An Information Theoretic Approach to Interaction-Grounded Learning

Xiaoyan Hu [1] **Farzan Farnia** [1] **Ho-fung Leung** [2]

Abstract

Reinforcement learning (RL) problems where the learner attempts to infer an unobserved reward from some feedback variables have been studied in several recent papers. The setting of *Interaction-Grounded Learning (IGL)* is an example of such feedback-based RL tasks where the learner optimizes the return by inferring latent binary rewards from the interaction with the environment. In the IGL setting, a relevant assumption used in the RL literature is that the feedback variable Y is conditionally independent of the context-action (X, A) given the latent reward R. In this work, we propose *Variational Information-based IGL (VI-IGL)* as an information-theoretic method to enforce the conditional independence assumption in the IGL-based RL problem. The VI-IGL framework learns a reward decoder using an information-based objective based on the conditional mutual information (MI) between (X, A) and Y. To estimate and optimize the information-based terms for the continuous random variables in the RL problem, VI-IGL leverages the variational representation of mutual information to obtain a min-max optimization problem. Also, we extend the VI-IGL framework to general f-Information measures leading to the generalized f-VI-IGL framework for the IGL-based RL problems. We present numerical results on several reinforcement learning settings indicating an improved performance compared to the existing IGL-based RL algorithm.

1. Introduction

In several applications of reinforcement learning (RL) algorithms, the involved agent lacks complete knowledge of the reward variable, e.g. in applications concerning brain-computer interface (BCI) (Schalk et al., 2004; Serrhini & Dargham, 2017) and recommender systems (Maghakian et al., 2023). In such RL settings, the lack of an explicit reward could lead to a challenging learning task where the learner needs to infer the unseen reward from observed feedback variables. The additional inference task for the reward variable could significantly raise the computational and statistical complexity of the RL problem. Due to the great importance of addressing such RL problems with a misspecified reward variable, they have been exclusively studied in several recent papers (Xie et al., 2021b; 2022; Maghakian et al., 2023).

To handle the challenges posed by a misspecified reward variable, Xie et al. (2021b; 2022) propose the *Interaction-Grounded Learning (IGL) framework*. According to the IGL framework, the agent observes a multidimensional *context vector* based on which she takes an *action*. Then, the environment generates a *latent* 0-1 *reward* and reveals a multidimensional *feedback vector* to the agent. The agent aims to maximize the (unobserved) return by inferring rewards from the interaction, a sub-task which needs to be solved based on the assumptions on the relationship between reward and feedback variables.

As a result, the key to addressing the IGL-based RL problem is a properly inferred *reward decoder* $\psi \in \Psi$, which maps a context-action-feedback tuple $(X, A, Y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ to a prediction of the posterior probability on the latent reward R. Given such a reward decoder, the optimal policy can be obtained using standard contextual bandit algorithms (Langford & Zhang, 2007; Dudík et al., 2014). However, such a reward decoder will be information-theoretically infeasible to learn without additional assumptions (Xie et al., 2022). Consequently, the existing works on the IGL setting (Xie et al., 2021b; 2022) make relevant assumptions on the statistical relationship between the random variables of context X, action A, feedback Y, and latent reward R. In particular, a sensible assumption on the connection between X, A, Y, R is the following conditional independence assumption proposed by Xie et al. (2021b) (The causal graph is given in Figure 1.):

Assumption 1 (Full conditional independence)**.** *For arbitrary (X, A, R, Y) tuple where R and Y are generated based on the context-action pair (X, A), the feedback Y*

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong [2]Independent Researcher. Correspondence to: Xiaoyan Hu <xyhu21@cse.cuhk.edu.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

is conditionally independent of X and A given the latent reward R, i.e., $Y \perp\!\!\!\perp X, A|R$.

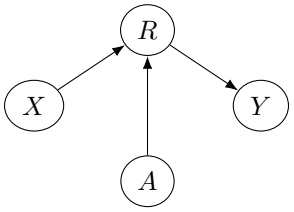


Figure 1. Causal graph of IGL under the full conditional independence assumption.

In the work of Xie et al. (2021b), a reward decoder $\psi : \mathcal{Y} \mapsto [0, 1]$ takes the feedback $Y \in \mathcal{Y}$ as input and outputs a prediction of the posterior distribution $\mathbb{P}(R = 1|Y)$. Their proposed approach performs a joint training of the policy and the decoder by maximizing the difference in the decoded return between the learned policy and a "bad" policy that is known to have a low (true) return. They show that a properly inferred reward decoder can be learned statistically efficiently when: (i) the full conditional independence assumption 1 *strictly* holds, and (ii) the distributions $\mathbb{P}(Y|R = 0)$ and $\mathbb{P}(Y|R = 1)$ of the feedback variable Y conditioned to the latent reward can be *well separated* (Xie et al., 2021b, Assumption 2). However, these conditions are quite restricted in practice, where the observation of the feedback variable is often under significant noise levels, e.g. in the BCI application. In such noisy settings, Assumption 1 may still hold under an independent noise from the discussed random variables or may not hold when the noise is correlated with the context or action variables. On the other hand, it can be much more difficult to distinguish between the feedback distributions conditioned to the latent reward. Consequently, the discussed IGL-based methods may no longer achieve optimal results under such noisy feedback conditions.

In this paper, we attempt to address the mentioned challenges in the IGL-based RL problem and propose *Variational Information-based IGL (VI-IGL)* as an information-theoretic approach to IGL-based RL tasks. The proposed VI-IGL methodology is based on the properties of information measures that allow measuring the dependence among random variables. According to these properties, Assumption 1 will hold, i.e., the feedback variable Y is conditionally independent of the context-action (X, A) given the latent reward R, *if and only if* the conditional mutual information (CMI) $I(Y; X, A|R)$ is zero. Therefore, we suggest an information bottleneck-based approach (Tishby et al., 2000) and propose to learn a reward decoder via the following information-based objective value where $\beta > 0$ is a tuning parameter and R_ψ is the random decoded reward from ψ:

$$\arg\min_{\psi \in \Psi} \{I(Y; X, A|R_\psi) - \beta \cdot I(X, A; R_\psi)\} \quad (1)$$

Intuitively, minimizing the first term $I(Y; X, A|R_\psi)$ ensures that the solved reward decoder satisfies the full conditional independence assumption. In addition, the second term $I(R_\psi; X, A)$ serves as a regularization term ruling out naive reward decoders.

Nevertheless, the objective function in (1) is challenging to optimize, since a first-order optimization of this objective requires estimating the value and derivatives of the MI for continuous random variables of the context X and the feedback Y. To handle this challenge, we leverage the variational representation of MI (Donsker & Varadhan, 1983; Nguyen et al., 2010) and cast Objective (1) as a min-max optimization problem that gradient-based algorithms can efficiently solve. Using the variational formulation of the information-based objective, we propose the Variational Information-based IGL (VI-IGL) minimax learning algorithm for solving the IGL-based RL problem. The VI-IGL method applies the standard gradient descent ascent algorithm to optimize the min-max optimization problem following the variational formulation of the problem.

We numerically evaluate the proposed VI-IGL method on several RL tasks. Our empirical results suggest that VI-IGL can perform better than the baseline IGL RL algorithm in the presence of a noisy feedback variable. The main contributions of this paper can be summarized as:

1. We propose an information-theoretic approach to the IGL-based RL problem, which learns a reward decoder by minimizing an information-based objective function.
2. To handle the challenges in estimating and optimizing (f-)MI for continuous random variables, we leverage the variational representation and formulate our objective as a min-max optimization problem, which can be solved via gradient-based optimization methods. We show that the optimal value can be sample-efficiently learned.
3. We extend the proposed approach to f-Variational Information-based IGL (f-VI-IGL), leading to a family of algorithms to solve the IGL-based RL task.
4. We provide empirical results indicating that f-VI-IGL performs successfully compared to existing IGL-based RL algorithms.

2. Related Works

Interaction-Grounded Learning (IGL). The framework of IGL is proposed by Xie et al. (2021b) to tackle learning scenarios without explicit reward. At each round, the agent observes a multidimensional context, takes an action, and then

the environment generates a latent 0-1 reward and outputs a multidimensional feedback. The agent aims to optimize the expected return by observing only the context-action-feedback tuple during the interaction. When the feedback is independent of both the context and the action given the latent reward (full conditional independence), Xie et al. show that the optimal policy can be sample-efficiently learned with additional assumptions. To relax the full conditional independence requirement, Xie et al. (2022) introduce Action-Inclusive IGL, where the feedback can depend on both the latent reward and the action. They propose a contrastive learning objective and show that the latent reward can be decoded under a symmetry-breaking procedure. Recently, Maghakian et al. (2023) apply the IGL paradigm with a multi-state latent reward to online recommender systems. Their proposed algorithm is able to learn personalized rewards and show empirical success.

Information-Theoretic Reinforcement Learning Algorithms. Reinforcement learning (RL) is a well-established framework for agents' decision-making in an unknown environment (Sutton & Barto, 2018). Several recent works focus on designing RL algorithms by exploiting the information-related structures in the learning setting. To perform exploration and sample-efficient learning, Russo and Van Roy (2014) propose information-directed sampling (IDS), where the agent takes actions that either with a small *regret* or yield large *information gain*, which is measured by the mutual information between the optimal action and the next observation. They show that IDS preserves numerous theoretical guarantees of Thompson sampling while offering strong performance in the face of more complex problems. In addition, information-theoretic approaches have been applied for *skills discovery* in machine learning contexts. Gregor, Rezende, and Wierstra (2016) introduce variational intrinsic control (VIC), which discovers useful and diverse behaviors (i.e., *options*) by maximizing the mutual information between the options and termination states. A setting that is close to our paper is using information-based methodology to learn reward functions in *inverse reinforcement learning* (IRL) (Ng & Russell, 2000). Levine, Popović, and Koltun (2011) propose to learn a cost function by maximizing the entropy between the corresponding optimal policy and human demonstrations. However, IGL is different from this setting, since it does not make any assumptions on the optimality of the observed behavior.

Estimation of Mutual Information (MI). Mutual information (MI) is a fundamental information-theoretic quantity that measures "the amount of information" between random variables. However, estimating MI in continuous settings is statistically and computationally challenging (Gao et al., 2015). Building upon the well-known characterization of the MI as the Kullback-Leibler (KL-) divergence (Kullback, 1997), recent works propose to use the variational representation of MI for its estimation and more generally for f- divergences (Nguyen et al., 2010; Belghazi et al., 2018; Molavipour et al., 2020). We note that estimating mutual information in high-dimensional setting is subject to the curse of dimensionality as discussed in the related papers (Paninski, 2003b; Poole et al., 2019; Song & Ermon, 2020). On the other hand, the neural net based variational estimator of mutual information seems to generalize well in practical numerical experiments. Studying the generalization properties of such deep variational estimators of information measures is an interesting subject for future studies. In addition, we note an extra challenge in our analysis is to estimate the conditional mutual information given a latent variable that has not been addressed in the mentioned related works.

