Reinforcement Learning from Multi-level and Episodic Human Feedback

Muhammad Qasim Elahi* ELAHI0@PURDUE.EDU
Somtochukwu Oguchienti* SOGUCHIE@PURDUE.EDU
Maheed H. Ahmed AHMED237@PURDUE.EDU
Mahsa Ghasemi MAHSA@PURDUE.EDU
Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

Designing an effective reward function has long been a challenge in reinforcement learning, particularly for complex tasks in unstructured environments. To address this, various learning paradigms have emerged that leverage different forms of human input to specify or refine the reward function. Reinforcement learning from human feedback is a prominent approach that utilizes human comparative feedback—expressed as a preference for one behavior over another—to tackle this problem. In contrast to comparative feedback, we explore multi-level human feedback, which is provided in the form of a score at the end of each episode. This type of feedback offers more coarse but informative signals about the underlying reward function than binary feedback. Additionally, it can handle non-Markovian rewards, as it is based on an entire episode's evaluation. We propose an algorithm to efficiently learn both the reward function and the optimal policy from this form of feedback. Moreover, we show that the proposed algorithm achieves sublinear regret and demonstrate its empirical effectiveness through extensive simulations.

Keywords: Categorical feedback, Preference-based reinforcement learning, Sparse reward

1. INTRODUCTION

Reinforcement Learning (RL), a framework for addressing the problem of decision-making under uncertainty, has proven to be highly effective in tackling complex real-world challenges. In the traditional RL setting, the agent interacts with the environment and receives a reward signal for each state-action pair. This paradigm has achieved huge successes in real-life applications ranging from games Mnih (2013); Mnih et al. (2015); Silver et al. (2016, 2018), manufacturing Wang and Usher (2005); Liu et al. (2022), medicine Zhao et al. (2011); Coronato et al. (2020).

This conventional setting assumes that a reward signal is readily available to the agent and accurately specifies the objective of the task. However, many real-world settings lack a clear reward signal for each state-action pair. For instance, in tasks like autonomous driving or stacking books on a shelf, defining a precise reward is complex, but it is feasible to evaluate success at the end of the task. Therefore, engineering a reward is very difficult and poses a significant challenge to RL. To address this challenge, Eysenbach et al. (2018) proposes methods that do not rely on predefined reward functions to learn diverse skills. Some earlier works explore Inverse Reinforcement Learning (IRL) as an approach to learn a reward function from demonstrations of successful trajectories Abbeel and Ng (2004); Ziebart et al. (2008). However, IRL depend heavily on high-quality

*Equal Contribution

demonstrations, which can be costly and difficult to obtain. Recently, Reinforcement Learning from Human Feedback (RLHF) has emerged as a promising approach, particularly in human-robot interaction settings, where human feedback replaces the need for explicit reward functions Knox and Stone (2008); MacGlashan et al. (2017). Within the framework of RLHF, Christiano et al. (2017) proposes pairwise comparison of trajectories while Rendle et al. (2012); Brown et al. (2019) leverage ranking of set of trajectories. Additionally, several techniques strategically combine both demonstrations and preferences to learn the reward function Bıyık et al. (2022). However, feedback from preferences is typically less informative than scalar reward functions, which poses a challenge for RLHF. Moreover, humans can be biased in their preferences, hence preferences may not capture the true reward signal.

Addressing the limitations of IRL and RLHF, Agarwal et al. (2019) proposes RL from sparse rewards where feedback is only provided at the end of a trajectory. This technique has shown potential in achieving optimal policies with low regret. However, the authors consider a setting where only binary human feedback scores are available Chatterji et al. (2021). In contrast, we examine a richer feedback structure, where the agent receives different levels of feedback, ranging from 0 to $K-1$, instead of a binary feedback. We propose an algorithm that leverages such categorical feedback to learn the underlying reward function. The main contributions of our work are as follows:

- We propose an online optimism-based algorithm that leverages multi-level and episodic feedback signals to estimate the underlying reward function and, consequently, learn the optimal policy for an episodic MDP setting.
- We analyze the proposed algorithm and show that it achieves sublinear regret in terms of the number of episodes under mild regularity assumptions.
- We show experimentally, using series of grid-world simulations, that our algorithm is able to learn an optimal policy from multilevel feedback.

2. RELATED WORK

Preference-based reinforcement learning (PbRL) has been extensively studied empirically, with diverse applications in domains such as games Ibarz et al. (2018); Wirth and Fürnkranz (2014); Runarsson and Lucas (2014), robotics Jain et al. (2013); Kupcsik et al. (2018) and training of large language models Stiennon et al. (2020); Ouyang et al. (2022). These empirical successes have motivated theoretical analysis of PbRL, providing provable guarantees for learning near-optimal policies across different settings. Pacchiano et al. (2021b); Zhan et al. (2023b); Novoseller et al. (2020) explore the online setting under linear or tabular MDPs for learning near-optimal policies from human preferences. In the offline setting, Shin et al. (2023); Zhan et al. (2023a); Zhu et al. (2023) show that near-optimal policies can be learned from collected human preference data under the maximum likelihood framework. Extending beyond tabular MDPs, Chen et al. (2022) explores PbRL with general function approximation. While Novoseller et al. (2020); Pacchiano et al. (2021b) assume an underlying utility function generating human preferences, Dudík et al. (2015); Wang et al. (2023) adopt more general preference models based on the von Neumann winner policy. More recent studies explore preference feedback via pairwise comparisons Novoseller et al. (2020); Wirth et al. (2016) or a score-based feedback Efroni et al. (2021); Chatterji et al. (2021).

