# PAC Apprenticeship Learning with Bayesian Active Inverse Reinforcement Learning

**Ondrej Bajgar, Dewi S.W. Gould, Jonathon Liu,**
**Alessandro Abate, Konstantinos Gatsis, Michael A. Osborne**

**Keywords:** inverse reinforcement learning, active learning, imitation learning, Bayesian methods

## Summary

As AI systems become increasingly autonomous, reliably aligning their decision-making to human preferences is essential. Inverse reinforcement learning (IRL) offers a promising approach to infer preferences from demonstrations. These preferences can then be used to produce an apprentice policy that performs well on the demonstrated task. However, in domains like autonomous driving or robotics, where errors can have serious consequences, we need not just good average performance but reliable policies with formal guarantees – yet obtaining sufficient human demonstrations for reliability guarantees can be costly. *Active* IRL addresses this challenge by strategically selecting the most informative scenarios for human demonstration. We introduce PAC-EIG, an information-theoretic acquisition function that directly targets probably-approximately-correct (PAC) guarantees for the learned policy – providing the first such theoretical guarantee for active IRL with noisy expert demonstrations. Our method maximises information gain about the regret of the apprentice policy, efficiently identifying states requiring further demonstration. We also present Reward-EIG as an alternative when learning the reward itself is the primary objective. Focusing on finite state-action spaces, we prove convergence bounds, illustrate failure modes of prior heuristic methods, and demonstrate our method's advantages experimentally.

## Contribution(s)

1. We formulate two principled information theoretic acquisition functions for active inverse reinforcement learning with Boltzmann rational demonstrations: Reward-EIG and PAC-EIG.
   **Context:** This gives a more principled alternative to previous, heuristic acquisition functions of Lopes et al. (2009), Brown et al. (2018), and Kweon et al. (2023).
2. For RegretEIG, we prove a lower bound on the expected number of steps of active learning needed to reach a probably-approximately-correct (PAC) apprentice policy.
   **Context:** This a first such proof for active IRL with an expert that is not perfectly rational. Metelli et al. (2021); Lindner et al. (2022) presented results for the, in many respects much simpler, case of perfectly optimal expert, focusing especially on transfer of a learnt reward to new environment dynamics.

# PAC Apprenticeship Learning with Bayesian Active Inverse Reinforcement Learning

**Ondrej Bajgar**[1], **Dewi S.W. Gould**[2], **Jonathon Liu**[3],
**Alessandro Abate**[1], **Konstantinos Gatsis**[4], **Michael A. Osborne**[1]
ondrej@bajgar.org

[1]**University of Oxford, United Kingdom**
[2]**Alan Turing Institute, United Kingdom**
[3]**Independent**
[4]**University of Southampton, United Kingdom**

## Abstract

As AI systems become increasingly autonomous, reliably aligning their decision-making to human preferences is essential. Inverse reinforcement learning (IRL) offers a promising approach to infer preferences from demonstrations. These preferences can then be used to produce an apprentice policy that performs well on the demonstrated task. However, in domains like autonomous driving or robotics, where errors can have serious consequences, we need not just good average performance but reliable policies with formal guarantees – yet obtaining sufficient human demonstrations for reliability guarantees can be costly. *Active* IRL addresses this challenge by strategically selecting the most informative scenarios for human demonstration. We introduce PAC-EIG, an information-theoretic acquisition function that directly targets probably-approximately-correct (PAC) guarantees for the learned policy – providing the first such theoretical guarantee for active IRL with noisy expert demonstrations. Our method maximises information gain about the regret of the apprentice policy, efficiently identifying states requiring further demonstration. We also present Reward-EIG as an alternative when learning the reward itself is the primary objective. Focusing on finite state-action spaces, we prove convergence bounds, illustrate failure modes of prior heuristic methods, and demonstrate our method's advantages experimentally.

## 1 Introduction

Stuart Russell suggested three principles for the development of beneficial artificial intelligence: its only objective is realizing human preferences, it is initially uncertain about these preferences, and its ultimate source of information about them is human behavior (Russell, 2019). *Apprenticeship learning* via Bayesian *inverse reinforcement learning* (IRL) can be understood as a possible operationalization of these principles: Bayesian IRL starts with a prior distribution over reward functions representing initial uncertainty about human preferences. It then combines this prior with *demonstration* data from a human expert acting approximately optimally with respect to the unknown reward, to produce a posterior distribution over rewards. In apprenticeship learning, this posterior over rewards is then used to produce a policy that should perform well with respect to the unknown reward function.

However, getting human demonstrations requires scarce human time. Also, many risky situations where we would wish AI systems to behave especially reliably may be rare in naturally occurring demonstration data. Bayesian active learning can help with both by giving queries to a human demonstrator that are likely to bring the most useful information about the reward.

Prior methods for active IRL each suffer from significant limitations: Metelli et al. (2021) provide a largely theoretical treatment assuming perfectly optimal expert demonstrations, which not only is a strong assumption, but also hinders identifiability. None of the methods that can address noisy demonstrations provide theoretical guarantees. Furthermore, most methods (Lopes et al., 2009; Brown et al., 2018; Metelli et al., 2021) query the expert for action annotations of particular isolated states. However, in domains such as autonomous driving with a high frequency of actions, it can be much more natural for the human to provide whole trajectories – say, to drive for a while in a simulator – than to annotate a large collection of unrelated snapshots. There is one previous paper on *active IRL with full trajectories* (Kweon et al., 2023) suggesting a heuristic acquisition function whose shortcomings can, however, completely prevent learning, as we will demonstrate. Instead, we propose using the principled tools of Bayesian active learning, formulate two methods that can query for full trajectories, and provide theoretical guarantees for one of them. While in this paper, we work in the setting of finite state and action spaces, the methods are designed to suitably generalize to continuous settings, which we plan in future work.

The article provides the following contributions:

1. We explain and demonstrate failure modes of existing heuristic methods for active IRL when the goal is to produce a well-performing apprentice policy. In particular, most previous methods are limited to querying for only a single state annotation, as opposed to whole trajectories. Furthermore, we show that the only prior method designed for querying whole trajectories can result in repeatedly querying a single uninformative state forever.
2. We propose PAC-EIG, an acquisition function based on expected information gain (EIG) that directly targets *probably approximately correct* (PAC) guarantees for the apprentice policy – providing the first such theoretical guarantee for active IRL with imperfect expert demonstrations.
3. We present Reward-EIG as an alternative when learning the reward itself is the primary objective.
4. We prove convergence bounds showing the expected number of expert demonstrations needed to achieve PAC guarantees.
5. We illustrate the performance of our methods in a set of gridworld experiments, demonstrating their effectiveness compared to prior heuristic approaches.

## 2 Task formulation

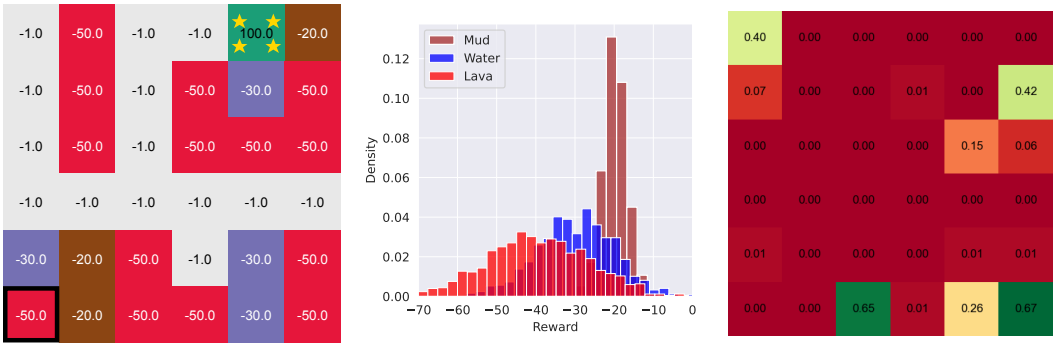
Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma, t_{\max}, \rho)$ be a parameterized Markov decision process (MDP), where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces respectively, $p : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the transition function where $\mathcal{P}(\mathcal{S})$ is a set of probability measures over $\mathcal{S}$, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is an (expected) reward function,[1] $\gamma \in (0, 1)$ is a discount rate, $t_{\max} \in \mathbb{N} \cup \{\infty\}$ is the time horizon, and $\rho$ is the initial state distribution. We assume the learner has full knowledge of the MDP except for the reward.

