Interaction-Grounded Learning with Action-inclusive Feedback

Tengyang Xie[*]
UIUC
tx10@illinois.edu

Akanksha Saran[*]
Microsoft Research, NYC
akanksha.saran@microsoft.com

Dylan J. Foster
Microsoft Research, New England
dylanfoster@microsoft.com

Lekan P. Molu
Microsoft Research, NYC
lekanmolu@microsoft.com

Ida Momennejad
Microsoft Research, NYC
idamo@microsoft.com

Nan Jiang
UIUC
nanjiang@illinois.edu

Paul Mineiro
Microsoft Research, NYC
pmineiro@microsoft.com

John Langford
Microsoft Research, NYC
jcl@microsoft.com

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., Grizou et al., 2014; Nguyen et al., 2021], including a recent formalization [Xie et al., 2021b] 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. [2021b] 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 [Katyal et al., 2014; Mishra and Gazzaley, 2015; Debettencourt et al., 2015; Muñoz-Moldes and Cleeremans, 2020; Akinola et al., 2020; Poole and Lee, 2021] and multimodal interactive feedback in HCI [Pantic and Rothkrantz, 2003; Vitense et al., 2003; Freeman et al., 2017; Mott et al., 2017; Duchowski, 2018; Noda, 2018; Saran et al., 2018, 2020; Cui et al., 2021; Gao et al.,

[*]equal contribution

Preprint. Under review.

2021], 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., Dann et al., 2018; Du et al., 2019; Misra et al., 2020], 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 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 [Bischl et al., 2021] and demonstrate the effectiveness of the proposed approach (Section 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 [Xie et al., 2021b], 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 $R(x, a) := \mathbb{E}_{x,a}[r]$ 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. [2021b] 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. [2021b] 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:

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

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 [Xie et al., 2021b].

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) $|\mathbb{E}_\mu\left[f^\star_{\bar{a}}(x, \bar{a})|\bar{a}, r = 1\right] - \mathbb{E}_\mu\left[f^\star_{\bar{a}}(x, \bar{a})|\bar{a}, r = 0\right]| \geq \Delta_{\mathcal{F}}$. 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

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

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.

Note that $\mathcal{L}$ is also similar to many popular contrastive losses [e.g., Wu et al., 2018; Chen et al., 2020] especially for the spectral contrastive loss [HaoChen et al., 2021].

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

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 $(f, \psi) \in \mathcal{F} \times \Psi$,

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

$f_{a,r}$ 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 similar intuition as the identifiability assumption of [Xie et al., 2021b, 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 $\rho_a^{\pi_{\mathsf{base}}}$, using the learned $\widehat{\psi}_a$, $\widehat{\rho}_a^{\pi^a_{\mathsf{base}}} = \frac{\sum_x d_0(x)\pi^a_{\mathsf{base}}(a|x)\widehat{\psi}_a(x,a)}{\sum_x d_0(x)\pi^a_{\mathsf{base}}(a|x)}$. If $\widehat{\psi}_a$ can efficiently decode the latent reward, then $\widehat{\rho}_a^{\pi^a_{\mathsf{base}}}$ converges to either $\rho_a^{\pi_{\mathsf{base}}}$ or $1 - \rho_a^{\pi_{\mathsf{base}}}$. 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 $R(x,a) = \pi^\star(a|x) \coloneqq \mathbb{1}(x = a)$. Then, we have, for any $\psi : \mathcal{Y} \to [0,1]$ (approach proposed by Xie et al. [2021b] 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 $\pi_1(a|x) \coloneqq \mathbb{1}(a = 1)$ for all $x \in \mathcal{X}$ and decoder $\psi_2(y) \coloneqq \mathbb{1}(y = 2)$ for all $y \in \mathcal{Y}$, then,

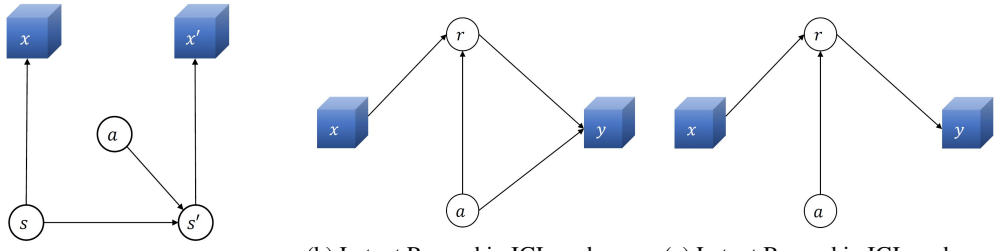
$$\begin{aligned}\mathcal{L}(\pi_1, \psi_2) &= \tfrac{1}{10}\textstyle\sum_{x=1}^{10}\sum_{a=1}^{10}\mathbb{1}(a=1)\psi_2((a+1)\%10) - \tfrac{1}{100}\sum_{x=1}^{10}\sum_{a=1}^{10}\psi_2((a+\mathbb{1}(x=a))\%10)\\ &= \psi_2(2) - \tfrac{1}{10}\textstyle\sum_{a=1}^{10}\psi_2(a) = 0.9 > \mathcal{L}(\pi^\star,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

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 [Misra et al., 2020]. 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.



(a) Latent State in Rich-Observation RL [Misra et al., 2020]

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

(c) Latent Reward in IGL under Assumption 1 [Xie et al., 2021b]

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

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:

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

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*

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

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 $\varepsilon_{\mathsf{CB}} \coloneqq \max_{\pi\in\Pi} \mathbb{E}_{d_0\times\pi}[r] - \mathbb{E}_{d_0\times\widehat{\pi}}[r] \leq o(1)$.

The notion in Definition 1 corresponds to the standard policy learning approaches in the contextual bandits literature [e.g., Langford and Zhang, 2007; Dudik et al., 2011; Agarwal et al., 2014], 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., Mohri et al., 2018]. 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,*

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

Similar to the performance of [Xie et al., 2021b], 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.

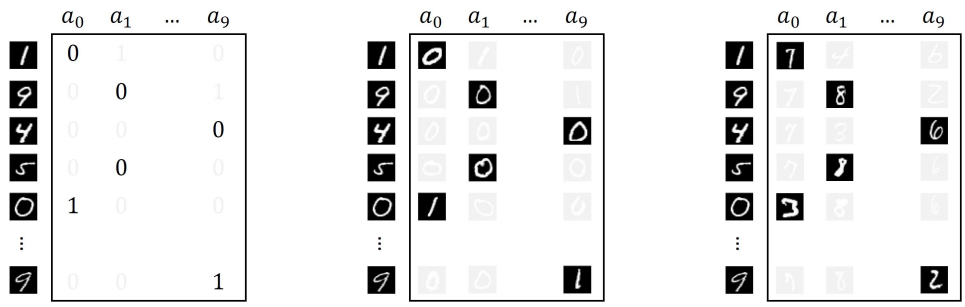
Lemma 4 (Breaking symmetry)**.** *Suppose Assumption 4 holds. For any $\bar{a} \in \mathcal{A}$, if we have $|\mathcal{D}| \geq \mathcal{O}\left(\frac{K^3 d_{\mathcal{F},\Psi}}{(\min\{\eta\Delta_{\mathcal{F}}, Kc_m\})^2}\right)$, then, with high probability, $\psi'_{\bar{a},1} \geq \psi'_{\bar{a},0}$.*

Combining these two lemmas above as well as the CB oracle establishes the proof of Theorem 2, and the 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. [2021b] 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



(a) Contextual Bandits (CB) (b) IGL (full CI) [Xie et al., 2021b] (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.

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 (Section 5.1) and low-dimensional feedback vectors for more than 200 OpenML CC-18 datasets (Section 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 uniformly 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 [LeCun et al., 1998] which is licensed under Attribution-Share Alike 3.0 license[2]. 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. [2021b] on the MNIST environment. This setting is similar to the one described in Section 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.

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

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 [Vanschoren et al., 2015; Casalicchio et al., 2019; Feurer et al., 2021; Bischl et al., 2021]. OpenML datasets are licensed under CC-BY license[3] and the platform and library are licensed under the BSD (3-Clause) license[4]. Similar to Section 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 and shown in Table 2.

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.

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 Torra et al. [2008]; Abdelmessih et al. [2010]; Reif et al. [2014]; Lorena et al. [2019]) for 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), and the K factors can be also possibly improved under some specific choice of function class (see discussion in Section 4). We then select N, $N/\sqrt{K}$ and 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 using 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 (Figure 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 [Lorena et al., 2019] (a measure of classification complexity that quantifies the informativeness of a given sample) to be the most predictive of relative performance (Figure 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 (Figure 5 and Figure 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 [Xie et al., 2021b], 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

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

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

naturally connects to latent state discovery in rich-observation reinforcement learning [e.g., Dann et al., 2018; Du et al., 2019; Misra et al., 2020]. 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 [Momennejad and Haynes, 2012, 2013; Belghazi et al., 2018; Shah and Peters, 2020]. Another example is a continually self-calibrating eye tracker, used by people with motor disabilities such as ALS [Hansen et al., 2004; Liu et al., 2010; Mott et al., 2017; Gao et al., 2021]. 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.

Acknowledgment

The authors thank Mark Rucker for sharing code to compute meta-properties of the OpenML Datasets. NJ acknowledges funding support from ARL Cooperative Agreement W911NF-17-2-0196, NSF IIS-2112471, NSF CAREER award, and Adobe Data Science Research Award.

References

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

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.

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.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 530–539. PMLR, 2018.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael Gomes Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. Openml benchmarking suites. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks*, 2021.

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.

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.

Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*, pages 604–626. PMLR, 2021.

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.

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.

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.

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

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.

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. *Journal of Machine Learning Research*, 22:1–5, 2021.

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

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. In *9th International Conference on Learning Representations, ICLR*, 2021.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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

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

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.

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.

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

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.

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

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

Benjamin Poole and Minwoo Lee. Towards intrinsic interactive reinforcement learning. *arXiv preprint arXiv:2112.01575*, 2021.

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.

Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8615–8621. IEEE, 2018.

Akanksha Saran, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Understanding teacher gaze patterns for robot learning. In *Conference on Robot Learning*, pages 1247–1258. PMLR, 2020.

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.

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

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.

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.

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.

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, 2021a.