IRL techniques assume the existence of expert demonstrations to learn a reward function for the agent Abbeel and Ng (2004). Hence, the policy derived from the learned reward function is inherently limited by the quality of the collected demonstrations. In practice, expert demonstrations are difficult to obtain due to the complexity of the environment or the demonstrator's inexperience, which can limit the quality of their demonstrations. This challenge motivated recent studies to address the problem of suboptimality in the human demonstrations Brown et al. (2019); Oguchienti and Ghasemi (2023); Chen et al. (2021). Brown et al. (2019) proposes T-REX algorithm to learn an optimal policy from trajectory rankings. Building on this, Brown et al. (2020); Chen et al. (2021) design a learning framework using noise-injected policies to learn a reward function. Assessing demonstrator expertise, Beliaev et al. (2022) explores an imitation learning setting to learn an optimal policy alongside the expertise level corresponding to each demonstrator. Similarly, Oguchienti and Ghasemi (2023) extends to an IRL setting to learn the expertise levels of each demonstrator and incorporates this parameter into the reward learning framework. Furthermore, Beliaev and Pedarsani (2024) models suboptimality in human demonstrations via their reward bias and action variance and proposes an algorithm to learn an optimal policy alongside these parameters.

Online RL focuses on techniques that achieve low regret for agent's interaction while balancing exploration and exploitation based on the principle of optimism in the face of uncertainty. Several authors have analyzed the sample complexity of online RL in episodic MDPs. For instance, Osband and Van Roy (2016); Azar et al. (2017); Pacchiano et al. (2021a) provide theoretical guarantees for regret bounds in this setting. Additionally, Auer and Ortner (2006); Auer et al. (2008) propose the upper confidence reinforcement learning (UCRL) algorithm for undiscounted MDPs, aimed at deriving regret bounds over a finite number of steps. Liu and Su (2020); Zhou et al. (2021) provide similar analysis for discounted MDP settings. Moreover, Azar et al. (2017) develop the upper confidence bound value iteration (UCB-VI) algorithm to address the problem of optimal exploration in finite-horizon MDPs. Beyond UCB-based approaches, Osband et al. (2013); Agrawal and Jia (2017) explore posterior sampling based RL (PSRL) algorithm that samples from the posterior distribution over MDPs and chooses the optimal actions based on the sampled MDP. Similar to Chatterji et al. (2021), our work adopts the UCB-based framework for non-Markovian rewards. However, Chatterji et al. (2021) assumes that the reward is drawn from an unknown binary logistic model. This model is inherently limited since the agent receives a score of 0 or 1 indicating whether a goal is achieved or not. To address this limitation, we propose a more informative feedback model based on a categorical distribution, specifically the softmax distribution. This approach allows the agent to receive a richer feedback based on the observed trajectory, overcoming the binary feedback constraint.

3. PRELIMINARIES

A Markov Decision Process (MDP) in an episodic setting is represented as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, H, \rho)$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathbb{P}(\cdot|s, a)$ represents the state transition distribution, and H is the horizon or length of trajectories. The initial state s_0 is sampled from a known distribution ρ. We assume that all state transition probabilities under all actions are known. A trajectory $\tau = (s_0, a_0, ..., a_{H-1}, s_H)$ consists of all states and actions for a given episode, and we denote the set of all possible trajectories by $\mathcal{T}$. Our formulation differs from the standard RL setting, where the agent receives a reward for every state-action pair, i.e., $R(s, a)$. Instead, the reward is a function of the entire trajectory, $R(\tau)$. A reward that is a function of the entire trajectory can capture non-Markovian tasks by incorporating historical context. We propose K-ary trajectory feedback labels, where the agent receives a categorical score rather than a binary score. K-ary feedback is easier for humans to provide, serves as an abstraction of real-valued rewards, and is more expressive than binary feedback as used in Chatterji et al. (2021). Let $\mathbf{w}_i^\star \in \mathbb{R}^d, \ \forall i \in \{0, 1, \dots, K-1\}$ be a set of K vectors unknown to the learner. Furthermore, every trajectory τ is associated with a known feature vector $\phi(\tau) \in \mathbb{R}^d$.

Assumption 1 (K-ary Categorical Feedback) *For any trajectory $\tau \in \mathcal{T}$, the agent receives categorical feedback $y_\tau \in \{0, 1, 2, ...K-1\}$ with the probability distribution as follows:*

$$P(y_\tau = i) = \frac{\exp(\mathbf{w}_i^{\star T} \phi(\tau))}{\sum_{j=0}^{K-1} \exp(\mathbf{w}_j^{\star T} \phi(\tau))},$$

where $i \in \{0, 1, 2, ...K-1\}$.

