# LEARNING THE PREFERENCES OF A LEARNING AGENT

**Karim Abdel Sadek**[*], **Mark Bedaywi**[*], **Rhys Gould**[*], **Stuart Russell**
University of California, Berkeley
{karimabdel,mark_bedaywi,rhys_gould,russell}@berkeley.edu

## ABSTRACT

For AI systems to be useful to humans, they must understand and act in accordance with our values and preferences. Since specifying preferences is a hard task, *inverse reinforcement learning* (IRL) aims to develop methods that allow for inferring preferences from observed behavior. However, IRL assumes the human to be approximately optimal. This is a big limitation in cases where the human themselves may be learning to act optimally in an environment. In this paper, we formalize the problem of *learning the preferences of a learning agent*: a predictor observes a learner acting online and tries to infer the underlying reward function being (initially suboptimally) optimized by the learner. We model the learner as either being no-regret, or as converging to an optimal Boltzmann policy over time. In each of these settings, we establish theoretical guarantees for various preference learning algorithms, or otherwise show that such guarantees are impossible.

## 1 INTRODUCTION

AI systems are human-level or superhuman in many disciplines, from playing games such as Go and chess (Silver et al., 2016) to predicting protein structures (Jumper et al., 2021). Reinforcement learning (RL) has been the underpinning of many such breakthroughs. In RL, much of the progress has been driven by having clean and well-defined *reward functions*, such as game scores or win/loss outcomes (Mnih et al., 2013; Silver et al., 2016). The ability to optimize for a reward signal has even been argued to be sufficient and necessary to get us all we may ever want out of AI systems (Sutton & Barto, 2018; Silver et al., 2021).

A successful approach to designing effective reward functions has been learning them from observed behavior, known as inverse reinforcement learning (IRL) (Ng et al., 2000; Abbeel & Ng, 2004). In this framework, we observe the actions of an agent in an environment and try to infer the underlying reward function that the agent is optimizing. A similar approach in spirit has allowed large language models to follow human instructions and optimize for humans' preferences (Christiano et al., 2017; Stiennon et al., 2020; Bai et al., 2022). IRL assumes that the agent of interest acts (approximately) optimally/. Yet, we argue that in practice the agent will likely be learning how to act optimally over the course of data collection, taking suboptimal actions initially.

In this paper, we formalize and address the problem of *learning the preferences of a learning agent*. We model this problem via an online learning formulation: the *learner* (or *human*) selects actions online while receiving rewards from an unknown reward function. In particular, we assume that the learner agent is either a no-regret learner, or is converging to an optimal Boltzmann policy. The *predictor* observes the actions taken by the learner and aims to infer the true reward function at each step. Note that recovering the exact reward function is not possible in IRL (Ng et al., 1999; 2000; Abbeel & Ng, 2004; Cao et al., 2021). We formalize different notions of error, which include being able to find a reward function that matches the optimal or the Boltzmann policy of the agent, or more simply minimizing a metric distance between rewards (Gleave et al., 2020; Skalse et al., 2023a).

Our problem setting captures real issues that arise in a variety of settings. For example, consider a recommendation system that tries to suggest movies to a user who is still learning which movies they like. Such a system will need to infer the user's true preferences while the user themselves is acting suboptimally to discover them (e.g. exploring a range of genres before deciding they enjoy thrillers

[*]Equal contribution, alphabetical order.

the most). This is a more pressing issue as we develop increasingly personalized AI assistants which will need to infer our preferences across tasks in which humans may not yet be optimal.

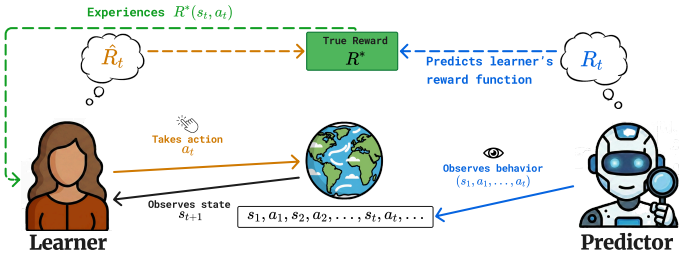


Figure 1: *Learning the preferences of a learning agent.* A learner interacts with an environment and learns to act optimally over time, with optimality measured by a ground-truth reward function $R^*$. The predictor observes only the learner's behavior $(s_1, a_1, \ldots, s_t, a_t)$ and aims to infer the preferences of the agent, producing reward estimates $R_1, \ldots, R_t$ (or $Q$-function estimates).

## 2 Related Work

**Inverse Reinforcement Learning** Specifying a reward function may be extremely challenging as AI systems are employed in increasingly complex tasks. Inverse Reinforcement Learning (IRL) (Russell, 1998; Ng et al., 2000) studies how a reward function can be recovered from data generated by an expert. However, IRL is underspecified (Jeon et al., 2020; Cao et al., 2021; Skalse et al., 2023b). An approach to the IRL problem is to use *apprenticeship learning* (Abbeel & Ng, 2004), where the task is to infer a reward function useful for finding a policy that matches the behavior of a human. Syed & Schapire (2007) generalizes this approach via a game-theoretic formulation, and shows that it is possible to find a policy that might even outperform the human's. Other relevant work tried to frame IRL as a Bayesian problem (Ramachandran & Amir, 2007), or to solely identify the entire set of plausible rewards under different conditions (Metelli et al., 2021; 2023).

**Modeling human feedback** We are not the first to note that modeling humans correctly is instrumental in correctly identifying the reward function. For example, observing behavior from uncertain humans (Laidlaw & Russell, 2021) or explicitly learning their suboptimality (Evans et al., 2016; Laidlaw & Dragan, 2022; Reddy et al., 2018) can improve reward inference. All of these works assume the human policy to be fixed over time. Chan et al. (2019) is the closest related work. They model the human as learning to play a bandit problem, while being assisted by a robot. Our work differs from theirs in several aspects. First, they only consider the bandit case, while we consider the sequential setting. They also assume the human to be noisily optimal or to be a greedy learner, which are stronger assumption than ours. Additionally, their results only concern assisting the human: there is no explicit learning about the reward function itself, as in our work.

## 3 Model and Problem Setting

We will consider a learner agent interacting with an environment, and a predictor agent who must determine the ground-truth reward function being optimized by the learner using only *observed behavior*. Our main focus will be proving guarantees for various such prediction strategies. We provide a general overview of the settings we consider in Table 1.

**Background and notation.** In general, we will consider a state space $\mathcal{S}$, an action space $\mathcal{A}$, a ground-truth reward function $R^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a transition distribution $\mu : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and a discount rate $\gamma \in [0, 1]$, with $\langle S, A, \mu, R^*, \gamma \rangle$ defining an MDP. For simplicity, in this work we always assume $\gamma = 1$. The *stateless* case is defined by the choice $|\mathcal{S}| = 1$ (letting us write $R^* : \mathcal{A} \to \mathbb{R}$), while in the *stateful* case we will generally assume $|\mathcal{S}| > 1$. Throughout the paper we will prove results in both stateless and stateful setups. For the stateful case, we will define the ground-truth action-value function recursively as $Q^*(s, a) := R^*(s, a) + \mathbb{E}_{\mu(s'|s,a)} [\max_{a' \in \mathcal{A}} Q^*(s', a')]$.

We will assume that all learning of the agent takes place within a single episode of finite length. In the stateful case, at time step $t$ the learner will have access to the information

$(s_1, a_1, R^*(s_1, a_1), \ldots, s_{t-1}, a_{t-1}, R^*(s_{t-1}, a_{t-1}), s_t)$ for taking their next action $a_t$. In the stateless case, this reduces to the information $(a_1, R^*(a_1), \ldots, a_{t-1}, R^*(a_{t-1}))$. We will denote the learner's state-visit count by $N_t(s) := |\{\tau \leq t : s_\tau = s\}|$, and $N_t(s, a)$ similarly. The predictor can only observe the learner's *behavior* $(s_1, a_1, \ldots, s_t, a_t)$ and must predict action-value functions $Q_{1:t}$ in the stateful case, or reward functions $R_{1:t}$ in the stateless case. To support the generality of our model, we also show in Section B that this setup is equivalent to predicting the reward/action-value function only at the end of the episode.