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

Appendix

A Detailed Proofs

Proof of Proposition 1. Over this proof, we follow the definition of Eq.(3) for each $(f, \psi) \in \mathcal{F} \times \Psi$. For any π, let $d_a^\pi \coloneqq \sum_x d_0(x)\pi(a|x)$ and $\rho_a^\pi \coloneqq {}^1\!/_{d_a^\pi} \sum_x d_0(x)\pi(a|x)\Pr(r=1|x,a)$, then

$$\begin{aligned}
\mathbb{E}_{\mu_a}\left[f(x)\psi(y)\right] &= \frac{1}{d_a^\mu}\sum_{x,y} d_0(x)\mu(a|x)f(x)\Pr(y|a,r=0)\psi(y)\Pr(r=0|x,a) \\
&\quad + \frac{1}{d_a^\mu}\sum_{x,y} d_0(x)\mu(a|x)f(x)\Pr(y|a,r=1)\psi(y)\Pr(r=1|x,a) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by } x \perp\!\!\!\perp y|r,a) \\
&\overset{\text{(a)}}{=} (1-\rho_a^\mu)f_{a,0}\psi_{a,0} + \rho_a^\mu f_{a,1}\psi_{a,1},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{\mu_a}\left[f(x)\right]\mathbb{E}_{\mu_a}\left[\psi(y)\right] &= \left(\frac{1}{d_a^\mu}\sum_x d_0(x)\mu(a|x)f(x)\right)\left(\frac{1}{d_a^\mu}\sum_{x,y} d_0(x)\mu(a|x)\Pr(y|a)\psi(y)\right) \\
&\overset{\text{(b)}}{=} \left((1-\rho_a^\mu)f_{a,0} + \rho_a^\mu f_{a,1}\right)\left((1-\rho_a^\mu)\psi_{a,0} + \rho_a^\mu\psi_{a,1}\right) \\
&= (1-\rho_a^\mu)^2 f_{a,0}\psi_{a,0} + \left(\rho_a^\mu\right)^2 f_{a,1}\psi_{a,1} + (1-\rho_a^\mu)\rho_a^\mu(f_{a,0}\psi_{a,1} + f_{a,1}\psi_{a,0}).
\end{aligned}$$

To see why (a) and (b) hold, we use the following argument. By definition of $f_{a,r}$ and $\psi_{a,r}$, we have

$$\begin{aligned}
f_{a,0} &= \sum_x \Pr(x|a,r=0)f_a(x,a) \\
&= \sum_x \frac{\Pr(x,a,r=0)f_a(x,a)}{\Pr(a,r=0)} \\
&= \frac{\sum_x d_0(x)\mu(a|x)f_a(x,a)\Pr(r=0|x,a)}{\sum_x d_0(x)\mu(a|x)\Pr(r=0|x,a)}. \\
\implies d_a^\mu(1-\rho_a^\mu)f_{a,0} &= \sum_x d_0(x)\mu(a|x)f_a(x,a)\Pr(r=0|x,a).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\frac{1}{d_a^\mu}\sum_{x,y} d_0(x)\mu(a|x)f_a(x,a)\Pr(y|a,r=0)\psi_a(y,a)\Pr(r=0|x,a) \\
&= \frac{1}{d_a^\mu}\sum_x d_0(x)\mu(a|x)f_a(x,a)\Pr(r=0|x,a)\sum_y \Pr(y|a,r=0)\psi_a(y,a) \\
&= \frac{\psi_{a,0}}{d_a^\mu}\sum_x d_0(x)\mu(a|x)f_a(x,a)\Pr(r=0|x,a) \\
&= (1-\rho_a^\mu)f_{a,0}\psi_{a,0}.
\end{aligned}$$

Similar procedure also induces the remaining terms of (a) and (b).

Therefore, combining the two equalities above, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mu_a}\left[f(x)\psi(y)\right] - \mathbb{E}_{\mu_a}\left[f(x)\right]\mathbb{E}_{\mu_a}\left[\psi(y)\right] \\
&= (1-\rho_a^\mu)\rho_a^\mu(f_{a,0}\psi_{a,0} + f_{a,1}\psi_{a,1}) - (1-\rho_a^\mu)\rho_a^\mu(f_{a,0}\psi_{a,1} + f_{a,1}\psi_{a,0}) \\
&= (1-\rho_a^\mu)\rho_a^\mu(f_{a,0}\psi_{a,0} + f_{a,1}\psi_{a,1} - f_{a,0}\psi_{a,1} - f_{a,1}\psi_{a,0}) \\
&= (1-\rho_a^\mu)\rho_a^\mu(f_{a,1} - f_{a,0})(\psi_{a,1} - \psi_{a,0}). \qquad (6)
\end{aligned}$$

This completes the proof. □

Lower bound of $\Delta_{\mathcal{F}}$ when $\mu = \pi_{\mathsf{unif}}$. For any $f \in \mathcal{F}$, let $V_f(\pi) \coloneqq \sum_x \mathbb{E}_{d_0 \times \pi}[f(x,a)] \in [0,1]$. Then,

$$
\begin{aligned}
&\mathbb{E}_\mu\left[f(x,\bar{a})|\bar{a}, r=1\right] - \mathbb{E}_\mu\left[f(x,\bar{a})|\bar{a}, r=0\right] \\
&= \frac{\sum_x \Pr(x,\bar{a},r=1) f(x,\bar{a})}{\sum_x \Pr(x,\bar{a},r=1)} - \frac{\sum_x \Pr(x,\bar{a},r=0) f(x,\bar{a})}{\sum_x \Pr(x,\bar{a},r=0)} \\
&= \frac{\sum_x d_0(x)\mu(\bar{a}|x)R(x,\bar{a})f(x,\bar{a})}{\sum_x d_0(x)\mu(\bar{a}|x)R(x,\bar{a})} - \frac{\sum_x d_0(x)\mu(\bar{a}|x)f(x,\bar{a})(1-R(x,\bar{a}))}{\sum_x d_0(x)\mu(\bar{a}|x)(1-R(x,\bar{a}))} \\
&= \frac{\sum_x d_0(x)R(x,\bar{a})f(x,\bar{a})}{\sum_x d_0(x)R(x,\bar{a})} - \frac{\sum_x d_0(x)f(x,\bar{a})(1-R(x,\bar{a}))}{\sum_x d_0(x)(1-R(x,\bar{a}))} \qquad (\mu = \pi_{\mathsf{unif}}) \\
&= \frac{\sum_x d_0(x)f(x,\bar{a})R(x,\bar{a})}{V(\pi_{\bar{a}})} - \frac{\sum_x d_0(x)f(x,\bar{a})(1-R(x,\bar{a}))}{1-V(\pi_{\bar{a}})} \\
&\qquad\qquad (\pi_a \text{ denotes the constant policy with action } \bar{a}) \\
&= \frac{\sum_x d_0(x)f(x,\bar{a})R(x,\bar{a}) - V_f(\pi_{\bar{a}})V(\pi_{\bar{a}})}{V(\pi_{\bar{a}})} \\
&\quad - \frac{\sum_x d_0(x)f(x,\bar{a})(1-R(x,\bar{a})) - V_f(\pi_{\bar{a}})(1-V(\pi_{\bar{a}}))}{1-V(\pi_{\bar{a}})} \\
&= \frac{\mathrm{Cov}_{\pi_{\bar{a}}}(f,R)}{V(\pi_{\bar{a}})} - \frac{\mathrm{Cov}_{\pi_{\bar{a}}}(f,1-R)}{1-V(\pi_{\bar{a}})} \\
&\overset{\text{(a)}}{=} \frac{\mathrm{Cov}_{\pi_{\bar{a}}}(f,R)}{V(\pi_{\bar{a}})} + \frac{\mathrm{Cov}_{\pi_{\bar{a}}}(f,R)}{1-V(\pi_{\bar{a}})} \\
&= \frac{\mathrm{Cov}_{\pi_{\bar{a}}}(f,R)}{V(\pi_{\bar{a}})(1-V(\pi_{\bar{a}}))} \\
&\implies \left|\mathbb{E}_\mu\left[f(x,\bar{a})|\bar{a}, r=1\right] - \mathbb{E}_\mu\left[f(x,\bar{a})|\,\bar{a}, r=0\right]\right| \geq 4|\mathrm{Cov}_{\pi_{\bar{a}}}(f,R)|,
\end{aligned}
$$

where (a) follows from

$$
\begin{aligned}
\mathrm{Cov}_{\pi_{\bar{a}}}(f, 1-R) &= \mathbb{E}_{\pi_{\bar{a}}}[(f(x,a) - V_f(\pi_{\bar{a}}))(1 - R(x,a) - 1 + V(\pi_{\bar{a}}))] \\
&= -\mathbb{E}_{\pi_{\bar{a}}}[(f(x,a) - V_f(\pi_{\bar{a}}))(R(x,a) - V(\pi_{\bar{a}}))] \\
&= -\mathrm{Cov}_{\pi_{\bar{a}}}(f,R).
\end{aligned}
$$

This completes the proof. □

Proof of Lemma 3. Let,

$$
(\widetilde{f}_{\bar{a}}, \widetilde{\psi}_{\bar{a}}) \leftarrow \operatorname*{argmax}_{(f,\psi)\in\mathcal{F}\times\Psi} \mathcal{L}_{\bar{a},\mu}(f,\psi).
$$

Over this proof, we define

$$
\Delta_{\mathcal{F}} = \max_{f\in\mathcal{F}}(f_{a,1} - f_{a,0}).
$$

Now, for $\mathcal{L}_{\bar{a},\mathcal{D}}(f,\psi)$ with any (f,ψ), we also have with probability at least $1-\delta$

$$
|\mathcal{L}_{\bar{a},\mathcal{D}}(f,\psi) - \mathcal{L}_{\bar{a},\mu}(f,\psi)| \leq \varepsilon_{\mathsf{stat},\bar{a}}.
$$

