Interaction-Grounded Learning with Action-inclusive Feedback

Anonymous Author(s)
Affiliation
Address
email

Abstract

Consider the problem setting of Interaction-Grounded Learning (IGL), in which a learner's goal is to optimally interact with the environment with no explicit reward to ground its policies. The agent observes a context vector, takes an action, and receives a feedback vector, using this information to effectively optimize a policy with respect to a latent reward function. Prior analyzed approaches fail when the feedback vector contains the action, which significantly limits IGL's success in many potential scenarios such as Brain-computer interface (BCI) or Human-computer interface (HCI) applications. We address this by creating an algorithm and analysis which allows IGL to work even when the feedback vector contains the action, encoded in any fashion. We provide theoretical guarantees and large-scale experiments based on supervised datasets to demonstrate the effectiveness of the new approach.

1 Introduction

Most real-world learning problems, such as BCI and HCI problems, are not tagged with rewards. Consequently, (biological and artificial) learners must infer rewards based on interactions with the environment, which reacts to the learner's actions by generating feedback, but does not provide any explicit reward signal. This paradigm has been previously studied by researchers [e.g., 1, 2], including a recent formalization [3] that proposed the term Interaction-Grounded Learning (IGL).

In IGL, the learning algorithm discovers a grounding for the feedback which implicitly discovers a reward function. An information-theoretic impossibility argument indicates additional assumptions are necessary to succeed. Xie et al. [3] proceed by assuming the action is conditionally independent of the feedback given the reward. However, this is unnatural in many settings such as neurofeedback in BCI [4, 5, 6, 7, 8] and multimodal interactive feedback in HCI [9, 10, 11, 12, 10, 13, 14, 15, 16, 17], where the action proceeds and thus influences the feedback. If you apply this prior approach to these settings, it will fail catastrophically because the requirement of conditional independence is essential to its function. This motivates the question:

> *Is it possible to do interaction-ground learning when the feedback has the full information of the action embedded in it?*

We propose a new approach to solve IGL, which we call action-inclusive IGL (AI-IGL), that allows the action to be incorporated into the feedback in arbitrary ways. We consider latent reward as playing the role of latent states, which can be further separated via a contrastive learning method (Section 3.1). Different from the typical latent state discovery in rich-observation reinforcement learning [e.g., 18, 19, 20], the IGL setting also requires identifying the semantic meaning of the latent reward states, which is addressed by a symmetry breaking procedure (Section 3.2). We analyze the theoretical properties of the proposed approach, and we prove that it is guaranteed to learn a near-optimal policy

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

as long as the feedback satisfies a weaker context conditional independence assumption. We also evaluate the proposed AI-IGL approach using large-scale experiments on Open-ML's supervised classification datasets [21] and demonstrate the effectiveness of the proposed approach (Sec. 5). Thus, our findings broaden the scope of applicability for IGL.

The paper proceeds as follows. In Section 2, we present the mathematical formulation for IGL. In Section 3, we present a contrastive learning perspective for grounding latent reward which helps to expand the applicability of IGL. In Section 4, we state the resulting algorithm AI-IGL. We provide experimental support for the technique in Section 5 using a diverse set of datasets. We conclude with discussion in Section 6.

2 Background

Interaction-Grounded Learning This paper studies the *Interaction-Grounded Learning* (IGL) setting [3], where the learner optimizes for a latent reward by interacting with the environment and associating ("grounding") observed feedback with the latent reward. At each time step, the learner receives an i.i.d. context x from context set $\mathcal{X}$ and distribution d_0. The learner then selects an action $a \in \mathcal{A}$ from a finite action set $|\mathcal{A}| = K$. The environment generates a latent binary reward $r \in \{0, 1\}$ (can be either deterministic or stochastic) and a feedback vector $y \in \mathcal{Y}$ conditional on (x, a), but only the feedback vector y is revealed to the learner. In this paper, we use $R(x, a) \coloneqq \mathbb{E}_{x,a}[r]$ to denote the expected (latent) reward after executing action a on context x. The space of context $\mathcal{X}$ and feedback vector $\mathcal{Y}$ can be arbitrarily large.

Throughout this paper, we use $\pi \in \Pi : \mathcal{X} \to \Delta(\mathcal{A})$ to denote a (stochastic) policy. The expected return of policy π is defined by $V(\pi) = \mathbb{E}_{(x,a)\sim d_0 \times \pi}[r]$. The learning goal of IGL is to find the optimal policy in the policy class, $\pi^\star = \operatorname{argmax}_{\pi \in \Pi} V(\pi)$, only from the observations of context-action-feedback tuples, (x, a, y). This paper mainly considers the batch setting, and we use μ to denote the behavior policy. In this paper, we also introduce value function classes and decoder classes. We assume the learner has access to a value function class $\mathcal{F}$ where $f \in \mathcal{F} : \mathcal{X} \times \mathcal{A} \to [0, 1]$ and reward decoder class $\psi \in \Psi : \mathcal{Y} \times \mathcal{A} \to [0, 1]$. We defer the assumptions we made on these classes to Section 3 for clarity.

We may hope to solve IGL without any additional assumptions. However, it is information-theoretically impossible without additional assumptions, even if the latent reward is decodable from (x, a, y), as demonstrated by the following example.

Example 1 (Hardness of assumption-free IGL)**.** *Suppose $y = (x, a)$ and suppose the reward is deterministic in (x, a). In this case, the latent reward r can be perfectly decoded from y. However, the learner receives no more information than (x, a) from the (x, a, y) tuple. Thus if Π contains at least 2 policies, for any environment where any IGL algorithm succeeds, we can construct another environment with the same observable statistics where that algorithm must fail.*

IGL with full conditional independence Example 1 demonstrates the need for further assumptions to succeed at IGL. Xie et al. [3] proposed an algorithm that leverages the following conditional independence assumption to facilitate grounding the feedback in the latent reward.

Assumption 1 (Full conditional independence)**.** *For arbitrary (x, a, r, y) tuples where r and y are generated conditional on the context x and action a, we assume the feedback vector y is conditionally independent of context x and action a given the latent reward r, i.e. $x, a \perp\!\!\!\perp y | r$.*

Xie et al. [3] introduce a reward decoder class $\psi \in \Psi : \mathcal{Y} \to [0, 1]$ for estimating $\mathbb{E}[r|y]$, which leads to the decoded return $V(\pi, \psi) \coloneqq \mathbb{E}_{(x,a)\sim d_0 \times \pi}[\psi(y)]$. They proved that it is possible to learn the best π and ψ jointly by optimizing the following proxy learning objective:

$$\operatorname{argmax}_{(\pi,\psi)\in\Pi\times\Psi} \mathcal{L}(\pi, \psi) \coloneqq V(\pi, \psi) - V(\pi_{\mathsf{base}}, \psi), \tag{1}$$

where π_{base} is a policy known to have low expected return. Over this paper, we use IGL (full CI) to denote the proposed algorithm in [3].

