end{aligned}$$

where the last step is by AM-GM inequality. For the second term, we know that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\underline{f}^\star(x_t,a_t) - \mathbb{E}_{y_t}[\widehat{r}_\sigma(x_t,y_t,a_t)]\right)^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{E}_{y_t|x_t,a_t}\left[G\left(h_{a_t}^\star(x_t,y_t), \frac{\theta}{\alpha}-\sigma,\sigma\right) - G\left(\widehat{h}_{a_t}(x_t,y_t), \frac{\theta}{\alpha}-\sigma,\sigma\right)\right]\right)^2\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{y_t|x_t,a_t}\left[\left(G\left(h_{a_t}^\star(x_t,y_t), \frac{\theta}{\alpha}-\sigma,\sigma\right) - G\left(\widehat{h}_{a_t}(x_t,y_t), \frac{\theta}{\alpha}-\sigma,\sigma\right)\right)^2\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}\mathbb{E}_{y|x_t,a}\left[\left(G\left(h_a^\star(x_t,y),\frac{\theta}{\alpha}-\sigma,\sigma\right)-G\left(\widehat{h}_a(x_t,y),\frac{\theta}{\alpha}-\sigma,\sigma\right)\right)^2\right]\right]\\
&\leq \frac{1}{\sigma^2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}\mathbb{E}_{y|x_t,a}\left[\left(h_a^\star(x_t,y)-\widehat{h}_a(x_t,y)\right)^2\right]\right]\\
&\leq \mathcal{O}\left(\frac{TK\log(|\mathcal{H}|T)}{\sigma^2 N}\right), \qquad (21)
\end{aligned}
$$

where the first inequality is from Jensen's inequality; the third inequality is from that $G(v,\beta,\sigma)$ is $\frac{1}{\sigma}$-Lipschitz in v; and the last inequality is by Eq. (18) with $\delta=\frac{1}{T}$. Combining Eqs. (19)-(21), we obtain that

$$
\mathbf{Reg}_{\mathsf{CB}} \leq \mathcal{O}\left(\gamma\mathbf{Reg}_{\mathsf{Sq}}+\frac{TK\gamma\log(|\mathcal{H}|T)}{\sigma^2 N}+\frac{TK}{\gamma}+N\right).
$$

Picking $N=\frac{1}{\sigma}\sqrt{TK\gamma\log(|\mathcal{H}|T)}$ and $\gamma=\min\left\{\sqrt{KT/\mathbf{Reg}_{\mathsf{Sq}}},\sigma^{-2/3}(KT)^{2/3}\log^{-1/3}(|\mathcal{H}|T)\right\}$, we obtain that

$$
\mathbf{Reg}_{\mathsf{CB}} \leq \mathcal{O}\left(\sqrt{KT\mathbf{Reg}_{\mathsf{Sq}}}+\sigma^{-2/3}(KT)^{2/3}\log^{1/3}(|\mathcal{H}|T)\right),
$$

which finishes the proof. □

D Omitted Details in Section 4

D.1 Prompt Used in Generating Dataset

The prompt we use in answer generation basically the question itself shown as follows. We replace "question" by x in our experiment.

Prompt

{question}

The prompt we use in feedback generation as follows. We replace "question" and "answer" by x and a respectively in our experiments. We replace "mood" by "satisfied" ("not satisfied") when the answer is generated by Qwen1.5-32B-Chat (Qwen1.5-0.5B-Chat).

Prompt

System: You are a user simulator. A user has been presented a Question and an Answer. Simulate the user's next statement. The user is {mood} with the Answer to the Question.
Question: {question}
Answer:{answer}

D.2 Prompt Used in Learning Inverse Kinematic Model

The prompt we use in learning the inverse kinematic model is as follows. We replace "question", "answer", and "feedback" by x, a, and y respectively in our experiments.

Prompt

System: You are a conversation evaluating agent. Given a User's Question, an Answer, and the User's Feedback: determine if the User's Feedback is consistent with Answer. Respond with Yes or No only.
User's Question: {question}
Answer:{answer}
User's Feedback:{feedback}

https://doi.org/10.52202/079017-3847

Respond with Yes or No only.

D.3 Prompt Used in Learning Policy

The prompt we use in learning the reward predictor is as follows. We replace "question" and "answer" by x and a respectively in our experiments.

Prompt

System: You are an Answer evaluating agent. Given a User's Question and an Answer: assess if the Answer is good. Respond with Yes or No only.
User's Question:{question}
Answer:{answer}
Respond with Yes or No only.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See abstract and Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 1 and Section 2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-3847

Justification: See Assumption 1, Assumption 2, Assumption 3, Assumption 4, Assumption 5, and the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Algorithm 1, Algorithm 2, and detailed descriptions in Section 4 and Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

https://doi.org/10.52202/079017-3847

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics. The research conducted in this paper conforms with it in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work is mostly theoretical, and we do not foresee any negative ethical or societal outcomes.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the used dataset and models and respect the licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

https://doi.org/10.52202/079017-3847

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.