This means

$$
\begin{aligned}
&\mathcal{L}_{\bar{a},\mathcal{D}}(\widetilde{f}_{\bar{a}}, \widetilde{\psi}_{\bar{a}}) \leq \mathcal{L}_{\bar{a},\mathcal{D}}(f_{\bar{a}}, \psi_{\bar{a}}) \\
&\implies \mathcal{L}_{\bar{a},\mu}(\widetilde{f}_{\bar{a}}, \widetilde{\psi}_{\bar{a}}) - 2\varepsilon_{\mathsf{stat},\bar{a}} \leq \mathcal{L}_{\bar{a},\mu}(f_{\bar{a}}, \psi_{\bar{a}}) \\
&\implies \underbrace{(1-\rho_a^\mu)\rho_a^\mu}_{=K-1/K^2,\ \text{as } \mu=\pi_{\mathsf{unif}}} \cdot(\widetilde{f}_{\bar{a},1} - \widetilde{f}_{\bar{a},0})(\widetilde{\psi}_{\bar{a},1} - \widetilde{\psi}_{\bar{a},0}) - 2\varepsilon_{\mathsf{stat},\bar{a}} \leq (1-\rho_a^\mu)\rho_a^\mu(f_{\bar{a},1} - f_{\bar{a},0})(\psi_{\bar{a},1} - \psi_{\bar{a},0}). \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Eq.(6))} \\
&\implies (f_{\bar{a},1} - f_{\bar{a},0})(\psi_{\bar{a},1} - \psi_{\bar{a},0}) \geq (\widetilde{f}_{\bar{a},1} - \widetilde{f}_{\bar{a},0})(\widetilde{\psi}_{\bar{a},1} - \widetilde{\psi}_{\bar{a},0}) - \frac{2K^2}{(K-1)}\varepsilon_{\mathsf{stat},\bar{a}}.
\end{aligned}
$$

This implies

$$|f_{\bar{a},1} - f_{\bar{a},0}| \geq \Delta_{\mathcal{F}} - \frac{2K^2}{(K-1)}\varepsilon_{\mathsf{stat},\bar{a}}$$
$$|\psi_{\bar{a},1} - \psi_{\bar{a},0}| \geq 1 - \frac{2K^2}{\Delta_{\mathcal{F}}(K-1)}\varepsilon_{\mathsf{stat},\bar{a}}$$
$$\implies |\psi^{\star}_{\bar{a},1} - \max\{\psi_{\bar{a},1}, \psi_{\bar{a},0}\}|, |\min\{\psi_{\bar{a},1}, \psi_{\bar{a},0}\} - \psi^{\star}_{\bar{a},0}| \leq \frac{2K^2}{\Delta_{\mathcal{F}}(K-1)}\varepsilon_{\mathsf{stat},\bar{a}}. \tag{7}$$

This completes the proof. □

Proof of Lemma 4. We now provide the proof for any fixed $\bar{a} \in \mathcal{A}$. We define $d^{\pi}_{\bar{a}} := \sum_x d_0(x)\pi(\bar{a}|x)$, and $R_{\psi}(x,a) = \mathbb{E}[\psi(y,a)|x,a]$ as the decoded reward by ψ for all $(x,a,\psi) \in \mathcal{X} \times \mathcal{A} \times \Psi$. Also, let $\rho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}}$, $\widehat{\rho}^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}}$, $\varrho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}}$, $\widehat{\varrho}^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}}$ be,

$$\rho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} = \frac{\overbrace{\sum_x d_0(x)\pi^{\bar{a}}_{\mathsf{base}}(\bar{a}|x)R(x,\bar{a})}^{=:\text{(I).pop}}}{\underbrace{\sum_x d_0(x)\pi^{\bar{a}}_{\mathsf{base}}(\bar{a}|x)}_{=:\text{(II).pop}}}, \qquad \widehat{\rho}^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} = \frac{\overbrace{\sum_{(x,a,y)\sim\mathcal{D}} \pi^{\bar{a}}_{\mathsf{base}}(a|x)R(x,a)\mathbb{1}(a=\bar{a})}^{=:\sum_{(x,a,y)\sim\mathcal{D}}\mathbb{1}(a=\bar{a})\cdot\text{(I).emp}}}{\underbrace{\sum_{(x,a,y)\sim\mathcal{D}} \pi^{\bar{a}}_{\mathsf{base}}(a|x)\mathbb{1}(a=\bar{a})}_{=:\sum_{(x,a,y)\sim\mathcal{D}}\mathbb{1}(a=\bar{a})\cdot\text{(II).emp}}} \tag{8}$$

$$\varrho^{\pi_{\mathsf{base}}}_{\bar{a}} = \frac{\overbrace{\sum_x d_0(x)\pi^{\bar{a}}_{\mathsf{base}}(\bar{a}|x)R_{\psi_{\bar{a}}}(x,\bar{a})}^{:=\text{(III).pop}}}{\sum_x d_0(x)\pi^{\bar{a}}_{\mathsf{base}}(\bar{a}|x)}, \qquad \widehat{\varrho}^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} = \frac{\overbrace{\sum_{(x,a,y)\sim\mathcal{D}} \pi^{\bar{a}}_{\mathsf{base}}(a|x)\psi_a(y,a)\mathbb{1}(a=\bar{a})}^{=:\sum_{(x,a,y)\sim\mathcal{D}}\mathbb{1}(a=\bar{a})\cdot\text{(III).emp}}}{\sum_{(x,a,y)\sim\mathcal{D}} \pi^{\bar{a}}_{\mathsf{base}}(a|x)\mathbb{1}(a=\bar{a})} \tag{9}$$

Over this section, we define $\varepsilon_{\mathsf{stat},\bar{a}}$ as,

$$\varepsilon_{\mathsf{stat},\bar{a}} := \mathcal{O}\left(\sqrt{\frac{Kd_{\mathcal{F},\Psi}}{|\mathcal{D}|}}\right) \geq \mathcal{O}\left(\sqrt{\frac{d_{\mathcal{F},\Psi}}{\sum_{(x,a,y)\sim\mathcal{D}}\mathbb{1}(a=\bar{a})}}\right).$$

(by [Xie et al., 2021a, Lemma A.1])

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

$$|\text{(I).pop} - \text{(I).emp}|, |\text{(II).pop} - \text{(II).emp}|, |\text{(III).pop} - \text{(III).emp}| \leq \varepsilon_{\mathsf{stat},\bar{a}} \tag{10}$$

by a standard concentration argument.

We now show when line 5-6 in Algorithm 1 correctly break the symmetry. By Assumption 4, we have $\rho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} \leq \frac{1}{2} - \eta$. In addition, by Eq.(10),

$$\left|\widehat{\rho}^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} - \rho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}}\right|, \left|\widehat{\varrho}^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} - \varrho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}}\right| \geq 2\varepsilon_{\mathsf{stat},\bar{a}} + \frac{2\varepsilon^2_{\mathsf{stat},\bar{a}}}{c_m - \varepsilon_{\mathsf{stat},\bar{a}}}. \tag{11}$$

Recall the definition of $\rho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}}$ and $\varrho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}}$ (Eq.(8) and Eq.(9)), we have

$$\rho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} - \varrho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} = \frac{\sum_x d_0(x)\pi^{\bar{a}}_{\mathsf{base}}(\bar{a}|x)\mathbb{E}[\psi^{\star}(y,\bar{a}) - \psi_{\bar{a}}(y,\bar{a})|x,\bar{a}]}{\sum_x d_0(x)\pi^{\bar{a}}_{\mathsf{base}}(\bar{a}|x)R_{\psi_{\bar{a}}}(x,\bar{a})}.$$

According to Eq.(7), if $\psi_{\bar{a},1} > \psi_{\bar{a},0}$, we have

$$\varrho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} \leq \rho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} + \frac{2K^2}{(K-1)}\varepsilon_{\mathsf{stat},\bar{a}}$$
$$\implies \widehat{\varrho}^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} \leq \rho^{\pi^{\bar{a}}_{\mathsf{base}}}_{\bar{a}} + 2\varepsilon_{\mathsf{stat},\bar{a}} + \frac{2\varepsilon^2_{\mathsf{stat},\bar{a}}}{c_m - \varepsilon_{\mathsf{stat},\bar{a}}}, \qquad \text{(by Eq.(11))}$$

otherwise,

$$\rho_{\bar{a}}^{\pi_{\mathsf{base}}^{\bar{a}}} - \varrho_{\bar{a}}^{\pi_{\mathsf{base}}^{\bar{a}}} = 1 - \frac{\sum_x d_0(x)\pi_{\mathsf{base}}(\bar{a}|x)\mathbb{E}[1-\psi_{\bar{a}}(y,\bar{a})-\psi^\star(y,\bar{a})|x,\bar{a}]}{\sum_x d_0(x)\pi_{\mathsf{base}}(\bar{a}|x)R_{\psi_{\bar{a}}}(x,\bar{a})}$$

$$\implies \varrho_{\bar{a}}^{\pi_{\mathsf{base}}} \geq 1 - \rho_{\bar{a}}^{\pi_{\mathsf{base}}} - \frac{2K^2}{\Delta_{\mathcal{F}}(K-1)}\varepsilon_{\mathsf{stat},\bar{a}}$$

$$\implies \widehat{\varrho}_{\bar{a}}^{\pi_{\mathsf{base}}^{\bar{a}}} \geq 1 - \rho_{\bar{a}}^{\pi_{\mathsf{base}}^{\bar{a}}} - \frac{2K^2}{\Delta_{\mathcal{F}}(K-1)}\varepsilon_{\mathsf{stat},\bar{a}} - 2\varepsilon_{\mathsf{stat},\bar{a}} - \frac{2\varepsilon_{\mathsf{stat},\bar{a}}^2}{c_m - \varepsilon_{\mathsf{stat},\bar{a}}}. \qquad \text{(by Eq.(11))}$$

To guarantee the correctness of the symmetry breaking step (line 5-6 in Algorithm 1), we need

1. If $f_{\bar{a},1} > f_{\bar{a},0}$, $\widehat{\varrho}_{\bar{a}}^{\pi_{\mathsf{base}}^{\bar{a}}} \leq \frac{1}{2}$,
2. Otherwise, $\widehat{\varrho}_{\bar{a}}^{\pi_{\mathsf{base}}^{\bar{a}}} > \frac{1}{2}$.