### 3.1 Modeling the Learner

If we wish to prove guarantees about our ability to predict the preferences of a learning agent, we must make some assumptions about the process by which this learner selects actions and learns from their experience. We will consider two models of the learner: (a) assume the agent achieves no-regret (Section 3.1.1), or (b) assume the agent behaves Boltzmann-rationally with respect to some estimate of reward/action-value that is updated over time (Section 3.1.2).

#### 3.1.1 No-Regret Learner

One approach to modeling the learner is to assume that they achieve no-regret.

**Stateless.** In the stateless case, the learner will have taken some actions $(a_1, \ldots, a_T)$, and we say that the learner is $f(T)$-no-regret (for some function $f : \mathbb{N} \to \mathbb{R}$) if

$$\mathrm{Reg}_L := \max_{a^*} \sum_{t=1}^{T} R^*(a^*) \; - \; \sum_{t=1}^{T} R^*(a_t) \; \leq \; f(T).$$

where $\mathrm{Reg}_L$ is the regret incurred by the learner. We say that the learner is *no-regret* if they are $f(T)$-no-regret for some $f(T) = o(T)$, i.e. if the average per-step regret vanishes as $T \to \infty$.

**Stateful.** For the stateful case, we will assume that the no-regret learner selects actions by sampling from a Markovian policy $\hat{p}_t(a_t|s_t)$ that satisfies

$$\mathrm{Reg}_L := \max_{\pi^*} \mathbb{E}_{\mu,\pi^*}\left[\sum_{t=1}^{T} R^*(s_t, a_t)\right] - \mathbb{E}_{\mu,\hat{p}_{1:T}}\left[\sum_{t=1}^{T} R^*(s_t, a_t)\right] \leq f(T)$$

for some $f(T) = o(T)$. Note that the time indexing of $\hat{p}_t$ allows for capturing the fact that the agent is learning over time. This notion of regret that replays the episode from the beginning when comparing against the optimal policy in hindsight—as opposed to external regret which appends the fixed action to the current trajectory—is known as policy regret (Arora et al., 2012).

#### 3.1.2 Boltzmann Rational Learner

We will also consider another approach to modeling the learner: assume that, over time, they converge to a Boltzmann rational policy (Luce et al., 1959; Ziebart et al., 2010).

**Stateless.** In the stateless case, we model the learner's action selection process as $a_t \sim \hat{p}_t$, where $\hat{p}_t(a) \propto \exp(\beta \hat{R}_t(a))$ for a rationality parameter $\beta \in [0, \infty)$, and where $\hat{R}_t$ models the learner's estimate of the reward function at time step $t$. To capture the fact that the agent is learning, we assume that $\hat{R}_t$ converges to the true reward function $R^*$ over time. Formally,

$$\sum_{t=1}^{T} \|\hat{R}_t - R^*\|_\infty \leq f(T) \tag{1}$$

for some $f(T) = O(T^\alpha)$ where $\alpha \in (0, 1)$ captures the rate of the agent's learning.

**Stateful.** In the stateful case, the learner receives $s_t$ at time step $t$ and selects their action as $a_t \sim \hat{p}_t(\cdot|s_t)$, with $\hat{p}_t(a|s) \propto \exp(\beta \hat{Q}_t(s, a))$ where $\hat{Q}_t$ models the learner's estimate of the action-value function at time step $t$. We model learning by assuming that this estimate $\hat{Q}_t$ satisfies:

$$\sum_{\tau=1:\, s_\tau=s}^{T} \|\hat{Q}_\tau(s, \cdot) - Q^*(s, \cdot)\|_\infty \leq f(N_T(s)) \tag{2}$$

for every $s \in \mathcal{S}$ and for some $f(N_T(s)) = O(N_T(s)^\alpha)$ where $\alpha \in (0, 1)$.

An important detail here is that Boltzmann rationality is not itself a model of learning or exploration. Instead, it is a model capturing the inability to optimize the ground-truth reward function. Boltzmann rationality could model exploration if $\beta \to \infty$ over time. This is the reason why the above is paired with the assumption that the learner's estimate of the reward/action-value converges over time.

### 3.2 Evaluation Measures

There are various reasonable performance measures that we could use to evaluate the quality of the predictor's estimates $R_{1:T} = (R_1, \ldots, R_T)$, (or $Q_{1:T} = (Q_1, \ldots, Q_T)$) for the true reward function $R^*$ (or $Q^*$). We will introduce the best-response distance (Section 3.2.1), the KL divergence between Boltzmann rational policies (Section 3.2.2), and norm-based measures (Section 3.2.3). These measures range from weak (best-response, which only requires matching the optimal action) to strong (norm-based, which require matching the full reward/action-value structure).

#### 3.2.1 Best-response distance

Intuitively, the best response distance between two reward functions measures whether both reward functions induce similar optimal policies. We will now make this concrete.

**Stateless.** Let $a^{R_t}$ denote the best action response given the reward estimate $R_t$. The stateless best-response distance between $R^*$ and a sequence of reward functions $R_{1:T} = (R_1, \ldots, R_T)$ is defined as:

$$\mathrm{D}_{\mathrm{BR}}(R^*, R_{1:T}) := \max_{a^* \in \mathcal{A}} \sum_{t=1}^{T} (R^*(a^*) - R^*(a^{R_t}))$$

This measures whether the optimal policies associated with the sequence $R_{1:T}$ also perform well under $R^*$, or in other words, whether they match the optimal policy $\pi^*$.

**Stateful.** Let $\pi_{1:T}$ denote the best response policies of the action-value estimates $Q_{1:T}$ at each respective time step $t$. We define the stateful best-response distance between $Q^*$ and $Q_{1:T}$ as:

$$\mathrm{D}_{\mathrm{BR}}(Q^*, Q_{1:T}) := \max_{\pi^*} \mathbb{E}_{\mu,\pi^*} \left[ \sum_{t=1}^{T} R^*(s_t, a_t) \right] - \mathbb{E}_{\mu,\pi_{1:T}} \left[ \sum_{t=1}^{T} R^*(s_t, a_t) \right]$$

#### 3.2.2 KL divergence between Boltzmann rational policies

**Stateless.** Given a reward function $R$, let $\pi^\beta(R)$ denote the associated Boltzmann rational policy, i.e. $\pi^\beta(R)(a) \propto \exp(\beta R(a))$. We can define the KL distance between $R^*$ and the sequence $R_{1:T}$:

$$\mathrm{D}^\beta_{\mathrm{KLBP}}(R^*, R_{1:T}) := \sum_{t=1}^{T} \mathrm{KL}(\pi^\beta(R_t) \mid\mid \pi^\beta(R^*)).$$

Note that this direction for the KL penalizes the predictor's induced policy for placing high probability on actions that are suboptimal under $R^*$.

**Stateful.** For the stateful case, we will similarly denote the Boltzmann rational policy associated with the action-value function $Q$ by $\pi^\beta(Q)$, with $\pi^\beta(Q)(a|s) \propto \exp(\beta Q(s, a))$. We can then define the KL distance between $Q^*$ and the sequence $Q_{1:T}$ as:

$$\mathrm{D}^\beta_{\mathrm{KLBP}}(Q^*, Q_{1:T}) := \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} v_t(s)\, \mathrm{KL}(\pi^\beta(Q_t)(\cdot|s) \mid\mid \pi^\beta(Q^*)(\cdot|s)).$$

for some appropriate state-wise weighting function $v_t$, e.g. $v_t(s) = \sqrt{N_{t-1}(s)/(t-1)}$. We argue that weighting by visit counts is fair; if the learner has only rarely entered a state $s$, we have less data available for predicting $Q^*(s, \cdot)$, and so it is fair to suppress the associated prediction errors.

#### 3.2.3 NORM-BASED MEASURES

We also consider norm-based measures of error. In particular, we will consider defining distances between reward functions based on the $\ell_2$ and $\ell_\infty$ norms.