3 A Contrastive-learning Perspective for Grounding Latent Reward

The existing work of interaction-grounded learning leverages the assumption of full conditional independence (Assumption 1), where the feedback vector only contains information from the latent

reward. The goal of this paper is to relax these constraining assumptions and broaden the scope of applicability for IGL. This paper focuses on the scenario where the action information is possibly embedded in the feedback vector, which is formalized by the following assumption.

Assumption 2 (Context Conditional Independence)**.** *For arbitrary (x, a, r, y) tuple where r and y are generated conditional on the context x and action a, we assume the feedback vector y is conditionally independent of context x given the latent reward r and action a. That is, $x \perp\!\!\!\perp y | a, r$.*

Assumption 2 allows the feedback vector y to be generated from both latent reward r and action a, differing from Assumption 1 which constrains the feedback vector y to be generated based only on the latent reward r. We discuss the implications at the end of this section.

3.1 Grounding Latent Reward via Contrastive Learning

In this section, we propose a contrastive learning objective for interaction-grounded learning, which further guides the design of our algorithm. We perform derivations with exact expectations for clarity and intuitions and provide finite-sample guarantees in Section 4.

Assumption 3 (Separability)**.** *For each $\bar{a} \in \mathcal{A}$, there exists an $(f^\star_{\bar{a}}, \psi^\star_{\bar{a}}) \in \mathcal{F} \times \Psi$, such that: 1) $\mathbb{E}\left[\psi^\star_{\bar{a}}(y, \bar{a}) | \bar{a}, r = 1\right] - \mathbb{E}\left[\psi^\star_{\bar{a}}(y, \bar{a}) | \bar{a}, r = 0\right] = 1$; 2) $\left|\mathbb{E}_\mu\left[f^\star_{\bar{a}}(x, \bar{a}) | \bar{a}, r = 1\right] - \mathbb{E}_\mu\left[f^\star_{\bar{a}}(x, \bar{a}) | \bar{a}, r = 0\right]\right| \geq \Delta_{\mathcal{F}}$. We also assume $1 - \Psi \subseteq \Psi$, where $1 - \Psi \coloneqq \{1 - \psi(\cdot, \cdot) : \psi \in \Psi\}$.*

Assumption 3 consists of two components. For Ψ, it is a realizability assumption that ensures that a perfect reward decoder is included in the function classes. Although this superficially appears unreasonably strong, note y is generated based upon a and the *realization* of r; therefore this is compatible with stochastic rewards. For $\mathcal{F}$, it ensures the expected predicted reward conditioned on the latent reward, having value $r \in \{0, 1\}$ and the action being a, is separable. When $\mu = \pi_{\mathsf{unif}}$, $\Delta_{\mathcal{F}}$ can be lower bounded by $\max_{f \in \mathcal{F}} 4 |\mathrm{Cov}_{\pi_{\bar{a}}}(f, R)|$ ($\pi_{\bar{a}}$ denotes the constant policy with action $\bar{a}$). Detailed proof of this argument can be found in Appendix A. One sufficient condition is that the expected latent reward $R(\cdot, \bar{a})$ has enough variance and $R \in \mathcal{F}$. The condition of $1 - \Psi \subseteq \Psi$ can be constructed easily via standard realizable classes. That is, if $\psi^\star_{\bar{a}} \in \Psi'$ for some classes Ψ', simply setting $\Psi \leftarrow \Psi' \bigcup (1 - \Psi')$ satisfies Assumption 3. Note that, this construction of Ψ only amplifies the size of Ψ' by a factor of 2.

Reward Prediction via Contrastive Learning We now formulate the following contrastive-learning objective for solving IGL. Suppose $\mu(a|x)$ is the behavior policy. We also abuse $\mu(x, a, y)$ to denote the data distribution, and $\mu_a(x, y)$, $\mu_a(x)$, $\mu_a(y)$ to denote the marginal distribution under action $a \in \mathcal{A}$. We construct an augmented data distribution for each $a \in \mathcal{A}$: $\widetilde{\mu}_a(x, y) \coloneqq \mu_a(x) \cdot \mu_a(y)$ (i.e., sampling x and y independently from μ_a).

Conceptually, for each action $a \in \mathcal{A}$, we consider tuples $(x, y) \sim \mu_a(x, y)$ and $(\widetilde{x}, \widetilde{y}) \sim \widetilde{\mu}_a(x, y)$. From Assumption 3, conditioned on (x, a) the optimal feedback decoder $\psi^\star(y, a)$ has mean equal to the optimal reward predictor $f^\star(x, a)$. Therefore we might seek an (f, ψ) pair which minimizes any consistent loss function, e.g., squared loss. However this is trivially achievable, e.g., by having both always predict 0. Therefore we formulate a contrastive-learning objective, where we maximize the loss between the predictor and decoder on the augmented data distribution. For each $a \in \mathcal{A}$, we solve the following objective

$$\operatorname*{argmin}_{(f_a, \psi_a) \in \mathcal{F} \times \Psi} \mathcal{L}_a(f_a, \psi_a) \coloneqq \mathbb{E}_{\mu_a}\left[\left(f_a(x, a) - \psi_a(y, a)\right)^2\right] - \mathbb{E}_{\widetilde{\mu}_a}\left[\left(f_a(\widetilde{x}, a) - \psi_a(\widetilde{y}, a)\right)^2\right]. \quad (2)$$

In the notation of equation (2), the a subscript indicates the (f_a, ψ_a) pair are optimal for action a. Note they are always evaluated at a, which we retain as an input for compatibility with the original function classes.