Assumption 2 (Bounded Parameters) *We assume that:*

- $\|\mathbf{w}_i^\star\|_2 \leq \frac{B}{K} \ \ \forall i \in 0, \cdots, K-1$ *and for some known* $B > 0$,
- $\|\phi(\tau)\|_2 \leq 1 \ \ \forall \tau \in \mathcal{T}$.

We make boundedness assumptions on the features and true weight parameter, which are commonly used in prior literature Chatterji et al. (2021); Zhu et al. (2023); Faury et al. (2020); Russac et al. (2021). Notice that this assumption holds without loss of generality for any finite features and finite true parameter with the appropriate choice of B. We denote the trajectories in the N episodes by $\{\tau^{(n)}\}_{n=1}^N$ and the corresponding categorical feedback received by $\{y^{(n)}\}_{n=1}^N$. Given the stochastic nature of the reward feedback, i.e., the categorical distribution above, we aim to maximize the expected value of this reward feedback i.e. R(τ) = $\mathbb{E}[y_\tau]$. This is more formally expressed in the equation below:

$$R(\tau) = \sum_{i=0}^{K-1} i P(y_\tau = i) = \sum_{i=0}^{K-1} i \frac{\exp(\mathbf{w}_i^{\star T} \phi(\tau))}{\sum_{j=0}^{K-1} \exp(\mathbf{w}_j^{\star T} \phi(\tau))}. \tag{1}$$

The value function of a policy $V^\pi(s)$ for any state $s \in \mathcal{S}$ is defined as:

$$V^\pi(s) := \mathbb{E}_{\substack{s_t \sim \mathbb{P}(.|s_{t-1}, a_{t-1}) \\ a_t \sim \pi(.|s_t)}} \big[R(\tau) \mid S_0 = s \big] = \mathbb{E}_{\tau \sim \mathbb{P}^\pi(.|s)}[R(\tau)],$$

The average value function for some initial state distribution ρ is defined as $V^\pi := \mathbb{E}_{s_0 \sim \rho}[V^\pi(s_0)]$. The optimal policy $\pi^\star$ is given by $\pi^\star \in \arg\max_{\pi \in \Pi} V^\pi$, where Π is the set of all possible policies, including non-Markovian policies as well. It is worth noting that, similar to Chatterji et al. (2021), in our setting, the optimal policy may be *non-Markovian* because of the non-Markovian reward. We use the notation $\pi^{(n)}$ to represent the policy used for the episode n. Our goal is to minimize the cumulative regret $\mathcal{CR}(N)$ of the agent over a fixed number of episodes N, which represents the cumulative suboptimality of the agent over the span of its learning/interaction.

$$\mathcal{CR}(N) := \sum_{n=1}^{N} V^{\pi^\star} - V^{\pi^{(n)}}. \tag{2}$$

4. LEARNING FROM MULTILEVEL FEEDBACK SCORES

In this section, we discuss the design of our proposed algorithm that leverages multilevel episodic feedback scores and present theoretical guarantees demonstrating that the cumulative regret for our algorithm exhibits sublinear scaling with respect to the number of episodes N. The details of proofs for lemmas 1, 2 and theorem 3 can be found in the supplementary section of the full paper here.

4.1. Algorithm Design and Analysis

In this subsection, we present the details of the algorithm design and provide our main theoretical results, including the cumulative regret guarantee. We propose Algorithm 1, which is an adaptation of the UCBVI algorithm Azar et al. (2017) to our setting with multilevel episodic feedback and known transition probabilities. The learner does not have access to the true weight vectors $\mathbf{w}_i^\star$, $\forall i \in \{0, 1, \ldots, K-1\}$. Therefore, an estimate for these weight vectors must be constructed. Instead of estimating the weights individually, these vectors are concatenated into a single vector $\mathbf{w}^\star \in \mathbb{R}^{Kd}$. The feature vectors for the trajectories $(\phi(\tau))$ are transformed into a set of vectors $\phi_i(\tau) \in \mathbb{R}^{Kd}$, $\forall i \in \{0, 1, \ldots, K-1\}$. All entries of the vector $\phi_i(\tau)$ are zeros except for those from index $i \times d$ to $i \times d + (d-1)$, which correspond to the entries of the feature vector $\phi(\tau)$. Based on Assumption 2, it follows that $||\mathbf{w}^\star||_2 \le B$ and $||\phi_i(\tau)||_2 \le 1$, $\forall i \in \{0, 1, \ldots, K-1\}$. The reward for a trajectory $R(\tau)$ can equivalently be expressed as follows:

$$R(\tau) = \sum_{i=0}^{K-1} iP(y_\tau = i) = \sum_{i=0}^{K-1} i \frac{\exp(\mathbf{w}^{\star T}\phi_i(\tau))}{\sum_{j=0}^{K-1}\exp(\mathbf{w}^{\star T}\phi_j(\tau))}.$$