**Stateless.** Given that for a finite action space we can view a reward function as a real-valued vector $R \in \mathbb{R}^{|\mathcal{A}|}$, we can define the norms $\|R\|_2 := \sqrt{\sum_{a\in\mathcal{A}} R(a)^2}$ and $\|R\|_\infty := \max_{a\in\mathcal{A}} |R(a)|$. We can then define the associated measures:

$$D_{\ell_2}(R^*, R_{1:T}) := \sum_{t=1}^{T} \|R^* - R_t\|_2, \qquad D_{\ell_\infty}(R^*, R_{1:T}) := \sum_{t=1}^{T} \|R^* - R_t\|_\infty$$

**Stateful.** The stateful case follows similarly, but we must take an appropriate weighting over states. We define:

$$D_{\ell_\infty}(Q^*, Q_{1:T}) := \sum_{t=1}^{T} \sum_{s\in\mathcal{S}} v_t(s)\, \|Q^*(s,\cdot) - Q_t(s,\cdot)\|_\infty$$

for an appropriate state-wise weighting $v_t$. For example, in Section 4.3.2 we establish performance guarantees for an empirical prediction strategy under the performance measure $D_{\ell_\infty}$ with weighting $v_t(s) = \sqrt{N_{t-1}(s)/(t-1)}$.

## 4 THEORETICAL RESULTS

| Learner | No-Regret Learner $o(T)$ | | | Boltzmann-Rational Learner | | |
|---|---|---|---|---|---|---|
| **Metric** | **BR** | **KL** | **$\ell_\infty$ norm** | **BR** | **KL** | **$\ell_\infty$ norm** |
| **Stateful** | ✓ Cor. 1 | × (Prop. 1) | × (Prop. 2) | ? Conj. 2 | ? Conj. 2 | ✓ Prop. 6 |
| **Stateless** | ✓ Prop. 3 | × (Prop. 1) | × (Prop. 2) | ? Conj. 1 | ? Conj. 1 | ✓ Prop. 5 |

Table 1: ✓ denotes positive results, × impossibility results, and ? open problems. The assumptions are on the environment (stateful vs. stateless) and learner model (no-regret or Boltzmann), under the evaluation metrics of best-response distance (BR), KL divergence between policies, and $\ell_\infty$ norm.

Now that we have formalized our problem setup, we can begin to prove guarantees for various prediction strategies. In Table 1, we summarize our results. We consider each combination of learner model and performance measure for both the stateless and stateful cases and either establish a guarantee, propose a conjecture, or prove that a good guarantee is impossible. We use assumptions that closely match ones previously considered in the literature. We discuss this in detail in Section 5.

### 4.1 MINIMIZING THE *KL-Boltzmann distance*

#### 4.1.1 LEARNER MODEL: *No-regret learner*

Here we will show that for a no-regret learner, there does not exist a generic guarantee on the cumulative $\ell^2$ distance better than $D_{\ell_2}(R^*, R_{1:T}) = \Theta(T)$. Since the $\ell_2$ distance is bounded by the KL, our impossibility result also implies an impossibility result for the KL-Boltzmann distance, meaning we can't hope to minimize the KL-Boltzmann distance when observing a no-regret learner.

**Lemma 1.** *Suppose a learner is no-regret. Given any algorithm $A$ predicting $R_{1:T}$, there exists an instance of the learner such that $D_{\ell_2}(R^*, R_{1:T}) = \Theta(T)$.* [proof]

This shows that minimizing the $\ell_2$ distance is impossible in our case. We can now show that having a large $\ell_2$ distance implies having a large KL distance.

**Proposition 1.** *Suppose a learner is no-regret. Given any algorithm $A$ predicting $R_{1:T}$, there exists an instance of the learner such that $\mathrm{D}^{\beta}_{\mathrm{KLBP}}(R^*, R_{1:T}) = \Theta(T)$.* [proof]

This impossibility result also immediately follows in the stateful case, since that is a strictly harder problem to solve. We can also show the same result for the $\ell_\infty$ distance via analogous arguments.

**Proposition 2.** *Suppose a learner is no-regret, and we are in the stateless (or stateful) case. Given any algorithm $A$ predicting $R_{1:T}$ (or $Q_{1:T}$), there exists an instance of the learner such that $D_{\ell_\infty}(R^*, R_{1:T}) = \Theta(T)$ (or $D_{\ell_\infty}(Q^*, Q_{1:T}) = \Theta(T)$).* [proof]

The takeaway from this section is clear: if you observe a no-regret learner, the best thing you can hope for is learning what the optimal action is. This is insufficient to achieve sub-linear prediction errors under richer evaluation measures, such as norm-based ones.

#### 4.1.2 LEARNER MODEL: *Boltzmann learner*

What can we hope to retrieve instead when the learner is Boltzmann rational? We conjecture that the following results may hold.

**Conjecture 1.** *In the stateless case with a Boltzmann rational learner (with rationality parameter $\beta$), there exists an algorithm $A$ producing a sequence of reward predictions $R_{1:T}$ such that*

$$\mathrm{D}^{\beta}_{\mathrm{KLBP}}(R^*, R_{1:T}) \leq O\left(\beta\sqrt{|\mathcal{A}|T}\right).$$

**Conjecture 2.** *In the stateful case with a Boltzmann rational learner, there exists an algorithm $A$ producing a sequence of action-value predictions $Q_{1:T}$ such that*

$$\mathrm{D}^{\beta}_{\mathrm{KLBP}}(Q^*, Q_{1:T}) \leq O\left(\beta\sqrt{T} \cdot \mathit{poly}(|\mathcal{S}|, |\mathcal{A}|)\right).$$

If these conjectures hold, they would also imply the corresponding results for the best-response measure, as a consequence of Prop. 4.

### 4.2 MINIMIZING THE *Best-response distance*

#### 4.2.1 LEARNER MODEL: *No-regret learner*

**Stateless case** It is easy to show that to minimize BR if the learner is no-regret it's easy: it suffices to predict the reward function that puts all mass on the action played by the learner at the previous time step.

**Proposition 3.** *In the stateless case with a learner whose regret is bounded by $f(t)$ at every iteration $t$, there exists an algorithm $A$ producing a sequence of reward predictions $R_{1:T}$ such that*

$$\mathrm{D}_{\mathrm{BR}}(R^*, R_{1:T}) \leq 1 + f(T).$$ [proof]

**Stateful case** By running the same simple proof as above, but playing the previous action the learner played at every state, we achieve a similar guarantee.

**Corollary 1.** *In the stateful case with a learner whose regret is bounded by $f(t)$ at every iteration $t$, there exists an algorithm $A$ producing a sequence of action-value predictions $Q_{1:T}$ such that*

$$\mathrm{D}_{\mathrm{BR}}(Q^*, Q_{1:T}) \leq |S| + f(T).$$

#### 4.2.2 LEARNER MODEL: *Boltzmann learner*

We will now show that minimizing the KL-Boltzmann distance is sufficient to induce bounds on the best-response distance. Hence, while we do not have an algorithm that minimizes the KL-Boltzmann, we show that finding one would suffice to minimize the best response distance

**Proposition 4.** *Suppose that there exists an algorithm $A$ predicting $R_{1:T}$ such that $\mathrm{D}^{\beta}_{\mathrm{KLBP}}(R^*, R_{1:T}) \leq f(T)$. Then, it follows that $\mathrm{D}_{\mathrm{BR}}(R^*, R_{1:T}) \leq O(f(T))$.* [proof]

### 4.3 MINIMIZING THE $\ell_\infty$ *norm*

#### 4.3.1 LEARNER MODEL: *No-regret learner*

It is impossible to prove a meaningful guarantee in this case, for the same reason that the $\ell^2$ distance (Section 4.1.1) fails; the same counterexample in that section applies here identically.

#### 4.3.2 LEARNER MODEL: *Boltzmann learner*

In the following, we will prove that for a Boltzmann-rational learner (Section 3.1.2), a prediction strategy that makes use of the empirical average achieves good guarantees in terms of the $\ell_\infty$ error.