We now show that $\mathcal{L}$ is equivalent to a spectral contrastive loss [e.g., 22, 23, 24].

$$\begin{aligned}
\mathcal{L}_a(f_a, \psi_a) &= \mathbb{E}_{\mu_a}\left[f_a(x, a)^2 - 2 f_a(x, a)\psi_a(y, a) + \psi_a(y, a)^2\right] \\
&\quad - \mathbb{E}_{\widetilde{\mu}_a}\left[f_a(x, a)^2 - 2 f_a(x, a)\psi_a(y, a) + \psi_a(y, a)^2\right] \\
&= -2\left(\mathbb{E}_{\mu_a}\left[f_a(x, a)\psi_a(y, a)\right] - \mathbb{E}_{\widetilde{\mu}_a}\left[f_a(x, a)\psi_a(y, a)\right]\right) \quad \textbf{(spectral contrastive loss)} \\
&= -2\left(\mathbb{E}_{\mu_a}\left[f_a(x, a)\psi_a(y, a)\right] - \mathbb{E}_{\mu_a}\left[f_a(x, a)\right]\mathbb{E}_{\mu_a}\left[\psi_a(y, a)\right]\right). \\
&\qquad\qquad (\widetilde{\mu}_a(x, y) = \mu_a(x) \cdot \mu_a(y))
\end{aligned}$$

Below we show that minimizing $\mathcal{L}_a(f_a, \psi_a)$ decodes the latent reward under Assumptions 2 and 3 up to a sign ambiguity. For simplicity, we introduce the notation of $f_{a,r}$ and $\psi_{a,r}$ for any $(f, \psi) \in \mathcal{F} \times \Psi$,

$$f_{a,r} \coloneqq \sum_x \Pr(x|a,r) f(x,a), \quad \psi_{a,r} \coloneqq \sum_y \Pr(y|a,r)\psi(y,a). \tag{3}$$

$f_{a,r}$ and $\psi_{a,r}$ are the expected predicted reward of $f(x,a)$ and decoded reward $\psi(y,a)$ (under the behavior policy μ for $f_{a,r}$) conditioned on the latent reward having value r and the action being a.

Proposition 1. *For any action $a \in \mathcal{A}$, if $\mu(a|x) > 0$ for all $x \in \mathcal{X}$ and Assumption 2 and 3 hold, and let $(\widehat{f}_a, \widehat{\psi}_a)$ be the solution of Eq.(2). Then, $|\widehat{f}_{a,1} - \widehat{f}_{a,0}| = \max_{f \in \mathcal{F}} |f_{a,1} - f_{a,0}|$ and $|\widehat{\psi}_{a,1} - \widehat{\psi}_{a,0}| = \max_{\psi \in \Psi} |\psi_{a,1} - \psi_{a,0}|$.*

Proof Sketch. For any policy π, we use $d_a^\pi \coloneqq \sum_x d_0(x)\pi(a|x)$ to denote the visitation occupancy of action a under policy π, and $\rho_a^\pi \coloneqq {}^1\!/_{d_a^\pi} \sum_x d_0(x)\pi(a|x)R(x,a)$ to denote the average reward received under executing action a. Then, by the context conditional independent assumption $x \perp\!\!\!\perp y | r, a$ (Assumption 2), we know

$$\begin{aligned}
\mathbb{E}_{\mu_a}\left[f(x)\psi(y)\right] &= (1-\rho_a^\mu) f_{a,0}\psi_{a,0} + \rho_a^\mu f_{a,1}\psi_{a,1} \\
\mathbb{E}_{\mu_a}\left[f(x)\right]\mathbb{E}_{\mu_a}\left[\psi(y)\right] &= (1-\rho_a^\mu)^2 f_{a,0}\psi_{a,0} + (\rho_a^\mu)^2 f_{a,1}\psi_{a,1} + (1-\rho_a^\mu)\rho_a^\mu (f_{a,0}\psi_{a,1} + f_{a,1}\psi_{a,0}) \\
\implies \mathcal{L}_a(f_a, \psi_a) &\propto -(f_{a,1} - f_{a,0})(\psi_{a,1} - \psi_{a,0}).
\end{aligned}$$

Therefore, separately maximizing $|f_{a,1} - f_{a,0}|$ and $|\psi_{a,1} - \psi_{a,0}|$ maximizes $\mathcal{L}_a(f_a, \psi_a)$. □

3.2 Symmetry Breaking

In the last section, we demonstrated that the latent reward could be decoded in a contrastive-learning manner up to a sign ambiguity. The following example demonstrates the ambiguity.

Example 2 (Why do we need extra information to identify the latent reward?)**.** *In the optimal solution of objective Eq.(2), both $\widehat{\psi}_a$ and $\widehat{\psi}'_a \coloneqq 1 - \widehat{\psi}_a$ yield the same value. It is information-theoretically difficult to distinguish which one of them is the correct solution without extra information. That is because, for any environment* ENV1*, there always exists a "symmetric" environment* ENV2*, where: 1) $R(x,a)$ of* ENV1 *is identical to $(1 - R(x,a))$ of* ENV2 *for all $(x,a) \in \mathcal{X} \times \mathcal{A}$; 2) the the conditional distribution of $y|r,a$ in* ENV1 *is identical to the conditional distribution of $y|1-r,a$ in the* ENV2 *for all $a \in \mathcal{A}$. In this example,* ENV1 *and* ENV2 *will always generate the identical distribution of feedback vector y after any $(x,a) \in \mathcal{X} \times \mathcal{A}$. However,* ENV1 *and* ENV2 *have the exactly opposite latent reward information.*

As we demonstrate in Example 2, the learner decoder from Eq.(2) could be corresponding to a symmetric pair of semantic meanings, and identifying them without extra information is information-theoretically impossible. The *symmetry breaking* procedure is one of the key challenges of interaction-grounded learning. To achieve symmetry breaking, we make the following assumption to ensure the identifiability of the latent reward.

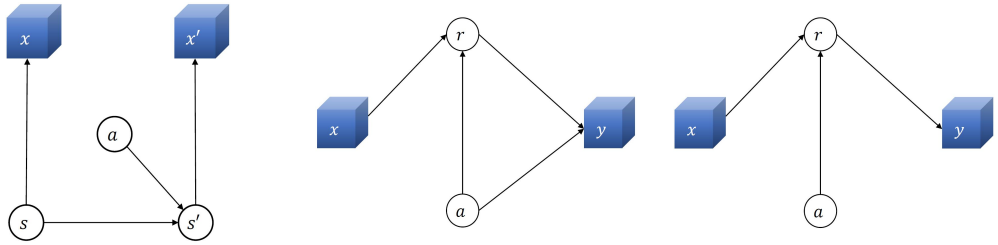
Assumption 4 (Baseline Policies)**.** *For each $a \in \mathcal{A}$, there exists a baseline policy π^a_{base}, such that,*

(a) π^a_{base} *satisfies* $\sum_x d_0(x)\pi(a|x) \geq c_m > 0$.

(b) $|\frac{1}{2} - \rho_a^{\pi^a_{\mathsf{base}}}| \geq \eta$, *where* $\rho_a^{\pi^a_{\mathsf{base}}} = \frac{\sum_x d_0(x)\pi^a_{\mathsf{base}}(a|x)R(x,a)}{\sum_x d_0(x)\pi^a_{\mathsf{base}}(a|x)}$.