3. Preliminaries

3.1. Interaction-Grounded Learning (IGL)

In the Interaction-Grounded Learning (IGL) paradigm, at each round, a multidimensional *context* $x \in \mathcal{X}$ is drawn from a distribution d_0 and is revealed to the agent. Upon observing x, the agent takes action $a \in \mathcal{A}$ from a finite action space. Let $\Delta_{\mathcal{S}}$ denote the probability simplex on space $\mathcal{S}$. Given the context-action pair (x, a), the environment generates a *latent and binary reward* $r \sim R(x, a) \in \Delta_{\{0,1\}}$ and returns a multidimensional *feedback* $y \in \mathcal{Y}$ to the agent. It can be seen that IGL recovers a contextual bandit (CB) problem (Langford & Zhang, 2007) if the reward is observed. Let $\pi \in \Pi : \mathcal{X} \to \Delta_{\mathcal{A}}$ denote any stochastic policy. The expected return of policy π is given by $V(\pi) := \mathbb{E}_{x\sim d_0}\mathbb{E}_{a\sim\pi(\cdot|x)}[\mu(x, a)]$, where $\mu(x, a)$ is the expected (latent) reward of any context-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$. We consider batch mode learning, where the agent has access to a dataset $\{(x_k, a_k, y_k)\}_{k=1}^K$ collected by the behavior policy $\pi_{\mathrm{b}} : \mathcal{X} \to \Delta_{\mathcal{A}}$, where $x_k \sim d_0, a_k \sim \pi(\cdot|x_k)$, and y_k is the stochastic feedback. The agent aims to learn the optimal policy, that is, $\pi^* := \arg\max_{\pi\in\Pi} V(\pi)$ while only observing the context-action-feedback tuple (x, a, y) at each round of interaction.

3.2. (f-)Conditional Mutual Information

The (f-)mutual information (MI) (Ali & Silvey, 1966) is a standard measure of dependence between random variables in information theory. Formally, let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function satisfying $f(1) = 0$. The f-MI (Csiszár, 1967) between a pair of random variables Z_1 and Z_2 is given by

$$I_f(Z_1; Z_2) := D_f(\mathbb{P}_{Z_1 Z_2} \| \mathbb{P}_{Z_2} \otimes \mathbb{P}_{Z_1}). \tag{2}$$

In this definition, $D_f(\mathbb{P}\|\mathbb{Q})$ denotes the f-divergence between distributions $\mathbb{P}$ and $\mathbb{Q}$ defined as

$$D_f(\mathbb{P}\|\mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right]$$

Note that the standard KL-based conditional mutual information, which is denoted by $I(Z_1; Z_2)$, is given by $f(x) = x \log x$. Another popular f-divergence is Pearson-χ^2 (Peason, 1900), where $f(x) = (x-1)^2$. An important property of f-MI is that two random variables Z_1, Z_2 are statistically independent if and only $I_f(Z_1; Z_2) = 0$, and hence dependence among between random variables can be measured via an f-mutual information.

Furthermore, the f-conditional MI (Csiszár, 1967) between a pair of random variables Z_1 and Z_2 when Z_3 is observed can be defined as

$$I_f(Z_1; Z_2|Z_3) := D_f(\mathbb{P}_{Z_1 Z_2|Z_3} \| \mathbb{P}_{Z_2|Z_3} \otimes \mathbb{P}_{Z_1|Z_3}). \quad (3)$$

Similarly, the standard KL-based conditional mutual information, denoted by $I(Z_1; Z_2|Z_3)$, is given by $f(x) = x \log x$. One useful property of the f-CMI is that, if Z_1 is conditionally independent of Z_2 given Z_3 then it holds that $I_f(Z_1; Z_2|Z_3) = 0$.

4. Variational Information-Based IGL

In this section, we derive an information-theoretic formulation for the IGL-based RL problem. As discussed earlier, in information theory, a standard measure of the (conditional) dependence between random variables is (conditional) mutual information (MI). Particularly, Assumption 1 (i.e., $Y \perp\!\!\!\perp X, A|R$) is equivalent to that the conditional MI between the context-action (X, A) and the feedback variable Y is zero given the latent reward R, i.e., $I(Y; X, A|R) = 0$. However, training a reward decoder by minimizing the conditional mutual information can either learn undesired causal structures or "overfit" to the feedback variable, which may underperform in the IGL setting. We defer the detailed discussion to Section 4.1. Hence, we propose an information-theoretic objective function to learn a reward decoder $\psi \in \Psi : \mathcal{X} \times \mathcal{A} \times \mathcal{Y} \to [0, 1]$

$$\arg\min_{\psi \in \Psi} \{I(Y; X, A|R_\psi) - \beta \cdot I(X, A; R_\psi)\} \quad (4)$$

where $\beta > 0$ is a tunable hyperparameter. In the optimization of the above objective function, minimizing the first term $I(Y; X, A|R_\psi)$ guides the reward decoder to satisfy the conditional independence assumption. Furthermore, as the feedback variable is often under significant noise levels in practice, the second term $I(X, A; R_\psi)$ will play the role of a regularization term improving the robustness of the learned reward decoder against the noisy feedback. (The detailed discussion can be found in Appendix A.)

To handle the continuous random variables of the context X and the feedback Y, we leverage the variational representation of the mutual information (Nguyen et al., 2010; Belghazi et al., 2018) and reduce (4) to the following *variational information-based IGL (VI-IGL)* optimization problem. Here, we first present a min-max formulation for the above information-based optimization problem and a sample complexity bound for the resulting RL algorithm. Later in Section 4.2, we explain the steps in the proof.

Theorem 1 (VI-IGL optimization problem)**.** *Objective (4)*

$$\arg\min_{\psi \in \Psi} \{I(Y; X, A|R_\psi) - \beta \cdot I(X, A; R_\psi)\}$$

is equivalent to the following optimization problem:

$$\begin{aligned} \arg\min_{\psi \in \Psi} \mathcal{L}(\psi) := \max_{G \in \mathcal{G}} \min_{T \in \mathcal{T}} \Big\{ & \mathbb{E}_{\mathbb{P}_{XAYR_\psi}}[G] \\ & - \mathbb{E}_{\mathbb{P}_{Y|R_\psi} \otimes \mathbb{P}_{XAR_\psi}}\left[e^G\right] \\ & - \beta \cdot \left(\mathbb{E}_{\mathbb{P}_{XAR_\psi}}[T] - \mathbb{E}_{\mathbb{P}_{XA} \otimes \mathbb{P}_{R_\psi}}\left[e^T\right] \right) \Big\} \end{aligned} \quad (5)$$

where $G \in \mathcal{G} : \mathcal{X} \times \mathcal{A} \times \mathcal{Y} \times \{0, 1\} \to \mathbb{R}$ and $T \in \mathcal{T} : \mathcal{X} \times \mathcal{A} \times \{0, 1\} \to \mathbb{R}$ are two function classes.

Algorithm 1 Variational Information-based IGL (VI-IGL)

Require: dataset $\mathcal{D}_{\text{train}} = \{(x_k, a_k, y_k)\}_{k=1}^K$, parameter $\beta > 0$, reward decoder class $\Psi = \{\psi_\theta\}_{\theta \in \Theta}$, estimators $\mathcal{G} = \{G_{\omega_1}\}_{\omega_1 \in \Omega_1}$ and $\mathcal{T} = \{T_{\omega_2}\}_{\omega_2 \in \Omega_2}$ of $I(Y; X, A|R_\psi)$ and $I(X, A; R_\psi)$, respectively.

1: **for** epoch $k = 1, 2, \cdots, K$ **do**
2: Sample a mini-batch $\mathcal{D}_{\text{mini}} \sim \mathcal{D}_{\text{train}}$.
3: Estimate the objective value $\widehat{\mathcal{L}}(\psi)$ given by Equation (5).
4: Update the parameters

$$\omega_1 \leftarrow \omega_1 + \eta \cdot \nabla_{\omega_1} \widehat{\mathcal{L}}(\psi)$$
$$\omega_2 \leftarrow \omega_2 - \frac{\eta}{\beta} \cdot \nabla_{\omega_2} \widehat{\mathcal{L}}(\psi)$$
$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \widehat{\mathcal{L}}(\psi)$$

where η is the learning rate.
5: **end for**
6: Train a policy π via an offline contextual bandit oracle.
7: **Output:** Policy π.

The inner level of the VI-IGL optimization problem minimizes over function class $\mathcal{T}$ to estimate $I(X, A; R_\psi)$, the medium level maximizes over function class $\mathcal{G}$ to estimate $I(Y; X, A|R_\psi)$, and the outer level minimizes over class Ψ to find the appropriate reward decoder.

Finally, as learning in IGL requires interaction with the environment, which can be expensive in practice, we provide theoretical guarantees for the VI-IGL optimization problem and show that the optimal objective value can be sample-efficiently learned. (The detailed proof can be found in Appendix B.)

Theorem 2 (Sample complexity)**.** *Consider a feedback-dependent reward decoder class Ψ such that $\psi(y) \in [c, 1 -$*

$c]$ for any $\psi \in \Psi$ and $y \in \mathcal{Y}$, where $c \in (0, \frac{1}{2})$. Suppose the function classes $\mathcal{T}$ and $\mathcal{G}$ are bounded by $B \leq \infty$. Then, for any $\delta \in (0, 1]$, given a dataset $\mathcal{D} = \{(x_k, a_k, y_k)\}_{k=1}^K$ collected by the behavior policy $\pi_b : \mathcal{X} \to \Delta_{\mathcal{A}}$, there exists an algorithm such that the solved reward decoder $\widehat{\psi}$ from the optimization problem (5) satisfies that

$$\left|\mathcal{L}(\widehat{\psi}) - \mathcal{L}^*\right| \leq \max\{1, \beta\} \cdot O\left(\frac{(1-c)^2}{c^2} \cdot \sqrt{\frac{\mathcal{C}(\mathcal{Y}_{\mathcal{P}}^{\epsilon}, B)}{K} \log\left(\frac{|\mathcal{Y}_{\mathcal{F}}^{\epsilon}| d_{\Psi,\mathcal{T},\mathcal{G}}}{\delta}\right)}\right) \tag{6}$$

where $\mathcal{L}^* := \min_{\psi \in \Psi} \mathcal{L}(\psi)$ is the optimal value, $\mathcal{Y}_{\mathcal{F}}^{\epsilon} \subset \mathcal{Y}$ is a ϵ-covering of the feedback space $\mathcal{Y}$ equipped with (pseudo-)metric $\rho(y, y') := \max_{F \in \mathcal{F}} |F(y) - F(y')|$ where $\mathcal{F} := \Psi \cup \{e^{G(x,a,\cdot,r)} : (x, a, r) \in \mathcal{X} \times \mathcal{A} \times \{0, 1\}\}_{G \in \mathcal{G}}$, $\mathcal{C}(\mathcal{Y}_{\mathcal{P}}^{\epsilon})$ is the capacity number defined in Equation (25) in the Appendix, and $d_{\Psi,\mathcal{T},\mathcal{G}}$ is the statistical complexity of the joint function classes Ψ, $\mathcal{G}$, and $\mathcal{T}$, with the parameter $\epsilon = K^{-1/2}$.

In practice, the functions T, G, and the reward decoder ψ are often overparameterized deep neural networks which enable expressing complex functions. As discussed in the related works, estimating MI with finite samples can be statistically and computationally challenging, and we note that exponential factors indeed show up in our analysis. Specifically, the capacity number $\mathcal{C}$ in the sample complexity bound (6) depends on the covering number of the function class and also scales with $O(e^{2B})$, where B is the upper bound of the function classes. In application to deep neural networks, the covering number in the above sample complexity bound can be prohibitively large. We note that this issue in theoretically bounding the generalization error and sample complexity of deep learning algorithms is well-recognized in the supervised learning literature and is considered an open problem (Zhang et al., 2021). Similar to the supervised learning setting, we observed satisfactory numerical results achieved by the proposed VI-IGL-learned function, which highlights the role of gradient-based optimization in the success of the algorithm. Proving a sample complexity bound that takes the role of the gradient-based optimization algorithm into account will be an interesting future direction to our analysis.