We define feasible set for the estimated weight vector as $\mathbf{W}_B := \{\mathbf{w} \in \mathbb{R}^{kd} \mid ||\mathbf{w}||_2 \le B\}$. For any episode, say $n \in [N]$, we construct a maximum likelihood estimate (MLE) $\widehat{\mathbf{w}}_n$ based on the sampled trajectories $\{\tau^{(i)}\}_{i=1}^n$ and the corresponding categorical feedback received as $\{y^{(i)}\}_{i=1}^n$:

$$\widehat{\mathbf{w}}_n = \operatorname*{argmin}_{\mathbf{w} \in \mathbf{W}_B} \ell^{(n)}(\mathbf{w}),$$

where the loss function $\ell^{(n)}(\mathbf{w})$ is defined as the negative log likelihood of the data samples, i.e., $\ell^{(n)}(\mathbf{w}) = -\frac{1}{n}\log P(y^{(1)}, \ldots, y^{(n)}|\tau^{(1)}, \ldots, \tau^{(n)}, \mathbf{w}) = -\frac{1}{n}\sum_{t=1}^n \log P(y^{(t)}|\tau^{(t)}, \mathbf{w})$. Based on Assumption 1, we have $P(y^{(t)}|\tau^{(t)}, \mathbf{w}) = \frac{\exp(\mathbf{w}^T\phi_{y^{(t)}}(\tau^{(t)}))}{\sum_{j=0}^{K-1}\exp(\mathbf{w}^T\phi_j(\tau^{(t)}))}$, which yields

$$\widehat{\mathbf{w}}_n = \operatorname*{argmin}_{\mathbf{w} \in \mathbf{W}_B} -\frac{1}{n}\sum_{t=1}^n \log \frac{\exp(\mathbf{w}^T\phi_{y^{(t)}}(\tau^{(t)}))}{\sum_{j=0}^{K-1}\exp(\mathbf{w}^T\phi_j(\tau^{(t)}))}. \tag{3}$$

Lemma 1 bounds the estimation error between the true weight vector and our MLE estimate. This result relies on the fact that the loss function $\ell^{(n)}(\mathbf{w})$ is strongly convex in the weight vector $\mathbf{w}$, and the entire proof is provided in the supplementary material. We use the notation $\lambda_{\text{min}}(A)$ to denote the smallest eigenvalue of matrix A.

Lemma 1 *For any episode $n \in [N]$, the following holds with probability at least $1-\delta$:*

$$\big|\big|\widehat{\mathbf{w}}_n - \mathbf{w}^\star\big|\big|_2 \le \frac{2}{\eta\lambda_{\text{min}}(\Sigma_{D_n})}\sqrt{\frac{C^2}{2n}\log\frac{4}{\delta}}\ ,$$

where $\Sigma_{D_n} = \frac{1}{nK^2}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{l=0}^{K-1}(\phi_j(\tau^{(i)}) - \phi_l(\tau^{(i)}))\,(\phi_j(\tau^{(i)}) - \phi_l(\tau^{(i)}))^T$, $\eta = \frac{\exp(-4B)}{2}$ *and* $C = \log\big(K\exp(2B)\big)$.

In this work, we assume that the matrix Σ_{D_n} is well-conditioned and invertible, which implies that $\lambda_{\text{min}}(\Sigma_{D_n})$ is bounded away from zero. In practice, one can use a regularized MLE estimator when the matrix Σ_{D_n} is non-invertible. Lemma 1 shows that the L_2-norm distance between the MLE estimate $\widehat{\mathbf{w}}_n$ and the true weight $\mathbf{w}^\star$ remains bounded with high probability. Additionally, the MLE estimate converges to the true weight with high probability for large enough episode numbers n. Since we don't have access to the true weight vector $\mathbf{w}^\star$, we do not have access to the reward function for a trajectory $R(\tau)$. Instead, using the MLE estimate $\widehat{\mathbf{w}}_n$, we can construct an estimate of the reward function as follows:

$$R(\widehat{\mathbf{w}}_n, \tau) = \sum_{i=0}^{K-1} i\widehat{P}(y_\tau = i) = \sum_{i=0}^{K-1} i\frac{\exp(\widehat{\mathbf{w}}_n^T\phi_i(\tau))}{\sum_{j=0}^{K-1}\exp(\widehat{\mathbf{w}}_n^T\phi_j(\tau))}.$$

Using the concentration result in Lemma 1 on the weight parameters, we show the concentration of the estimated reward, that is, $R(\widehat{\mathbf{w}}_n, \tau)$ converges to the true reward function $R(\tau)$, as defined in Equation 1, with a high probability for sufficiently large n. This result is crucial since the estimated reward function serves as a proxy for the true reward function for the next stage of the algorithm that learns the policy.