To instantiate Assumption 4 in practice, we provide the following simple example of π_{base} that satisfies Assumption 4. Suppose $\pi_{\mathsf{base}} = \pi_{\mathsf{unif}}$ (uniformly random policy), and we have "all constant policies are bad", i.e., $V(\pi_{\bar{a}} = \mathbb{1}(a = \bar{a})) < {}^1\!/_2 - \eta$ for all $\bar{a} \in \mathcal{A}$. Then it is easy to verify that $c_m = {}^1\!/_K$ and $\rho_{\bar{a}}^{\pi_{\mathsf{base}}} \leq {}^1\!/_2 - \eta$ for all $\bar{a} \in \mathcal{A}$.

Note that π^a_{base} can be different over actions. Intuitively, Assumption 4(a) is saying that the total probability of π^a_{base} selecting action a (over all context $x \in \mathcal{X}$) is at least c_m. This condition ensures that π^a_{base} has enough visitation to action a and makes symmetry breaking possible. Assumption 4(b) states that if we only consider the reward obtained from taking action a, π^a_{base} is known to be either "sufficiently bad" or "sufficiently good". Note the directionality of the extremeness of π^a_{base} must be known, e.g., a policy which has a unknown reward of either 0 or 1 is not usable. This condition follows



(a) Latent State in Rich-Observation RL [20]

(b) Latent Reward in IGL under Assumption 2 (this paper)

(c) Latent Reward in IGL under Assumption 1 [3]

Figure 1: Causal graphs of interaction-grounded learning under different assumptions as well as rich-observation reinforcement Learning.

a similar intuition as the identifiability assumption of [3, Assumption 2] and breaks the symmetry. For example, consider the ENV1 and ENV2 introduced in Example 2, $\rho^{\pi}_{a,\mathsf{ENV1}} = 1 - \rho^{\pi}_{a,\mathsf{ENV2}}$ for any policy π. To separating ENV1 and ENV2 using some policy π, $\rho^{\pi}_{a,\mathsf{ENV1}}$ and $\rho^{\pi}_{a,\mathsf{ENV2}}$ require to have a non-zero gap, which leads to Assumption 4(b).

The effectiveness of symmetry breaking under Assumption 4 can be summarized as below: we conduct the following estimation of $\rho^{\pi_{\mathsf{base}}}_{a}$, using the learned $\widehat{\psi}_a$, $\widehat{\rho}^{\pi^a_{\mathsf{base}}}_{a} = \frac{\sum_x d_0(x)\pi^a_{\mathsf{base}}(a|x)\widehat{\psi}_a(x,a)}{\sum_x d_0(x)\pi^a_{\mathsf{base}}(a|x)}$. If $\widehat{\psi}_a$ can efficiently decode the latent reward, then $\widehat{\rho}^{\pi^a_{\mathsf{base}}}_{a}$ converges to either $\rho^{\pi_{\mathsf{base}}}_{a}$ or $1 - \rho^{\pi_{\mathsf{base}}}_{a}$. Therefore, applying Assumption 4(b) breaks the symmetry.

3.3 Comparison to Full CI

When we have the context conditional independence, it is easy to verify the failure of optimizing the original IGL objective Eq.(1) by the following example.

Example 3 (Failure of the original IGL objective under Assumption 2)**.** *Let $\mathcal{X} = \mathcal{A} = \{1, 2, \ldots, 10\}$ and feedback vector is generated by $y = (a + R(x, a)) \mod 10$ (we use $\%$ to denote $\mod$ in the following part). We also assume $d_0(x) = \pi_{\mathsf{base}}(a|x) = 1/10$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $R(x, a) = \pi^\star(a|x) := \mathbb{1}(x = a)$. Then, we have, for any $\psi : \mathcal{Y} \to [0, 1]$ (approach proposed by Xie et al. [3] assumes the reward decoder only takes feedback vector y as the input),*

$$\begin{aligned}\mathcal{L}(\pi^\star, \psi) &= \tfrac{1}{10}\textstyle\sum_{x=1}^{10}\sum_{a=1}^{10} \mathbb{1}(x = a)\psi((a + 1)\%10) - \tfrac{1}{100}\sum_{x=1}^{10}\sum_{a=1}^{10} \psi((a + \mathbb{1}(x = a))\%10)\\ &= \tfrac{1}{10}\textstyle\sum_{a=1}^{10} \psi(a) - \tfrac{1}{10}\sum_{a=1}^{10} \psi(a) = 0.\end{aligned}$$

On the other hand, consider the constant policy $\pi_1(a|x) := \mathbb{1}(a = 1)$ for all $x \in \mathcal{X}$ and decoder $\psi_2(y) := \mathbb{1}(y = 2)$ for all $y \in \mathcal{Y}$, then,

$$\begin{aligned}\mathcal{L}(\pi_1, \psi_2) &= \tfrac{1}{10}\textstyle\sum_{x=1}^{10}\sum_{a=1}^{10} \mathbb{1}(a = 1)\psi_2((a + 1)\%10) - \tfrac{1}{100}\sum_{x=1}^{10}\sum_{a=1}^{10} \psi_2((a + \mathbb{1}(x = a))\%10)\\ &= \psi_2(2) - \tfrac{1}{10}\textstyle\sum_{a=1}^{10} \psi_2(a) = 0.9 > \mathcal{L}(\pi^\star, \psi), \quad \forall \psi \in \Psi.\end{aligned}$$

This implies that maximizing the original IGL objective Eq.(1) could not always converge to $\pi^\star$ when we only have the context conditional independence.

This example indicates optimizing a combined contrastive objective with a single symmetry-breaking policy is insufficient to succeed in our $x \perp\!\!\!\perp y | r, a$ case. Our current approach corresponds to optimizing a contrastive objective and breaking symmetry for each action separately rather than simultaneously.

3.4 Viewing Latent Reward as a Latent State

Our approach is motivated by latent state discovery in Rich-Observation RL [20]. Figure 1 compares the causal graphs of Rich-Observation RL, IGL with context conditional independence, and IGL with full conditional independence. In Rich-Observation RL, a contrastive learning objective is used to discover latent states; whereas in IGL a contrastive learning objective is used to discover latent rewards. In this manner we view latent rewards analogously to latent states.