4.1. Minimizing Conditional MI with Regularization

In this section, we present the detailed derivation of our information-theoretic objective (4). Recall that we aim to learn a reward decoder $\psi \in \Psi : \mathcal{X} \times \mathcal{A} \times \mathcal{Y} \to [0, 1]$ which minimizes the dependence measure $I(Y; X, A|R_\psi)$. Here, $R_\psi \sim \text{Bernoulli}(\psi(X, A, Y))$ is the decoded 0-1 reward. However, minimizing only $I(Y; X, A|R_\psi)$ can be problematic. On one hand, a reward decoder satisfying $(Y; X, A|R_\psi) = 0$ may correspond to a causal structure such as $X \to R_\psi \to Y$ or $A \to R_\psi \to Y$, which performs poorly in IGL as the latent reward often depends on the context-action pair (X, A). On the other hand, note that the chain rule of MI results in the following identity

$$\begin{aligned} &I(Y; X, A|R_\psi) \\ =&I(Y; R_\psi|X, A) - I(Y; R_\psi) + I(Y; X, A) \end{aligned} \tag{7}$$

As a result of the above information-theoretic identity, training to minimize only $I(Y; X, A|R_\psi)$ may result in a context-action-dependent reward decoder $\psi : \mathcal{X} \times \mathcal{A} \to [0, 1]$, i.e., $I(Y; R_\psi|X, A) = 0$, that "over-fits" to the feedback Y to maximize $I(Y; R_\psi)$, and hence may underperform under a noisy feedback variable. One such example is given as follows.

Example 1 (Unregularized objective leads to overfitting)**.** *By the expansion (7), for a reward decoder ψ such that $I(Y; R_\psi|X, A) = 0$, it holds that $I(Y; X, A|R_\psi) = 0$ if $I(Y; R_\psi) = I(Y; X, A)$, i.e.,*

$$H(Y|R_\psi) = H(Y|X, A)$$

There could exist multiple reward decoders satisfying the condition. One case is that if the action-context pairs can be partitioned into two disjoint sets, i.e., $\mathcal{X} \times \mathcal{A} = \mathcal{S}_0 \cup \mathcal{S}_1$ where $\mathcal{S}_0 \cap \mathcal{S}_1 = \emptyset$, such that the conditional entropy $H(Y|x, a)$ is the same for any (x, a) in the same subset, then the reward decoder R_ψ assigning value 0 (and 1) to all the context-action pairs in $\mathcal{S}_0$ (and $\mathcal{S}_1$) satisfies $I(Y; R_\psi|X, A) = 0$ and $I(Y; X, A|R_\psi) = 0$.

To address this issue, we propose the regularized information-based IGL objective (4) where $\beta > 0$ is a tunable parameter:

$$\arg\min_{\psi \in \Psi} \{I(Y; X, A|R_\psi) - \beta \cdot I(X, A; R_\psi)\} \tag{4}$$

To gain intuition on why Objective (4) can be robust against noisy feedback, note that

$$I(X, A; R_\psi) = H(R_\psi) - H(R_\psi|X, A)$$

where H is the Shannon entropy. Thus, Objective (4) encourages the reward decoder to remain unchanged to the context-action (X, A) to minimize $H(R_\psi|X, A)$. Hence, the noises present in the feedback variable Y cannot significantly affect the accuracy of the optimized reward decoder. On the other hand, we can show that in the noiseless setting, including the regularization term does not (greatly) affect the quality of the optimized reward decoder with a proper selection of β. (The detailed proof can be found in Appendix C.)

Theorem 3 (Regularization (almost) ensures conditional independence)**.** *Assume the reward decoder class Ψ admits realizability assumption, i.e., there exists $\tilde{\psi} \in \Psi$ such*

that $I(Y; X, A|R_{\tilde{\psi}}) = 0$. Then, under Assumption 1, any reward decoder optimizing Objective (4) satisfies that

$$I(Y; X, A|R_\psi) \leq \beta \cdot (\log 2 - I(Y; R)) \tag{8}$$

where R is the true latent binary reward and $I(Y; R) \leq \log 2$. Particularly, when Ψ is feedback-dependent, a reward decoder $\psi : \mathcal{Y} \to [0, 1]$ attains the minimum if and only if $I(Y; X, A|R_\psi) = 0$.

In other words, for a feedback-dependent reward decoder class, the optimized reward decoder is guaranteed to satisfy the conditional independence assumption *regardless of* the selection of β. For reward decoder class that also depends on the context-action, the learned reward decoder violates Assumption 1 by at most (a multiplicative of) β.

As demonstrated by our numerical results in Section 6.2, introducing this regularizer not only helps to handle a noisy feedback variable, but also results in a more consistent algorithm performance under lower noise levels.

4.2. Leveraging Variational Representation to Solve Information-based Objective

While the previous sub-section introduces an information-theoretic objective to address the IGL-based RL problem, optimizing (4) in complex environments can be highly challenging. The primary challenge to solve (4) is that it requires estimating MI among continuous random variables of the context X and the feedback Y, which is widely recognized as a statistically and computationally difficult problem (Paninski, 2003a). To derive a tractable optimization problem, we utilize the variational representation of the KL-divergence, which reduces the evaluation and estimation of MI to an optimization task.

Proposition 1 (Donsker-Varadhan representation (Donsker & Varadhan, 1983))**.** *Let $\mathbb{P}, \mathbb{Q} \in \Delta_{\mathcal{S}}$ be two probability distributions on space $\mathcal{S}$. Then,*

$$D_{\text{KL}}(\mathbb{P}\|\mathbb{Q}) = \sup_{T \in \mathcal{T}} \left\{ \mathbb{E}_{s\sim\mathbb{P}}[T(s)] - \mathbb{E}_{s\sim\mathbb{Q}}\left[e^{T(s)}\right] \right\}$$

where the supremum is taken over all functions T such that the two expectations are finite.

Recall that the MI between random variables $Z_1 \in \mathcal{Z}_1$ and $Z_2 \in \mathcal{Z}_2$ is the KL-divergence between their joint distribution $\mathbb{P}_{Z_1 Z_2}$ and the product of their marginal distributions $\mathbb{P}_{Z_1} \otimes \mathbb{P}_{Z_2}$, i.e., $I(Z_1; Z_2) = D_{\text{KL}}(\mathbb{P}_{Z_1 Z_2} \| \mathbb{P}_{Z_1} \otimes \mathbb{P}_{Z_2})$. Proposition 1 enables us to estimate $I(Z_1; Z_2)$ through optimizing over a class of function $T : \mathcal{Z}_1 \times \mathcal{Z}_2 \to \mathbb{R}$. Therefore, directly applying Proposition 1 to Objective (4) results in the VI-IGL optimization problem in Theorem 1.

$$\arg\min_{\psi\in\Psi} \max_{G\in\mathcal{G}} \min_{T\in\mathcal{T}} \left\{ \mathbb{E}_{\mathbb{P}_{XAYR_\psi}}[G] - \mathbb{E}_{\mathbb{P}_{Y|R_\psi}\otimes\mathbb{P}_{XAR_\psi}}\left[e^G\right] - \beta \cdot \left(\mathbb{E}_{\mathbb{P}_{XAR_\psi}}[T] - \mathbb{E}_{\mathbb{P}_{XA}\otimes\mathbb{P}_{R_\psi}}\left[e^T\right] \right) \right\}$$

5. Extension of VI-IGL to General f-divergences

Algorithm 2 f-Variational Information-based IGL (f-VI-IGL)

Require: dataset $\mathcal{D}_{\text{train}} = \{(x_k, a_k, y_k)\}_{k=1}^K$, parameter $\beta > 0$, reward decoder class $\Psi = \{\psi_\theta\}_{\theta\in\Theta}$, estimators $\mathcal{G} = \{G_{\omega_1}\}_{\omega_1\in\Omega_1}$ and $\mathcal{T} = \{T_{\omega_2}\}_{\omega_2\in\Omega_2}$ of $I_{f_1}(Y; X, A|R_\psi)$ and $I_{f_2}(X, A; R_\psi)$, respectively.

1: **for** epoch $k = 1, 2, \cdots, K$ **do**

2: Sample a mini-batch $\mathcal{D}_{\text{mini}} \sim \mathcal{D}_{\text{train}}$.

3: Construct datasets with distributions $\mathbb{P}_Y \otimes \mathbb{P}_{R_{\psi_\theta}}$ and $\mathbb{P}_{Y|R_{\psi_\theta}} \otimes \mathbb{P}_{XAR_{\psi_\theta}}$ using $\mathcal{D}_{\text{mini}}$ (See the algorithm description).

4: Estimate the f-MI terms

$$\widehat{I}_{f_1}(X, A; R_{\psi_\theta}) \leftarrow \mathbb{E}_{\mathbb{P}_{XAR_{\psi_\theta}}}[T] - \mathbb{E}_{\mathbb{P}_{XA}\otimes\mathbb{P}_{R_{\psi_\theta}}}[f_1^*(T)]$$

$$\widehat{I}_{f_2}(Y; X, A|R_{\psi_\theta}) \leftarrow \mathbb{E}_{\mathbb{P}_{XAYR_{\psi_\theta}}}[G] - \mathbb{E}_{\mathbb{P}_{Y|R_{\psi_\theta}}\otimes\mathbb{P}_{XAR_{\psi_\theta}}}[f_2^*(G)]$$

5: Update the parameters

$$\omega_1 \leftarrow \omega_1 + \eta \cdot \nabla_{\omega_1} \left\{ \widehat{I}_{f_1}(X, A; R_\psi) \right\}$$

$$\omega_2 \leftarrow \omega_2 + \eta \cdot \nabla_{\omega_2} \left\{ \widehat{I}_{f_2}(X, A; R_\psi) \right\}$$

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \left\{ \widehat{I}_{f_1}(Y; X, A|R_{\psi_\theta}) - \beta \cdot \widehat{I}_{f_2}(X, A; R_{\psi_\theta}) \right\}$$

where η is the learning rate.

6: **end for**

7: Select between ψ_θ and its opposite counterpart $1 - \psi_\theta$ based on their decoded returns of π_b.

8: Train a policy π via an offline contextual bandit oracle.

9: **Output:** Policy π.

5.1. The Extended f-Variational Information-based IGL

In this section, we first propose an extended version of the information-based objective in (4) and the VI-IGL optimization problem (5). Recall that f-mutual information defined in Equation (3) generalizes the standard KL-divergence-based MI to a general f-divergence-based MI. Therefore, we can extend the standard MI-based IGL objective (4) to the following f-MI-based IGL objective:

$$\psi^* := \arg\min_{\psi\in\Psi} \{ I_{f_1}(Y; X, A|R_\psi) - \beta \cdot I_{f_2}(X, A; R_\psi) \} \tag{9}$$

where f_1 and f_2 are two f-divergences. Note that Objective (4) is a special case of the above formulation by selecting $f_1(x) = f_2(x) = x \log x$ to obtain the standard KL-based mutual information. Similar to the VI-IGL problem formulation, to derive a tractable optimization problem corresponding to the above task, we adopt the variational representation of f-divergences (Proposition 2 in Appendix D). We propose the following min-max optimization problem to solve Objective (9).