**Stateless case** Let $R_t$, $\hat{R}_t$, and $R^*$ denote the predictor's, learner's, and ground-truth reward function respectively. Recall that a Boltzmann-rational learner is modeled as selecting actions via $a_t \sim \pi^\beta(\hat{R}_t) =: \hat{p}_t$, and is also assumed to satisfy the $f$-guarantee of Equation (1) for some appropriate $f$. The prediction $R_t$ can only depend on $a_{<t} \sim \hat{p}$, where $\hat{p}(a_{<t}) \equiv \prod_{\tau=1}^{t-1} \hat{p}_\tau(a_\tau)$. We will assume that $R_t$ and $R^*$ are $\sigma$-normalized in the following sense:

**Definition 1** (stateless $\sigma$-normalized)**.** *We say that $R : \mathcal{A} \to \mathbb{R}$ is $\sigma$-normalized (for some $\sigma \in \mathbb{R}$) if and only if* $\sum_{a\in\mathcal{A}} R(a) = \sigma$

The goal of this section will be proving guarantees about the averaging strategy for $R_t$, defined as:

**Definition 2** (stateless $\sigma$-averaging strategy)**.** *Define the averaging strategy as selecting the $\sigma$-normalized $R_t$ that satisfies $\pi^\beta(R_t)(a) = \frac{N_{t-1}(a)}{t-1} =: p_t(a)$ at time step $t > 1$. Such $R_t$ exists for all $t > 1$ and is unique (by $\sigma$-normalization), with the explicit form*

$$R_t(a) = \frac{1}{\beta}\left(\log p_t(a) - \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} \log p_t(a')\right) + \frac{\sigma}{|\mathcal{A}|}$$

We will find that our results are independent of $\sigma$ and hence will refer to this strategy as simply the "averaging strategy" from now on. We can now prove our main result in the stateless case:

**Proposition 5.** *Suppose that the predictor $R_t$ follows the $\sigma-$averaging strategy, and that the learner $\hat{R}_t$ satisfies an $f$-guarantee (Equation* (1)*). Then, with probability $1-\epsilon$, the cumulative $\ell_\infty$ prediction error can be bounded as:*

$$D_{\ell_\infty}(R^*, R_{t_e:T}) = \sum_{t=t_e}^{T} \|R_t - R^*\|_\infty \le \frac{2}{\beta}\overbrace{\sum_{t=t_e}^{T}\frac{1}{\kappa_t}\sqrt{\frac{2\log(2|\mathcal{A}|(T-1)/\epsilon)}{t-1}}}^{\sim O\left(\sqrt{T\log(|\mathcal{A}|T/\epsilon)}\right)} + \sum_{t=t_e}^{T}\frac{1}{\kappa_t}\frac{f(t-1)}{t-1}$$

*defining $t_e := \min(t : N_{t-1}(a) > 0 \;\forall\; a \in \mathcal{A})$ and $\kappa_t := \min_{a\in\mathcal{A}} \min(p_t(a), p^*(a))$ (where $\kappa_t > 0$ for $t \ge t_e$), with $f$ capturing the learning rate of a Boltzmann learner (see Section 3.1.2).* [proof]

**Discussion.** As an example, suppose the learner satisfies a $\sqrt{t}$-guarantee. Then the above becomes:

$$D_{\ell_\infty}(R^*, R_{t_e:T}) \le \underbrace{O\left(\sqrt{T\log(|\mathcal{A}|T/\epsilon)}\right)}_{\text{predictor-learner error}} + \underbrace{O\left(\sqrt{T}\right)}_{\text{learner-oracle error}}$$

i.e., in this case, the error in predicting the learner's estimate ($\sum_t \|R_t - \hat{R}_t\|_\infty$) dominates the error in the learner's estimate of the true reward function ($\sum_t \|\hat{R}_t - R^*\|_\infty$). The assumption that our reward functions are $\sigma$-normalized is helpful for identifiability reasons; otherwise, the averaging strategy $R_t$ (Definition 2) is only unique up to translation, which makes the value of measure $\|R_t - R^*\|_\infty$ ambiguous without fixing a unique $R_t$ via $\sigma$-normalization.

**Stateful case** Let $Q_t$, $\hat{Q}_t$, and $Q^*$ denote the predictor's, learner's, and ground-truth action-value functions respectively. The learner is modeled as selecting actions via $a_t \sim \pi^\beta(\hat{Q}_t(s_t, \cdot)) =: \hat{p}_t(\cdot|s_t)$, and is also assumed to satisfy the $f$-guarantee of Equation (2). The prediction $Q_t$ at time step $t$ can only depend on $(s_{<t}, a_{<t}) \sim \hat{\mu}$, with $\hat{\mu}$ capturing the learner-environment interaction:

$$\hat{\mu}(s_{<t}, a_{<t}) = \prod_{\tau=1}^{t-1} \mu(s_\tau|s_{\tau-1}, a_{\tau-1})\hat{p}_\tau(a_\tau|s_\tau)$$

for the environment's transition distribution $\mu(s_\tau|s_{\tau-1}, a_{\tau-1})$ (and $\mu(s_1|s_0, a_0) \equiv \mu(s_1)$). Similarly to the stateless case, we will assume $Q_t$ and $Q^*$ are $\sigma$-normalized in the following sense:

**Definition 3** (stateful $\sigma$-normalized)**.** *We say that $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is $\sigma$-normalized (for some $\sigma : \mathcal{S} \to \mathbb{R}$) iff $Q$ satisfies $\sum_{a \in \mathcal{A}} Q(s, a) = \sigma(s), \quad \forall\, s \in \mathcal{S}$*

In the stateful case, we define the averaging strategy as follows:

**Definition 4** (stateful averaging strategy)**.** *Define the averaging strategy as selecting the $\sigma$-normalized $Q_t$ that satisfies $\pi^\beta(Q_t)(a|s) = \frac{N_{t-1}(s,a)}{N_{t-1}(s)} =: p_t(a|s)$ at time step $t > t_e(s)$. Such $Q_t$ exists for all $t > t_e(s)$ and is unique (by $\sigma$-normalization), with the explicit form*

$$Q_t(s, a) = \frac{1}{\beta}\left(\log p_t(a|s) - \frac{1}{|\mathcal{A}|}\sum_{a' \in \mathcal{A}} \log p_t(a'|s)\right) + \frac{1}{|\mathcal{A}|}$$

We will again find that our results are independent of $\sigma$. The results for the stateful case follow an identical structure to the stateless case, with analogous intermediate lemmas and a final proposition:

**Proposition 6.** *Suppose that the predictor $Q_t$ follows the averaging strategy (Definition 4), and that the learner $\hat{Q}_t$ satisfies the $f$-guarantee (Equation (2)). Then, with probability $1 - \epsilon$, the visit-weighted cumulative prediction error can be bounded as:*

$$\sum_{s \in \mathcal{S}} \sum_{t=t_e(s)}^{T} v_t(s) \left\|Q_t(s, \cdot) - Q^*(s, \cdot)\right\|_\infty \leq \frac{2}{\beta} \overbrace{\sum_{s \in \mathcal{S}} \sum_{t=t_e(s)}^{T} \frac{1}{\kappa_t(s)} \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|(T-1)/\epsilon)}{t-1}}}^{\sim O\left(|\mathcal{S}|\sqrt{T\log(|\mathcal{S}||\mathcal{A}|T/\epsilon)}\right)} + \sum_{s \in \mathcal{S}} \sum_{t=t_e(s)}^{T} \frac{1}{\kappa_t(s)} \frac{f(N_{t-1}(s))}{\sqrt{(t-1)N_{t-1}(s)}}$$

*where $v_t(s) := \sqrt{N_{t-1}(s)/(t-1)}$ and $t_e(s) := \min(t : N_{t-1}(s, a) > 0 \;\forall\; a \in \mathcal{A})$, with $f$ capturing the learning rate of a Boltzmann learner as described in Section 3.1.2.* [proof]

**Discussion.** Suppose that the learner learns at a rate of $f(N_t(s)) \propto \sqrt{N_t(s)}$. Then we have:

$$\sum_{s \in \mathcal{S}} \sum_{t=t_e(s)}^{T} v_t(s) \left\|Q_t(s, \cdot) - Q^*(s, \cdot)\right\|_\infty \leq \underbrace{O\left(|\mathcal{S}|\sqrt{T\log(|\mathcal{S}||\mathcal{A}|T/\epsilon)}\right)}_{\text{predictor-learner error}} + \underbrace{O\left(|\mathcal{S}|\sqrt{T}\right)}_{\text{learner-oracle error}}$$

Only states $s$ that are eventually explored during the $T$-length horizon will contribute to this cumulative error (by the definition of $t_e(s)$), i.e. we do not necessarily require our MDP to be ergodic.

## 5 Conclusion

In this paper, we have formalized the problem of learning the preferences of an agent that themselves is learning to act optimally. We showed that for a no-regret learner, we are unable to say anything *in general* about the structure of their preferences other than what their preferred action is in each state. This is an intuitive result: we need the learner to somewhat indicate their preferences for suboptimal actions if we want to have any hope of learning their full preference structure, and no-regret does not generally guarantee this. Under alternative assumptions on the learner (e.g. Boltzmann rationality), we are able to obtain general guarantees on the $\ell_\infty$ error between the ground-truth and predicted reward function based on how often different states were visited under the learner's policy.