Identifying latent states up to a permutation is completely acceptable in Rich-Observations RL, as the resulting imputed MDP orders policies identically to the true MDP. However in IGL the latent

states have scalar values associated with them (rewards), and identifying up to a permutation is not sufficient to order policies correctly. Thus, we require the additional step of symmetry breaking.

4 Main Algorithm

This section instantiates the algorithm for IGL with action-inclusive feedback, following the concept we introduced in Section 3. For simplicity, we select the uniformly random policy π_{unif} as behavior policy μ in this section. That choice of μ can be further relaxed using an appropriate importance weight. We first introduce the empirical estimation of $\mathcal{L}_a$ (defined in Eq.(2)) as follows. For any $\bar{a} \in \mathcal{A}$, we define the following empirical estimation of the spectral contrastive loss:

$$\mathcal{L}_{\bar{a},\mathcal{D}}(f,\psi) = \mathbb{E}_{\mathcal{D}}\left[f(x)\psi(y)\mathbb{1}(a=\bar{a})\right] - \mathbb{E}_{\mathcal{D}}\left[f(x)\mathbb{1}(a=\bar{a})\right]\mathbb{E}_{\mathcal{D}}\left[\psi(y)\mathbb{1}(a=\bar{a})\right]. \tag{4}$$

Using this definition, Algorithm 1 instantiates a version of the IGL algorithm with action-inclusive feedback. Without loss of generality, we also assume $\rho_a^{\pi_{\mathsf{base}}^a} \leq 1/2$ for Assumption 4(b) for all $a \in \mathcal{A}$ in this section. The case of $\rho_a^{\pi_{\mathsf{base}}^a} > 1/2$ for some action a can be addressed by modifying the symmetry-breaking step properly in Algorithm 1.

Algorithm 1 Action-inclusive IGL (AI-IGL)

Input: Batch data $\mathcal{D}$ generated by $\mu = \pi_{\mathsf{unif}}$. baseline policy $\pi_{\mathsf{base}}^{a\in\mathcal{A}}$.

1: Initialize policy π_1 as the uniform policy.
2: **for** $\bar{a} \in \mathcal{A}$ **do**
3: Obtain $(f_{\bar{a}}, \psi_{\bar{a}})$ by ▷ *Latent State (Reward) Discovery*

$$(f_{\bar{a}}, \psi_{\bar{a}}) \leftarrow \operatorname*{argmax}_{(f,\psi)\in\mathcal{F}\times\Psi} \mathcal{L}_{\bar{a},\mathcal{D}}(f,\psi), \tag{5}$$

where $\mathcal{L}_{\bar{a},\mathcal{D}}(f,\psi)$ is defined in Eq.(4)

4: Compute $\widehat{\varrho}_{\bar{a}}^{\pi_{\mathsf{base}}^{\bar{a}}}$ by $\widehat{\varrho}_{\bar{a}}^{\pi_{\mathsf{base}}^{\bar{a}}} = \frac{\sum_{(x,a,y)\in\mathcal{D}} \pi_{\mathsf{base}}^{\bar{a}}(a|x)\psi_{\bar{a}}(x,\bar{a})\mathbb{1}(a=\bar{a})}{\sum_{(x,a,y)\in\mathcal{D}} \pi_{\mathsf{base}}^{\bar{a}}(a|x)\mathbb{1}(a=\bar{a})}$. ▷ *Symmetry Breaking*
5: **if** $\widehat{\varrho}_{\bar{a}}^{\pi_{\mathsf{base}}} > \frac{1}{2}$ **then** $\psi'_{\bar{a}} \leftarrow (1-\psi_{\bar{a}})$.
6: **else** $\psi'_{\bar{a}} \leftarrow \psi_{\bar{a}}$.
7: **end if**
8: **end for**
9: Generate decoded contextual bandits dataset $\mathcal{D}_{\mathsf{CB}} \leftarrow \{(x,a,\psi'_a(y,a),\mu(a|x)) : (x,a,y)\in\mathcal{D}\}$.
10: Output policy $\widehat{\pi}(x) \leftarrow \mathsf{CB}(\mathcal{D}_{\mathsf{CB}})$, where CB denotes an offline contextual bandit oracle.

At a high level, Algorithm 1 has two separate components, latent state (reward) discovery (line 3) and symmetry breaking (line 4-7), for each action in $\mathcal{A}$.

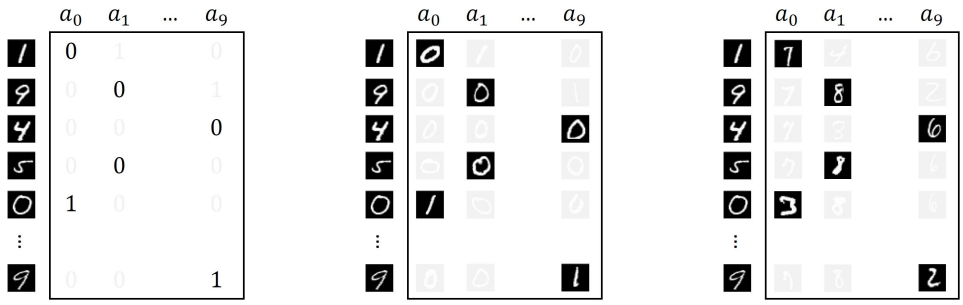
Theoretical guarantees In Algorithm 1, the output policy $\widehat{\pi}$ is obtained by calling an offline contextual bandits oracle (CB). We now formally define this oracle and its expected property.

Definition 1 (Offline contextual bandits oracle)**.** An algorithm CB is called an *offline contextual bandit oracle* if for any dataset $\mathcal{D} = \{(x_i, a_i, r_i, \mu(a_i|x_i))\}_{i=1}^{|\mathcal{D}|}$ ($x_i \sim d_0, a_i \sim \mu$, and r_i is the reward determined by (x_i, a_i)) and any policy class Π, the policy $\widehat{\pi}$ produced by $\mathsf{CB}(\mathcal{D})$ satisfies $\varepsilon_{\mathsf{CB}} := \max_{\pi\in\Pi} \mathbb{E}_{d_0\times\pi}[r] - \mathbb{E}_{d_0\times\widehat{\pi}}[r] \leq o(1)$.

The notion in Definition 1 corresponds to the standard policy learning approaches in the contextual bandits literature [e.g., 25, 26, 27], and typically leads to $\varepsilon_{\mathsf{CB}} = \sqrt{K^{\log|\Pi|/\delta}/|\mathcal{D}|}$. We now provide the theoretical analysis of Algorithm 1. In this paper, we use $d_{\mathcal{F},\Psi}$ to denote the joint statistical complexity of the class of $\mathcal{F}$ and Ψ. For example, if the function classes are finite, we have $d_{\mathcal{F},\Psi} = \mathcal{O}(\log|\mathcal{F}||\Psi|/\delta)$, and δ is the failure probability. The infinite function classes can be addressed by some advanced methods such as covering number or Rademacher complexity [see, e.g., 28]. The following theorem provides the performance guarantee of the output policy of Algorithm 1.