Theorem 4 (f-VI-IGL optimization problem)**.** *Let f_1 and f_2 be functions satisfying the requirements in Proposition 2 and we denote by f_1^* and f_2^* their Fenchel conjugate, respectively. Objective (9) is equivalent to the following min-max optimization problem*

$$\min_{\psi\in\Psi}\max_{G\in\mathcal{G}}\min_{T\in\mathcal{T}}\Big\{\mathbb{E}_{\mathbb{P}_{XAYR_\psi}}[G]-\mathbb{E}_{\mathbb{P}_{Y|R_\psi}\otimes\mathbb{P}_{XAR_\psi}}[f_1^*(G)] - \beta\cdot(\mathbb{E}_{\mathbb{P}_{XAR_\psi}}[T])-\mathbb{E}_{\mathbb{P}_{XA}\otimes\mathbb{P}_{R_\psi}}[f_2^*(T)])\Big\} \tag{10}$$

where $G \in \mathcal{G} : \mathcal{X}\times\mathcal{A}\times\mathcal{Y}\times\{0,1\}\to\mathbb{R}$ and $T\in\mathcal{T}$: $\mathcal{X}\times\mathcal{A}\times\{0,1\}\to\mathbb{R}$.

Similarly, we derive the sample complexity for the above optimization problem in Theorem 5 in the Appendix.

5.2. Algorithm Description

Here, we present f-VI-IGL Algorithm 2 as an optimization method to solve the f-VI-IGL optimization problem (10) for continuous random variables of the context X and the feedback Y. The algorithm optimizes over three function classes $\mathcal{G}, \mathcal{T}$, and Ψ. Specifically, function class $\Psi = \{\psi_\theta\}_{\theta\in\Theta}$ consists of the reward decoders parameterized by $\theta\in\Theta$. Function class $\mathcal{G} = \{G_{\omega_1}\}$ parameterized by $\omega_1\in\Omega_1$ is the estimator of f_1-MI $I_{f_1}(Y;X,A|R_{\psi_\theta})$. In addition, function class $\mathcal{T} = \{T_{\omega_2}\}_{\omega_2\in\Omega_2}$ parameterized by $\omega_2\in\Omega_2$ is the estimator of f_2-MI $I_{f_2}(X,A;R_{\psi_\theta})$. We focus on learning in the batch mode, where the algorithm has access to an offline dataset $\mathcal{D}_{\text{train}} = \{(x_t,a_t,y_t)\}_{t=1}^T$ consisting of the context-action-feedback tuples, which is collected by the behavior policy π_b interacting with the environment.

At each epoch, f-VI-IGL first uses a mini-batch of data to estimate the value of Objective (10) (Lines 2-4). One difficulty is that estimating $I_{f_1}(Y;X,A|R_{\psi_\theta})$ requires sampling $(x,a,y)\sim\mathbb{P}_{R_{\psi_\theta}}\otimes\mathbb{P}_{Y|R_{\psi_\theta}}\otimes\mathbb{P}_{XA|R_{\psi_\theta}}$, where $\mathbb{P}_{Y|R_{\psi_\theta}}$ and $\mathbb{P}_{XA|R_{\psi_\theta}}$ can be intractable for continuous random variables of the context X and the feedback Y. To address the problem, we first augment each sample (x_t,a_t,y_t) N times to obtain $\{(x_t,a_t,y_t,r_t^i)\}_{i=1}^N$, where $r_t^i\sim\text{Bernoulli}(\psi_\theta(x_t,a_t,y_t))$ and N is a small positive integer (e.g. 5 in our experiments). To sample, e.g., the feedback $y\sim\mathbb{P}_{Y|R_{\psi_\theta}=1}$, we randomly sample a data point from $\{(x_t,a_t,y_t,r_t^j): j\in[N], r_t^j=1\}_{t=1}^T$, i.e., the "augmented" data points whose random decoded reward is 1. Given the estimated objective value, we alternatively update the parameters for the f-MI estimators and the reward decoder (Line 5). At the end of the training, we use the learned reward decoder ψ_θ to train a policy via an offline contextual bandit oracle (Langford & Zhang, 2007; Dudik et al., 2011). However, note that in Objective (9), both the optimal reward decoder ϕ^* and its opposite counterpart $1-\phi^*$ may attain the minimum simultaneously (while only one of them aligns is consistent with the true latent reward). Hence, we use the data-driven collector (Xie et al., 2021b) and select the reward decoder (between the learned reward decoder ψ_θ and its opposite counterpart $1-\psi_\theta$) that gives a decoded return of π_b lower than 0.5.[1]

6. Empirical Results

In this section, we numerically evaluate the f-VI-IGL algorithm on the number-guessing task (Xie et al., 2021b) with noisy feedback, the details of which are described in the following.

Number-guessing task with noisy feedback. In the standard setting, a *random* image x_t (context), whose corresponding number is denoted by $l_{x_t}\in\{0,1,\cdots,9\}$, is drawn from the MNIST dataset (Lecun et al., 1998) at the beginning of each round t. Upon observing x_t, the learner selects $a_t\in\{0,1,\cdots,9\}$ as the predicted number of x_t (action). The latent binary reward $r_t = \mathbb{1}[a_t = l_{x_t}]$ is the correctness of the prediction label. Then, a *random* image of digit $r_t\in\{0,1\}$ is revealed to the learner (feedback). In many real-world scenarios, the observation of the feedback variable is often under significant noise level, e.g., in the BCI application. To simulate these cases, we consider four types of noisy feedback. Specifically, with a small probability, the feedback is replaced with: 1) *independent noises* (`I`): a random image of letter "t" (*True*) when the guess is correct or a random image of letter "f" (*False*) when the guess is wrong, which is sampled from the EMNIST Letter dataset (Cohen et al., 2017), 2) *action-inclusive noises* (`A`): a random image of digit $(a_t+6\cdot r_t-3)$ mod 10, 3) *context-inclusive noises* (`C`): a random image of digit $(l_{x_t}+6\cdot r_t-3)$ mod 10, 4) *context-action-inclusive noises* (`C-A`): a random image of digit $(l_{x_t}+a_t+6\cdot r_t-3)$ mod 10. An example is given in Table 1. Note that the full conditional independence assumption does not strictly hold as the feedback is also affected by the context-action pair (except for the independent noises).

Data collection. We focus on the batch learning mode, where a training dataset $\mathcal{D}_{\text{train}} = \{(x_k,a_k,y_k)\}_{k=1}^K$ is collected by the uniform behavior policy using the *training set*. In all the experiments, the training dataset contains $60,000$ samples, i.e., $K = 60,000$. The output (linear) policy is

[1] Following the previous works (Xie et al., 2021b; 2022), we assume the behavior policy has a low (true) return.

$X =$ 	I	A	C	C−A
Noisy Y ($A = 5$)				
Noisy Y ($A = 6$)				

Table 1. An example of the noisy feedback: The context is a random image of digit "5", i.e., $l_x = 5$. The rows show the noisy feedback when the guess is digit "5" (correct) and "6" (wrong), respectively. The columns show the cases for each type of noise (`I`: independent, `A`: action-inclusive, `C`: context-inclusive, `C−A`: context-action-inclusive).

evaluated on a test dataset $\mathcal{D}_{\text{test}}$ containing $10,000$ samples of context, which is randomly collected from the *test set*. Additional experimental details are provided in Appendix G.

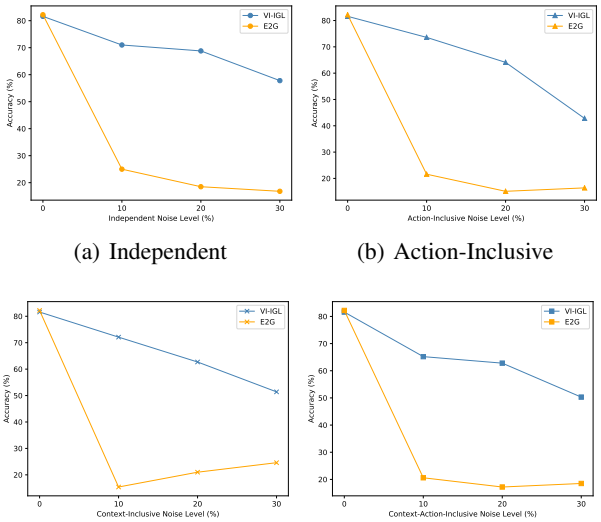
6.1. Robustness to Noises

In this section, we show that VI-IGL optimizing the standard MI-based Objective (4) is more robust to the noisy feedback than the previous IGL-based E2G algorithm (Xie et al., 2021b). We compare the accuracy of the output policy under different noise levels ($10\%, 20\%, 30\%$), and the results are summarized in Figure 2. (The detailed data can be found in Appendix F.1) In the noiseless setting, VI-IGL achieves a comparable performance ($(81.6 \pm 7.9)\%$) to E2G ($(82.2 \pm 4.3)\%$). However, VI-IGL (blue lines) significantly outperforms E2G (orange lines) *in all noisy settings and across all noise levels*.

Why previous IGL method fails. Recall that solving an appropriate reward decoder in the previous IGL method is given by (Xie et al., 2021b, Assumption 2), which states that there exists a reward decoder that well distinguishes between the feedback (distribution) generated from a latent reward of 0 and the one generated from a latent reward of 1. When additional noises present in the feedback, these two distributions can be quite similar. For example, for context-inclusive noises, a latent reward of 0 can also generate an image of digit "1" ($l_{x_t} = 4$ and $r_t = 0$). Hence, the condition easily fails and the performance degrades.

6.2. Necessity of Regularization

In this section, we show that including the regularization term $I(X, A; R_\psi)$ in Objective (4) helps achieve a more consistent algorithm performance. We compare the algorithm performance when optimizing the unregularized objective ($\beta = 0$) and the regularized objective ($\beta = 10$). Note that the case of $\beta = 0$ corresponds to minimizing only $I(Y; X, A|R_\psi)$. The results are summarized in Table 2. (Results for other selections of β can be found in Appendix F.2.) The results show that regularization significantly improves the performance.



(a) Independent (b) Action-Inclusive

(c) Context-Inclusive (d) Context-Action-Inclusive

Figure 2. Policy accuracy under different noise level: Our VI-IGL algorithm outperforms batch E2G (Xie et al., 2021b) in all noisy environments and across all noise levels. The results are averaged over 16 trials.

Methods	VI-IGL ($\beta = 0$)	VI-IGL ($\beta = 10$)
I	60.6 ± 15.5	$\mathbf{71.0 \pm 16.2}$
A	64.8 ± 17.9	$\mathbf{73.6 \pm 16.0}$
C	52.4 ± 25.3	$\mathbf{72.1 \pm 10.2}$
C−A	63.4 ± 16.4	$\mathbf{65.2 \pm 16.3}$
N	55.7 ± 27.3	$\mathbf{81.6 \pm 7.9}$

Table 2. Noise level=0.1. The results are averaged over 16 trials (`I`: independent, `A`: action-inclusive, `C`: context-inclusive, `C−A`: context-action-inclusive, `N`: noiseless setting).