We assume we are initially uncertain about the reward $r$, and our initial knowledge is captured by a prior distribution $p(r)$ over rewards, which is a distribution over $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ – a space of vectors representing the reward associated with each state-action pair (there may be fewer reward parameters than $|\mathcal{S}||\mathcal{A}|$, but that can be seen as a special case). We also have access to an expert that, given an initial state $s_0$ of the MDP, can produce a trajectory $\tau_i = \big((s_0^i, a_0^i), \ldots, (s_{n_i}^i, a_{n_i}^i)\big)$, where $s_0^i \sim \rho$, $s_{t+1} \sim p(\cdot|s_t, a_t)$, and

$$\pi^E(a_t|s_t) = \frac{\exp(\beta Q^*(s_t, a_t))}{\sum_{a' \in \mathcal{A}} \exp(\beta Q^*(s_t, a'))}\,, \tag{1}$$

which is called a *Boltzmann-rational* policy, given the optimal Q-function $Q^*$ and a coefficient $\beta$ expressing how close to optimal the expert behaviour is (where $\beta = 0$ corresponds to fully random

[1] Our formulation permits the reward to be stochastic. However, our expert model (1) depends on the rewards only via the optimal Q-function, which in turn depends only on the expected reward. Thus, the demonstrations can only ever give us information about the expectation. Throughout the paper, the learnt reward function can be interpreted either as modeling a deterministic reward, or an expectation of a stochastic reward.




(a) Ground-truth rewards. (b) Current belief over rewards. (c) Reward EIG of each initial state.

Figure 1: Illustration of the active IRL task. (a) shows a gridworld and its true rewards. The lower left corner has a "jail" state with negative reward from which an agent cannot leave. The starred green state is the terminal "goal" state with a large positive reward. The brown, blue, and red states are "mud", "water", and "lava" type states respectively, whose rewards are unknown to the IRL agent. The IRL agent tries to learn the rewards of these three state types from expert demonstrations. (b) shows current distributions over the rewards of the "mud", "water", and "lava" state types respectively, at some particular step of the active learning process. These learned reward distributions are used to calculate an acquisition function (here the reward EIG) of obtaining another expert demonstration starting from each given state, shown in (c). In this case, a demonstration starting in the bottom right state gives the most information about the unknown reward parameters.

behaviour and $\beta \to +\infty$ would yield the optimal policy). We assume $\beta$ is known as is usual in related IRL literature (Ramachandran & Amir, 2007; Chan & van der Schaar, 2021; Kweon et al., 2023; Bajgar et al., 2024). We will also denote by $\pi_r^{\mathrm{E}}$ the hypothetical expert policy that would correspond to a reward $r$.

The task of *Bayesian active inverse reinforcement learning* is to sequentially query the expert to provide demonstrations from initial states $\xi_1, \ldots, \xi_N \in \mathcal{S}$ to gain maximum information about the unknown reward.[2] We start with a (possibly empty) set of expert trajectories $\mathcal{D}_0$ and then, at each step of active learning, we choose an initial state $\xi_i$ for the MDP, from which we get the corresponding expert trajectory $\tau_i$. We then update our demonstration dataset to $\mathcal{D}_i = \mathcal{D}_{i-1} \cup \{\tau_i\}$, and the distribution over rewards to $p(r|\mathcal{D}_i)$, which we again use to select the most informative initial state $\xi_{i+1}$ in the next step. We repeat until we exhaust our limited demonstration budget $N$.

This can be done with one of two possible objectives in mind.

The first, which we call the *reward-learning objective*, is relevant when our primary interest is in the reward itself, e.g. when using IRL to understand the motivations of mice in a maze (Ashwood et al., 2022) or the preferences of drivers (Huang et al., 2022). In the active setting, we operationalize this objective as trying to minimize the entropy of the posterior distribution over rewards, once all expert demonstrations have been observed. This is equivalent to maximizing the log likelihood of the true parameter value in expectation, or to maximizing the mutual information between the demonstrations and the reward. Figure 1 illustrates Active IRL with this objective.

The second possible objective, which we term the *apprenticeship-learning objective*, uses the final posterior $p(r|\mathcal{D}_N)$ to produce an *apprentice policy* that should perform well in the MDP. One option may be to optimize the expected return of the apprentice policy, i.e. solve for $\mathrm{argmax}_\pi \mathbb{E}_{r|\mathcal{D}_N}[\mathbb{E}_\tau[\sum_{s_t,a_t\in\tau} \gamma^t r(s_t,a_t)]]$, where $\tau$ is a trajectory with $s_0 \sim \rho$, $s_{t+1} \sim p(\cdot|s_t,a_t)$ and $a_t = \pi(s_t)$. The argmax can be resolved by solving the forward planning problem for finding the optimal policy for the expected reward with respect to the learner's current posterior over rewards

[2] Since the queries $\xi_i$ in this paper are limited to the choice of the initial state $s_0$, $\xi$ and $s_0$ are used somewhat interchangeably throughout the paper.

(e.g. using generalized policy iteration (Sutton & Barto, 2018)). Going forward, we generally assume a deterministic apprentice policy, i.e. the class of policies we search over is the set of mappings $\pi : \mathcal{S} \to \mathcal{A}$, though stochastic policies could easily be accommodated as well. The apprentice policy will thus be distinct from the stochastic (Boltzmann rational) expert policy and with enough knowledge can have higher expected return, since the expert gives non-zero probability to sub-optimal actions.

However, maximising expected return may not be sufficient in safety-critical domains. We may instead require a *reliable* apprentice policy that performs well with high probability – formally, one that is $\epsilon$-$\delta$-*probably approximately correct* (PAC). This means finding an apprentice policy $\pi^{\text{A}}$ such that the probability (with respect to the reward posterior) of the expected return (with respect to initial state and transition distributions) being at least $G^* - \epsilon$ is at least $1 - \delta$, where $G^*$ is the expected return of the optimal policy.

These objectives are often closely connected – learning about the reward function enables improving the apprentice policy. However, espcially in the active setting, they can come apart – for instance, once we know an action $a$ leads to lower return than $a'$ in a particular state, we may no longer need to gather further information about rewards in these states for the apprenticeship learning objective as we already know to choose the better action, while the reward-learning objective may motivate further queries to reduce the reward uncertainty.

Stemming from a common inspiration in Bayesian active learning, we will present an acquisition function tailored to each of these objectives.

**Notation** By $V_r^\pi$ we denote the state-value function of policy $\pi$ with respect to reward $r$. $V_r^*$ is then the value function of the optimal policy with respect to $r$. A lack of subscript, as in $V^*$, indicates (optimal) value with respect to the true reward (since the true reward is not known by the learner, this generally needs to be treated as random variable). By $G_r(\tau)$ we denote the return of trajectory $\tau$ with respect to $r$. By $R_r^\pi(s_0)$ we denote the regret of policy $\pi$ starting from state $s_0$, i.e. $R_r^\pi(s_0) := V_r^*(s_0) - V_r^\pi(s_0)$, and $R_r^\pi := \mathbb{E}_{s_0 \sim \rho} R_r^\pi(s_0)$. We also call *immediate regret* the quantity $R_{\pi,r}^*(s) = V^*(s) - Q^*(s, \pi(s))$ and also denote by $R_{\pi,r}^*(s,a) = \max\{0, Q^*(s,a) - Q^*(s,\pi(s))\}$ the immediate regret relative to action $a$ in state $s$. You can also find a table with the notation used in this paper in Appendix F.