Theorem 2. *Suppose Assumptions 2, 3 and 4 hold. Let $\widehat{\pi}$ be the output policy of Algorithm 1 and $\Delta_{\mathcal{F}} = \max_{f\in\mathcal{F}} |f_{a,1} - f_{a,0}|$. If we have $|\mathcal{D}| \geq \mathcal{O}\left(\frac{K^3 d_{\mathcal{F},\Psi}}{(\min\{\eta\Delta_{\mathcal{F}}, Kc_m\})^2}\right)$, then, with high probability,*

$$V(\pi^\star) - V(\pi) \leq \mathcal{O}\left(\frac{1}{\Delta_{\mathcal{F}}}\sqrt{\frac{K^3 d_{\mathcal{F},\Psi}}{|\mathcal{D}|}}\right) + \varepsilon_{\mathsf{CB}}.$$



(a) Contextual Bandits (CB) (b) IGL (full CI) [3] (c) AI-IGL

Figure 2: Different learning approaches based on the MNIST dataset. The gray number/image denotes the unobserved reward/feedback vector. **Figure 2(a):** In the contextual bandits setting, the exact reward information on the selected action can be observed. **Figure 2(b):** In IGL (full CI), the feedback vector is generated only based on the latent reward. **Figure 2(c):** In AI-IGL, the feedback vector can be generated based on both latent reward and selected action.

Similar to the performance of [3], the learned is guaranteed to converge in the right direction only after we have sufficient data for the symmetric breaking. The dependence on K in Theorem 2 can be improved as different action has a separate learning procedure. For example, if we consider $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_K$ and $\Psi = \Psi_1 \times \Psi_2 \times \cdots \times \Psi_K$, where $\mathcal{F}_a$ and Ψ_a are independent components that is only corresponding to action a (this is a common setup for linear approximated reinforcement learning approaches with discrete action space). If $\mathcal{F}_a$ and Ψ_a are identical copies of K separate classes, we know $\log|\mathcal{F}| = K \log|\mathcal{F}_a|$ and $\log|\Psi| = K \log|\Psi_a|$, which leads a $\sqrt{K}$ improvement.

We now provide the proof sketch of Theorem 2 and we defer the detailed proof to Appendix A.

Proof Sketch. The proof of Theorem 2 consists of two different components—discovering latent reward and breaking the symmetry, which are formalized by the following lemma.

Lemma 3 (Discovering latent reward)**.** *Suppose Assumptions 2 and 3 hold, and let* $(f_{\bar{a}}, \psi_{\bar{a}})$ *be obtained by Eq.(5). Then, with high probability, we have* $|\psi_{\bar{a},1} - \psi_{\bar{a},0}| \geq \left(1 - \mathcal{O}\left(\frac{1}{\Delta_{\mathcal{F}}}\sqrt{\frac{K^3 d_{\mathcal{F},\Psi}}{|\mathcal{D}|}}\right)\right)$.

Lemma 3 ensures that the learned decoder on Eq.(5) correctly separates the latent reward. In particular since ψ ranges over $[0,1]$, Lemma 3 ensures $\max\left(\Pr(\psi(y,a)=r), \Pr(\psi(y,a)=1-r)\right) > 1-o(1)$ under the behaviour policy. Thus, if we can break symmetry, we can use ψ to generate a reward signal and reduce to ordinary contextual bandit learning. The following lemma guarantees the correctness of the symmetry-breaking step.

Lemma 4 (Breaking symmetry)**.** *Suppose Assumptions 4 hold. For any* $\bar{a} \in \mathcal{A}$*, if we have* $|\mathcal{D}| \geq \mathcal{O}\left(\frac{K^3 d_{\mathcal{F},\Psi}}{(\min\{\eta\Delta_{\mathcal{F}}, Kc_m\})^2}\right)$*, then, with high probability,* $\psi'_{\bar{a},1} \geq \psi'_{\bar{a},0}$.

Combining these two lemmas above as well as the CB oracle establishes the proof of Theorem 2, and their detailed proof can be found in Appendix A. □

5 Empirical Evaluations

In this section, we provide empirical evaluations in simulated environments created using supervised classification datasets.

The task is depicted in Figure 2. We evaluate our approach by comparing: (1) CB: Contextual Bandits with exact reward, (2) IGL (full CI): The method proposed by Xie et al. [3] which assumes the feedback vector contains no information about the context and reward, and (3) AI-IGL: The proposed method which assumes that the feedback vector could contain information about the action but is conditionally independent of the context given the reward. Note that contextual bandits (CB) is a skyline compared to both IGL (full CI) and AI-IGL, since it need not disambiguate the latent reward.

All methods use logistic regression with a linear representation. At test time, each method takes the argmax of the policy. We provide details of setting up the experiment in Appendix B.

We evaluate our approach on different environments with action-inclusive feedback generated using supervised classification datasets. We generate high-dimensional feedback vectors for the MNIST classification dataset (Sec. 5.1) and low-dimensional feedback vectors for more than 200 OpenML CC-18 datasets (Sec. 5.2). Following the practical instantiation of Assumption 4 (see Section 3.2), we know if the dataset has a balanced action distribution (no action belongs to more than 50% of the samples), selecting random policy π_{unif} as π^a_{base} for all $a \in \mathcal{A}$ satisfies Assumption 4. Therefore, in this section, our experiments are all based on the dataset with a balanced action distribution, and we select $\pi^a_{\mathsf{base}} = \pi_{\mathsf{unif}}$ for all $a \in \mathcal{A}$.

5.1 Experiments on the MNIST dataset

The environment for this experiment is generated from the supervised classification MNIST dataset [29] which is licensed under Attribution-Share Alike 3.0 license[1]. At each time step, the context x_t is generated uniformly at random. Then the learner selects an action $a_t \in \{0, \ldots, 9\}$ as the predicted label of x_t. The binary reward r is the correctness of the prediction label a_t. The high-dimensional feedback vector y_t is an image of the digit $(a_t + 6r - 3) \mod 10$. An example is shown in Figure 2. Our results are averaged over 20 trials.