6.3. Ablation Experiments

6.3.1. Selection of f-divergences. Recall that in Objective (9), we use f_1 and f_2 as general measures of $I(Y; X, A|R_\psi)$ and $I(X, A; R)$, respectively. We analyze how the selection of f-divergences affects the performance. We test three pairs of f_1-f_2: (i) KL-KL: both f_1 and f_2 are KL divergence, i.e., $f_1(x) = f_2(x) = x \log x$ (this case corresponds to Objective (4)), (ii) χ^2-χ^2: both f_1 and f_2 are Pearson-χ^2 divergence, i.e., $f_1(x) = f_2(x) = (x-1)^2$, and (iii) χ^2-KL: $f_1(x) = x \log x$ is KL divergence and $f_2(x) = (x-1)^2$ is Pearson-χ^2 divergence. Note that in the last case, the objective value, i.e., $I_{\chi^2}(Y; X, A|R_\psi) - \beta \cdot I(X, A; R_\psi)$, upper bounds the value of Objective (4).[2] We summarize the

[2] By the inequality $\log \le x - 1$, we have that $D_{\text{KL}}(\mathbb{P}\|\mathbb{Q}) = \mathbb{E}_{\mathbb{P}}[\log(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}})] \le \mathbb{E}_{\mathbb{P}}[(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} - 1)] = \mathbb{E}_{\mathbb{Q}}[(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}})^2] - 1 = D_\chi^2(\mathbb{P}\|\mathbb{Q})$.

results in Table 3 for a feedback-dependent reward decoder and $\beta = 10$. The results show that different f-divergences benefit from different types of noises.

f_1-f_2	KL-KL	χ^2-χ^2	χ^2-KL
I	71.0 ± 16.2	72.7 ± 17.4	$\mathbf{74.1 \pm 12.7}$
A	$\mathbf{73.6 \pm 16.0}$	65.3 ± 11.1	71.5 ± 16.7
C	72.1 ± 10.2	$\mathbf{76.2 \pm 11.5}$	69.4 ± 11.5
C-A	65.2 ± 16.3	$\mathbf{70.4 \pm 15.5}$	64.6 ± 14.7
N	$\mathbf{81.6 \pm 7.9}$	74.8 ± 13.3	77.3 ± 11.1

Table 3. Selection of f-divergences: The results are averaged over 16 trials (I: independent, A: action-inclusive, C: context-inclusive, C-A: context-action-inclusive, N: noiseless setting).

6.3.2. Input of reward decoder. We empirically analyze how the input of the reward decoder affects the actual performance. Particularly, we consider two types of input: (i) feedback Y and (ii) context-action-feedback (X, A, Y). We present the results in Table 4 for $\beta = 10$ and KL-KL divergence measure. The results show that in all cases, using a feedback-dependent reward decoder class leads to better performance than a context-action-feedback-dependent reward decoder class.

Input	Y	(X, A, Y)
I	$\mathbf{71.0 \pm 16.2}$	42.0 ± 24.1
A	$\mathbf{73.6 \pm 16.0}$	49.6 ± 24.4
C	$\mathbf{69.3 \pm 10.9}$	46.2 ± 16.9
C-A	$\mathbf{72.1 \pm 10.2}$	59.4 ± 17.4
N	$\mathbf{81.6 \pm 7.9}$	62.5 ± 19.1

Table 4. Input of Reward Decoder: The results are averaged over 16 trials (I: independent, A: action-inclusive, C: context-inclusive, C-A: context-action-inclusive, N: noiseless setting).

7. Discussion and Future Work

Regarding the limitations of our methodology and analysis, we observed that the variance of the numerical performance could be considerably large in some experiments. We hypothesize that this issue could be related to jointly training multiple networks and model initialization, also reported by Xie et al. (2021b), which could be an interesting topic for future studies. Furthermore, a relevant direction for future exploration is to consider non-information-theoretic dependence measures, e.g., Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) and Wasserstein distances (Villani et al., 2009), to enforce the IGL assumption. Another potential extension of our studied IGL problem is to relax the conditional independence assumption (Assumption 1). Such an extension could follow (Xie et al., 2022)'s idea on the Action-Inclusive IGL (AI-IGL), where the feedback may also be affected by the action.

Acknowledgments

The work of Farzan Farnia is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and is partially supported by a CUHK Direct Research Grant with CUHK Project No. 4055164. Also, the authors would like to thank the anonymous reviewers for their constructive feedback and suggestions.

Impact Statement

This paper presents work whose goal is to advance the field of Reinforcement Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966. ISSN 00359246. URL http://www.jstor.org/stable/2984279.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/belghazi18a.html.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. Emnist: an extension of mnist to handwritten letters, 2017.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

Donsker, M. and Varadhan, S. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36 (2):183–212, March 1983. ISSN 0010-3640. doi: 10.1002/cpa.3160360204.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits, 2011.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statistical*

Science, 29(4), nov 2014. doi: 10.1214/14-sts500. URL https://doi.org/10.1214%2F14-sts500.

Gao, S., Ver Steeg, G., and Galstyan, A. Efficient Estimation of Mutual Information for Strongly Dependent Variables. In Lebanon, G. and Vishwanathan, S. V. N. (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 277–286, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/gao15.html.

Gregor, K., Rezende, D. J., and Wierstra, D. Variational intrinsic control, 2016.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Kullback, S. *Information theory and statistics*. Courier Corporation, 1997.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/file/4b04a686b0ad13dce35fa99fa4161c65-Paper.pdf.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Levine, S., Popovic, Z., and Koltun, V. Nonlinear inverse reinforcement learning with gaussian processes. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/c51ce410c124a10e0db5e4b97fc2af39-Paper.pdf.

Maghakian, J., Mineiro, P., Panaganti, K., Rucker, M., Saran, A., and Tan, C. Personalized reward learning with interaction-grounded learning (IGL). In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=wGvzQWFyUB.

Molavipour, S., Bassi, G., and Skoglund, M. Conditional mutual information neural estimator. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5025–5029, 2020. doi: 10.1109/ICASSP40776.2020.9053422.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theor.*, 56(11):5847–5861, nov 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2068870. URL https://doi.org/10.1109/TIT.2010.2068870.

Paninski, L. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, jun 2003a. ISSN 0899-7667. doi: 10.1162/089976603321780272. URL https://doi.org/10.1162/089976603321780272.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003b.

Peason, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. doi: 10.1080/14786440009463897. URL https://doi.org/10.1080/14786440009463897.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. On variational bounds of mutual information. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/301ad0e3bd5cb1627a2044908a42fdc2-Paper.pdf.

Schalk, G., McFarland, D., Hinterberger, T., Birbaumer, N., and Wolpaw, J. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004. doi: 10.1109/TBME.2004.827072.

Serrhini, M. and Dargham, A. Toward incorporating bio-signals in online education case of assessing student attention with bci. In Rocha, Á., Serrhini, M., and Felgueiras, C. (eds.), *Europe and MENA Cooperation Advances in Information and Communication Technologies*, pp. 135–146, Cham, 2017. Springer International Publishing. ISBN 978-3-319-46568-5.

Song, J. and Ermon, S. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1x62TNtDS.

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

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method, 2000.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021a. URL https://openreview.net/forum?id=MjNFN44NbZm.

Xie, T., Langford, J., Mineiro, P., and Momennejad, I. Interaction-grounded learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11414–11423. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/xie21e.html.

Xie, T., Saran, A., Foster, D. J., Molu, L. P., Momennejad, I., Jiang, N., Mineiro, P., and Langford, J. Interaction-grounded learning with action-inclusive feedback. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=pBz3h8VibKY.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A. Proof of Theorem 1: The VI-IGL Optimization Problem

Proof. The theorem is a direct application of Proposition 1. The optimization problem possesses three levels: (i) the inner level minimizes over function class $T \in \mathcal{T}$ to estimate $I(X, A; R_\psi) = D_{\mathrm{KL}}(\mathbb{P}_{XAR_\psi} || \mathbb{P}_{XA} \otimes \mathbb{P}_{R_\psi}) = \sup_{T\in\mathcal{T}}\{\mathbb{E}_{\mathbb{P}_{XAR_\psi}}[T] - \mathbb{E}_{\mathbb{P}_{XA}\otimes\mathbb{P}_{R_\psi}}[e^T]\}$, (ii) the medium level maximizes over function class $G \in \mathcal{G}$ to estimate $I(Y; X, A|R_\psi) = D_{\mathrm{KL}}(\mathbb{P}_{XAYR_\psi} || \mathbb{P}_{Y|R_\psi} \otimes \mathbb{P}_{XAR_\psi}) = \sup_{G\in\mathcal{G}}\{\mathbb{E}_{\mathbb{P}_{XAYR_\psi}}[G] - \mathbb{E}_{\mathbb{P}_{Y|R_\psi}\otimes\mathbb{P}_{XAR_\psi}}[e^G]\}$, and (iii) the outer level finds the desired reward decoder. □

B. Proof of Theorem 2: Sample Complexity of VI-IGL Optimization Problem

Recall that the optimization problem (5) is minimizing

$$\mathcal{L}(\psi) = \max_{G\in\mathcal{G}} \min_{T\in\mathcal{T}} \left\{ \mathbb{E}_{\mathbb{P}_{XAYR_\psi}}[G] - \mathbb{E}_{\mathbb{P}_{Y|R_\psi}\otimes\mathbb{P}_{XA|R_\psi}}\left[e^G\right] - \beta \cdot \left(\mathbb{E}_{\mathbb{P}_{XAR_\psi}}[T] - \mathbb{E}_{\mathbb{P}_{XA}\otimes\mathbb{P}_{R_\psi}}\left[e^T\right] \right) \right\} \tag{11}$$

over the reward decoder $\psi \in \Psi$. In the offline setting, the learner has access to a K-size dataset $\mathcal{D} = \{(x_k, a_k, y_k)\}_{k=1}^K$ collected by the behavior policy $\pi_{\mathrm{b}} : \mathcal{X} \to \Delta_{\mathcal{A}}$. Particularly, at round k, a context $x_k \sim d_0$ is drawn from the context distribution. The behavior policy returns $a_k \sim \pi_{\mathrm{b}}(\cdot|x_k)$ and receives feedback y_k from the environment.

The algorithm constructs the empirical objective $\widehat{\mathcal{L}}(\psi)$ from the dataset for any reward decoder $\psi \in \Psi$ and outputs the minimizer $\widehat{\psi} = \arg\min_{\psi\in\Psi} \widehat{\mathcal{L}}(\psi)$. To show Theorem 2, it suffices to show that

$$\left|\mathcal{L}(\psi) - \widehat{\mathcal{L}}(\psi)\right| \le \epsilon + \max\{1, \beta\} \cdot O\left(\frac{(1-c)^2}{c^2} \cdot \sum_{y\in\mathcal{Y}_{\mathcal{F}}^\epsilon : \sigma_{\pi_{\mathrm{b}}}^\epsilon(y) > 0} \sqrt{\frac{e^{2B} + B^2}{K \cdot \sigma_{\pi_{\mathrm{b}}}^\epsilon(y)} \log\left(\frac{|\mathcal{Y}_{\mathcal{F}}^\epsilon| d_{\Psi,\mathcal{T},\mathcal{G}}}{\delta} \right)} \right)$$

for any reward decoder $\psi \in \Psi$ with high probability, where the parameters $\mathcal{Y}_{\mathcal{F}}^\epsilon$, $\sigma_{\pi_{\mathrm{b}}}^\epsilon$, and $d_{\Psi,\mathcal{T},\mathcal{G}}$ are specified in the proof. Once obtained, we set the parameter $\epsilon = K^{-1/2}$ and invoke the following inequality

$$\mathcal{L}(\widehat{\psi}) - \mathcal{L}(\psi^*) \le \left|\mathcal{L}(\widehat{\psi}) - \widehat{\mathcal{L}}(\widehat{\psi})\right| + \underbrace{\widehat{\mathcal{L}}(\widehat{\psi}) - \widehat{\mathcal{L}}(\psi^*)}_{\le 0} + \left|\widehat{\mathcal{L}}(\psi^*) - \mathcal{L}(\psi^*)\right| \tag{12}$$

to conclude the proof. Since the optimization problem over function classes $\mathcal{G}$ and $\mathcal{T}$ are decoupled, we define

$$\mathcal{L}_1(\psi) = \max_{G\in\mathcal{G}} \left\{ \mathbb{E}_{\mathbb{P}_{XAYR_\psi}}[G] - \mathbb{E}_{\mathbb{P}_{Y|R_\psi}\otimes\mathbb{P}_{XA|R_\psi}}\left[e^G\right] \right\} \tag{13}$$

$$\mathcal{L}_2(\psi) = \max_{T\in\mathcal{T}} \left\{ \mathbb{E}_{\mathbb{P}_{XAR_\psi}}[T] - \mathbb{E}_{\mathbb{P}_{XA}\otimes\mathbb{P}_{R_\psi}}\left[e^T\right] \right\} \tag{14}$$

Hence, we have that $\mathcal{L}(\psi) = \mathcal{L}_1(\psi) - \beta \cdot \mathcal{L}_2(\psi)$.