## 3 Related work

IRL was first introduced by Russell (1998), preceded by the closely related problem of *inverse optimal control* formulated by Kalman (1964). See Arora & Doshi (2021) and Adams et al. (2022) for recent reviews of the already extensive literature on IRL. In our work we build upon the Bayesian formulation of the problem introduced by Ramachandran & Amir (2007).

We will now summarize prior work on *active* IRL in particular. We first describe a number of methods that query for single state annotations (which can be cast into our framework from Section 2 as trajectories of length one), and then describe the one previous method which queries for whole trajectories. Lastly, we review a few other works for setups not directly comparable to ours.

### 3.1 Active learning with single action annotations

The concept of active IRL was first introduced by Lopes et al. (2009). The authors propose an acquisition function equal to the entropy of the posterior predictive distribution about the Boltzmann expert policy, i.e. they query a state maximizing $\alpha_n^{\text{Lopes}}(s) = H(\Pi_s | \mathcal{D}_n)$ where $\Pi_s$ is the vector of expert action probabilities in state $s$ (according to the posterior predictive distribution).

An issue with this approach is that $H(\Pi_s|\mathcal{D}_n)$ does not take into account the effect of improved knowledge on the apprentice policy. For example, we may know the optimal action in a particular state, but with high uncertainty about the exact action probabilities, while another state may have

uncertainty about the optimal action, but lower entropy about exact probabilities of actions. Then, $\alpha_n^{\text{Lopes}}(s)$ would prioritize the latter, which may be suboptimal from the apprenticeship learning perspective. See Appendix A for a full example.

Brown et al. (2018) query the expert by maximizing the $\delta$-value-at-risk of the policy loss (i.e. regret) of the current apprentice policy starting from the given initial state, computed as

$$\alpha_n^{\text{Brown}}(s) = \text{VaR}_\delta \left( V^{\pi^*}(s) - V^{\pi^A}(s) | \mathcal{D}_n \right) . \tag{2}$$

This is a risk-aware approach: the states with a high risk of the apprentice action being much worse than the expert's action are queried. A limitation of this approach is that regret attributed to some initial state $s$ may be due to a choice made further along the trajectory where an expert query would be more informative as shown in Appendix A.

### 3.2 Active learning with full trajectories

Kweon et al. (2023) query full trajectories with a starting state $s_0$ chosen to maximize

$$\alpha_n^{\text{Kweon}}(s_0) = \mathbb{E}_{\tau \sim \hat{\pi}_E^{\mathcal{D}_n}} \left[ \sum_{s_t \in \tau} \tilde{\alpha}_n(s_t) | s_0 \right], \tag{3}$$

where

$$\tilde{\alpha}_n(s) := H(\hat{\pi}_E^{\mathcal{D}_n}(a|s)) := \sum_a -\hat{\pi}_E^{\mathcal{D}_n}(a|s) \log \hat{\pi}_E^{\mathcal{D}_n}(a|s),$$

is the entropy of $\hat{\pi}_E^{\mathcal{D}_n}$, the posterior predictive distribution over the expert actions at state $s$, estimated from demonstration data $\mathcal{D}_n$.

However, note that this action entropy can remain high even in states where we have perfect knowledge, but multiple actions are equally good, so the Boltzmann rational policy chooses them with equal probabilities, resulting in high action entropy. However, querying for extra demonstrations in such states will bring no useful knowledge. In fact, this can result in learning getting completely stuck, sometimes right at the beginning, preventing *any* learning from taking place. This is the case in the jail environment in Figure 1, and we show this in Section 6.

### 3.3 Other settings

Instead of querying at arbitrary states, Losey & O'Malley (2018) and Lindner et al. (2022) synthesize a policy that explores the environment to produce a trajectory which subsequently gets annotated by the expert. We instead let the expert produce the trajectory. Buening et al. (2024) query full trajectories in the context of IRL, where the active component arises in the choice of a transition function from a set of transition functions at each step. Büning et al. (2022) also query full trajectories in a different context involving two cooperating autonomous agents. In Sadigh et al. (2017), the expert is asked to provide a relative preference between two sample trajectories synthesized by the algorithm. While this generally provides less information per query than our formulation, it is a useful alternative for situations where providing high-quality demonstrations is difficult for humans.

On the side of theoretical sample complexity of (active) IRL, all prior work assumes a perfectly rational expert policy, which is a stronger assumption than our Boltzmann rationality. In particular, seeing each state once is enough to determine the optimal policy. The first lower bound on the complexity of IRL was given by Komanduru & Honorio (2021) for the case of a $\beta$-separable finite set of candidate rewards. Metelli et al. (2021), Lindner et al. (2022), and Metelli et al. (2023) then focus on recovering a feasible reward set in settings where also the transition dynamics are only estimated, and address the problem of the transferability of the learnt reward to environments with different dynamics.

# 4 Method

We propose PAC-EIG, an acquisition function based on expected information gain (EIG) that aims to produce a probably approximately correct (PAC) apprentice policy. Our approach builds on principled Bayesian experimental design (Rainforth et al., 2023) to identify initial states for expert demonstrations that will yield information about the immediate regret of the apprentice policy. The intuition is that knowing about the regret in various states allows us to identify high-regret states where the apprentice policy can be improved. We then also show how EIG can be adapted to the reward-learning objective, resulting in the Reward-EIG acquisition function.

## 4.1 PAC-EIG: Information Gain for Reliable Policies

Our goal is to produce an apprentice policy that is probably approximately correct – that is, with high probability $(1-\delta)$, the policy's regret is bounded by $\epsilon$. To achieve this efficiently, we need to identify states where the current apprentice policy might be making poor decisions. To this end, we define the *immediate regret* of an apprentice policy $\pi^A$ in state $s$ as:

$$R^*_{\pi^A,r}(s) = V^*_r(s) - Q^*_r(s, \pi^A(s)) \tag{4}$$

which captures how much value we lose by following the apprentice policy in state $s$ compared to optimal behavior. This can be decomposed per action as $R^*_{\pi^A,r}(s,a) = \max\{0, Q^*_r(s,a) - Q^*_r(s,\pi^A(s))\}$, representing the regret relative to choosing a particular alternative action $a$. This regret is unknown to us, so we need to treat it as a random variable.

For computational tractability, we discretize the immediate regret into a ternary variable $E_{s,a}$ tracking state- and action-wise correctness as follows:

$$E^{\pi^{\mathrm{A}}}_{s,a} = \begin{cases} \text{“strongly approximately correct”} & \text{if } R^*_{\pi^A,r}(s,a) < \epsilon(1-\gamma)/2 \\ \text{“weakly approximately correct”} & \text{if } \epsilon(1-\gamma)/2 \le R^*_{\pi^A,r}(s,a) \le \epsilon(1-\gamma) \\ \text{“not approximately correct”} & \text{if } R^*_{\pi^A,r}(s,a) > \epsilon(1-\gamma) \end{cases} \tag{5}$$

Our acquisition function then maximizes the expected information gain about these discretized regret values:

$$\alpha^{\text{PAC-EIG}}_n(s_0) := I(\tau; E^{\pi^{\mathrm{A}}} | s_0, \mathcal{D}_n) \tag{6}$$

where $E^{\pi^{\mathrm{A}}} = (E^{\pi^{\mathrm{A}}}_{s,a})_{s\in\mathcal{S}, a\in\mathcal{A}}$ represents the discretized regret across all state-action pairs, and $\tau$ is the expert trajectory starting from $s_0$. Note that if the apprentice policy is approximately correct in all states with probability at least $1-\delta$ in the immediate regret sense, then we also satisfy the PAC criterion globally.