Algorithm	Policy accuracy for action-inclusive feedback (%)	Policy accuracy for action-exclusive feedback (%)
CB	$\mathbf{87.64 \pm 0.25}$	$\mathbf{87.64 \pm 0.25}$
IGL (full CI)	9.18 ± 2.10	86.13 ± 0.52
AI-IGL	$\mathbf{83.63 \pm 1.25}$	$\mathbf{86.14 \pm 0.94}$

Table 1: Results in the MNIST environment with high-dimensional action-inclusive and action-exclusive feedback. Average and standard error reported over 20 trials for each algorithm.

To highlight that the proposed algorithm still operates well under conditions where the feedback vector does not include action information, we also perform experiments under the setting introduced by Xie et al. [3] on the MNIST environment. This setting is similar to the one described in Sec. 5.1 except the feedback vector is the image of the digit r instead of $a + r$. We find that under this action-exclusive feedback setting, the new proposed algorithm AI-IGL works as well as the IGL (full CI). This signifies that our algorithm, which incorporates the presence of action information in the feedback vector, does not hurt performance in cases when the action information is missing from the feedback vector.

5.2 Large-scale Experiments with OpenML CC-18 Datasets

To verify that our proposed algorithm scales to a variety of tasks, we evaluate performance on datasets from the publicly available OpenML Curated Classification Benchmarking Suite [30, 31, 32, 21]. OpenML datasets are licensed under CC-BY license[2] and the platform and library are licensed under the BSD (3-Clause) license[3]. Similar to Sec. 5.1, at each time step, the context x_t is generated uniformly at random. Again, the learner selects an action $a_t \in \{0, \ldots, K-1\}$ as the predicted label of x_t (where K is the total number of actions available in the environment). The binary reward r_t is the correctness of the prediction label a_t. The feedback y_t is a two dimensional vector (a_t, r_t). Each dataset has a different sample size (N) and a different set of available actions (K). We sample datasets with 3 or more actions, and with a balanced action distribution (no action belongs to more than 50% of the samples) to satisfy Assumption 4. We use 90% of the data for training and the remaining 10% for evaluation. The results are averaged over 20 trials.

Analysis of Dataset Properties in relation to AI-IGL's performance To further understand under which conditions AI-IGL succeeds, we analyzed its performance against 12 different features (based on findings from Lorena et al. [33], Torra et al. [34], Reif et al. [35], Abdelmessih et al. [36]) for

[1] https://creativecommons.org/licenses/by-sa/3.0/

[2] https://creativecommons.org/licenses/by/4.0/

[3] https://opensource.org/licenses/BSD-3-Clause

Dataset Criteria	Dataset Count	Constant Action	CB Policy Accuracy (%)	IGL (full CI) Policy Accuracy (%)	Performance w.r.t CB	AI-IGL Policy Accuracy (%)	Performance w.r.t CB
$K \geq 3$	271	25.28 ± 2.96	$\mathbf{57.98 \pm 5.65}$	15.65 ± 2.30	0.30 ± 0.04	$\mathbf{35.74 \pm 1.45}$	$\mathbf{0.59 \pm 0.02}$
$K \geq, N \geq 70000$	83	22.57 ± 3.14	$\mathbf{58.41 \pm 5.04}$	11.91 ± 1.80	0.22 ± 0.04	$\mathbf{50.11 \pm 2.98}$	$\mathbf{0.79 \pm 0.03}$

Table 2: Results in the OpenML environments with two-dimensional action-inclusive feedback. Average and standard error reported over 20 trials for each algorithm. 'Performance w.r.t. CB' reports the ratio of an IGL method's policy accuracy over CB policy accuracy.

each OpenML dataset (more details in Appendix B.3). We consider 3 additional features based on the following criteria. Compared to the typical CB guarantee, AI-IGL needs one more K factor in its theoretical guarantees (Theorem 2). Since K factors can be improved under some specific choice of function class (see discussion in Section 4), we select all of N, $N/\sqrt{K}$, N/K as the additional features of the dataset for predicting the relative performance of AI-IGL.

We use a binary random forest classifier to predict the success of AI-IGL's performance relative to CB. If the relative performance is ≥ 0.7, we label it as a success. We find that with all 271 datasets with $K \geq 3$, 10-fold cross-validation with 100 trees gives an average F1-score of 0.85. We find N/K to be the most predictive feature of AI-IGL's relative performance (Fig. 4(a)). It can alone predict it's performance with an average F1 score of 0.79 under the same experimental setup. However, for datasets with a small value of N/K (≤ 1000), there is high variability in relative performance. Using such a subset of datasets, we find maximum Fisher discriminant [33] (a measure of classification complexity that quantifies the informativeness of a given sample) to be the most predictive of relative performance (Fig 4(b)). Such datasets are representative of realistic interaction datasets with small sample sizes. More details on the experimental analysis are in Appendix B.3 (Fig. 5 and Fig. 6). This finding makes it possible to predict, for a given dataset, whether AI-IGL can match CB performance. It can also help researchers improve the design of novel applications of IGL, e.g., in HCI and BCI, by ensuring the resulting dataset's features are amenable to high performance.

6 Discussion

We have presented a new approach to solving Interaction-Grounded Learning, in which an agent learns to interact with the environment in the absence of any explicit reward signals. Compared to a prior solution [3], the proposed AI-IGL approach removes the assumption of conditional independence of actions and the feedback vector by treating the latent reward as a latent state. It thereby provably solves IGL for action-inclusive feedback vectors.

By viewing the feedback as containing an action-(latent) reward pair which is an unobserved latent space, we propose latent reward learning using a contrastive learning approach. This solution concept naturally connects to latent state discovery in rich-observation reinforcement learning [e.g., 18, 19, 20]. On the other hand, different from rich-observation RL, the problem of IGL also contains a unique challenge in identifying the semantic meaning of the decoded class, which is addressed by a symmetry-breaking procedure. In this work, we focus on binary latent rewards, for which symmetry breaking is possible using one policy (per action). Breaking symmetry in more general latent reward spaces is a topic for future work. A possible negative societal impact of this work can be performance instability, especially with inappropriate use of the techniques in risk-sensitive applications.

Barring intentional misuse, we envision several potential benefits of the proposed approach. The proposed algorithm broadens the scope of IGL's feasibility for real-world applications (where action signals are included in the feedback). Imagine an agent being trained to interpret the brain signals of a user to control a prosthetic arm. The brain's response (feedback vector) to an action is a neural signal that may contain information about the action itself. This is so prevalent in neuroimaging that fMRI studies routinely use specialized techniques to orthogonalize different information (e.g. action and reward) within the same signal [37, 38, 39, 40]. Another example is a continually self-calibrating eye tracker, used by people with motor disabilities such as ALS [15, 17, 41, 42]. A learning agent adapting to the ability of such users can encounter feedback directly influenced by the calibration correction action. The proposed approach is a stepping stone on the path to solving IGL in such complex settings, overcoming the need for explicit rewards as well as explicit separation of action and feedback information.