That requires,

$$\frac{2K^2}{\Delta_{\mathcal{F}}(K-1)}\varepsilon_{\mathsf{stat},\bar{a}} + 2\varepsilon_{\mathsf{stat},\bar{a}} + \frac{2\varepsilon_{\mathsf{stat},\bar{a}}^2}{c_m - \varepsilon_{\mathsf{stat},\bar{a}}} \leq \eta.$$

This can be induced by

$$\varepsilon_{\mathsf{stat},\bar{a}} \leq \mathcal{O}\left(\min\left\{\eta\Delta_{\mathcal{F}}/K, c_m\right\}\right) \quad \left(\Longleftarrow \quad |\mathcal{D}| \geq \mathcal{O}\left(\frac{K^3 d_{\mathcal{F},\Psi}}{(\min\{\eta\Delta_{\mathcal{F}}, Kc_m\})^2}\right)\right). \qquad \square$$

Proof of Theorem 2. Combining Lemma 3 and Lemma 4, we know for any $\bar{a} \in \mathcal{A}$

$$|\psi_{\bar{a},1}^\star - \psi_{\bar{a},1}|, |\psi_{\bar{a},0} - \psi_{\bar{a},0}^\star| \leq \frac{2K^2}{\Delta_{\mathcal{F}}(K-1)}\varepsilon_{\mathsf{stat},\bar{a}}.$$

Then,

$$\begin{aligned}
&V(\pi^\star) - V(\widehat{\pi}) \\
&= \mathbb{E}_{d_0\times\pi^\star}[r] - \mathbb{E}_{d_0\times\widehat{\pi}}[r] \\
&= \mathbb{E}_{(x,a,y)\sim d_0\times\pi^\star}[\psi^\star(y,a)] - \mathbb{E}_{(x,a,y)\sim d_0\times\widehat{\pi}}[\psi^\star(y,a)] && \text{(by Assumption 3)} \\
&= \mathbb{E}_{(x,a,y)\sim d_0\times\pi^\star}[\psi_a'(y,a)] - \mathbb{E}_{(x,a,y)\sim d_0\times\widehat{\pi}}[\psi_a'(y,a)] \\
&\quad + \left|\mathbb{E}_{(x,a,y)\sim d_0\times\pi^\star}[\psi^\star(y,a) - \psi_a'(y,a)]\right| + \left|\mathbb{E}_{(x,a,y)\sim d_0\times\widehat{\pi}}[\psi^\star(y,a) - \psi_a'(y,a)]\right| \\
&\leq \mathbb{E}_{(x,a,y)\sim d_0\times\pi^\star}[\psi_a'(y,a)] - \mathbb{E}_{(x,a,y)\sim d_0\times\widehat{\pi}}[\psi_a'(y,a)] \\
&\quad + (V(\pi^\star) + V(\widehat{\pi}))\max_{a\in\mathcal{A}}|\psi_{\bar{a},1}^\star - \psi_{\bar{a},1}| + (2 - V(\pi^\star) - V(\widehat{\pi}))\max_{a\in\mathcal{A}}|\psi_{\bar{a},0} - \psi_{\bar{a},1}^\star| \\
&\leq \mathbb{E}_{(x,a,y)\sim d_0\times\pi^\star}[\psi_a'(y,a)] - \mathbb{E}_{(x,a,y)\sim d_0\times\pi^\star}[\psi_a'(y,a)] + \frac{4K^2}{\Delta_{\mathcal{F}}(K-1)}\varepsilon_{\mathsf{stat},\bar{a}} \\
&\leq \varepsilon_{\mathsf{CB}} + \frac{4K^3}{\Delta_{\mathcal{F}}(K-1)}\varepsilon_{\mathsf{stat},\bar{a}}. && \text{(by the property of CB oracle in Definition 1)}
\end{aligned}$$

This completes the proof. $\square$

B Additional Experimental Details

This section provides additional details on our implementation. The experiments on the MNIST dataset were conducted using a Google Colab, which was based on Intel Xeon CPU (2.30GHz) and 12 GB memory. Large scale experiments on the OpenML datasets were conducted on CPU instances of an internal cluster. No GPU was used. The prototype codes were built over Python, PyTorch, and Vowpal Wabbit. With the single process in the setup above, a single trial of each experiment took less than 30 minutes to finish. Each experiment runs for 10 epochs over the dataset it is being evaluated on. The data is shuffled only once before the training begins so each algorithms views the data in the same order in each epoch.

B.1 Results for Individual OpenML datasets

For each individual OpenML dataset, we report the mean performance and standard errors for the best contact action policy, contextual bandit (CB), IGL assuming conditional independence of feedback on context and action given the latent reward (full CI) [Xie et al., 2021b], and IGL with conditional independence of feedback on context given reward (AI-IGL) in Table 3. The best constant action policy is computed based on the highest percentage of samples in the entire dataset with the same action. We do not report the standard error for it.

Table 3: Average accuracy and standard errors for different algorithms on individual OpenML datasets.

Dataset ID	N	K	Best Constant Action Policy (%)	CB (%)	IGL (full CI) (%)	AI-IGL (%)
6	20000	26	4.07	53.90±0.53	3.54±0.18	11.40±1.45
7	226	24	25.22	34.35±2.70	5.65±1.69	8.26±2.26
9	205	6	32.68	42.38±2.56	16.90±2.98	20.95±3.25
11	625	3	46.08	85.79±1.07	32.38±4.33	25.00±6.60
12	2000	10	10.00	87.10±0.79	10.17±1.04	32.17±4.25
14	2000	10	10.00	67.75±0.77	10.98±0.68	20.47±3.15
16	2000	10	10.00	87.37±0.42	9.85±0.73	48.52±4.34
18	2000	10	10.00	59.13±0.94	9.25±0.71	26.48±3.48
20	2000	10	10.00	92.25±0.49	9.82±0.75	43.35±4.35
22	2000	10	10.00	71.55±0.88	9.88±0.81	23.85±3.64
23	1473	3	42.70	48.85±0.92	38.61±1.72	31.79±1.66
26	12960	5	33.33	89.86±0.23	19.66±3.31	77.34±2.83
28	5620	10	10.18	93.99±0.20	10.30±0.75	66.49±4.47
32	10992	10	10.41	88.10±0.22	9.62±0.59	74.47±3.01
35	366	6	30.60	95.00±0.82	12.16±1.62	46.62±6.06
36	2310	7	14.29	87.27±0.64	13.35±1.00	63.35±4.70
39	336	8	42.56	71.76±2.20	11.47±2.69	23.53±4.86
41	214	6	35.51	49.55±2.87	14.55±2.96	22.05±3.68
42	683	19	13.47	60.22±1.89	4.57±0.70	20.36±2.46
48	151	3	34.44	42.19±3.03	27.19±2.09	31.25±2.30
54	846	4	25.77	50.12±1.56	23.88±0.91	28.94±2.19
60	5000	3	33.84	86.02±0.27	33.85±0.48	68.82±4.89
61	150	3	33.33	89.00±1.90	33.33±2.87	37.67±6.67
62	101	7	40.59	75.00±4.40	10.91±3.99	23.18±5.32
74	1000000	26	4.08	41.96±0.30	3.87±0.59	33.08±4.14
75	1000000	7	32.33	60.04±0.09	15.91±3.47	58.35±0.28
78	1000000	10	10.06	83.61±0.05	10.07±0.12	80.28±0.07
115	1000000	10	10.04	93.85±0.03	9.81±0.99	92.96±0.05
116	1000000	6	42.27	67.41±0.06	16.20±1.90	66.11±0.06
117	1000000	6	42.27	66.79±0.08	19.84±5.02	65.11±0.10
118	1000000	10	10.04	76.10±0.07	9.58±0.82	72.91±0.02
119	55296	3	42.78	48.76±0.17	33.59±1.75	40.31±1.91
123	1000000	10	10.17	93.02±0.04	11.90±1.77	90.73±0.02
127	1000000	10	10.46	80.56±0.06	9.48±2.12	75.76±0.14
129	1000000	6	30.46	97.47±0.01	11.11±2.06	96.54±0.01
130	1000000	7	14.36	80.83±0.08	14.63±1.01	75.56±0.27
133	137781	7	35.04	59.08±0.14	15.86±2.34	47.74±3.06
134	1000000	19	13.33	88.33±0.05	5.64±1.05	83.17±0.04
141	1000000	4	25.81	58.92±0.03	22.28±1.56	58.59±0.17
147	1000000	3	33.78	83.32±0.10	33.22±0.14	82.46±0.07
148	1000000	7	39.62	93.46±0.05	16.52±2.19	92.25±0.01
149	1455525	10	44.97	61.58±0.07	4.11±1.93	46.16±6.88
150	581012	7	48.76	71.13±0.06	8.42±2.06	51.13±4.08
154	1000000	10	10.08	54.97±0.10	9.88±0.41	53.95±0.12
156	1000000	5	30.01	45.51±0.25	25.14±2.56	42.81±0.16
157	1000000	5	30.01	47.69±0.19	22.62±1.77	46.54±0.36
158	1000000	5	30.01	47.84±0.16	22.23±2.11	46.98±0.07
159	1000000	5	30.01	29.89±0.09	18.77±1.09	12.02±2.21
160	1000000	5	30.01	30.52±0.06	16.17±1.79	13.51±1.70
163	32	3	40.62	38.75±4.50	25.00±4.33	35.00±5.70
171	339	21	24.78	23.82±2.13	3.53±1.11	7.21±2.14
180	110393	7	46.82	64.34±0.11	10.34±1.86	40.44±3.94
181	1484	10	31.20	47.05±1.22	6.31±1.17	21.07±2.09
182	6430	6	23.81	82.20±0.36	18.99±1.40	76.87±0.88
183	4177	28	16.50	19.38±0.50	2.79±0.73	6.60±1.40
184	28056	18	16.23	23.94±0.32	5.71±0.63	6.95±0.83
187	178	3	39.89	93.06±1.41	30.56±2.65	37.50±6.65
188	736	5	29.08	49.59±1.06	18.85±1.46	21.96±3.81
247	1000000	26	4.08	42.91±0.28	3.08±0.48	26.60±1.03
248	1000000	7	32.36	54.00±0.06	8.53±3.26	51.59±0.33
250	1000000	10	10.05	82.52±0.05	9.81±0.33	78.84±0.10
252	1000000	10	10.04	96.43±0.01	9.63±0.59	95.62±0.04
253	1000000	6	42.31	65.27±0.05	12.10±1.77	50.18±11.17
254	1000000	10	10.03	74.63±0.04	11.11±1.23	72.07±0.09

Table 3: Average accuracy and standard errors for different algorithms on individual OpenML datasets.