While we are aiming for a $\epsilon$-$\delta$-PAC policy as the final output, for intermediate steps, our algorithm uses an apprentice policy $\pi^{\mathrm{A}}(s) = \text{argmax}_a \mathbb{P}[Q^*(s,\pi^*(s)) - Q^*(s,a) < \epsilon(1-\gamma)/2 | \mathcal{D}_n],\ \forall s \in \mathcal{S}$. Once the probability of this state-wise criterion on immediate regret holding in every state is higher than $1-\delta$, this is a sufficient condition for this policy being $\epsilon/2$-$\delta$-PAC (though it is not a necessary condition – an $\epsilon/2$-$\delta$-PAC will usually be found earlier). The reason for this choice will become clear in Section 5. Especially if we care about gradually tightening the bound (rather than just reaching a particular bound) we may wish to dynamically set $\epsilon$ in each step to just below the threshold for which the condition is currently satisfied.

## 4.2 Computing PAC-EIG

To compute PAC-EIG in practice, we leverage our Bayesian IRL posterior over Q-values. Given $\mathcal{Q}_n$, a set of $M$ samples from $p(Q^*|\mathcal{D}_n)$, we can:

1. For each Q-value sample, compute the discretized regret values $E^{\pi^{\mathrm{A}}}_{s,a}$ for all state-action pairs. Note that multiple Q-value samples may map to the same discretized configuration $E^{\pi^{\mathrm{A}}}$, so there are at most $M_E \le M$ distinct values of $E^{\pi^{\mathrm{A}}}$.

2. Given a Q-value sample $Q_i^*$, sample expert trajectories $\tau$ starting from $s_0$ using the Boltzmann policy corresponding to $Q_i^*$.
3. Estimate the expected information gain using the standard Monte Carlo estimator as

$$\alpha_n^{\text{PAC-EIG}}(s_0) \approx \frac{1}{M} \sum_{i=1}^{M} \left[ \log p(\tau^{(i)} | E^{\pi^{\text{A}},(i)}, s_0) - \log p(\tau^{(i)} | s_0) \right] \tag{7}$$

where the trajectory probability given $E^{\pi^{\text{A}}}$ can be computed as $p(\tau | E^{\pi^{\text{A}}}, s_0) = \prod_{(s_t,a_t)\in\tau} p(a_t | s_t, E)$ (omitting the transition probabilities since they would cancel out in the log-ratio). To compute $p(a_t | s_t, E^{\pi^{\text{A}}})$, we average the expert action probabilities over all Q-value samples that map to the same discretized configuration: $p(a_t | s_t, E^{\pi^{\text{A}}}) = \frac{1}{|\mathcal{Q}_E^{\pi^{\text{A}}}|} \sum_{Q^* \in \mathcal{Q}_E} p(a_t | s_t, Q^*)$, where $\mathcal{Q}_E$ denotes the set of Q-value samples corresponding to the discretized regret configuration $E^{\pi^{\text{A}}}$.

See Appendix C for additional implementation details.

### 4.3 Reward EIG: When Learning the Reward is the Goal

While our primary focus above has been on producing reliable apprentice policies, in some applications the reward function itself is of intrinsic interest – for instance, when using IRL to understand animal behavior (Ashwood et al., 2022) or human preferences (Huang et al., 2022). For these cases, we can still use the EIG framework, but instead maximize the expected information gain about the reward:

$$\alpha_n^{\text{Reward-EIG}}(s_0) := I(\tau; r | s_0, \mathcal{D}_n) \tag{8}$$

where $\tau$ is treated as a random variable representing the expert's trajectory that would be produced starting from $s_0$.

This acquisition function aims to reduce posterior uncertainty about the reward parameters, which may query different states than PAC-EIG. For example, it might seek to precisely estimate reward values in states that the apprentice already knows to avoid, whereas PAC-EIG would consider such queries unnecessary.

The reward EIG can be computed as:

$$\alpha_n^{\text{Reward-EIG}}(s_0) = \mathbb{E}_{r|\mathcal{D}_n} \left[ \mathbb{E}_{\tau|r,s_0} [\log p(\tau | r, s_0) - \log p(\tau | s_0; \mathcal{D}_n)] \right] \tag{9}$$

where the inner expectation is tractable to compute from the Q-values that are usually obtained as a byproduct of a Bayesian IRL algorithm.

See Appendix D for a detailed discussion of alternative acquisition functions and the theoretical connections between different formulations.

## 5 Producing a PAC Policy

Our PAC-EIG acquisition function is designed to efficiently produce a probably-approximately-correct (PAC) apprentice policy. We will show that PAC-EIG leads to such a policy by establishing bounds on the expected number of expert demonstrations needed. The analysis proceeds through three key steps (with formal results and proofs in Appendix B):

1. If no apprentice policy satisfies the PAC condition, there must exist a state where there is a significant chance that any apprentice policy makes a significantly suboptimal choice – specifically:

$$\mathbb{P}_{r|\mathcal{D}_n} \left[ V_r^*(s) - Q_r^*(s, \pi^{\text{A}}(s)) \geq (1-\gamma)\epsilon \right] \geq \frac{\delta}{|\mathcal{S}|} \qquad \text{(Lemma 2).}$$

2. In such a state, there is both a chance that the apprentice policy $\pi^{\mathrm{A}}$ is close to optimal and that it is significantly suboptimal (as defined in step 1). Since these two options would result in a sufficiently different expert policies in this state, we can gain a lower-bounded expected amount of information by observing the expert in that state.
3. Since we gain at least this minimum information per query while the PAC condition is unmet, and our initial uncertainty is finite, we must eventually achieve the PAC condition. The number of steps is bounded by the ratio of initial entropy to the minimum information gain per step.

These insights translate into the following two theorems:

**Theorem 1.** *For $\epsilon > 0$ and $\delta \in (0, \frac{1}{2}]$, assume that no policy $\pi$ is $(\epsilon, \delta)$-probably-approximately-correct, i.e., $\mathbb{P}[R_r^\pi \geq \epsilon] > \delta$, $\forall \pi$. Then, there exists a state $s \in \mathcal{S}$ such that observing a new expert demonstration at $s$ has an expected information gain with respect to the variable $E^{\pi^A}$ (Eq. 5) of at least*

$$EIG_{min}(\epsilon, \delta) = \frac{\delta e^{-\beta(1-\gamma)\epsilon}(1 - e^{-\beta(1-\gamma)\epsilon/2})^2}{4|\mathcal{A}|^5|\mathcal{S}|}. \tag{10}$$

Then, we can translate this into the following result on the expected number of steps to reach the PAC criterion:

**Theorem 2.** *Let $h_{max}$ be an upper bound on the entropy of the prior distribution over $E$, the PAC-EIG discretized regret values $E^{\pi^A}$ aggregated across all apprentice policies $\pi^A$. Then, the expected number of steps needed to reach the PAC condition is upper bounded by*

$$\frac{h_{max}}{EIG_{min}(\epsilon, \delta)} = \frac{4h_{max}|\mathcal{A}|^5|\mathcal{S}|}{\delta e^{-\beta(1-\gamma)\epsilon}(1 - e^{-\beta(1-\gamma)\epsilon/2})^2}. \tag{11}$$

For the ternary discretization used in PAC-EIG, $h_{\max} \leq \log(3)|\mathcal{S}||\mathcal{A}|^2$, giving a concrete bound on sample complexity.