Lemma 2 *For any possible trajectory* $\tau \in \mathcal{T}$*, the following holds with probability at least* $1-\delta$*:*

$$\Big|R(\widehat{\mathbf{w}}_n, \tau) - R(\tau)\Big| \leq \frac{4K\exp(4B)}{\eta\lambda_{\text{min}}(\Sigma_{D_n})}\sqrt{\frac{C^2}{2n}\log\frac{4}{\delta}}\,,$$

where $\Sigma_{D_n} = \frac{1}{nK^2}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{l=0}^{K-1}(\phi_j(\tau^{(i)}) - \phi_l(\tau^{(i)}))\,(\phi_j(\tau^{(i)}) - \phi_l(\tau^{(i)}))^T$, $\eta = \frac{\exp(-4B)}{2}$ *and* $C = \log\big(K\exp(2B)\big)$.

$$\overline{R}(\widehat{\mathbf{w}}_n, \tau) = \min\left(R(\widehat{\mathbf{w}}_n, \tau) + \frac{4K\exp(4B)}{\eta\lambda_{\text{min}}(\Sigma_{D_n})}\sqrt{\frac{C^2}{2n}\log\frac{4}{\delta}},\ K-1\right) \tag{4}$$

The concentration result in Lemma 2 shows that the estimate of the reward function using the MLE estimate $\widehat{\mathbf{w}}_n$ is close to the true reward function. Building upon the principle of optimism in the face of uncertainty—an established framework for online learning algorithms Lattimore and Szepesvári (2020); Chatterji et al. (2021); Faury et al. (2020)—we utilize this concentration result to define the optimistic reward function, as expressed in Equation 4, denoted by $\overline{R}(\widehat{\mathbf{w}}_n, \tau)$. The minimum function in Equation 4 ensures that the optimistic estimate does not go beyond $K-1$, which is the highest level of feedback. By applying Lemma 2, we observe that $R(\tau) \leq \overline{R}(\widehat{\mathbf{w}}_n, \tau)$ with at least $1-\delta$ probability. It's important to note that the width of the confidence interval in Lemma 2 scales exponentially with respect to the parameter B, similar to the findings in Chatterji et al. (2021). However, in practice, this width can be large due to the constants involved. In our experiments section, to avoid scaling with respect to these constants, we use a confidence interval width of $\mathcal{O}(\frac{1}{\sqrt{n}})$.

Our proposed Algorithm 1 involves two key steps: (1) computing the MLE estimate $\widehat{\mathbf{w}}_n$ using the sampled trajectory and corresponding feedback labels, and (2) computing the policy $\pi^{(n)}$

Algorithm 1: K-UCBVI with Multilevel Episodic Feedback and Known Transition Probabilities.

```
for $n = 1, \cdots, N$ do
    if $n > 1$ then
        $$\pi^{(n)} \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s_0 \sim \rho, \tau \sim \mathbb{P}^{\pi}(.|s_0)}[\overline{R}(\widehat{\mathbf{w}}_n, \tau)] \quad (5)$$
    else
        set $\pi^{(1)}(.|s)$ to be uniform distribution over the action set for all the states.
    end
    Observe the trajectory $\tau^{(n)} \sim \mathbb{P}^{\pi^{(n)}}$ and the corresponding feedback $y^{(n)}$.
    Calculate $\widehat{\mathbf{w}}_n$ by solving Equation (3).
end
```

by solving the optimization in Equation 5. Note that the solution to the optimization problem in Equation 5 can generally be a non-Markovian policy, as our reward function depends on the entire trajectory, which may not be Markovian in general. To address this issue, in the next subsection, we propose limiting the search to Markovian policies. This restriction simplifies the problem and allows us to leverage a variety of RL algorithms, including policy gradient methods such as the REINFORCE algorithm—a Monte Carlo variant of policy gradients. By focusing on Markovian policies, we avoid solving the optimization problem in Equation 5 directly and instead approximate the solution empirically. While one could extend the search to non-Markovian policies by employing memory-based strategies—such as using the entire history, a window of recent history, or a low-dimensional embedding of the history (e.g., through recurrent neural networks)—such approaches often become computationally intractable for Markov Decision Processes (MDPs) with long horizons.

We prove a high-probability regret bound for our proposed Algorithm using the concentration results in Lemma 2, and show that the regret is sublinear in the number of episodes N. The regret bound is stated below, with the detailed regret analysis provided in the supplementary material.

Theorem 3 *For any $\delta \in (0, 1]$, under Assumptions 1 and 2, the cumulative regret of K-UCBVI is upper bounded with probability at least $1 - \delta$ as follows:*

$$\mathcal{CR}(N) \leq \frac{16K\exp(4B)}{\eta\lambda}\sqrt{\frac{NC^2}{2}\log\frac{12N}{\delta}} + 4K\sqrt{N\log\left(\frac{18N\log N}{\delta}\right)},$$

where $\lambda = \min_{i\in\{1,2,3,\ldots,N\}} \lambda_{\min}(\Sigma_{D_n})$.

Theorem 3 shows that the regret of the proposed algorithm scales as $\mathcal{O}\left(\sqrt{N}\log\frac{N\log N}{\delta}\right)$ with respect to the number of episodes N, with a probability of at least $1 - \delta$, which establishes the sublinearity of the regret. The regret scaling of our algorithm with respect to the number of episodes is similar to that in Chatterji et al. (2021), which considers binary feedback scores. There is also an exponential dependence on the parameter B, similar to the term κ in the regret bound presented in Chatterji et al. (2021). This issue of exponential scaling of regret bounds with respect to the bound parameter has also been discussed in Faury et al. (2020).