Dataset ID	N	K	Best Constant Action Policy (%)	CB (%)	IGL (full CI) (%)	AI-IGL (%)
255	55296	3	42.62	52.14±0.17	32.04±2.04	41.94±1.93
261	1000000	10	10.45	77.09±0.07	9.23±1.58	73.85±0.14
263	1000000	6	30.46	97.50±0.01	17.02±2.75	96.54±0.01
265	137781	7	35.40	53.82±0.14	15.45±2.60	38.90±2.25
268	1000000	4	25.75	57.85±0.11	24.47±0.60	57.02±0.16
271	1000000	3	33.84	88.70±0.04	33.14±0.07	87.99±0.12
272	1000000	7	39.65	93.41±0.05	17.01±3.15	91.99±0.05
285	194	8	30.93	31.00±2.36	12.25±2.25	19.75±2.68
300	7797	26	3.85	80.51±0.46	3.44±0.27	30.18±2.08
307	990	11	9.09	32.63±1.06	10.20±1.05	11.67±1.24
313	531	48	10.36	8.52±1.08	1.20±0.38	2.31±0.63
327	105	6	41.90	51.82±2.41	20.45±3.33	20.91±4.12
328	105	6	41.90	50.45±3.77	13.64±3.55	25.45±3.61
329	160	3	40.62	47.50±2.00	29.06±2.91	32.50±3.94
338	155	4	31.61	41.88±2.03	20.62±2.30	21.25±2.44
339	36	3	33.33	78.75±4.06	30.00±3.35	31.25±4.96
340	52	3	44.23	49.17±5.33	31.67±2.86	35.00±5.26
342	52	3	46.15	69.17±3.78	25.00±4.93	37.50±5.12
372	10108	46	28.47	26.67±0.59	1.38±0.21	12.17±1.05
375	9961	9	16.20	90.36±0.25	10.85±0.89	72.55±3.19
377	600	6	16.67	64.58±2.41	12.42±1.62	21.75±3.10
382	7019	8	27.61	36.33±0.69	10.42±0.72	20.98±1.52
383	690	10	23.19	70.94±1.61	10.43±1.17	25.00±3.73
385	927	7	37.97	90.16±1.16	15.48±2.25	27.31±4.33
386	913	10	17.20	56.79±1.08	10.76±0.99	14.67±2.17
387	414	9	31.88	70.48±2.09	7.62±1.57	19.52±2.89
388	204	6	44.61	67.62±2.93	18.10±3.13	29.52±5.76
389	2463	17	20.54	65.22±0.75	6.09±0.90	24.64±2.33
390	9558	44	7.28	47.65±0.47	2.25±0.11	15.37±0.97
391	1504	13	40.43	61.95±1.10	4.64±0.97	19.74±3.08
392	1003	10	19.34	66.34±1.14	8.61±0.58	21.78±2.74
393	3075	6	29.43	83.07±0.49	15.83±1.36	37.74±4.94
394	918	10	16.23	61.74±1.13	9.89±0.89	14.78±2.23
395	1657	25	22.39	47.74±0.83	2.35±0.31	13.40±2.00
396	3204	6	29.43	81.59±0.49	15.67±1.49	35.42±4.69
397	313	8	29.71	61.72±1.70	14.69±1.98	25.94±3.34
398	1560	20	21.86	55.80±1.31	3.27±0.58	18.17±2.81
399	11162	10	14.52	65.26±0.30	10.44±0.39	33.83±2.58
400	878	10	27.68	79.43±1.16	7.78±1.03	23.41±4.39
401	1050	10	15.71	59.90±1.03	9.05±0.86	15.52±2.27
452	285	7	41.40	35.69±2.32	16.38±2.70	18.10±2.42
457	27	4	44.44	51.67±6.86	30.00±7.42	23.33±5.82
458	841	4	37.69	99.59±0.15	23.35±2.67	59.82±8.04
460	379	4	37.20	50.53±1.64	21.71±1.74	30.13±2.37
468	72	6	16.67	48.75±4.83	20.62±3.22	18.12±4.28
469	797	6	19.45	19.31±0.91	16.75±1.01	16.88±1.07
473	2796	6	24.32	63.55±1.21	15.73±1.30	24.27±3.77
475	400	4	25.00	33.25±1.67	24.38±1.73	26.38±2.12
554	70000	10	11.25	89.92±0.10	9.98±0.29	84.17±0.12
679	1024	4	39.45	40.24±1.66	27.67±2.36	23.98±3.03
685	130	5	20.00	33.85±3.00	20.38±2.68	20.38±3.14
694	310	9	13.23	28.23±1.89	8.71±1.71	10.81±2.25
952	214	6	35.51	50.91±2.66	16.14±3.36	20.91±3.19
1041	3468	10	11.04	82.03±0.63	9.60±0.57	43.08±3.18
1044	10936	3	38.97	50.41±0.37	32.98±1.29	43.37±1.46
1079	95	5	28.42	46.50±5.49	17.50±2.90	16.50±3.62
1080	113	5	45.13	32.50±2.88	15.00±3.36	17.50±2.42
1081	89	4	48.31	38.33±3.89	23.33±2.93	23.89±4.03
1083	214	7	32.24	28.41±3.20	14.32±2.52	19.32±3.08
1088	383	9	40.47	47.56±3.16	10.00±1.88	12.56±2.23
1102	96	9	47.92	60.50±3.35	17.50±4.29	20.00±4.06
1106	190	14	15.79	22.63±2.91	7.89±1.21	6.05±1.36
1109	96	11	23.96	43.50±4.08	11.00±1.99	19.00±4.06
1115	151	3	34.44	47.19±2.71	28.12±2.14	29.69±1.82
1177	1000000	22	24.07	49.84±0.16	6.52±2.14	41.14±0.99
1183	1000000	6	23.84	80.99±0.13	23.14±3.79	76.32±0.21
1185	1000000	3	40.11	92.69±0.04	33.58±2.51	92.60±0.02
1186	1000000	5	28.98	46.57±0.17	17.83±2.19	46.59±0.12
1209	1000000	11	9.14	30.53±0.22	9.14±0.55	27.91±1.06
1214	1000000	9	16.08	77.46±0.09	8.53±0.90	74.37±0.35
1233	945	7	14.81	35.58±1.48	14.42±0.65	16.11±1.30
1378	1000000	26	4.08	34.15±0.36	3.98±0.44	20.64±4.44
1379	1000000	26	4.08	20.26±0.57	3.64±0.19	5.09±0.38
1380	1000000	26	4.08	13.97±0.46	3.79±0.16	4.01±1.18
1381	1000000	26	4.08	39.52±0.34	3.85±0.40	26.77±6.91
1382	1000000	26	4.08	34.60±0.34	4.02±0.43	26.70±2.34

Table 3: Average accuracy and standard errors for different algorithms on individual OpenML datasets.