**Extension to trajectory queries.** While the theorems establish a lower bound for single-state queries, this naturally extends to a trajectory-based version of our PAC-EIG acquisition function. When querying for a trajectory starting from state $s_0$, the information gained is at least as large as querying any single state visited along the trajectory. In practice, trajectories typically visit multiple informative states, potentially providing substantially more information than the theoretical lower bound suggests. However, an improved theoretical bound would need to build on additional assumptions about the environment and the prior.[3]

## 6 Experiments

We evaluated the performance of the two proposed acquisition functions in a set of gridworld experiments with respect to both objectives introduced earlier: the reward learning objective, measured by the entropy of the posterior distribution over rewards, and the apprenticeship learning objective, measured by the regret. We also track the posterior distribution over regrets which directly relates to the PAC criterion.

We evaluate across three types of environment:

1. **Structured gridworld**: Features fewer reward parameters than states. It includes a known goal state with a reward of +100, neutral states with a reward of -1, and three obstacle types with unknown negative rewards with a uniform prior between -100 and 0 independently for each obstacle type. This was meant as an illustrative example (Figure 1) and a counterexample to the only prior method designed for collecting full trajectories, *action entropy* (Kweon et al., 2023), by including a jail state where all actions are equivalent and which always gets selected by this baseline, thus preventing any useful learning.

[3]For example, the environment may terminate after a single step, forcing the EIG from a trajectory to that of a single state.

2. **10x10 random gridworld with 2 initial states**: Each state has a random reward drawn from the prior, $\mathcal{N}(0, 3)$, with only two possible initial states to test the ability of methods to recognize only relevant parts of the state space. Here we use $\beta = 4$ so the expert behaves closely to optimal.
3. **8x8 random gridworld with fully uniform initial states**: Each state has a random reward drawn from the prior, $\mathcal{N}(0, 3)$, and the initial state distribution is uniform across all states. We use $\beta = 2$ so the expert is fairly stochastic.

We evaluate both the setting where each query results in a full expert trajectory, where we compare against the only prior method (Kweon et al., 2023) as well as random sampling, and the setting where each query results in a single state-action annotation, where we also evaluate against the methods by Lopes et al. (2009) and Brown et al. (2018).

Each experiment type was run with with 16 different random reward functions, different terminal states (except for the jail environment) and different 2 initial states in the 10x10 environment. The plots display the mean and the standard error across these 16 random instances.

### 6.1 Results

Results on the simple environment from Figure 1 illustrate a crucial failure mode of the *action entropy* (Kweon et al., 2023) acquisition function – it always queries the jail state and thus fails to learn anything useful, while both reward and regret EIG learn an optimal policy within 10 steps in all 16 instances with similar posterior entropies.

Figure 3 shows the results on the 10x10 gridworld with 2 random initial states and single-state annotations, Figure 4 shows the results for querying single-state annotations on the 8x8 gridworld with a uniform initial state distribution, and Figure 5 results on the 8x8 environment when querying trajectories of maximum length 5.

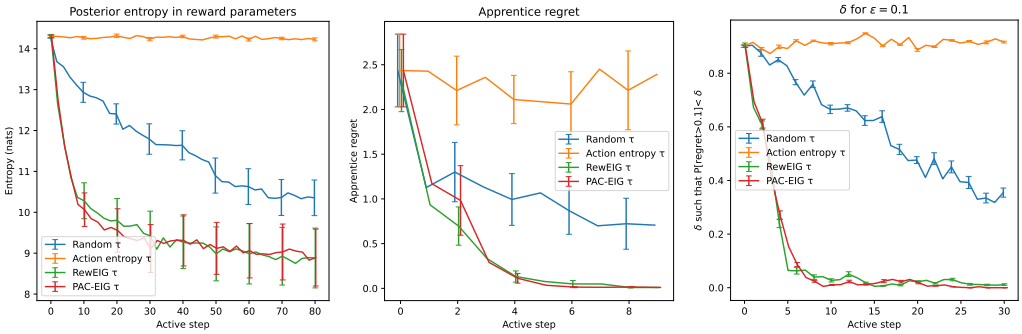
In the case of only 2 initial states on the 10x10 gridworld, we can see our regret-focused PAC-EIG acquisition function to do much better in terms of both actual and posterior regret, reaching a zero true regret, as well as 0.1-0.1-PAC apprentice policy, by step 50. ActiveVaR remains competitive with RewardEIG in terms of the reward-learning objective (entropy of the rewardposterior), but both fall behind PAC-EIG in terms of true and posterior regret.

On the gridworld with fully uniform initial states, we observe that both our information theoretic acquisition functions result in lower posterior reward entropy *and* lower regret than prior methods except for ActiveVaR, which seems to roughly match their performance. Interestingly, the reward-based acquisition function and the regret-based ones seem to perform similarly well on both objectives, suggesting that there is a strong correlation between learning about the reward and learning about the apprentice regret in this environment with uniform initial states.

The action entropy acquisition function still stops yielding significant improvements after about step 50 – it again gets stuck querying states that have high action entropy due to multiple actions being similarly good, even if these states do not yield any more information.

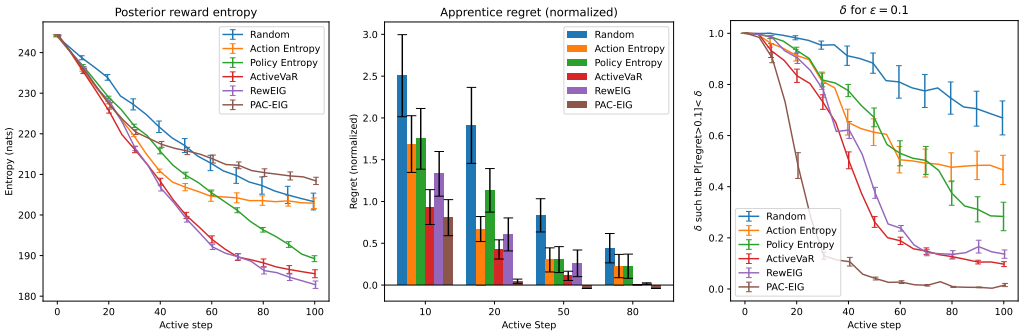
## 7 Discussion and conclusion

In this paper we have proposed new acquisition functions for active IRL, each geared toward one of two possible objectives: learning about an unknown reward function, or producing a well-performing apprentice policy. We have shown that across a set of gridworld experiments, our acquisition functions outperform or at least match prior methods on their respective objective. Furthermore, our immediate-regret EIG acquisition function is a first acquisition function with a regret bound in our setting. While we have so far tested the methods only in finite state spaces, both of them were constructed to generalize also to continuous spaces, which will be addressed in future work.




(a) Posterior reward entropy (b) True regret (c) Posterior prob. of regret $> 0.1$

Figure 2: Results of the experiments on the environment with 3 cell types and a jail state with full-trajectory demonstrations.




(a) Posterior reward entropy (b) True regret (c) Posterior prob. of regret $> 0.1$

Figure 3: Results of the experiments with single state annotations (i.e. $|\tau| = 1$) on the 10x10 fully random gridworld with two initial states. In the barplot (b), results with zero regret are visualized below the horizontal axis to make their presence clearer.

## Impact statement

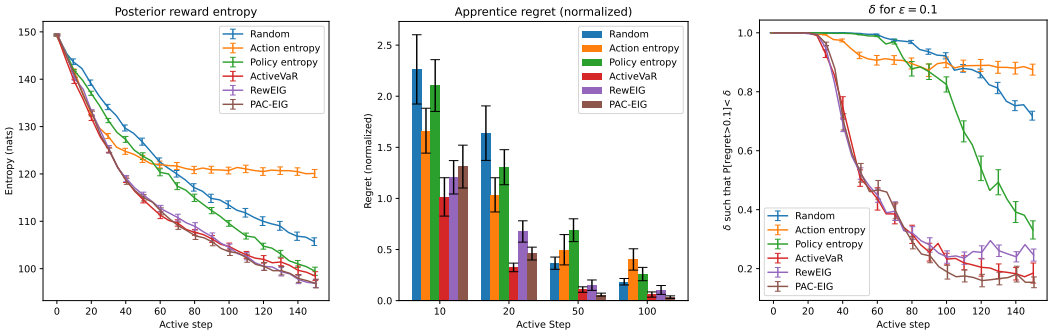
Through this paper, we hope to contribute to more effective and reliable learning of human preferences and values by AI systems, which aims to improve their alignment and facilitate their beneficial use.