**Choices of learner models and evaluation measures.** Making the right assumptions is fundamental if we want our insights to be relevant to practice. One ambition is to have learner models that capture the bounded rationality of humans, which we note is something that the Boltzmann rationality model does not respect: for a learner to be perfectly Boltzmann, they would need to perfectly estimate the reward/action-value function. We leave the extension of our results to such learner models for future work. We note that our norm-based errors require that the full preference structure is captured, as intended for our downstream application. For example, while the best-response distance is a manageable metric to optimize for, it only guarantees that we identify the best action, which might

not be very useful in practice. For example, we may instead want to recover reward functions that are *robust*: ones that if optimized in a different environment, would still lead to a desired behavior.

**Future work.** Plenty of interesting questions remain open. For example, can we find efficient algorithms that minimize the KL-Boltzmann distance defined in Section 3.2.2? Are our bounds tight? Do our results cleanly translate to the case of stochastic rewards? What could we gain if we were able to act through the expert, and not only observe, as considered in Chan et al. (2019)?

Ultimately, we believe that AI systems will be deployed in increasingly high-stakes scenarios. In such cases, humans may not be perfect actors, and developing methods that can learn from a multitude of data collected from suboptimal & learning humans will be of increasing importance. We hope that our work lays solid foundations for the study of these challenges.

## 6 ACKNOWLEDGMENTS

We thank Nika Hagthalab and Annie Ulichney for helpful feedback on drafts.

This work was supported by a gift from Open Philanthropy (now Coefficent Giving) to the Center for Human-Compatible AI (CHAI) at UC Berkeley. Karim Abdel Sadek and Mark Bedawyi are supported by the Cooperative AI PhD Fellowship.

## REFERENCES

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

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan. The assistive multi-armed bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 354–363, 2019. doi: 10.1109/HRI.2019.8673234.

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.

Owain Evans, Andreas Stuhlmüller, and Noah Goodman. Learning the preferences of ignorant, inconsistent agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. *arXiv preprint arXiv:2006.13900*, 2020.

Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Cassidy Laidlaw and Anca Dragan. The boltzmann policy distribution: Accounting for systematic suboptimality in human models. *arXiv preprint arXiv:2204.10759*, 2022.

Cassidy Laidlaw and Stuart Russell. Uncertain decisions facilitate better preference learning. *Advances in Neural Information Processing Systems*, 34:15070–15083, 2021.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Alberto Maria Metelli, Giorgia Ramponi, Alessandro Concetti, and Marcello Restelli. Provably efficient learning of transferable rewards. In *International Conference on Machine Learning*, pp. 7665–7676. PMLR, 2021.

Alberto Maria Metelli, Filippo Lazzati, and Marcello Restelli. Towards theoretical understanding of inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 24555–24591. PMLR, 2023.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pp. 2586–2591, 2007.

Sid Reddy, Anca Dragan, and Sergey Levine. Where do you think you're going?: Inferring beliefs about dynamics from behavior. *Advances in Neural Information Processing Systems*, 31, 2018.

Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 101–103, 1998.

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, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial intelligence*, 299:103535, 2021.

Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. Starc: A general framework for quantifying differences between reward functions. *arXiv preprint arXiv:2309.15257*, 2023a.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, pp. 32033–32058. PMLR, 2023b.

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.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. *Advances in neural information processing systems*, 20, 2007.

Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. 2010.

# A PROOFS FOR MAIN RESULTS

## A.1 PROOF OF LEMMA 1

*Proof.* Suppose the learner always plays action $a_t = a_1$ for all $t \in [T]$. We can see that this learner is no-regret & optimal under at least two possible choices of ground-truth reward functions: $R^* = R_1^* := \frac{1}{|\mathcal{A}|}(1, 1, \ldots, 1)$ and $R^* = R_2^* := (1, 0, \ldots, 0)$ (and in fact, under any reward function $R^*$ such that $R^*(a_1) \geq R^*(a)$ for all $a \in \mathcal{A}$). This means that in the worst-case of an adversarial choice for $R^*$, the best prediction strategy will always incur linear error since

$$\begin{aligned}\min_{R_{1:T}} \max_{R^*} D_{\ell_2}(R^*, R_{1:T}) &\geq \min_{R_{1:T}} \max_{R^* \in \{R_1^*, R_2^*\}} D_{\ell_2}(R^*, R_{1:T}) \\ &\geq \frac{1}{2} \min_{R_{1:T}} \left(D_{\ell_2}(R_1^*, R_{1:T}) + D_{\ell_2}(R_2^*, R_{1:T})\right) \\ &\geq \frac{T}{2} \|R_1^* - R_2^*\|_2 = \Theta(T)\end{aligned}$$

where in the third line we have made use of the triangle inequality. □

## A.2 PROOF OF PROP. 1

We will need the following result to prove Prop. 1.

**Lemma 2.** *Given two probability vectors $p, q$ it follows that*

$$\mathrm{KL}(p\|q) \;\geq\; \frac{1}{2}\,\|p - q\|_2^2.$$

*Proof.* Pinsker's inequality gives $\mathrm{TV}(p, q) \;\leq\; \sqrt{\frac{\mathrm{KL}(p\|q)}{2}}$, where $\mathrm{TV}(p, q) = \frac{1}{2}\|p - q\|_1$. Rearranging, we have $\mathrm{KL}(p\|q) \;\geq\; 2\,\mathrm{TV}(p, q)^2 \;=\; \frac{1}{2}\,\|p - q\|_1^2$.. Since $\ell_1 > \ell_2$ always,

$$\mathrm{KL}(p\|q) \;\geq\; \frac{1}{2}\,\|p - q\|_1^2 \;\geq\; \frac{1}{2}\,\|p - q\|_2^2.$$

as desired. □

Prop. 1 then follows directly by applying Lemma 2 to the impossibility result in Lemma 1.

## A.3 PROOF OF PROP. 2

*Proof.* Follows directly from Lemma 1, since the $\ell_2$ is always bounded by the $\ell_\infty$. □

## A.4 PROOF OF PROP. 3

*Proof.* The idea is simple. The algorithm plays an arbitrary action $b \in \mathcal{A}$ at time step 1. Then, at each iteration, the previous action $a_{t-1}$ is maintained. The algorithm picks the reward function $R_t(a) = \mathbb{1}(a = a_{t-1})$. Then, the best response under reward $R_t$ is to play action $a_{t-1}$. This means regret is

$$\begin{aligned}\mathrm{D}_{\mathrm{BR}}(R^*, R_{1:T}) &= (\max_a R^*(a) - R^*(b)) + \sum_{t=2}^{T} (\max_a R^*(a) - R^*(a_{t-1})) \\ &\leq (1 - 0) + \sum_{t=1}^{T-1} (\max_a R^*(a) - R^*(a_t)) \\ &\leq 1 + f(T).\end{aligned}$$

By playing the reward that puts all the weight on the previous action, we are able to match the learner's regret up to a constant additive factor of 1. □

### A.5 PROOF OF THE LEMMA FROM SYED & SCHAPIRE (2007)

*Proof.* Because $R$ is the minimax solution, it must satisfy, for all other reward functions $R'$ and all possible policies $\pi \in \Pi$,

$$\mathbb{E}[R(\pi) - R(\pi_E)] \leq R'(\pi) - R'(\pi_E).$$

In particular, it must hold for the true underlying reward function $R^*$, and for the optimal policy $\pi_R = \arg\max_{\pi \in \Pi} \mathbb{E}[R(\pi)]$ under $R$,

$$\mathbb{E}[R(\pi_R) - R(\pi_E)] \leq R^*(\pi_R) - R^*(\pi_E).$$

By definition, the left hand side is greater than zero. Therefore,

$$R^*(\pi_E) \geq R^*(\pi_R),$$

as claimed. □

### A.6 PROOF OF PROP. 4

*Proof.* First, use Prop. 7, and note that we have that an algorithm that only predicts a single $R_t$ such that $T \cdot \mathrm{D}^{\beta}_{\mathrm{KLBP}}(R^*, R_t) \leq f(T)$. Now, let's show that if $\mathrm{D}^{\beta}_{\mathrm{KLBP}}(R^*, R_t) \leq \epsilon$, then we can bound $\mathrm{D}_{\mathrm{BR}}(R^*, R_{1:T})$.