Dataset ID	N	K	Best Constant Action Policy (%)	CB (%)	IGL (full CI) (%)	AI-IGL (%)
1383	1000000	26	4.08	29.49±0.20	4.08±0.28	16.45±5.71
1384	1000000	26	4.08	41.35±0.20	3.86±0.14	22.29±6.30
1385	1000000	26	4.08	37.62±0.14	3.92±0.50	21.30±5.19
1386	1000000	26	4.08	33.98±0.28	4.43±0.22	30.04±0.80
1387	1000000	24	24.14	74.78±0.05	2.85±1.01	66.79±0.07
1388	1000000	24	24.14	55.58±0.05	4.55±0.95	49.54±0.88
1389	1000000	24	24.14	44.84±0.11	3.43±0.50	34.55±1.30
1390	1000000	24	24.14	77.74±0.11	1.94±0.42	71.71±0.14
1391	1000000	24	24.14	74.80±0.05	2.80±0.79	67.01±0.31
1392	1000000	24	24.14	69.23±0.07	6.82±2.16	62.90±0.54
1393	1000000	7	32.36	46.65±0.12	20.86±4.63	36.27±1.42
1394	1000000	7	32.36	35.02±0.05	15.67±3.38	22.66±2.35
1395	1000000	7	32.36	33.36±0.07	7.12±2.60	4.77±2.63
1396	1000000	7	32.36	55.51±0.20	17.11±2.66	53.13±0.18
1397	1000000	7	32.36	46.56±0.13	14.76±4.15	37.56±1.30
1398	1000000	7	32.36	37.47±0.10	8.33±2.70	21.45±1.93
1399	1000000	7	32.36	53.28±0.09	12.66±2.06	49.73±0.46
1400	549796	7	32.43	53.06±0.05	20.55±2.25	46.79±0.41
1401	1000000	7	32.36	46.72±0.11	11.69±2.61	37.31±0.35
1413	150	3	33.33	88.67±2.63	31.00±2.22	35.33±7.57
1457	1500	50	2.00	11.60±0.64	1.90±0.32	2.30±0.31
1459	10218	10	13.86	32.69±0.39	8.52±0.57	9.77±1.22
1465	106	6	20.75	24.55±3.02	13.64±1.76	14.55±2.60
1466	2126	10	27.23	99.77±0.08	7.89±1.31	62.93±5.20
1468	1080	9	11.11	87.41±0.84	10.32±0.85	41.94±3.52
1472	768	37	9.64	13.90±1.14	3.83±0.73	2.73±0.78
1475	6118	6	41.75	46.27±0.45	17.03±1.63	24.70±2.06
1476	13910	6	21.63	94.80±0.25	17.51±1.00	87.76±1.11
1477	13910	6	21.63	96.46±0.18	15.03±1.34	84.52±1.91
1478	10299	6	18.88	95.02±0.29	17.29±1.27	91.06±2.23
1481	28056	18	16.23	23.34±0.31	5.33±0.61	7.39±1.05
1482	340	30	4.71	10.59±1.05	3.24±0.65	3.24±0.62
1483	164860	11	33.05	38.96±0.18	7.54±1.30	17.07±1.56
1491	1600	100	1.00	6.81±0.50	0.88±0.14	1.66±0.24
1492	1600	100	1.00	2.47±0.30	0.75±0.18	1.06±0.24
1493	1599	100	1.00	6.31±0.36	0.88±0.17	1.28±0.24
1497	5456	4	40.41	66.40±0.52	24.53±3.06	44.01±4.83
1499	210	3	33.33	70.95±3.87	32.62±1.66	25.95±4.74
1500	210	3	33.33	75.71±2.84	32.86±1.90	27.86±4.63
1501	1593	10	10.17	80.28±0.62	9.12±0.63	30.31±3.21
1503	263256	10	10.06	10.13±0.03	10.03±0.05	10.03±0.05
1508	403	5	32.01	67.68±1.44	19.39±2.34	24.63±3.92
1509	149332	22	14.73	23.68±0.41	5.07±0.67	7.76±0.43
1512	200	5	28.00	25.00±1.66	20.75±2.07	21.00±1.82
1513	123	5	39.02	31.54±3.12	19.23±3.24	26.92±3.46
1514	360	10	10.00	65.56±2.37	8.33±1.08	24.72±3.01
1515	571	20	10.51	18.28±1.29	6.29±0.73	7.07±1.21
1516	88	4	38.64	58.89±4.24	21.11±3.13	16.11±3.47
1517	47	5	42.55	31.00±4.12	17.00±3.54	18.00±5.08
1518	47	4	42.55	39.00±4.79	13.00±2.56	21.00±4.58
1520	164	5	28.66	42.94±2.28	19.41±2.73	22.35±4.27
1523	215	3	46.05	82.95±1.84	32.95±2.18	37.50±7.14
1525	5456	4	40.41	74.57±0.56	23.72±3.21	39.73±4.93
1548	2500	3	46.92	47.08±0.86	35.46±3.66	36.28±1.41
1549	742	8	22.10	18.20±0.99	11.87±0.98	12.80±1.31
1551	400	8	27.75	16.50±1.64	14.12±1.72	12.50±1.26
1552	1100	5	27.73	29.23±0.89	19.86±1.63	21.86±1.69
1553	700	3	35.00	38.71±1.18	30.43±1.06	34.93±1.16
1554	500	5	38.40	32.20±1.31	20.40±2.49	21.00±2.17
1555	976	8	23.57	15.26±1.09	11.43±0.79	12.60±0.83
1568	12958	4	33.34	89.77±0.25	23.83±2.77	83.15±1.89
1596	580943	7	48.76	69.45±0.07	18.54±3.50	36.97±2.49
4552	5665	102	8.88	51.48±1.22	0.39±0.12	15.04±1.60
40927	60000	10	10.00	30.01±0.69	9.98±0.17	13.24±1.30
40966	1080	8	13.89	47.27±2.21	13.52±1.14	22.18±2.34
40971	1000	30	8.00	10.40±0.71	4.45±0.63	4.75±0.83
40979	2000	10	10.00	92.40±0.46	10.32±0.97	38.15±3.68
40982	1941	7	34.67	64.08±0.61	13.31±1.97	31.62±3.78
40984	2310	7	14.29	79.94±0.81	16.15±1.25	50.71±4.97
40985	45781	20	6.35	6.28±0.13	5.05±0.22	5.33±0.33
40996	70000	10	10.00	83.51±0.14	9.61±0.45	80.84±0.16
41002	5880	3	44.47	97.57±0.16	32.84±1.91	83.48±5.57
41003	5880	3	49.61	87.84±0.40	34.97±2.99	65.33±6.96
41004	4704	3	44.20	97.44±0.18	31.80±1.46	92.30±3.63
41039	131600	47	2.13	51.52±0.18	2.25±0.08	16.85±0.86
41081	44557	10	18.82	15.47±0.80	8.84±0.44	10.35±0.38

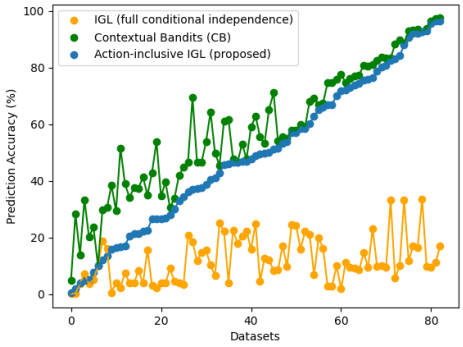
Table 3: Average accuracy and standard errors for different algorithms on individual OpenML datasets.

Dataset ID	N	K	Best Constant Action Policy (%)	CB (%)	IGL (full CI) (%)	AI-IGL (%)
41082	9298	10	16.7	91.12±0.20	11.21±0.82	77.66±2.29
41083	400	40	2.50	4.88±1.33	2.88±0.69	2.00±0.45
41084	575	20	8.35	29.91±1.41	4.14±0.59	6.29±0.59
41163	10000	5	20.49	87.03±0.45	19.44±0.55	66.98±4.69
41164	8237	7	23.39	54.04±0.37	14.51±0.52	43.50±2.91
41165	10000	10	10.43	27.56±0.45	10.08±0.32	10.49±0.42
41166	58310	10	21.96	53.11±0.44	11.71±1.12	45.36±1.32
41167	416188	355	0.59	28.28±0.26	0.27±0.02	2.06±0.17
41168	83733	4	46.01	62.84±0.15	24.93±2.63	48.94±3.76
41169	65196	100	6.14	19.49±0.24	0.65±0.05	7.34±0.61
41511	150	3	33.33	84.67±3.09	34.33±2.59	52.00±7.14
41568	150	3	33.33	85.00±2.98	32.33±1.90	44.67±7.59
41583	150	3	33.33	82.67±3.00	36.67±1.91	39.67±8.47
41919	527	4	39.47	50.57±1.34	23.40±1.99	31.51±4.68
41939	218	5	38.53	35.23±2.31	23.64±3.21	23.64±2.22
41950	150	3	33.33	80.00±2.67	32.33±1.84	25.67±6.62
41960	523590	144	25.08	38.46±0.79	0.56±0.14	15.89±2.20
41972	9144	8	44.29	63.16±1.64	12.44±3.11	26.07±4.36
41981	296	14	30.74	26.33±1.65	4.83±1.09	12.00±1.77
41982	70000	10	10.00	75.86±0.12	10.11±0.25	70.14±0.14
41986	51839	43	5.79	87.55±0.17	3.23±0.29	50.86±1.25
41988	51839	43	5.79	89.76±0.14	2.97±0.31	56.77±1.73
41989	51839	43	5.79	88.43±0.17	1.61±0.16	45.18±1.15
41990	51839	43	5.79	15.89±0.13	1.54±0.16	8.34±0.42
41991	270912	49	2.58	53.96±0.13	2.17±0.08	26.62±0.65
41997	150	3	33.33	89.00±2.07	30.00±1.73	42.33±8.44
42003	150	3	33.33	82.67±3.00	34.67±2.19	16.67±5.30
42011	150	3	33.33	81.67±2.90	33.67±3.18	27.33±7.87
42016	150	3	33.33	83.67±2.96	32.00±2.39	32.67±7.66
42021	150	3	33.33	82.00±3.17	36.00±3.14	37.00±6.77
42026	150	3	33.33	83.00±3.28	31.33±2.27	38.33±7.01
42031	150	3	33.33	84.67±2.83	33.67±2.69	27.33±6.53
42036	150	3	33.33	85.67±2.68	37.00±2.23	44.67±7.57
42041	150	3	33.33	83.33±2.81	29.33±2.08	41.33±8.06
42046	150	3	33.33	86.67±2.58	32.67±3.09	30.33±6.90
42051	150	3	33.33	82.67±3.03	30.00±2.33	29.33±7.19
42056	150	3	33.33	80.33±3.35	33.00±2.52	28.67±7.29
42066	150	3	33.33	82.00±3.50	31.67±2.20	40.00±5.85
42071	150	3	33.33	83.33±2.56	29.67±2.60	43.67±6.95
42098	150	3	33.33	84.00±2.92	29.67±2.33	37.00±7.31
42140	9927	10	19.10	13.87±0.77	9.21±0.80	10.69±0.54
42141	49644	10	19.10	15.84±0.86	11.59±0.89	11.03±0.54
42186	150	3	33.33	82.33±2.92	32.67±2.26	28.00±7.27
42261	150	3	33.33	81.67±3.30	33.33±2.94	24.00±5.88
42345	70340	3	48.88	61.16±0.14	22.11±4.71	45.74±1.31
42396	108000	1000	0.10	4.78±0.09	0.11±0.01	0.42±0.03
42468	830000	5	20.22	68.18±0.14	21.15±1.19	60.11±0.39
42532	2778	10	27.29	45.90±0.91	9.95±0.78	21.89±2.99
42544	265	8	17.74	42.04±1.68	14.81±1.76	15.00±2.31
42585	344	3	44.19	75.14±1.62	39.43±2.45	26.57±6.01
42700	150	3	33.33	87.33±2.70	33.67±1.53	22.00±6.15
42718	1000000	4	25.75	57.78±0.10	24.35±1.67	57.13±0.07
42793	75	4	40.00	71.25±3.07	23.12±3.22	34.38±5.44
43859	150	3	33.33	84.33±2.99	30.67±2.47	32.33±7.25
43875	150	3	33.33	85.33±2.34	32.00±2.34	42.33±7.34

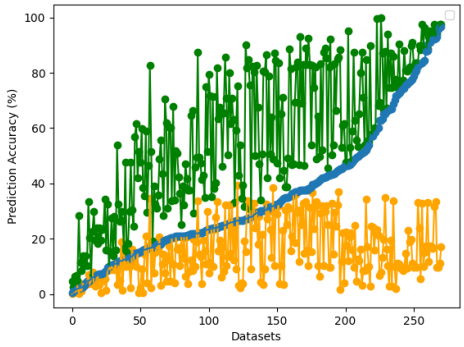
B.2 Ablation Analysis on OpenML Datasets

This section provides an ablation study based on the results of OpenML datasets in Table 2 and Figure 3. Both Table 2 and Figure 3 compares the performance of AI-IGL, CB, and IGL (full CI) on all OpenML datasets with balanced action distributions, and that with large sample size (sample size $\geq$ MNIST). We can observe that AI-IGL almost always outperforms IGL (full CI). We now demonstrate that the performance of AI-IGL is affected by: 1) the label noise; 2) the size of datasets (expected from theory), by the following observation.