4.2. Practical Approximation Using the REINFORCE Algorithm

In this subsection, we propose an alternative to the optimization in Equation 5, which is not easy to solve in general. We focus on limiting our search to Markovian policies where the action depends only on the current state, and we use a REINFORCE-style algorithm Agarwal et al. (2019) to compute the optimal policy. In general, the deterministic class of policies will not be differentiable, so we can't employ the gradient ascent approach. This motivates us to consider policy classes that are stochastic, which permit differentiability. We restrict ourselves to the class of stochastic Markovian policies parameterized by some $\theta \in \mathbb{R}^d$. Consider a parametric class of policies $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^d\}$. The objective is to maximize the value function over the class of policies, i.e., $\max_{\theta\in\mathbb{R}^d} V^{\pi_\theta}$. We specifically consider a softmax set of policies (Definition 4), which can model all the stationary (Markovian) policies where the action depends only on the current state. For any realizable trajectory ($\tau \in \mathcal{T}$), we use the notation $Pr_\rho^{\pi_\theta}(\tau)$ for the probability of realizing trajectory τ under policy π_θ and initial state distribution ρ. Also, we have $Pr_\rho^{\pi_\theta}(\tau) = \rho(s_0)\pi_\theta(a_0|s_0)P(s_1|s_0,a_0)\pi_\theta(a_1|s_1)\dots P(s_H|s_{H-1},a_{H-1})$.

Definition 4 (Softmax Policies) *The softmax policy is an explicit tabular representation of a policy defined as:*

$$\pi_\theta(a|s) = \frac{\exp\theta_{a,s}}{\sum_{a'\in\mathcal{A}}\exp\theta_{a',s}}.$$

The parameter space is $\Theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. The closure of the set of softmax policies contains all stationary and stochastic policies. Instead of the optimization problem in Equation 5, we solve

$$\arg\max_{\theta\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \overline{V}_n^{\pi_\theta} := \mathbb{E}_{\tau\sim Pr_\rho^{\pi_\theta}}\left[\overline{R}(\widehat{\mathbf{w}}_n,\tau)\right]$$

$\overline{V}_n^{\pi_\theta}$ is the optimistic value function at episode n. The gradient of the optimistic value function of policy π_θ with respect to the parameter vector θ is given by:

$$\nabla\overline{V}_n^{\pi_\theta} = \mathbb{E}_{\tau\sim Pr_\rho^{\pi_\theta}}\left[\overline{R}(\widehat{\mathbf{w}}_n,\tau)\sum_{t=0}^{H}\nabla\log(\pi_\theta(a_t|s_t))\right].$$

This follows from the fact that $\nabla\overline{V}_n^{\pi_\theta} = \sum_{\tau\in\mathcal{T}}\overline{R}(\widehat{\mathbf{w}}_n,\tau)Pr_\rho^{\pi_\theta}(\tau)\nabla\log Pr_\rho^{\pi_\theta}(\tau)$. Using the definition of $Pr_\rho^{\pi_\theta}(\tau)$, we obtain the final result $\nabla\overline{V}_n^{\pi_\theta} = \mathbb{E}_{\tau\sim Pr_\rho^{\pi_\theta}}\left[\overline{R}(\widehat{\mathbf{w}}_n,\tau)\sum_{t=0}^{H}\nabla\log(\pi_\theta(a_t|s_t))\right]$.

Remark 5 *For the softmax class of policies where* $\pi_\theta(a_t|s_t) = \frac{\exp\theta_{a_t,s_t}}{\sum_{a'\in\mathcal{A}}\exp\theta_{a'_t,s_t}}$, *the partial derivatives* $\frac{\partial\log(\pi_\theta(a_t|s_t))}{\partial\theta_{a,s}} = \mathbb{1}_{s_t=s}[\mathbb{1}_{a_t=a} - \pi_\theta(a_t|s_t)]$. *This property allows the gradient* $\nabla\log(\pi_{\theta_t}(a_t|s_t)$ *to be computed efficiently.*

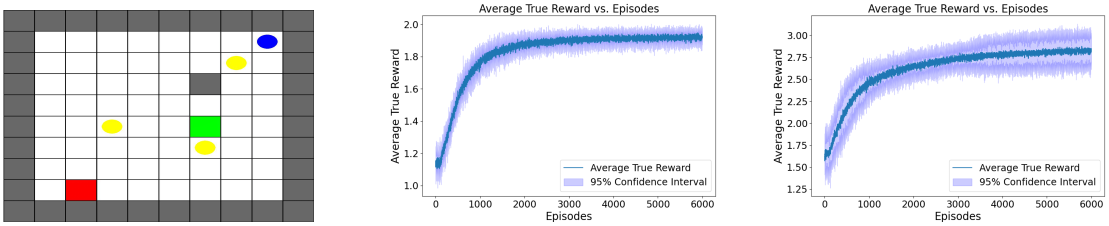
In this practical implementation, the objective is to maximize the optimistic value function with respect to the parameter vector θ, which parameterizes the policy. We use the stochastic gradient ascent algorithm for this purpose. At each step n of gradient ascent, we sample a fixed number of trajectories using the transition probabilities and the policy π_{θ_i}. Then, we compute the average policy gradient over these samples and update the parameter vector as $\theta_{i+1} = \theta_i + \eta\nabla\overline{V}_n^{\pi_{\theta_i}}$ until some termination criteria are met, such as when the change in the parameter vector becomes small.