Define $a^* = \arg\max_a R^*(a)$, and let $p^*(a) := \pi^{\beta}(R^*)(a)$ and $p_t(a) = \pi^{\beta}(R_t)(a)$. Note that

$$\max_{a^* \in \mathcal{A}} (R^*(a^*) - R^*(a^{R_t})) = \frac{1}{\beta} \log \frac{p^*(a^*)}{p^*(a^{R_t})}$$

.

for each $t$. We note that the TV distance is bounded by the KL, which implies that

$$TV(p_t, p^*) \leq \sqrt{\mathrm{KL}(p_t, p^*)/2} = \delta$$

It is easy to see that for all actions, $|p_t(a) - p^*(a)| \leq TV(p_t, p^*) \leq \delta$. Note that now for all actions, $p^*(a) \geq p_t(a) - \delta$ and $p^*(a) \leq p_t(a) + \delta$. This implies by definition that $p^*(a^*) \leq p_t(a^{R_t}) + \delta$ and $p^*(a^{R_t}) \geq p_t(a^{R_t}) - \delta$. So we can note that if $p_t(a^{R_t}) > \delta$, then

$$\frac{p^*(a^*)}{p^*(a^{R_t})} \leq \frac{p_t(a^{R_t}) + \delta}{p_t(a^{R_t}) - \delta}$$

Additionally, note that if $\delta < \frac{1}{m}$ (where $m = |\mathcal{A}|$), then

$$\frac{p^*(a^*)}{p^*(a^{R_t})} \leq \frac{\frac{1}{m} + \delta}{\frac{1}{m} - \delta} = \frac{1 + m\delta}{1 - m\delta}$$

Thus finally inducing that

$$\mathrm{D}_{\mathrm{BR}}(R^*, R_{1:T}) \leq \frac{T}{\beta} \log \frac{1 + m\delta}{1 - m\delta}$$

This shows that if the KL is bounded, the best response distance is also bounded when we predict a reward at the end. By applying Prop. 8, we are done. □

### A.7 PROOF OF PROP. 5

We will now work towards proving a bound on $D_{\ell_\infty}(R^*, R_{1:T})$ (Prop. 5), starting with two lemmas:

**Lemma 3.** *Suppose that $R_t$ follows an averaging strategy (Definition 2). Then with probability $1 - \epsilon$,*

$$\|p_t - \bar{p}_t\|_\infty < \sqrt{\frac{2 \log(2|\mathcal{A}|(T-1)/\epsilon)}{t-1}}$$

*for $t > 1$, defining the time-averaged learner policy:*

$$\bar{p}_t(a) := \mathbb{E}_{\hat{p}(a_{<t})}[p_t(a)] = \frac{1}{t-1} \sum_{\tau=1}^{t-1} \hat{p}_\tau(a)$$

*Proof.* First note that we can write

$$p_t(a) - \bar{p}_t(a) = \frac{1}{t-1} \underbrace{\sum_{\tau=1}^{t-1} (1[a = a_\tau] - \hat{p}_\tau(a))}_{=: M^a_{t-1}}$$

where $M^a_{t-1}$ defines a Martingale (i.e., $\mathbb{E}_{\hat{p}}[M^a_{t+1} \mid M^a_1, \ldots, M^a_t] = M_t$) and also satisfies $|M^a_{t+1} - M^a_t| \leq 1$ for all $t$, allowing us to apply the Azuma-Hoeffding inequality:

$$\mathbb{P}_{\hat{p}(a_{<t})}(|M^a_{t-1}| \geq \delta_t) \leq 2\exp(-\delta_t^2/2(t-1))$$
$$\implies \mathbb{P}_{\hat{p}(a_{<t})}(|p_t(a) - \bar{p}_t(a)| \geq \delta_t) \leq 2\exp(-(t-1)\delta_t^2/2)$$

For an $\ell^\infty$ guarantee, we need this bound to hold across all $a \in \mathcal{A}$. Applying a union bound,

$$\begin{aligned}
\mathbb{P}_{\hat{p}(a_{<t})}(|p_t(a) - \bar{p}_t(a)| < \delta_t \ \ \forall\, a \in \mathcal{A} \ \ \forall\, t \in [2,T]) &= 1 - \mathbb{P}_{\hat{p}(a_{<t})}(\exists\, a \in \mathcal{A},\, t \in [2,T] : |p_t(a) - \bar{p}_t(a)| \geq \delta_t) \\
&\geq 1 - \sum_{a \in \mathcal{A}} \sum_{t=2}^{T} \mathbb{P}_{\hat{p}(a_{<t})}(|p_t(a) - \bar{p}_t(a)| \geq \delta_t) \\
&\geq 1 - 2|\mathcal{A}| \sum_{t=2}^{T} \exp(-(t-1)\delta_t^2/2)
\end{aligned}$$

$$\implies \mathbb{P}_{\hat{p}(a_{<t})}\left( |p_t(a) - \bar{p}_t(a)| < \sqrt{\frac{2\log(2|\mathcal{A}|(T-1)/\epsilon)}{t-1}} \ \ \forall\, a \in \mathcal{A} \ \ \forall\, t \in [2,T] \right) \geq 1 - \epsilon$$

And so, with probability $1 - \epsilon$:

$$\|p_t - \bar{p}_t\|_\infty < \sqrt{\frac{2\log(2|\mathcal{A}|(T-1)/\epsilon)}{t-1}}$$

as required. □

**Lemma 4.** *Suppose that the learner's $\hat{R}_t$ satisfies the $f$-guarantee of Equation* (1)*. Then for $t > 1$ and $\bar{p}_t$ as defined in the statement of Lemma 3, we can bound*

$$\|\bar{p}_t - p^*\|_\infty \leq \frac{\beta}{2} \frac{f(t-1)}{t-1} \tag{3}$$

*Proof.* Note that $\hat{p}_\tau(a) = \text{softmax}(\beta \hat{R}_\tau)(a)$ and $p^*(a) = \text{softmax}(\beta R^*)(a)$. Since softmax is Lipschitz continuous under $\ell^\infty$ with constant $1/2$, we have that

$$\|\hat{p}_\tau - p^*\|_\infty \leq \frac{\beta}{2} \|\hat{R}_\tau - R^*\|_\infty$$

allowing us to bound

$$\begin{aligned}
\|\bar{p}_t - p^*\|_\infty &\leq \frac{1}{t-1} \sum_{\tau=1}^{t-1} \|\hat{p}_\tau - p^*\|_\infty \\
&= \frac{\beta}{2} \frac{1}{t-1} \sum_{\tau=1}^{t-1} \|\hat{R}_\tau - R^*\|_\infty \\
&\leq \frac{\beta}{2} \frac{f(t-1)}{t-1}
\end{aligned}$$

where we have made use of the stateless $f$-guarantee (Equation (1)) which says that:

$$\sum_{t=1}^{T} \|\hat{R}_t - R^*\|_\infty \leq f(T)$$

□

We are now ready to prove Prop. 5.