## References

Stephen Adams, Tyler Cody, and Peter A. Beling. A survey of inverse reinforcement learning. *Artificial Intelligence Review*, February 2022. ISSN 0269-2821, 1573-7462. DOI: 10.1007/s10462-021-10108-x. URL https://link.springer.com/10.1007/s10462-021-10108-x.

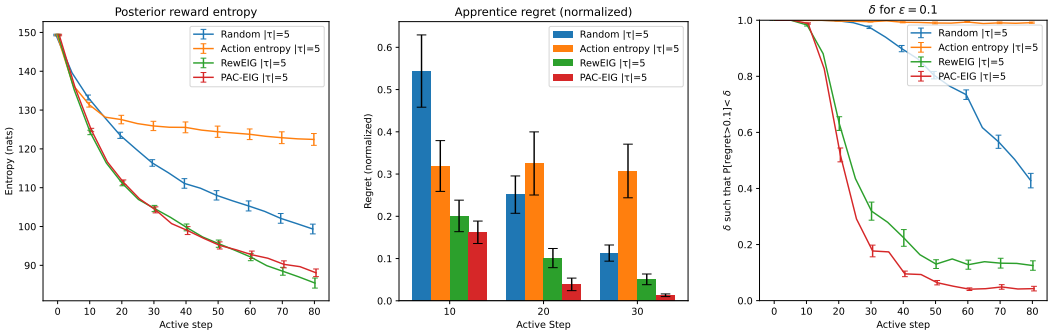
Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, August 2021. ISSN 00043702. DOI: 10.1016/j.artint.2021.103500. URL https://linkinghub.elsevier.com/retrieve/pii/S0004370221000515.

Zoe Ashwood, Aditi Jha, and Jonathan W. Pillow. Dynamic Inverse Reinforcement Learning for Characterizing Animal Behavior. *Advances in Neural Information Processing Systems*, 35:29663–29676, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/bf215fa7fe70a38c5e967e59c44a99d0-Abstract-Conference.html.




(a) Posterior reward entropy (b) True regret (c) Posterior prob. of regret $> 0.1$

Figure 4: Results of the experiments with single state annotations (i.e. $|\tau| = 1$) on the 8x8 fully random gridworld.




(a) Posterior reward entropy (b) True regret (c) Posterior prob. of regret $> 0.1$

Figure 5: Results of the experiments with expert trajectories of maximum length $|\tau| = 5$ on the 8x8 fully random gridworld.

Ondrej Bajgar, Konstantinos Gatsis, Alessandro Abate, and Michael A. Osborne. Walking the Values in Bayesian Inverse Reinforcement Learning. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Daniel S. Brown, Yuchen Cui, and Scott Niekum. Risk-Aware Active Inverse Reinforcement Learning. In *Proceedings of The 2nd Conference on Robot Learning*, pp. 362–372. PMLR, October 2018. URL https://proceedings.mlr.press/v87/brown18a.html. ISSN: 2640-3498.

Thomas Kleine Buening, Victor Villin, and Christos Dimitrakakis. Environment Design for Inverse Reinforcement Learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 24808–24828. PMLR, July 2024. URL https://proceedings.mlr.press/v235/kleine-buening24a.html. ISSN: 2640-3498.

Thomas Kleine Büning, Anne-Marie George, and Christos Dimitrakakis. Interactive Inverse Reinforcement Learning for Cooperative Games. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 2393–2413. PMLR, June 2022. URL https://proceedings.mlr.press/v162/buning22a.html. ISSN: 2640-3498.

Alex J Chan and Mihaela van der Schaar. Scalable Bayesian Inverse Reinforcement Learning. *ICLR 2021*, 2021.

Simon Duane, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, September 1987. DOI: 10.1016/0370-2693(87)91197-X.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

Zhiyu Huang, Jingda Wu, and Chen Lv. Driving Behavior Modeling Using Naturalistic Human Driving Data With Inverse Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10239–10251, August 2022. ISSN 1558-0016. DOI: 10.1109/TITS.2021.3088935. URL https://ieeexplore.ieee.org/abstract/document/9460807. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

R. E. Kalman. When Is a Linear Control System Optimal? *Journal of Basic Engineering*, 86(1): 51–60, March 1964. ISSN 0021-9223. DOI: 10.1115/1.3653115. URL https://doi.org/10.1115/1.3653115.

Abi Komanduru and Jean Honorio. A Lower Bound for the Sample Complexity of Inverse Reinforcement Learning. 2021.

Sehee Kweon, Himchan Hwang, and Frank C. Park. Trajectory-Based Active Inverse Reinforcement Learning for Learning from Demonstration. In *2023 23rd International Conference on Control, Automation and Systems (ICCAS)*, pp. 1807–1812, October 2023. DOI: 10.23919/ICCAS59377.2023.10316798. URL https://ieeexplore.ieee.org/document/10316798. ISSN: 2642-3901.

David Lindner, Andreas Krause, and Giorgia Ramponi. Active Exploration for Inverse Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35:5843–5853, December 2022. URL https://proceedings.neurips.cc/paper/2022/hash/26d01e5ed42d8dcedd6aa0e3e99cffc4-Abstract-Conference.html.

Manuel Lopes, Francisco Melo, and Luis Montesano. Active learning for reward estimation in inverse reinforcement learning. In Wray Buntine, Marko Grobelnik, Dunja Mladenić, and John Shawe-Taylor (eds.), *Machine learning and knowledge discovery in databases*, pp. 31–46, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-04174-7.

Dylan P. Losey and Marcia K. O'Malley. Including Uncertainty when Learning from Human Corrections. In *Proceedings of The 2nd Conference on Robot Learning*, pp. 123–132. PMLR, October 2018. URL https://proceedings.mlr.press/v87/losey18a.html. ISSN: 2640-3498.

Alberto Maria Metelli, Giorgia Ramponi, Alessandro Concetti, and Marcello Restelli. Provably Efficient Learning of Transferable Rewards. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7665–7676. PMLR, July 2021. URL https://proceedings.mlr.press/v139/metelli21a.html. ISSN: 2640-3498.

Alberto Maria Metelli, Filippo Lazzati, and Marcello Restelli. Towards Theoretical Understanding of Inverse Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 24555–24591. PMLR, July 2023. URL https://proceedings.mlr.press/v202/metelli23a.html. ISSN: 2640-3498.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian Experimental Design, February 2023. URL http://arxiv.org/abs/2302.14545. arXiv:2302.14545 [cs, stat].

Deepak Ramachandran and Eyal Amir. Bayesian Inverse Reinforcement Learning. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 2007.

Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 101–103, Madison Wisconsin USA, July 1998. ACM. ISBN 978-1-58113-057-7. DOI: 10.1145/279943.279964. URL https://dl.acm.org/doi/10.1145/279943.279964.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Random House, 2019.

Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems XIII*. Robotics: Science and Systems Foundation, July 2017. ISBN 978-0-9923747-3-0. DOI: 10.15607/RSS.2017.XIII.053. URL http://www.roboticsproceedings.org/rss13/p53.pdf.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Zi Wang and Stefanie Jegelka. Max-value Entropy Search for Efficient Bayesian Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 12, 2017.