5. EXPERIMENTS

This section presents the experiments conducted to evaluate the optimistic learning algorithm from multi-level feedback. We test the proposed algorithm on various grid-world environments and demonstrate empirically that the agent is able to learn an optimal policy from multi-level feedback. Specifically, we showcase the results from an 8×8 grid-world environment with the feedback level $K = 4$ and 6. The practical implementation of the algorithm, as used in the experiments, is detailed in Algorithm 2 of the supplementary material.

5.1. Simulation Setting

We describe the simulation setup for an 8×8 grid-world depicted in Figure 1(a) with the given state configurations. The goal of the agent, represented with the blue circle, is to collect the coins indicated by the three yellow circles and then reach the goal state depicted by the green square. In addition, the agent should avoid the absorbing danger state represented by the red square where feedback is the lowest. Walls are represented as gray cells. We set the horizon $H = 50$. The agent can move to any cell on the grid-world with actions—up, down, left, or right. The agent can move to the intended cell with a probability 0.91 and to the remaining three cells with a probability of 0.03 each.



(a) **(b)** **(c)**

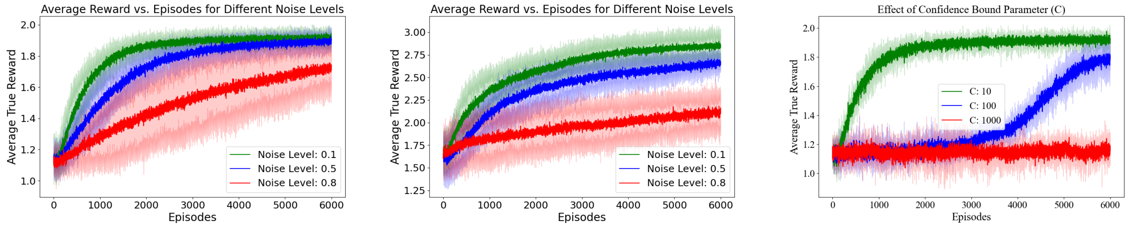
Figure 1: (a) 8×8 grid-world environment with the danger state (red cell), wall state (gray cells) and goal state (green cell) depicted. (b) Plot of the average true reward against the number of episodes for $K = 4$. (c) Plot of the average true reward against the number of episodes for $K = 6$.

The true reward model, parameterized by $\mathbf{w}^\star$ is unknown to the agent. To generate a realistic $\mathbf{w}^\star$, we design a rule-based reward mechanism and learn a corresponding $\mathbf{w}^\star$ that defines how well the agent achieves the goal. Particularly, for any feedback level, we assign a reward of 0 if the agent reaches the danger state at the end of the episode. Otherwise, the agent gets an additional level of feedback for each collected coin with the highest feedback assigned if all coins have been collected and the agent reached the goal. We define the features $\phi(\tau)$ as a vector with dimension $(4 + c)$ where c is the number of coins. The elements of the vector are the Manhattan distance between the final state at the end of the horizon s_H^τ and the goal state G, the Manhattan distance between s_H^τ and the danger state, an indicator function that represents whether or not the agent is at the goal state, an indicator function that represents whether or not the agent is at the danger state, and an indicator function for each coin determining whether or not it was picked. Given a softmax policy π_θ parameterized by θ, we train and update θ using the REINFORCE algorithm. Empirically, we approximate the expectation in the gradient of the value function by sampling 50 trajectories. At convergence, we sample a trajectory and the corresponding feedback from $\mathbf{w}^\star$ and update $\widehat{\mathbf{w}}$ using the projected gradient descent step. We repeat the experiment over 20 independent runs and 6000 episodes. The plots of the average true reward against the number of episodes with a confidence interval of ±2 standard deviations are presented in Figures 1(a) and (b).

5.2. Simulation Results

Human feedback is inherently prone to bias, and preference models like the Bradley-Terry model Bradley and Terry (1952) and the categorical model proposed in this work, may not accurately represent the feedback. Hence, we empirically analyze the robustness of the proposed algorithm to model misspecification. To do this, we introduce perturbations in the form of noise modeled as a uniform distribution over the feedback levels. We define the noisy feedback as a convex combination of the true feedback and varying levels of noise. Figure 2(a) and (b) illustrate the impact of different noise levels on the performance of the learned policy. For instance, at a noise level of 0.1, 10% of the feedback is sampled from the noise distribution, while 90% is sampled from the true feedback. It can be observed that incorporating noise affects the rate of convergence of the policy with slower convergence rates as the noise level increases.

As stated in section 4, it should be noted that in the experiments, we set the confidence term in the optimistic reward function to $(C \times \frac{1}{\sqrt{n}})$ where C is confidence bound parameter and n is the number of samples or episodes. For our experiments, we set $C = 10$. Figure 2(c) compares the effects of different values of C on the learned policy.