The details of the algorithm is given as follows. We consider a feedback-dependent reward decoder class, where $\psi(y) \in [c, 1-c]$ is the decoded probability given by $\psi \in \Psi$ that the feedback $y \in \mathcal{Y}$ is associated with a latent reward of 1. For convenience, we define $\psi_1(y) := \psi(y)$ and $\psi_0(y) := 1 - \psi(y)$, where the subscript is decoded binary reward. The algorithm computes the empirical counterpart of $\mathcal{L}_1(\psi)$ and $\mathcal{L}_2(\psi)$ as follows.

$$\widehat{\mathcal{L}}_1(\psi) = \max_{G\in\mathcal{G}} \left\{ \frac{1}{K} \sum_{k=1}^K \sum_{r=0,1} \psi_r(y_k) \cdot \left(G(x_k, a_k, y_k, r) - \mathbb{E}_{\widehat{\mathbb{P}}_{Y^\epsilon|R_\psi=r}}\left[e^{G(x_k,a_k,\tilde{y},r)}\right] \right) \right\} \tag{15}$$

$$\widehat{\mathcal{L}}_2(\psi) = \max_{T\in\mathcal{T}} \left\{ \frac{1}{K} \sum_{k=1}^K \sum_{r=0,1} \left(\psi_r(y_k) \cdot T(x_k, a_k, r) - \widehat{p}_\psi(r) \cdot e^{T(x_k,a_k,r)} \right) \right\} \tag{16}$$

where $\tilde{y} \sim \widehat{\mathbb{P}}_{Y^\epsilon|R_\psi=r}$ in $\widehat{\mathcal{L}}_1(\psi)$ is the empirical estimation of $\mathbb{P}_{Y|R_\psi=r}$ constructed from the dataset (see details in the proof) and $\widehat{p}_\psi(r) := \frac{1}{K}\sum_{k=1}^K \psi_1(y_k)$. In the proof, we aim to bound the estimation errors $|\mathcal{L}_1(\psi) - \widehat{\mathcal{L}}_1(\psi)|$ and $|\mathcal{L}_2(\psi) - \widehat{\mathcal{L}}_2(\psi)|$.

Proof. Fix a reward decoder $\psi \in \Psi$.

Step 1. Bounding $|\mathcal{L}_2(\psi) - \widehat{\mathcal{L}}_2(\psi)|$. Recall that $\mathcal{T}$ is bounded by B. Hence, with probability at least $1-\delta$ and applying a union bound over the function classes $\mathcal{T}$, the estimation errors are bounded by

$$\left|\mathbb{E}_{\mathbb{P}_{XAR_\psi}}[T] - \frac{1}{K}\sum_{k=1}^{K}\sum_{r=0,1}\psi_r(y_k)\cdot T(x_k,a_k,r)\right| \le O\left(\sqrt{\frac{B^2}{K}\log\left(\frac{d_{\mathcal{T}}}{\delta}\right)}\right) \tag{17}$$

$$\left|\mathbb{E}_{\mathbb{P}_{XA}\otimes\mathbb{P}_{R_\psi}}\left[e^T\right] - \frac{1}{K}\sum_{k=1}^{K}\sum_{r=0,1}\widehat{p}_\psi(r)\cdot e^{T(x_k,a_k,r)}\right| \le O\left(\sqrt{\frac{e^{2B}}{K}\log\left(\frac{d_{\mathcal{T}}}{\delta}\right)}\right) \tag{18}$$

for any function $T\in\mathcal{T}$, where $d_{\mathcal{T}}$ is the statistical complexity of the function class $\mathcal{T}$. Particularly, if $\mathcal{T}$ is finite, we have that $d_{\mathcal{T}} = |\mathcal{T}|$.

Step 2. Bounding $|\mathcal{L}_1(\psi) - \widehat{\mathcal{L}}_1(\psi)|$. Fix a function $G\in\mathcal{G}$. Recall that G is bounded by B. Hence, with probability at least $1-\delta$, we have that

$$\left|\mathbb{E}_{\mathbb{P}_{XAYR_\psi}}[G] - \frac{1}{K}\sum_{k=1}^{K}\sum_{r=0,1}\psi_r(y_k)\cdot G(x_k,a_k,y_k,r)\right| \le O\left(\sqrt{\frac{B^2}{K}\log\left(\frac{1}{\delta}\right)}\right) \tag{19}$$

The challenge is to analyze the estimation error

$$\left|\mathbb{E}_{\mathbb{P}_{Y|R_\psi}\otimes\mathbb{P}_{XAR_\psi}}\left[e^G\right] - \frac{1}{K}\sum_{k=1}^{K}\sum_{r=0,1}\psi_r(y_k)\cdot\mathbb{E}_{\widehat{\mathbb{P}}_{Y^\epsilon|r}}\left[e^{G(x_k,a_k,\tilde{y},r)}\right]\right| \tag{20}$$

To handle the continuous feedback space, we first introduce the notion of ϵ-covering, which results in a finite "clusterings" of the feedback and yields nice statistical properties.

Definition 1 (ϵ-covering)**.** *Let $\mathcal{G}\subset\{\mathcal{Y}\to\mathbb{R}\}$ denote a function class. A (finite) set $\mathcal{Y}^\epsilon_{\mathcal{G}}\subset\mathcal{Y}$ is said to be ϵ-covering the space $\mathcal{Y}$ with respect to function class $\mathcal{G}$ if for any $y\in\mathcal{Y}$, there exists $y^\epsilon\in\mathcal{Y}^\epsilon_{\mathcal{G}}$ such that $\max_{G\in\mathcal{G}}|G(y)-G(y^\epsilon)|\le\epsilon$. Further, we denote by $|\mathcal{Y}^\epsilon_{\mathcal{G}}|$ the ϵ-covering number.*

Remark 1. *Definition 1 is a classic ϵ-covering of space $\mathcal{Y}$ equipped with (pseudo-)metric $\rho(y,y') = \max_{G\in\mathcal{G}}|G(y)-G(y')|$.[3] For example, if the class $\mathcal{G}$ includes α-Lipschitz functions, i.e., $|G(y)-G(y')|\le\alpha\cdot\|y-y'\|_2$, then $\mathcal{Y}^\epsilon_{\mathcal{G}}$ is an $(\frac{\epsilon}{\alpha})$-covering of $\mathcal{Y}$ equipped with metric $\rho(y,y')=\|y-y'\|_2$.*

In the following, we denote by $\mathcal{Y}^\epsilon_{\mathcal{F}}$ an ϵ-covering with respect to the joint function class $\mathcal{F} := \Psi\cup\{e^{G(x,a,\cdot,r)} : (x,a,r)\in\mathcal{X}\times\mathcal{A}\times\{0,1\}\}_{G\in\mathcal{G}}$ and let $s:\mathcal{Y}\mapsto\mathcal{Y}^\epsilon_{\mathcal{F}}$ be a mapping from any $y\in\mathcal{Y}$ to $\mathcal{Y}^\epsilon_{\mathcal{F}}$ such that $\max_{F\in\mathcal{F}}|F(y)-F(y^\epsilon)|\le\epsilon$. Let $\sigma_{\pi_b}\in\Delta_{\mathcal{Y}}$ be the feedback distribution induced by the behavior policy π_b. We denote by $\sigma^\epsilon_{\pi_b}$ the corresponding distribution on the ϵ-covering $\mathcal{Y}^\epsilon_{\mathcal{F}}$. Specifically, the mass at any $y^\epsilon\in\mathcal{Y}^\epsilon_{\mathcal{F}}$ is given by $\sigma^\epsilon_{\pi_b}(y^\epsilon) := \int_{y:s(y)=y^\epsilon}\mathrm{d}\sigma_{\pi_b}(y)$. Further, the reward decoder $\psi\in\Psi$ induces posterior distributions $\mathbb{P}_{Y|R_\psi}$ and $\mathbb{P}_{Y^\epsilon|R_\psi}$ conditioned to the decoded reward on $\mathcal{Y}$ and $\mathcal{Y}^\epsilon_{\mathcal{F}}$, respective. By Definition 1, the expectation with respect to the distribution $\mathbb{P}_{Y|R_\psi}$ can be well estimated by the expectation computed from $\mathbb{P}_{Y^\epsilon|R_\psi}$.

Sub-Step 2.1. Construction of $\widehat{\mathbb{P}}_{Y^\epsilon|r}$. The construction of $\widehat{\mathbb{P}}_{Y^\epsilon|r}$, which involves: 1) computing the empirical feedback distribution $\sigma^\epsilon_{\pi_b} := \frac{1}{K}\sum_{k=1}^{K}\mathbb{1}[s(y)=y^\epsilon]$ for any $y^\epsilon\in\mathcal{Y}^\epsilon_{\mathcal{F}}$, and 2) utilizing Bayes rules to estimate the posterior distribution by

$$\widehat{\mathbb{P}}_{Y^\epsilon|R_\psi=r}(y^\epsilon) := \frac{\widehat{\sigma}^\epsilon_{\pi_b}(y^\epsilon)\cdot\psi_r(y^\epsilon)}{\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}}\widehat{\sigma}^\epsilon_{\pi_b}(y)\cdot\psi_r(y)} \tag{21}$$

[3] To show ρ is indeed a metric, note that 1) $\rho(y,y') = \rho(y',y)\ge 0$ $\rho(y,y)=0$ for any $y,y'\in\mathcal{Y}$ and 2) $\rho(y,y')\le\rho(y,y'')+\rho(y'',y')$ for any $y,y',y''\in\mathcal{Y}$.

for any $y^\epsilon \in \mathcal{Y}^\epsilon_{\mathcal{F}}$. Hence, the error (20) can be further written as

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbb{P}_{Y|R_\psi}\otimes\mathbb{P}_{XAR_\psi}}\left[e^G\right] - \frac{1}{K}\sum_{k=1}^{K}\sum_{r=0,1}\psi_r(y_k)\cdot\mathbb{E}_{\widehat{\mathbb{P}}_{Y^\epsilon|R_\psi}}\left[e^{G(x_k,a_k,\tilde{y},r)-1}\right]\right| \\
\leq &\underbrace{\left|\mathbb{E}_{\mathbb{P}_{Y|R_\psi}\otimes\mathbb{P}_{XAR_\psi}}\left[e^G\right] - \mathbb{E}_{\mathbb{P}_{Y^\epsilon|R_\psi}\otimes\mathbb{P}_{XAR_\psi}}\left[e^G\right]\right|}_{\leq\epsilon} \\
&+ \underbrace{\left|\mathbb{E}_{\mathbb{P}_{Y^\epsilon|R_\psi}\otimes\mathbb{P}_{XAR_\psi}}\left[e^G\right] - \frac{1}{K}\sum_{k=1}^{K}\sum_{r=0,1}\psi_r(y_k)\cdot\mathbb{E}_{\mathbb{P}_{Y^\epsilon|R_\psi=r}}\left[e^G\right]\right|}_{\text{concentration error}} \\
&+ \left|\frac{1}{K}\sum_{k=1}^{K}\sum_{r=0,1}\psi_r(y_k)\cdot\left(\mathbb{E}_{\mathbb{P}_{Y^\epsilon|R_\psi=r}}\left[e^{G(x_k,a_k,\tilde{y},r)}\right] - \mathbb{E}_{\widehat{\mathbb{P}}_{Y^\epsilon|R_\psi=r}}\left[e^{G(x_k,a_k,\tilde{y},r)}\right]\right)\right|
\end{aligned} \tag{22}$$

Observe that the last term is bounded by

$$\max_r \left\|\mathbb{P}_{Y^\epsilon|R_\psi=r} - \widehat{\mathbb{P}}_{Y^\epsilon|R_\psi=r}\right\|_1 \cdot e^{B-1}$$

for any $(\psi, G) \in \Psi \times \mathcal{G}$ and $(x, a) \in \mathcal{X} \times \mathcal{A}$. It remains to bound the error $\|\mathbb{P}_{Y^\epsilon|R_\psi=r} - \widehat{\mathbb{P}}_{Y^\epsilon|R_\psi=r}\|_1$.