At first, it is easy to notice that the performance of AI-IGL is significantly benefited from the sample size, by comparing with the performance of AI-IGL between all datasets and datasets with a large sample size (also suggested in Table 2), which follows the prediction of Theorem 2. On the other hand, we study how the label noise affects the performance of AI-IGL. To ablate the effect from the sample size, we only study the results in the dataset with large sample size. In this case, we



(a) Datasets with $K \geq 3$, $N \geq 70000$



(b) Datasets with $K \geq 3$, no constraints on N

Figure 3: Average performance on datasets with balanced action distributions from the OpenML benchmarking suite. For legibility of the figure, we do not include standard errors here. Standard errors for individual datasets are reported in Table 3. K is the size of the action set and N is the sample size.

consider the performance of CB to indicate the label noise, i.e., datasets with high CB performance means small label noise, and vice versa. We can observe from Figure 3(a) that the performance gap between CB and AI-IGL increases considerably with the label noise increasing (i.e., CB performance decreasing). This suggests that AI-IGL is more sensitive to the label noise compared with CB. It is unclear if this is an information-theoretical difficulty due to the setting without explicit reward, and we leave further investigation along this line as future work.

B.3 Analyzing OpenML Dataset Meta-Properties with respect to AI-IGL's Performance

In order to better understand which features of a dataset make them amenable to high IGL performance, we systematically analyzed the features of OpenML datasets. Specifically, we measured a total of 15 features/meta-properties for all OpenML datasets and compared the accuracy of AI-IGL's success relative to CB. We first explain the first 12, and explanation for the additional 3 measures follow.

We collected a total of 49 features following the findings of Lorena et al. [2019]; Torra et al. [2008]; Reif et al. [2014]; Abdelmessih et al. [2010]. Based on preliminary analysis. We narrowed these features down to the following 12 features for further investigation of AI-IGL on OpenML datasets, because they had the least correlations with one another:

- Accuracy of the 1-nearest neighbor classifier on the dataset (1nn_accuracy) [Reif et al., 2014; Abdelmessih et al., 2010]
- Best single decision node accuracy created using the feature attribute with the highest information gain (best_node_accuracy) [Reif et al., 2014; Abdelmessih et al., 2010]
- The number of 0/1 features when the dataset is one-hot encoded (feature_onehot_count) [Lorena et al., 2019]
- Ratio of sample dimensionality by sample count (instance_per_feature) [Lorena et al., 2019]
- The ratio of the class distribution entropy and the maximum entropy for the uniform distribution over classes (class_entropy_N)
- Maximum Fisher discriminant ratio (max_fisher_discrim) [Lorena et al., 2019]
- The percentage of values in the feature matrix that are non-zero (max_single_feature_eff) [Lorena et al., 2019]
- Mutual information mean (mutual_XY_info_mean) [Torra et al., 2008; Reif et al., 2014]
- Naive Bayes accuracy (naive_bayes_accuracy) [Reif et al., 2014; Abdelmessih et al., 2010]
- Noise to signal ratio (noise_signal_ratio) [Torra et al., 2008]

- Number of principal components needed to represent 95% of data variability (pca_dims_95) [Lorena et al., 2019]
- Percent of data variance explained by the top principal component (pca_top_1_percent) [Lorena et al., 2019]

In addition to these 12 features, we added three additional features as follows. 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), four additional features of the dataset were used to predict the relative performance of AI-IGL:

- N (n)
- $N/\sqrt{K}$ (n_by_sqrt_k)
- N/K (n_by_k)

We used 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. In Table 4, we report the F1 scores for the success and failure classes defined in this way for a binary random forest classifier with 100 trees (evaluated using 10-fold cross validation on all 271 OpenML datasets). Based on further analysis of the importance weight for each feature computed using the information gain metric (Figure 4(a)), we created smaller subsets of the datasets to further analyze datasets representative of realistic interaction datasets with small sample sizes. We also report the average F1 scores for random forest classifiers trained and evaluated in a similar manner in rows 2,3 in Table 4.

Dataset properties	AI-IGL accuracy$< 70\%$	AI-IGL accuracy$\geq 70\%$ (Success)	Average F1 score (both classes)
$K \geq 3$	0.91	0.83	0.87
$N/K \leq 1000$	0.95	0.46	0.71
$N/K \leq 200$	0.97	0.56	0.77

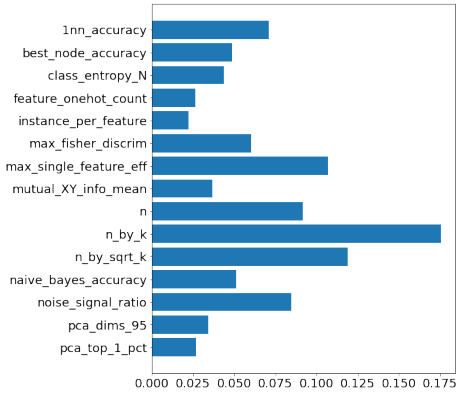
Table 4: F1 scores of binary random forest classifiers predicting the success of AI-IGL relative to CB on different datasets using curated features/meta-properties.

We also computed the feature importance for each of the random forest classifiers to determine which feature is most predictive of AI-IGL's success relative to CB. Feature importance was computed using the information gain metric. We used python's scikit-learn library to implement the random forest classifiers as well as feature importance values.

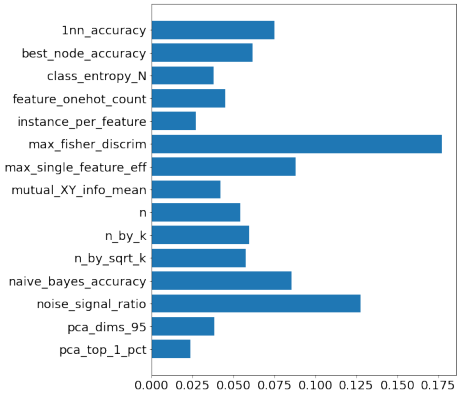
We present feature importance plots for three different subsets of datasets in Figure 4. We find N/K to be the most predictive feature of AI-IGL's relative performance (Figure 4(a)). It can alone predict its performance with an average F1 score of 0.79 under the same experimental setup. However, for datasets with a small value of N/K (≤ 1000, ≤ 200), there is high variability in relative performance. Using such a subset of datasets, we find maximum Fisher discriminant [Lorena et al., 2019] (a measure of classification complexity that quantifies the informativeness of a given sample) to be the most predictive of relative performance (Figure 4(b), Figure 4(c)).

This finding identifies clear measures of small and large datasets that can predict whether AI-IGL can match CB performance for any given dataset. 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.

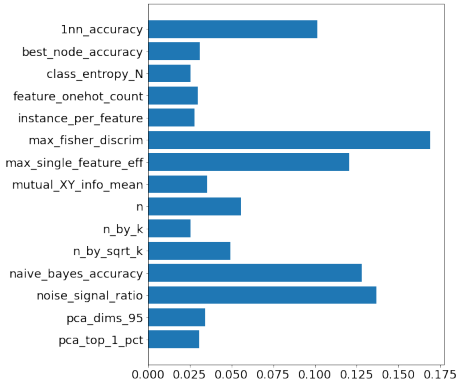
For the features considered relatively more important than others as shown in Figure 4(a), 4(b) and 4(c)), we also present how the relative performance of AI-IGL varies for different values of such features. Visualizations for different features across all 271 datasets are shown in Figure 5, across the subset of 155 datasets with $N/K \leq 1000$ are shown in Figure 6, and across the subset of 128 datasets with $N/K \leq 200$ are shown in Figure 7.



(a) Feature importance for 271 datasets with $K \geq 3$)



(b) Feature importance for 155 datasets with $N/K \leq 1000$



(c) Feature importance for 128 datasets with $N/K \leq 200$

Figure 4: Feature importance over dataset meta-features used to classify the success of AI-IGL. We find that for across all datasets N/K is the most informative feature (Figure 4(a)), whereas maximum Fisher discriminant ratio is the most important feature over datasets with smaller values of N/K (Figure 4(b) and 4(c)).