(a) **(b)** **(c)**

Figure 2: (a) Impact of varying noisy feedback on the learned policy $(K = 4)$. (b) Impact of varying noisy feedback on the learned policy $(K = 6)$. (c) Effect of the confidence bound parameter on the learned policy $(K = 4)$.

6. CONCLUSION AND FUTURE WORK

In this paper, we consider a multi-level feedback mechanism where the agent receives a score at the end of each episode. We design an optimistic algorithm for learning an optimal policy from these feedback scores and prove that, under mild assumptions, the algorithm achieves cumulative regret that scales sublinearly with respect to the number of episodes. We perform experiments on various grid-world environments to demonstrate that the algorithm learns an optimal policy. Furthermore, we conduct perturbation analysis on the feedback model to assess the impact of noise on the performance of the policy. As future work, we propose to build upon our quantized feedback model—which discretizes continuous reward signals into finite set of levels—by exploring information-theoretic approaches to understand how such quantization can affect the optimality of the learned policy. Additionally, we aim to investigate alternative preference models and incorporate more robust analysis techniques to address challenges such as model misspecification.

7. ACKNOWLEDGMENTS

This work was supported in part by the National Aeronautics and Space Administration under Grant 80NSSC24M0070.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

Rishabh Agarwal, Chen Liang, Dale Schuurmans, and Mohammad Norouzi. Learning to generalize from sparse and underspecified rewards. In *International Conference on Machine Learning*, pages 130–140. PMLR, 2019.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 30, 2017.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems*, 19, 2006.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Mark Beliaev and Ramtin Pedarsani. Inverse reinforcement learning by estimating expertise of demonstrators. *arXiv preprint arXiv:2402.01886*, 2024.

Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, and Ramtin Pedarsani. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*, pages 1732–1748. PMLR, 2022.

Erdem Bıyık, Dylan P Losey, Malayandi Palan, Nicholas C Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *The International Journal of Robotics Research*, 41(1): 45–67, 2022.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International Conference on Machine Learning*, pages 783–792. PMLR, 2019.

Daniel S Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on Robot Learning*, pages 330–359. PMLR, 2020.

Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. *Advances in Neural Information Processing Systems*, 34:3401–3412, 2021.

Letian Chen, Rohan Paleja, and Matthew Gombolay. Learning from suboptimal demonstration via self-supervised reward regression. In *Conference on Robot Learning*, pages 1262–1277. PMLR, 2021.

Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793. PMLR, 2022.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial intelligence in medicine*, 109: 101964, 2020.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587. PMLR, 2015.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7288–7295, 2021.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

Ashesh Jain, Brian Wojcik, Thorsten Joachims, and Ashutosh Saxena. Learning trajectory preferences for manipulators via iterative improvement. *Advances in Neural Information Processing Systems*, 26, 2013.

W Bradley Knox and Peter Stone. Tamer: Training an agent manually via evaluative reinforcement. In *2008 7th IEEE international conference on development and learning*, pages 292–297. IEEE, 2008.

Andras Kupcsik, David Hsu, and Wee Sun Lee. Learning dynamic robot-to-human object handover from human feedback. *Robotics Research: Volume 1*, pages 161–176, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Renke Liu, Rajesh Piplani, and Carlos Toro. Deep reinforcement learning for dynamic scheduling of a flexible job shop. *International Journal of Production Research*, 60(13):4049–4069, 2022.

Shuang Liu and Hao Su. Regret bounds for discounted mdps. *arXiv preprint arXiv:2002.05138*, 2020.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, pages 2285–2294. PMLR, 2017.

Volodymyr Mnih. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 1029–1038. PMLR, 2020.

Somtochukwu Oguchienti and Mahsa Ghasemi. Inverse reinforcement learning with learning and leveraging demonstrators' varying expertise levels. In *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE, 2023.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. Towards tractable optimism in model-based reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 1413–1423. PMLR, 2021a.

Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021b.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

Thomas Philip Runarsson and Simon M Lucas. Preference learning for move prediction and evaluation function approximation in othello. *IEEE Transactions on Computational Intelligence and AI in Games*, 6(3):300–313, 2014.

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-concordant analysis of generalized linear bandits with forgetting. In *International Conference on Artificial Intelligence and Statistics*, pages 658–666. PMLR, 2021.

Daniel Shin, Anca D Dragan, and Daniel S Brown. Benchmarks and algorithms for offline preference-based reward learning. *arXiv preprint arXiv:2301.01392*, 2023.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Yi-Chi Wang and John M Usher. Application of reinforcement learning for agent-based production scheduling. *Engineering applications of artificial intelligence*, 18(1):73–82, 2005.

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is RLHF more difficult than standard RL? *arXiv preprint arXiv:2306.14111*, 2023.

Christian Wirth and Johannes Fürnkranz. On learning from game annotations. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(3):304–316, 2014.

Christian Wirth, Johannes Fürnkranz, and Gerhard Neumann. Model-free preference-based reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. *arXiv preprint arXiv:2305.14816*, 2023a.

Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Provable reward-agnostic preference-based reinforcement learning. *arXiv preprint arXiv:2305.18505*, 2023b.

Yufan Zhao, Donglin Zeng, Mark A Socinski, and Michael R Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.