Sub-Step 2.2. Bounding $\|\mathbb{P}_{Y^\epsilon|R_\psi=r} - \widehat{\mathbb{P}}_{Y^\epsilon|R_\psi=r}\|_1$. To start with, by (Xie et al., 2021a, Lemma A.1), with probability at least $1-\delta$ and applying union bound over $\mathcal{Y}^\epsilon_{\mathcal{F}}$, it holds that

$$\left|\widehat{\sigma}^\epsilon_{\pi_b}(y^\epsilon) - \sigma^\epsilon_{\pi_b}(y^\epsilon)\right| \leq O\left(\sqrt{\frac{1}{K\cdot\sigma^\epsilon_{\pi_b}(y^\epsilon)}\log\left(\frac{|\mathcal{Y}^\epsilon_{\mathcal{F}}|}{\delta}\right)}\right)$$

for any $y^\epsilon \in \mathcal{Y}^\epsilon_{\mathcal{F}}$. Recall that ψ_r is bounded between $[c, 1-c]$ where $0 < c < \frac{1}{2}$. We have that

$$\begin{aligned}
&\left|\mathbb{P}_{Y^\epsilon|r}(y^\epsilon) - \widehat{\mathbb{P}}_{Y^\epsilon|r}(y^\epsilon)\right| \\
=&\left|\frac{\widehat{\sigma}^\epsilon_{\pi_b}(y^\epsilon)\cdot\psi_r(y^\epsilon)}{\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}}\widehat{\sigma}^\epsilon_{\pi_b}(y)\cdot\psi_r(y)} - \frac{\sigma^\epsilon_{\pi_b}(y^\epsilon)\cdot\psi_r(y^\epsilon)}{\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}}\sigma^\epsilon_{\pi_b}(y)\cdot\psi_r(y)}\right| \\
=&\left|\frac{\left(\widehat{\sigma}^\epsilon_{\pi_b}(y^\epsilon) - \sigma^\epsilon_{\pi_b}(y^\epsilon)\right)\cdot\psi_r(y^\epsilon)}{\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}}\widehat{\sigma}^\epsilon_{\pi_b}(y)\cdot\psi_r(y)} + \frac{\sigma^\epsilon_{\pi_b}(y^\epsilon)\cdot\psi_r(y^\epsilon)}{\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}}\widehat{\sigma}^\epsilon_{\pi_b}(y)\cdot\psi_r(y)} - \frac{\sigma^\epsilon_{\pi_b}(y^\epsilon)\cdot\psi_r(y^\epsilon)}{\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}}\sigma^\epsilon_{\pi_b}(y)\cdot\psi_r(y)}\right| \\
\leq&O\left(\frac{1-c}{c}\sqrt{\frac{1}{K\cdot\sigma^\epsilon_{\pi_b}(y^\epsilon)}\log\left(\frac{|\mathcal{Y}^\epsilon_{\mathcal{F}}|}{\delta}\right)} + \frac{(1-c)^2}{c^2}\cdot\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}}\sqrt{\frac{1}{K\cdot\sigma^\epsilon_{\pi_b}(y)}\log\left(\frac{|\mathcal{Y}^\epsilon_{\mathcal{F}}|}{\delta}\right)}\right) \\
\leq&O\left(\frac{(1-c)^2}{c^2}\cdot\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}:\sigma^\epsilon_{\pi_b}(y)>0}\sqrt{\frac{1}{K\cdot\sigma^\epsilon_{\pi_b}(y)}\log\left(\frac{|\mathcal{Y}^\epsilon_{\mathcal{F}}|}{\delta}\right)}\right)
\end{aligned}$$

where the second inequality holds by the fact that $\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}}\widehat{\sigma}^\epsilon_{\pi_b}(y)\cdot\psi_r(y)$ and $\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}}\sigma^\epsilon_{\pi_b}(y)\cdot\psi_r(y)$ are bounded between $[c, 1-c]$. Hence, for any $r \in \{0, 1\}$, it holds that

$$\left\|\mathbb{P}_{Y^\epsilon|R_\psi=r} - \widehat{\mathbb{P}}_{Y^\epsilon|R_\psi=r}\right\|_1 \leq O\left(\frac{(1-c)^2}{c^2}\cdot\sum_{y\in\mathcal{Y}^\epsilon_{\mathcal{F}}:\sigma^\epsilon_{\pi_b}(y)>0}\sqrt{\frac{|\mathcal{Y}^\epsilon_{\mathcal{F}}|^2}{K\cdot\sigma^\epsilon_{\pi_b}(y)}\log\left(\frac{|\mathcal{Y}^\epsilon_{\mathcal{F}}|}{\delta}\right)}\right) \tag{23}$$

Therefore, combining Inequalities (22)∼(23) and applying a union bound over $G \in \mathcal{G}$ yields,

$$\begin{aligned}&\left|\mathbb{E}_{\mathbb{P}_{Y|R_\psi}\otimes\mathbb{P}_{XAR_\psi}}\left[e^G\right] - \frac{1}{K}\sum_{k=1}^{K}\sum_{r=0,1}\psi_r(y_k)\cdot\mathbb{E}_{\widehat{\mathbb{P}}_{Y^\epsilon|r}}\left[e^{G(x_k,a_k,\tilde{y},r)}\right]\right| \\ \leq& O\left(\frac{(1-c)^2}{c^2}\cdot\sum_{y\in\mathcal{Y}_{\mathcal{F}}^\epsilon:\sigma_{\pi_b}^\epsilon(y)>0}\sqrt{\frac{e^{2B}|\mathcal{Y}_{\mathcal{F}}^\epsilon|^2}{K\cdot\sigma_{\pi_b}^\epsilon(y)}\log\left(\frac{|\mathcal{Y}_{\mathcal{F}}^\epsilon|d_{\mathcal{G}}}{\delta}\right)}\right)\end{aligned} \tag{24}$$

where $d_{\mathcal{G}}$ is the statistical complexity of the function class $\mathcal{G}$, with $d_{\mathcal{G}} = |\mathcal{G}|$ for finite class $\mathcal{G}$.

Step 3. Putting everything together. Combining Inequalities (17)∼(19) and (24) and applying a union bound over the reward decoder class Ψ, we have that

$$\left|\mathcal{L}(\psi) - \widehat{\mathcal{L}}(\psi)\right| \leq \epsilon + \max\{1,\beta\}\cdot O\left(\frac{(1-c)^2}{c^2}\cdot\sum_{y\in\mathcal{Y}_{\mathcal{F}}^\epsilon:\sigma_{\pi_b}^\epsilon(y)>0}\sqrt{\frac{(e^{2B}+B^2)|\mathcal{Y}_{\mathcal{F}}^\epsilon|^2}{K\cdot\sigma_{\pi_b}^\epsilon(y)}\log\left(\frac{|\mathcal{Y}_{\mathcal{F}}^\epsilon|d_{\Psi,\mathcal{T},\mathcal{G}}}{\delta}\right)}\right)$$

for any $\psi \in \Psi$, where we denote by $d_{\Psi,\mathcal{T},\mathcal{G}}$ the statistical complexity of the joint function classes Ψ, $\mathcal{T}$, and $\mathcal{G}$, with $d_{\Psi,\mathcal{T},\mathcal{G}} = |\Psi||\mathcal{T}||\mathcal{G}|$ for finite classes. Define the capacity number

$$\mathcal{C}(\mathcal{Y}_{\mathcal{P}}^\epsilon, B) := \sum_{y\in\mathcal{Y}_{\mathcal{F}}^\epsilon:\sigma_{\pi_b}^\epsilon(y)>0}\frac{(e^{2B}+B^2)|\mathcal{Y}_{\mathcal{F}}^\epsilon|^3}{\sigma_{\pi_b}^\epsilon(y)} \tag{25}$$

By Cauchy-Schwartz inequality, we further have

$$\left|\mathcal{L}(\psi) - \widehat{\mathcal{L}}(\psi)\right| \leq \epsilon + \max\{1,\beta\}\cdot O\left(\frac{(1-c)^2}{c^2}\cdot\sqrt{\frac{\mathcal{C}(\mathcal{Y}_{\mathcal{P}}^\epsilon, B)}{K}\log\left(\frac{|\mathcal{Y}_{\mathcal{F}}^\epsilon|d_{\Psi,\mathcal{T},\mathcal{G}}}{\delta}\right)}\right)$$

Set $\epsilon = K^{-1/2}$ and we conclude the proof. □

C. Proof of Theorem 3: Regularization (Almost) Ensures Conditional Independence

Proof. Under the realizability assumption, there exists either (i) a context-action-dependent reward decoder $\tilde{\psi} : \mathcal{X}\times\mathcal{A}\to[0,1]$ or (ii) a feedback-dependent reward decoder $\tilde{\psi} : \mathcal{Y}\to[0,1]$ such that $I(Y;X,A|R_{\tilde{\psi}}) = 0$.[4] We first show that

$$I(X,A;R_{\tilde{\psi}}) = I(Y;X,A) = I(Y;R) \tag{26}$$

holds for both cases, where R is the true latent reward.

Case (i). Note that by the chain rules of CMI, we derive

$$0 = I(Y;X,A|R_{\tilde{\psi}}) = I(Y;X,A) + \underbrace{I(Y;R_{\tilde{\psi}}|X,A)}_{=0} - I(Y;R_{\tilde{\psi}})$$

where the second term $I(Y;R_{\tilde{\psi}}|X,A)$ on the RHS is zero as R_ψ is context-action-dependent. Hence, we have that $I(Y;R_{\tilde{\psi}}) = I(Y;X,A)$. Further, note that the following Markov chain holds:

$$(X,A)\to R_{\tilde{\psi}}\to Y$$

By the data processing inequality, we derive that $I(X,A;R_{\tilde{\psi}}) \leq I(Y;X,A)$ and the equality holds *if and only if* $I(Y;R_{\tilde{\psi}}) = I(Y;X,A)$. Therefore, we have that $I(X,A;R_{\tilde{\psi}}) = I(Y;X,A)$. Since the true latent reward R is context-action-dependent, following the same analysis, we have

$$I(Y;X,A) = I(Y;R) = I(X,A;R) \tag{27}$$

[4]Note that if a reward decoder depends on the (X,A,Y) tuple, it can be regarded as case (i).

Combining the analysis above, Equation (26) is proved.