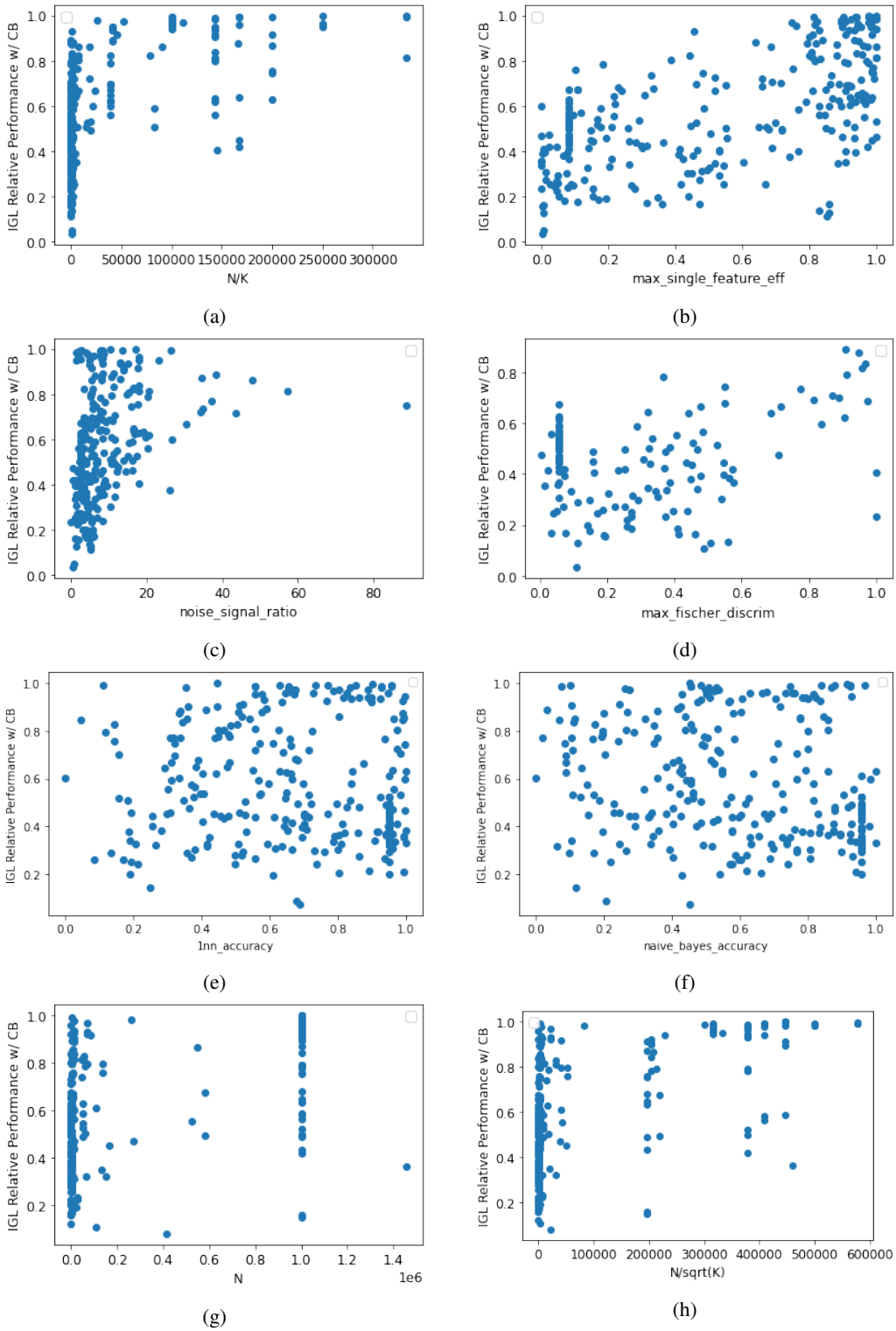


Figure 5: AI-IGL's relative performance w.r.t. CB versus different feature values for all 271 datasets with $K \geq 3$. N/K is the most important feature for predicting AI-IGL's relative performance as shown in Figure 4(a). Here we observe that the variability in AI-IGL's relative performance decreases as N/K increases. Datasets with a smaller values of N/K are analyzed separately (discussed in Figure 6 and Figure 7).

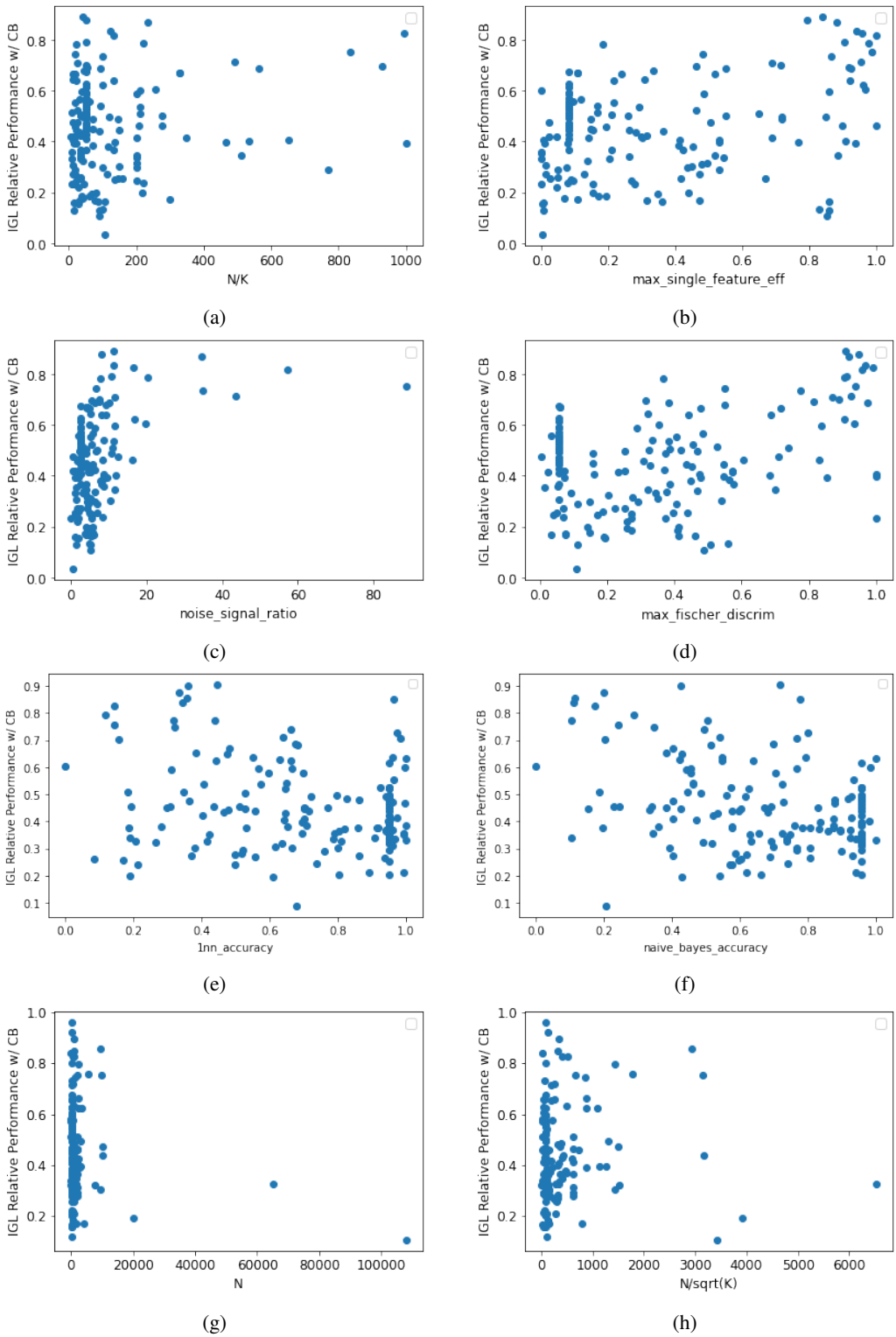


Figure 6: AI-IGL's relative performance w.r.t. CB versus different feature values for 155 datasets with $K \geq 3$ and $N/K \leq 1000$. N/K itself is not the most important feature (Fig. 4(b)) for this subset of datasets, but the maximum Fisher discriminant ratio is which shows an approximately linear trend (Figure 6(d)).

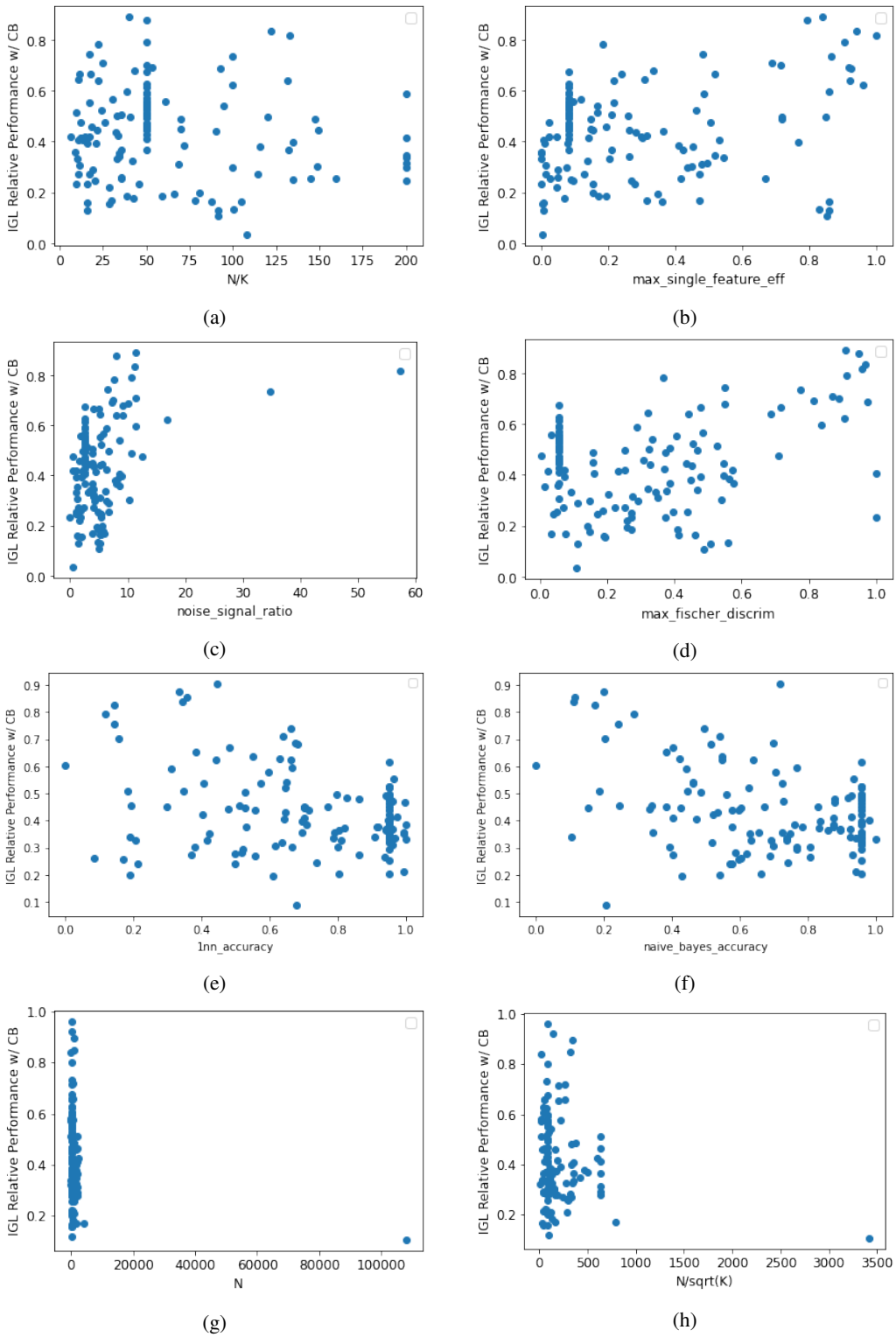


Figure 7: AI-IGL's relative performance w.r.t. CB versus different feature values for 128 datasets with $K \geq 3$ and $N/K \leq 200$. N/K itself has a small importance value (Figure 4(c)) for this subset of datasets. The maximum Fisher discriminant ratio is again the most important feature for this subset of datasets (showing an approximately linear trend Figure 7(d)) to predict the relative performance of AI-IGL.