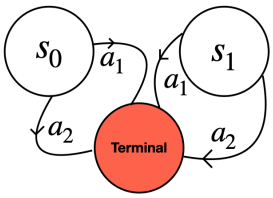



Figure 6: Two-state environment designed to illustrate a failure mode of Lopes et al. (2009)

# Supplementary Materials

*The following content was not necessarily subject to peer review.*

## A Failure modes of prior methods

For each of the three prior methods for active IRL, we will now present an example of a simple environment where the method makes a clearly suboptimal choice with respect to at least one of the two objectives.

**Policy entropy (Lopes et al., 2009)** As a reminder, the policy entropy acquisition function is $\alpha^{\text{Lopes}} = H(\pi^E)$, i.e. the entropy of the expert policy with respect to the current posterior over rewards (which induces a posterior over expert action probabilities). Consider an environment with two states $s_{0,1}$ each with two actions $a_{1,2}$ as shown in Figure 6. We aim to illustrate a scenario where $\alpha^{\text{Lopes}}$ can misallocate budget, from the point of view of the apprenticeship learning objective, by focusing on states where the optimal action is already known, rather than those where crucial information about optimality is missing.

To demonstrate this effect, we define a *discrete prior distribution over rewards*. This uncertainty in rewards will, in turn, induce a prior over the possible action probabilities for an optimal policy. Suppose in state $s_1$ we have strong prior knowledge that $a_1$ is the optimal action; however, we are uncertain about the exact reward obtained by taking $a_1$

$$P(R_{s_1,a_1} = 5) = 0.5\,, \quad P(R_{s_1,a_1} = 7) = 0.5\,, \tag{12}$$

and

$$P(R_{s_1,a_2} = 1) = 1.0\,. \tag{13}$$

An optimal apprentice policy will always choose $a_1$ in state $s_1$. Despite this certainty, the uncertainty in the exact reward for $a_1$ means there is still uncertainty regarding the precise probability an optimal policy would assign to $a_1$, which leads to a high measure of policy uncertainty (as measured by $\alpha^{\text{Lopes}}$).

For state $s_0$, we set priors

$$P(R_{s_0,a_1} = 2) = 0.1\,, \quad P(R_{s_0,a_1} = 3) = 0.9\,, \tag{14}$$

and

$$P(R_{s_0,a_2} = 2) = 0.1\,, \quad P(R_{s_0,a_2} = 3) = 0.9\,. \tag{15}$$

such that the optimal action is *uncertain*. In this state, the learner faces true ambiguity about the best action, and it is therefore the state that a good active IRL method focused on improving the apprentice policy should query. However, since the acquisition function of Lopes et al. (2009) is formulated using entropy of possible actions probabilities, examples of this type could have $\alpha^{\text{Lopes}}(s_0) < \alpha^{\text{Lopes}}(s_1)$, resulting in an inefficient use of budget. For example, given inverse temperature $\beta = 2$, we obtain values

$$\alpha^{\text{Lopes}}(s_0) = 0.860\,, \quad \alpha^{\text{Lopes}}(s_1) = 1.0\,, \tag{16}$$

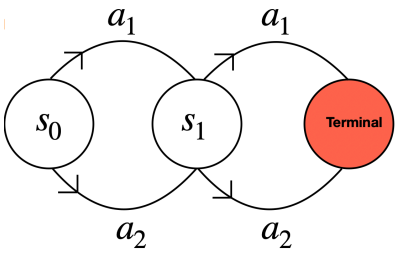



Figure 7: Two state environment to demonstrate a failure mode of Brown et al. (2018).

so this acquisition function would query $s_1$, where the policy *already knows which action is optimal*, rather than $s_0$ where there is key information to be gained. By contrast, assuming single-state queries, our PAC acquisition function would choose $s_0$, since in state the regret is already known to be 0 for the current apprentice policy so there is no regret information to be gained there.

**ActiveVaR (Brown et al., 2018)** This acquisition function for getting single-state annotations is equal to a particular quantile of the posterior distribution over regret of the apprentice policy starting from that state. In this example, we use a 0.9 quantile, though the example is robust with respect to the exact value. Consider an environment with two states labelled $s_{0,1}$ and two actions $a_{1,2}$ as shown in Figure 7. Both actions in state $s_0$ lead to $s_1$, one with reward $+2$ and the second with $-2$ (but we do not know which is which). In $s_1$, both actions lead to a terminal state, and give a reward of $-10$ and $+10$. Since the potential downside of any policy is maximal at $s_0$ ($-12$), the acquisition function would query $s_0$. On the other hand, querying state $s_1$ to distinguish the $\pm 10$ rewards would yield a greater reduction in expected regret.

To get even more concrete, consider an intermediate policy which *knows* the absolute values of all the rewards, but not the relative signs: (i.e. $(+2,-2)$ and $(-2,+2)$ are equally likely for $(r(s_0,a_1),r(s_0,a_2))$, as are $(+10,-10)$ and $(-10,+10)$ for $(r(s_1,a_1),r(s_1,a_2))$. We can easily compute

$$\alpha^{\text{Brown}}(s_0) = 2+\gamma 10\,, \quad \alpha^{\text{Brown}}(s_1) = 10\,, \tag{17}$$

so for a sufficiently large discount factor, state $s_0$ would be queried by this acquisition function. We can compute the reduction in expected regret after querying each of these states. The initial expected total regret for any apprentice policy, averaged over a uniform initial state distribution is

$$\mathbb{E}_r[R_{\pi^\wedge,r}] = \frac{1}{2}\big(V^*(s_0)-V^\pi(s_0)\big)+\frac{1}{2}\big(V^*(s_1)-V^\pi(s_1)\big) = \frac{1}{2}(2+\gamma 10-0)+\frac{1}{2}(10-0) = 6+10\gamma\,.$$

If we query the expert at $s_0$ and as a result switch the apprentice action at $s_0$ to the optimal one (which means we get a reward of 2 in $s_0$ while still getting an expectation of 0 in $s_1$, we get an expected regret of

$$\frac{1}{2}\big((2+\gamma 10)-2\big)+\frac{1}{2}\big(10-0\big) = 5+5\gamma$$

while if we query at $s_1$ we get an expected regret of

$$\frac{1}{2}\big((2+\gamma 10)-\gamma 10\big)+\frac{1}{2}\big(10-10\big) = 1\,.$$

We therefore observe that whilst Brown et al. (2018) would query $s_0$, querying $s_1$ yields a greater reduction in expected regret and also better tightening of the PAC condition (achieving it for any $\epsilon > 1$ and any $\delta$). For a sufficiently high $\epsilon$, our PAC-EIG acquisition function would correctly query $s_1$ (though for a small $\epsilon$, it would recognize that both states need to be queried to satisfy the PAC criterion and would be indifferent between them).

**Action Entropy (Kweon et al., 2023)** In its single-state-annotation version, the value of this acquisition function in a state is equal to the entropy of the posterior predictive distribution over expert actions in that state. Consider a situation where the learner has perfect knowledge of the action values in a particular state, but at least three actions in this state are equivalent (result in the same reward and next state distribution) and tied as optimal. Then the expert Boltzmann policy will assign uniform probabilities among these actions and due to high action entropy this state will always be queried in favour of any where one of two actions is best, but the learner does not know which of the two (assuming all other actions are known to be strongly suboptimal and thus unlikely to get queried). We offer an example of this in Figure 1 which renders this acquisition function useless since it will only ever query the jail cell without gaining any information (and this is the case in both the single-state-annotation version, and a trajectory based version), while both our acquisition functions keep gathering useful information and improving the both the posterior reward entropy and the regret of the apprentice policy (including the PAC bounds).