*Proof of Prop. 5.* Write $p_t(a) := \pi^\beta(R_t)(a)$, $\hat{p}_t(a) := \pi^\beta(\hat{R}_t)(a)$, and $p^*(a) = \pi^\beta(R^*)(a)$ for the Boltzmann policies associated with the predictor, learner, and ground-truth reward functions respectively. Using the expression for $R_t$ in Definition 2 (and using an analogous expression for $R^*$), we can write

$$R_t(a) - R^*(a) = \frac{1}{\beta}\left(\log\frac{p_t(a)}{p^*(a)} - \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}}\log\frac{p_t(a')}{p^*(a')}\right)$$

which lets us write

$$\begin{aligned}
\|R_t - R^*\|_\infty &= \max_{a\in\mathcal{A}}\frac{1}{\beta}\left|\log\frac{p_t(a)}{p^*(a)} - \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}}\log\frac{p_t(a')}{p^*(a')}\right| \\
&\le \max_{a\in\mathcal{A}}\frac{1}{\beta}\left(\left|\log\frac{p_t(a)}{p^*(a)}\right| + \left|\frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}}\log\frac{p_t(a')}{p^*(a')}\right|\right) \\
&\le \max_{a\in\mathcal{A}}\frac{2}{\beta}\left|\log\frac{p_t(a)}{p^*(a)}\right| \\
&\le \max_{a\in\mathcal{A}}\frac{2}{\beta}\frac{|p_t(a) - p^*(a)|}{\min(p_t(a), p^*(a))} \\
&\le \frac{2}{\kappa_t\beta}\|p_t - p^*\|_\infty
\end{aligned}$$

for $t \ge t_e$, where in the second-to-last line we have made use of

$$\left|\log\frac{p_t(a)}{p^*(a)}\right| \le \frac{|p_t(a) - p^*(a)|}{\min(p_t(a), p^*(a))}$$

via the mean-value theorem. We can bound $\|p_t - p^*\|_\infty$ by decomposing

$$p_t(a) - p^*(a) = \underbrace{(p_t(a) - \bar{p}_t(a))}_{\text{predictor-learner error}} + \underbrace{(\bar{p}_t(a) - p^*(a))}_{\text{learner-oracle error}} \tag{4}$$

Using the previous lemmas, with probablity $1 - \epsilon$:

$$\begin{aligned}
\|R_t - R^*\|_\infty &\le \frac{2}{\kappa_t\beta}(\underbrace{\|p_t - \bar{p}_t\|_\infty}_{\text{Lemma 3}} + \underbrace{\|\bar{p}_t - p^*\|_\infty}_{\text{Lemma 4}}) \\
&< \frac{2}{\kappa_t\beta}\sqrt{\frac{2\log(2|\mathcal{A}|(T-1)/\epsilon)}{t-1}} + \frac{1}{\kappa_t}\frac{f(t-1)}{t-1}
\end{aligned}$$

Summing over $t \in [t_e, T]$ gives us the required result. □

The assumption of a Boltzmann-rational learner is important for being able to prove things about $\|R_t - R^*\|_\infty$ since it ensures that the learner will always place a non-zero probability on each action (with a fixed $\beta$), giving the predictor an opportunity to understand the value of $\hat{R}_t(a)$ (and hence $R^*(a)$) across all actions $a \in \mathcal{A}$. In contrast, a no-regret learner has no such guarantees about exploration, since in principle they can instantly select the optimal action $a^*$ and commit to it for all time, making it impossible for $R_t$ to learn about the reward associated with any action other than $a^*$.

In summing over $t \in [t_e, T]$, we are only considering the prediction error after the learner has already explored each action. As a result, for very large $|\mathcal{A}|$, this guarantee may lose its usefulness due to ignoring the potentially long, non-negligible duration $t < t_e$ when the learner is still to explore all actions. In this case, a different proof strategy may be necessary for proving guarantees on these earlier errors.

### A.8 PROOF OF PROP. 6

The following follows a similar structure to the stateless case (Section A.7).

**Lemma 5.** *Suppose that $Q_t$ follows an averaging strategy (Definition 4). Then with probability $1-\epsilon$,*

$$\|p_t(\cdot|s) - \bar{p}_t(\cdot|s)\|_\infty < \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|(T-1)/\epsilon)}{N_{t-1}(s)}} \qquad \forall\, s \in \mathcal{S} : N_{t-1}(s) > 0$$

*for $t > 1$, defining the state-wise time-averaged learner policy:*

$$\bar{p}_t(a|s) := \frac{1}{N_{t-1}(s)} \sum_{i=1}^{N_{t-1}(s)} \hat{p}_{\tau_i(s)}(a|s)$$

*where $\tau_i(s)$ is the time step corresponding to the $i$th visit of state $s$.*

*Proof.* The proof mostly follows the structure of the stateless case (Lemma 3). See that

$$p_t(a|s) - \bar{p}_t(a|s) = \frac{1}{N_{t-1}(s)} \underbrace{\sum_{i=1}^{N_{t-1}(s)} (1[a_{\tau_i(s)} = a] - \hat{p}_{\tau_i(s)}(a|s))}_{=: M^{(s,a)}_{N_{t-1}(s)}}$$

where $M_n^{(s,a)}$ defines a Martingale (i.e., $\mathbb{E}_{\hat{\mu}}[M_{n+1}^{(s,a)} \mid M_1^{(s,a)}, \ldots, M_n^{(s,a)}] = M_n^{(s,a)}$) that also satisfies $|M_{n+1}^{(s,a)} - M_n^{(s,a)}| \leq 1$ for all $n$, allowing us to apply the Azuma-Hoeffding inequality to find:

$$\mathbb{P}_{\hat{\mu}(s_{<t},a_{<t})}(|p_t(a|s) - \bar{p}_t(a|s)| \geq \delta_n \mid N_{t-1}(s) = n) \leq 2\exp(-n\delta_n^2/2)$$

$$\implies \mathbb{P}_{\hat{\mu}(s_{<t},a_{<t})}\left(|p_t(a|s) - \bar{p}_t(a|s)| \geq \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|(T-1)/\epsilon)}{N_{t-1}(s)}} \,\middle|\, N_{t-1}(s) = n\right) \leq \frac{\epsilon}{|\mathcal{S}||\mathcal{A}|(T-1)}$$

Applying a union bound, we find that:

$$\mathbb{P}_{\hat{\mu}(s_{<t},a_{<t})}\left(|p_t(a|s) - \bar{p}_t(a|s)| < \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|(T-1)/\epsilon)}{N_{t-1}(s)}} \;\; \forall\, s \in \mathcal{S} \;\; \forall\, a \in \mathcal{A} \;\; \forall\, t \in [2,T] \,\middle|\, N\right) \geq 1-\epsilon$$

As a result, with probability $1-\epsilon$,

$$\|p_t(\cdot|s) - \bar{p}_t(\cdot|s)\|_\infty < \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|(T-1)/\epsilon)}{N_{t-1}(s)}}$$

as required. □

**Lemma 6.** *Suppose that the learner's $\hat{Q}_t$ satisfies the $f$-guarantee of Equation* (2)*. Then for $t > 1$ and $\bar{p}_t$ as defined in the statement of Lemma 5, we can bound*

$$\|\bar{p}_t(\cdot|s) - p^*(\cdot|s)\|_\infty \leq \frac{\beta}{2}\frac{f(N_{t-1}(s))}{N_{t-1}(s)} \tag{5}$$

*Proof.* Note that $\hat{p}_\tau(a|s) = \text{softmax}(\beta\hat{Q}_\tau(s,\cdot))(a)$ and $p^*(a|s) = \text{softmax}(\beta Q^*(s,\cdot))(a)$, and hence by the Lipschitz inequality:

$$\|\hat{p}_\tau(\cdot|s) - p^*(\cdot|s)\|_\infty \leq \frac{\beta}{2}\|\hat{Q}_\tau(s,\cdot) - Q^*(s,\cdot)\|_\infty$$

$$\begin{aligned}
\implies \|\bar{p}_t(\cdot|s) - p^*(\cdot|s)\|_\infty &\leq \frac{1}{N_{t-1}(s)} \sum_{i=1}^{N_{t-1}(s)} \|\hat{p}_{\tau_i(s)}(\cdot|s) - p^*(\cdot|s)\|_\infty \\
&\leq \frac{\beta}{2N_{t-1}(s)} \sum_{i=1}^{N_{t-1}(s)} \|\hat{Q}_{\tau_i(s)}(s,\cdot) - Q^*(s,\cdot)\|_\infty \qquad (6) \\
&\leq \frac{\beta}{2}\frac{f(N_{t-1}(s))}{N_{t-1}(s)} \qquad (7)
\end{aligned}$$

where in the final line we have used the stateful $f$-guarantee (Equation (2)) which tells us that:

$$\sum_{i=1}^{N_T(s)} \|\hat{Q}_{\tau_i(s)}(s,\cdot) - Q^*(s,\cdot)\|_\infty \leq f(N_T(s))$$

□

We are now equipped to prove Prop. 6.