References

[1] Khanh X Nguyen, Dipendra Misra, Robert Schapire, Miroslav Dudík, and Patrick Shafto. Interactive learning from activity description. In *International Conference on Machine Learning*, pages 8096–8108. PMLR, 2021.

[2] Jonathan Grizou, Inaki Iturrate, Luis Montesano, Pierre-Yves Oudeyer, and Manuel Lopes. Calibration-free bci based control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

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

[4] Kapil D Katyal, Matthew S Johannes, Spencer Kellis, Tyson Aflalo, Christian Klaes, Timothy G McGee, Matthew P Para, Ying Shi, Brian Lee, Kelsie Pejsa, et al. A collaborative bci approach to autonomous control of a prosthetic limb system. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1479–1482. IEEE, 2014.

[5] Jyoti Mishra and Adam Gazzaley. Closed-loop cognition: the next frontier arrives. *Trends in cognitive sciences*, 19(5):242–243, 2015.

[6] Megan T Debettencourt, Jonathan D Cohen, Ray F Lee, Kenneth A Norman, and Nicholas B Turk-Browne. Closed-loop training of attention with real-time brain imaging. *Nature neuroscience*, 18(3):470–475, 2015.

[7] Santiago Muñoz-Moldes and Axel Cleeremans. Delineating implicit and explicit processes in neurofeedback learning. *Neuroscience & Biobehavioral Reviews*, 118:681–688, 2020.

[8] Iretiayo Akinola, Zizhao Wang, Junyao Shi, Xiaomin He, Pawan Lapborisuth, Jingxi Xu, David Watkins-Valls, Paul Sajda, and Peter Allen. Accelerated robot learning via human brain signals. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 3799–3805. IEEE, 2020.

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

[10] Katherine Taken Smith. Marketing via smart speakers: what should alexa say? *Journal of Strategic Marketing*, 28(4):350–365, 2020.

[11] Holly S Vitense, Julie A Jacko, and V Kathlene Emery. Multimodal feedback: an assessment of performance and mental workload. *Ergonomics*, 46(1-3):68–87, 2003.

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

[13] Kenichiro Noda. Google home: smart speaker as environmental control unit. *Disability and rehabilitation: assistive technology*, 13(7):674–675, 2018.

[14] Radhika Garg and Subhasree Sengupta. He is just like me: A study of the long-term use of smart speakers by parents and children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–24, 2020.

[15] Jensen Gao, Siddharth Reddy, Glen Berseth, Nicholas Hardy, Nikhilesh Natraj, Karunesh Ganguly, Anca D Dragan, and Sergey Levine. X2t: Training an x-to-text typing interface with online learning from user feedback. *arXiv preprint arXiv:2203.02072*, 2022.

[16] Andrew T Duchowski. Gaze-based interaction: A 30 year retrospective. *Computers & Graphics*, 73:59–69, 2018.

[17] Martez E Mott, Shane Williams, Jacob O Wobbrock, and Meredith Ringel Morris. Improving dwell-based gaze typing with dynamic, cascading dwell times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2558–2570, 2017.

[18] Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, 31, 2018.

[19] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

[20] Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

[21] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Frank Hutter, Michel Lang, Rafael G Mantovani, Jan N van Rijn, and Joaquin Vanschoren. Openml benchmarking suites. *arXiv preprint arXiv:1708.03731*, 2017.

[22] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[23] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[24] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.

[25] John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

[26] Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178, 2011.

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

[28] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[30] Joaquin Vanschoren, Jan N van Rijn, B Bischl, G Casalicchio, M Lang, and M Feurer. Openml: a networked science platform for machine learning. In *ICML 2015 MLOSS Workshop*, volume 3, 2015.

[31] Matthias Feurer, Jan N van Rijn, Arlind Kadra, Pieter Gijsbers, Neeratyoy Mallik, Sahithya Ravi, Andreas Müller, Joaquin Vanschoren, and Frank Hutter. Openml-python: an extensible python api for openml. *arXiv preprint arXiv:1911.02490*, 2019.

[32] Giuseppe Casalicchio, Jakob Bossek, Michel Lang, Dominik Kirchhoff, Pascal Kerschke, Benjamin Hofner, Heidi Seibold, Joaquin Vanschoren, and Bernd Bischl. Openml: An r package to connect to the machine learning platform openml. *Computational Statistics*, 34(3): 977–991, 2019.

[33] Ana C Lorena, Luís PF Garcia, Jens Lehmann, Marcilio CP Souto, and Tin Kam Ho. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, 52(5):1–34, 2019.

[34] Vicenç Torra, Yasuo Narukawa, and Toho Gakuen. Modeling decisions for artificial intelligence. *International Journal of Intelligent Systems*, 23(2):113, 2008.

[35] Matthias Reif, Faisal Shafait, Markus Goldstein, Thomas Breuel, and Andreas Dengel. Automatic classifier selection for non-experts. *Pattern Analysis and Applications*, 17(1):83–96, 2014.

[36] Sarah Daniel Abdelmessih, Faisal Shafait, Matthias Reif, and Markus Goldstein. Landmarking for meta-learning using rapidminer. In *RapidMiner community meeting and conference*. Citeseer, 2010.

[37] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[38] Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

[39] Ida Momennejad and John-Dylan Haynes. Human anterior prefrontal cortex encodes the 'what'and 'when'of future intentions. *Neuroimage*, 61(1):139–148, 2012.

[40] Ida Momennejad and John-Dylan Haynes. Encoding of prospective tasks in the human prefrontal cortex under varying task loads. *Journal of Neuroscience*, 33(44):17342–17349, 2013.

[41] John Paulin Hansen, Kristian Tørning, Anders Sewerin Johansen, Kenji Itoh, and Hirotaka Aoki. Gaze typing compared with input by head and hand. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, pages 131–138, 2004.

[42] Shuo Samuel Liu, Andrew Rawicz, Teng Ma, Cheng Zhang, Kyle Lin, Siavash Rezaei, and Eion Wu. An eye-gaze tracking and human computer interface system for people with als and other locked-in diseases. *CMBES Proceedings*, 33, 2010.

[43] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative social impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumption 2, 3, and 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 4 and Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 5 and Appendix B.
   (b) Did you mention the license of the assets? [Yes] See Section 5.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]