Case (ii). Note that the following Markov chain holds for any feedback-dependent reward decoder:

$$(X, A) \to Y \to R_{\tilde{\psi}}$$

Then, the data processing inequality implies that $I(X, A; R_{\tilde{\psi}}) \leq I(Y; X, A)$ and the equality holds *if and only if* $I(Y; X, A|R_{\tilde{\psi}}) = 0$. Therefore, we derive $I(X, A; R_{\tilde{\psi}}) = I(Y; X, A) = I(Y; R)$, where the last equality holds by Equation (27). This also implies that for feedback-dependent reward decoder class, it holds that

$$\min_{\psi \in \Psi} \{I(Y; X, A|R_\psi) - \beta \cdot I(X, A; R_\psi)\} \geq -\beta \cdot I(Y; R)$$

Therefore, when Ψ is feedback-dependent, a reward decoder $\psi : \mathcal{Y} \to [0, 1]$ attains the minimum if and only if $I(Y; X, A|R_\psi) = 0$.

Once Equation (26) is obtained, we have that

$$\min_{\psi \in \Psi} \{I(Y; X, A|R_\psi) - \beta \cdot I(X, A; R_\psi)\} \leq I(Y; X, A|R_{\tilde{\psi}}) - \beta \cdot I(X, A; R_{\tilde{\psi}}) = -\beta \cdot I(Y; R)$$

Let ψ^* denote any reward decoder optimizing Objective (4). Rearranging the terms proves

$$I(Y; X, A|R_{\psi^*}) \leq \beta \cdot (I(X, A; R_{\psi^*}) - I(Y; R)) \leq \beta \cdot (\log 2 - I(Y; R))$$

where the second inequality holds by the fact that R^*_ψ is a 0-1 random variable. Therefore, we conclude the proof. □

D. The Variational Representation of f-divergences

Proposition 2 (Variational representation of f-divergences (Nguyen et al., 2010)). *Let $f : \mathbb{R}_+ \mapsto \mathbb{R}$ be a convex, lower-semicontinuous function satisfying $f(1) = 0$. Consider $\mathbb{P}, \mathbb{Q} \in \Delta_{\mathcal{S}}$ as two probability distributions on space $\mathcal{S}$. Then,*

$$D_f(\mathbb{P}\|\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right] \geq \sup_{T \in \mathcal{T}} \{\mathbb{E}_{s\sim\mathbb{P}}[T(s)] - \mathbb{E}_{s\sim\mathbb{Q}}[f^*(T(s))]\}$$

where $\mathcal{T} \subseteq \{T : \mathcal{S} \mapsto \mathbb{R}\}$ is any class of functions and $f^*(z) := \sup_{u\in\mathbb{R}}\{u \cdot z - f(u)\}$ for any $z \in \mathbb{R}_+$ is the Fenchel conjugate.

E. Sample Complexity of Optimization Problem (9)

Theorem 5 (Sample complexity of f-VI-IGL). *Consider a feedback-dependent reward decoder class Ψ such that $\psi(y) \in [c, 1 - c]$ for any $\psi \in \Psi$ and $y \in \mathcal{Y}$, where $c \in (0, \frac{1}{2})$. Suppose the functions $|G|, |T|, |f_1^*(G)|, |f_2^*(T)| \leq B^* \leq \infty$ are bounded. Then, for any $\delta \in (0, 1]$, given a dataset $\mathcal{D} = \{(x_k, a_k, y_k)\}_{k=1}^K$ collected by the behavior policy $\pi_{\mathsf{b}} : \mathcal{X} \to \Delta_{\mathcal{A}}$, there exists an algorithm such that the solved reward decoder $\widehat{\psi}$ from the optimization problem (10) satisfies that $|\mathcal{L}_f(\widehat{\psi}) - \mathcal{L}_f^*|$ is bounded by*

$$\max\{1, \beta\} \cdot O\left(\frac{(1-c)^2}{c^2} \cdot \sum_{y \in \mathcal{Y}^\epsilon_\Xi : \sigma^\epsilon_{\pi_{\mathsf{b}}}(y) > 0} \sqrt{\frac{(B^*)^2 |\mathcal{Y}^\epsilon_\Xi|^2}{K \cdot \sigma^\epsilon_{\pi_{\mathsf{b}}}(y)} \log\left(\frac{|\mathcal{Y}^\epsilon_\Xi| d_{\Psi,\mathcal{T},\mathcal{G}}}{\delta}\right)}\right)$$

where $\mathcal{L}_f^*$ is the optimal value to the optimization problem (10), parameters $\sigma^\epsilon_{\pi_{\mathsf{b}}}$ and $d_{\Psi,\mathcal{T},\mathcal{G}}$ are defined in the Appendix B, and $\mathcal{Y}^\epsilon_\Xi$ is ϵ-covering of feedback space $\mathcal{Y}$ with respect to the joint function class $\Xi := \Psi \cup \{f^*(G(x, a, \cdot, r)) : (x, a, r) \in \mathcal{X} \times \mathcal{A} \times \{0, 1\}\}_{G\in\mathcal{G}}$.

Proof. The proof follows the exact same analysis in Appendix B, with $f^*(x) = \exp(x - 1)$ as a special case. □

F. Additional Experimental Results

F.1. Robustness to Noises

This section provides the detailed data in Section 6.1. We compare the performance of E2G (Xie et al., 2021b) and the VI-IGL algorithm 1 in the number-guessing task. We report both the policy accuracy and the standard deviation. The results

are averaged over 16 trials. Specifically, Table 5 corresponds to the noiseless setting and Tables 6∼8 show the results under three noise levels $(0.1, 0.2, 0.3)$. We use the results for $\beta = 10$ to plot Figure 2.

In addition, the results show that as the noise level increases, our regularized objective (4) with $\beta = 10$ attains more consistent performance than the unregularized one (i.e., only minimizing $I(X, A; Y|R_\psi)$, which is shown by $\beta = 0$), in terms of both accuracy and the standard deviation. This reinforces the necessity to include the regularization term.

Methods	VI-IGL ($\beta = 0$)	VI-IGL ($\beta = 10$)	E2G
N	55.7 ± 27.3	$\mathbf{81.6 \pm 7.9}$	$\mathbf{82.2 \pm 4.3}$

Table 5. Noiseless setting. The results are averaged over 16 trials.

Methods	VI-IGL ($\beta = 0$)	VI-IGL ($\beta = 10$)	E2G
I	60.6 ± 15.5	$\mathbf{71.0 \pm 16.2}$	25.0 ± 18.4
A	64.8 ± 17.9	$\mathbf{73.6 \pm 16.0}$	21.6 ± 12.4
C	52.4 ± 25.3	$\mathbf{72.1 \pm 10.2}$	15.4 ± 14.3
C−A	63.4 ± 16.4	$\mathbf{65.2 \pm 16.3}$	20.6 ± 14.8

Table 6. Noise level=0.1. The results are averaged over 16 trials (I: independent, A: action-inclusive, C: context-inclusive, C−A: context-action-inclusive).

Methods	VI-IGL ($\beta = 0$)	VI-IGL ($\beta = 10$)	E2G
I	54.6 ± 23.5	$\mathbf{68.8 \pm 16.2}$	18.5 ± 13.1
A	43.9 ± 23.5	$\mathbf{64.1 \pm 18.7}$	15.1 ± 11.7
C	49.0 ± 25.9	$\mathbf{62.7 \pm 21.6}$	21.0 ± 13.3
C−A	57.6 ± 24.3	$\mathbf{62.8 \pm 23.9}$	17.2 ± 13.7

Table 7. Noise level=0.2. The results are averaged over 16 trials (I: independent, A: action-inclusive, C: context-inclusive, C−A: context-action-inclusive).

Methods	VI-IGL ($\beta = 0$)	VI-IGL ($\beta = 10$)	E2G
I	56.9 ± 21.4	$\mathbf{57.8 \pm 19.0}$	16.8 ± 12.7
A	33.6 ± 22.3	$\mathbf{42.8 \pm 17.2}$	16.4 ± 12.6
C	50.5 ± 21.2	$\mathbf{51.4 \pm 18.3}$	24.6 ± 16.4
C−A	$\mathbf{50.6 \pm 22.6}$	$\mathbf{50.3 \pm 18.1}$	18.5 ± 14.7

Table 8. Noise level=0.3. The results are averaged over 16 trials (I: independent, A: action-inclusive, C: context-inclusive, C−A: context-action-inclusive).

F.2. Value of Parameter β

This section provides the detailed data in Section 6.2.

Tables 9 and 10 show the results for the noiseless setting and the noisy settings (with noise level 0.1), respectively. In contrast, the performance of the unregularized objective significantly degrades.

β	0	5	10	15	20
N	55.7 ± 27.3	69.7 ± 15.6	$\mathbf{81.6 \pm 7.9}$	79.1 ± 9.1	72.7 ± 16.4

Table 9. Value of Parameter β: Noiseless environment. The results are averaged over 16 trials.

β	0	5	10	15	20
I	60.6 ± 15.5	59.5 ± 24.2	$\mathbf{71.0 \pm 16.2}$	$\mathbf{71.7 \pm 19.3}$	63.0 ± 18.7
A	64.8 ± 17.9	68.5 ± 18.8	$\mathbf{73.6 \pm 16.0}$	67.7 ± 20.6	58.2 ± 19.5
C	52.4 ± 25.3	61.4 ± 13.9	$\mathbf{72.1 \pm 10.2}$	63.1 ± 21.1	58.0 ± 26.3
C−A	63.4 ± 16.4	$\mathbf{68.7 \pm 20.8}$	65.2 ± 16.3	60.2 ± 15.0	62.6 ± 11.7

Table 10. Value of Parameter β: Noise level$= 0.1$. The results are averaged over 16 trials (I: independent, A: action-inclusive, C: context-inclusive, C−A: context-action-inclusive).

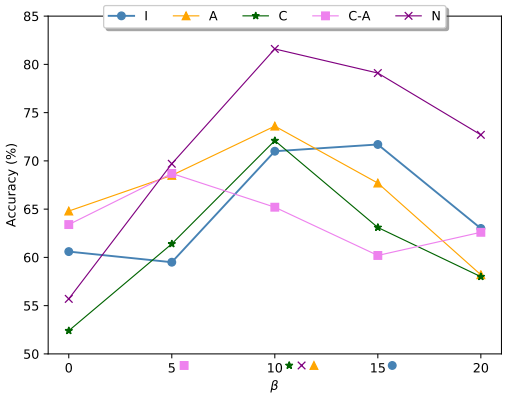


Figure 3. Policy accuracy for different β: All noisy settings have level$= 0.1$. The optimal selections are marked beside the value. The results show the necessity of regularization. The results are averaged over 16 trials.

G. Additional experimental details

For the f-variational estimators (functions T and G), the reward decoder ψ, and the linear policy π, we use a 2-layer fully-connected network to process each input image (i.e., the context or the feedback). Then, the concatenated inputs go through an additional linear layer and the final value is output. The same network structures are used to implement the reward decoder and the policy of the previous IGL algorithm (Xie et al., 2021b). In each experiment, we train the f-VI-IGL algorithm for $1,000$ epochs with a batch size of 600. Particularly, we alternatively update the parameters of the f-MI estimators and the reward decoders (i.e., 500 epochs of training for each). To stabilize the training, we clip the gradient norm to be no greater than 1 and use an exponential moving average (EMA) with a rate of 0.99. For the previous IGL method, we follow the experimental details provided in the work of Xie et al. (Xie et al., 2021b, Appendix C) and train the algorithm for 10 epochs over the entire training datasets.