*Proof of Prop. 6.* Similarly to the stateless case, we will write $p_t(a|s) := \pi^\beta(Q_t(s,\cdot))(a)$, $\hat{p}_t(a|s) := \pi^\beta(\hat{Q}_t(s,\cdot))(a)$, and $p^*(a|s) = \pi^\beta(Q^*(s,\cdot))(a)$ for the Boltzmann policies associated with the predictor, learner, and ground-truth action-value functions respectively. Analogously to the stateless case, using the explicit expression for $Q_t$ in Definition 4 (and using an analogous expression for $Q^*$), we can write

$$\begin{aligned}
\|Q_t(s,\cdot) - Q^*(s,\cdot)\|_\infty = \max_{a\in\mathcal{A}} |Q_t(s,a) - Q^*(s,a)| &\leq \max_{a\in\mathcal{A}} \frac{2}{\beta} \left|\log \frac{p_t(a|s)}{p^*(a|s)}\right| \\
&\leq \frac{2}{\kappa_t(s)\beta} \max_{a\in\mathcal{A}} |p_t(a|s) - p^*(a|s)| \\
&= \frac{2}{\kappa_t(s)\beta} \|p_t(\cdot|s) - p^*(\cdot|s)\|_\infty
\end{aligned}$$

for any $t \geq t_e(s)$. Then, by decomposing

$$p_t(a|s) - p^*(a|s) = (p_t(a|s) - \bar{p}_t(a|s)) + (\bar{p}_t(a|s) - p^*(a|s)) \tag{8}$$

we have that with probability $1-\epsilon$,

$$\begin{aligned}
\|Q_t(s,\cdot) - Q^*(s,\cdot)\|_\infty &\leq \frac{2}{\kappa_t(s)\beta} (\underbrace{\|p_t(\cdot|s) - \bar{p}_t(\cdot|s)\|_\infty}_{\text{Lemma 5}} + \underbrace{\|\bar{p}_t(\cdot|s) - p^*(\cdot|s)\|_\infty}_{\text{Lemma 6}}) \\
&\leq \frac{2}{\kappa_t(s)\beta} \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|(T-1)/\epsilon)}{N_{t-1}(s)}} + \frac{1}{\kappa_t(s)} \frac{f(N_{t-1}(s))}{N_{t-1}(s)}
\end{aligned}$$

Summing over $t \in [t_e(s), T]$ gives us the required result. □

## B Robustness of the Model

After introducing our model, one might wonder why predicting a reward/action-value function at each time step is necessary, rather than just predicting a final reward/action-value function at the end of the episode. We can prove that this doesn't matter. An algorithm that predicts a reward at the end of the interaction can be used to solve the problem of predicting a reward at each step and vice versa. *Importantly, this holds regardless of the choice of learner model and the choice of distance function.*

Fix a distance function $d$ and a function $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$. We say an algorithm $A$ that outputs a single reward estimate $R$ after $T$ rounds is $f(T)$-no-regret if, for the ground-truth reward $R^*$,

$$T\, d(R, R^*) \leq f(T).$$

Likewise, an algorithm $B$ that outputs an estimate $R_t$ on each round $t \in \{1, \ldots, T\}$ is $f(T)$-no-regret if

$$\sum_{t=1}^{T} d(R_t, R^*) \leq f(T).$$

Any algorithm $A$ that predicts a reward estimate at every step with regret $f(T)$ can be converted to an algorithm $B$ that predicts a reward at the end with no loss in regret, if we are okay with in expectation guarantees by simply picking a reward at random.

If we are not okay with randomized guarantees, there also exists a deterministic algorithm $B$ that holds for any convex distance function $d$, by simply averaging all of the outputted rewards. Distance functions, e.g. any norm, tend to be convex, so this is a natural assumption to be making.

**Proposition 7** (Per-step to final reduction)**.** *Under distance function $d$, for any learner model, if there exists an algorithm $A$ that predicts rewards at every iteration and is $f(T)$ no-regret, then there exists a randomized algorithm $B$ that predicts a single reward at the end and is $f(T)$ no-regret.*

*If the distance function $d$ is convex there also exists a deterministic algorithm $B'$ that predicts a single reward at the end and is $f(T)$ no-regret.*

*Proof.* Suppose you have access to an algorithm that predicts a sequence of rewards $R_t$ at every iteration satisfying the guarantee

$$\sum_{t=1}^{T} d(R_t, R^*) \leq f(T).$$

We will give two algorithms that take the sequence of rewards $R_1, \ldots, R_T$ and output a single reward at the end.

Consider the algorithm that simply samples some time step uniformly at random $s \sim \text{Unif}([T])$, and predicts $R_s$. The expected distance of this procedure is:

$$\mathbb{E}[Td(R_s, R^*)] = T\mathbb{E}[d(R_s, R^*)] = T \cdot \frac{1}{T} \sum_{t=1}^{T} d(R_t, R^*) = \sum_{t=1}^{T} d(R_t, R^*) \leq f(T).$$

To construct a deterministic algorithm when $d$ is convex, we use Jensen's inequality. In particular, consider the algorithm that simply outputs the average of the $R_t$. This has regret:

$$\begin{aligned} f(T) &\geq \sum_{t=1}^{T} d(R_t, R^*) \\ &= T \cdot \frac{1}{T} \sum_{t=1}^{T} d(R_t, R^*) \\ &\geq T d\left( \frac{1}{T} \sum_{t=1}^{T} R_t, R^* \right). \end{aligned}$$

□

While most distance functions are convex, the best-response distance (c.f. Section 3.2.1) is not. Despite this, there still exists a deterministic reduction that suffers a regret of at most $|\mathcal{A}|\, f(T)$ regret.

**Proposition 8** (Best-response per-step to final)**.** *Under the best-response distance, for any learner model, if there exists an algorithm $A$ that predicts rewards at every iteration and is $f(T)$ no-regret, then there exists a deterministic algorithm $B$ that predicts a single reward at the end and is $|\mathcal{A}|\, f(T)$ no-regret.*

*Proof.* Suppose you have access to an algorithm that predicts a sequence of rewards $R_t$ at every iteration satisfying the guarantee

$$\sum_{t=1}^{T} R^*(a^*) - R^*(a^{R_t}) \leq f(T),$$

where recall $a^{R_t}$ is the best action under reward $R_t$. Let $a$ be the most common best action under the sequence of rewards. Necessarily, it must be played at least $T/\,|\mathcal{A}|$ times. Therefore,

$$\frac{T}{|\mathcal{A}|} \left( R^*(a^*) - R^*(a) \right) \leq f(T).$$

Let $R$ be the reward that places a reward of 1 on $a$ and a reward of 0 otherwise. Predicting this reward at the end results in a final best-response regret of

$$T(R^*(a^*) - R^*(a^R)) \leq |\mathcal{A}|\, f(T),$$

as claimed. □

We can go the other direction too. Any algorithm that outputs a single reward at the end suffering a regret of at most $f(T)$ can be converted to an algorithm that outputs a reward at every step by running the algorithm at every step. This suffers a regret of $\sum_{t=1}^{T} f(t)/t$.

**Proposition 9.** *[Final to per-step] Suppose there exists an algorithm $B$ that is $f(T)$ no-regret, for non-decreasing $f$, and predicts a single reward at the end. Then there exists an algorithm $A$ that is $f(T)\log(T)$ no-regret and predicts a reward at every time step.*

*If regret is sublinear by a polynomial factor, i.e. $f(T) \in \mathcal{O}(T^{\alpha})$ for some $\alpha \in [0, 1)$, then this can be improved to $1 + \alpha^{-1}(f(T) - 1)$.*

*Proof.* Suppose that there exists an algorithm that predicts a final reward $R$ satisfying the guarantee

$$Td(R, R^*) \leq f(T).$$

To construct a low regret sequence of rewards at every iteration that only depend on observations up until that time step, consider the procedure that just calls this algorithm at every time step $t$. By assumption $d(R_t, R^*) \leq f(t)/t$. Summing over all time steps, this means total regret is

$$\sum_{t=1}^{T} d(R_t, R^*) \leq \sum_{t=1}^{T} f(t)/t \leq f(T) \sum_{t=1}^{T} 1/t \leq f(T)\log(T),$$

where we use the fact that regret $f$ never decreases.

When regret is sublinear by a polynomial factor, that is when regret is $f(T) \in \mathcal{O}(T^{\alpha})$ for $\alpha < 1$,

$$\begin{aligned}\sum_{t=1}^{T} f(t)/t &= \sum_{t=1}^{T} t^{\alpha-1} \\ &\leq 1 + \int_{1}^{T} t^{\alpha-1} dt \\ &= 1 + \frac{T^{\alpha} - 1}{\alpha}.\end{aligned}$$

□