For the single-state-annotation version, an even stronger counter example applies: if we allow querying terminal states in any environment that has them, the method always queries these, since actions have no effect, so a Boltzmann rational policy would be uniform and thus have maximum entropy.

## B Theoretical Analysis

We will establish an upper bound on the expected number of expert demonstrations needed to find a policy satisfying an $\epsilon$-$\delta$-PAC criterion using the PAC-EIG acquisition function. As outlined in Section 4, we will assume that during the learning process, the apprentice policy is one maximizing the probability of being $\epsilon(1-\gamma)/2$-approximately correct in each state in the immediate regret sense, i.e.

$$\pi^{\mathrm{A}}(s) = \mathrm{argmax}_a \mathbb{P}[Q*(s,\pi^*(s)) - Q^*(s,a) < \epsilon(1-\gamma)/2 | \mathcal{D}_n],\ \forall s \in \mathcal{S} \tag{18}$$

. The proof strategy proceeds in three steps:

1. First, we show that if a policy has a significant (overall) regret, there must exist a state where the policy's action is significantly suboptimal in terms of immediate regret. (Lemma 1)
2. Building on this, we prove that if a policy is not $(\epsilon, \delta)$-PAC, then there exists a state where the difference between optimal and apprentice policy's optimal Q-values is lower-bounded by a function of $\epsilon$ with probability at least $\delta/|\mathcal{S}|$.
3. Finally, we show that in such cases, observing an expert demonstration from an appropriately chosen initial state provides a guaranteed minimum amount of information about whether the policy is approximately correct. Since we can only gain a finite amount of information (bounded by the entropy of our prior), this leads to a bound on the number of demonstrations needed.

We begin with our first lemma, which connects overall policy regret to statewise immediate regret, i.e. differences in optimal Q-values:

**Lemma 1.** *Let $\pi$ be any policy, $r$ any reward function, and*

$$R_r^\pi = \mathbb{E}_{s_0 \sim \rho_0}\left[V_r^*(s_0) - V_r^\pi(s_0)\right] \geq 0,$$

*the regret of that policy. Then there exists a state $s \in \mathcal{S}$ such that*

$$R^*_{\pi,r}(s) := Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi(s)) \geq (1-\gamma) R_r^\pi.$$

*Proof.* Let us define

$$\Delta_Q = \max_{s \in \mathcal{S}} \left[Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi(s))\right].$$

We will prove the lemma by showing that $R_r^\pi \leq \Delta_Q/(1-\gamma)$.

Since $Q_r^\pi(s, \pi(s)) \leq Q_r^*(s, \pi(s))$ (because $Q_r^*$ is the optimal Q-function), we have

$$\begin{aligned} V_r^*(s) - V_r^\pi(s) &= Q_r^*(s, \pi_r^*(s)) - Q_r^\pi(s, \pi(s)) \\ &\geq Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi(s)) \\ &\geq 0\,. \end{aligned}$$

Using the Bellman equation, for any state $s \in \mathcal{S}$ we can write

$$\begin{aligned} V_r^*(s) - V_r^\pi(s) &= Q_r^*(s, \pi_r^*(s)) - Q_r^\pi(s, \pi(s)) \\ &= Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi(s)) + Q_r^*(s, \pi(s)) - Q_r^\pi(s, \pi(s)) \\ &\leq \Delta_Q + \big(r(s, \pi(s)) + \gamma \mathbb{E}_{s'|s,\pi(s)}[V_r^*(s')]\big) - \big(r(s, \pi(s)) + \gamma \mathbb{E}_{s'|s,\pi(s)} V_r^\pi(s')]\big) \\ &= \Delta_Q + \gamma \mathbb{E}_{s'|s,\pi(s)}[V_r^*(s') - V_r^\pi(s')] \\ &\leq \Delta_Q + \gamma \max_{s'}[V_r^*(s') - V_r^\pi(s')]. \end{aligned}$$

Here, the first equality just replaces state values by the corresponding Q-values, the second line adds and subtracts the same term, the third line uses the definition of $\Delta_Q$ for the first term and expands the latter two Q-values using the Bellman equation, the fourth just cancels out the repeated reward term. The final inequality follows because the expectation over next states is bounded by the maximum.

Since this inequality holds for all $s \in \mathcal{S}$, it holds also for the state maximizing the left-hand side, so we get

$$\max_s [V_r^*(s) - V_r^\pi(s)] \leq \Delta_Q + \gamma \max_{s'} [V_r^*(s') - V_r^\pi(s')]\,, \tag{19}$$

which can be readily rearranged into

$$\max_s [V_r^*(s) - V_r^\pi(s)] \leq \frac{\Delta_Q}{1-\gamma}.$$

Thus

$$R_r^\pi = \mathbb{E}_{s_0 \sim \rho_0}\left[V_r^*(s_0) - V_r^\pi(s_0)\right] \leq \max_s [V_r^*(s) - V_r^\pi(s)] \tag{20}$$

$$\leq \frac{\Delta_Q}{1-\gamma} = \frac{1}{1-\gamma} \max_{s \in \mathcal{S}} \left[Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi(s))\right], \tag{21}$$

which completes the proof. □

**Lemma 2.** *Let $\pi$ be the apprentice policy at step $n$. For any $\delta \in (0, \frac{1}{2}]$, let $R_{n,\delta}^\pi$ be the $(1-\delta)$-quantile of the regret distribution with respect to the current posterior distribution over rewards, i.e., $R_{n,\delta}^\pi$ satisfies*

$$\mathbb{P}_{r|\mathcal{D}_n}[R_r^\pi \geq R_{n,\delta}^\pi] = \delta.$$

*Then, there exists a state $s \in \mathcal{S}$ such that*

$$\mathbb{P}_{r|\mathcal{D}_n}\left[Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi(s)) \geq (1-\gamma) R_{n,\delta}^\pi\right] \geq \frac{\delta}{|\mathcal{S}|}.$$

*Proof.* Let us define the set of reward functions under which $\pi$ has high regret:

$$\mathcal{H} = \left\{r : R_r^\pi \geq R_{n,\delta}^\pi\right\}.$$

By definition of the quantile $R_{n,\delta}^\pi$, we have

$$\mathbb{P}_{r|\mathcal{D}_n}[r \in \mathcal{H}] = \delta.$$

For each $r \in \mathcal{H}$, applying Lemma 1, we know there exists a state $s_r \in \mathcal{S}$ such that

$$Q_r^*(s_r, \pi_r^*(s_r)) - Q_r^*(s_r, \pi(s_r)) \geq (1-\gamma) R_r^\pi \geq (1-\gamma) R_{n,\delta}^\pi.$$

Let $\mathcal{S}_r$ be the set of such states for a reward function $r \in \mathcal{H}$ and $\mathcal{S}_{\mathcal{H}}$ the collection of such states across all rewards $r \in \mathcal{H}$. Since the state space $\mathcal{S}$ is finite with cardinality $|\mathcal{S}|$, by the pigeonhole principle, there must exist at least one state $s \in \mathcal{S}$ such that

$$\mathbb{P}_{r|\mathcal{D}_n}\left[r \in \mathcal{H} \text{ and } s \in \mathcal{S}_r\right] \geq \frac{\delta}{|\mathcal{S}|}.$$

For this state $s$, whenever $r \in \mathcal{H}$ and $s \in \mathcal{S}_r$, we have

$$Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi(s)) \geq (1-\gamma) R_{n,\delta}^{\pi}.$$

Therefore,

$$\mathbb{P}_{r|\mathcal{D}_n}\left[Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi(s)) \geq (1-\gamma) R_{n,\delta}^{\pi}\right] \geq \frac{\delta}{|\mathcal{S}|},$$

which completes the proof. □

This lemma extends our previous result to the probabilistic setting of Bayesian IRL. While Lemma 1 showed that high regret implies the existence of a state with poor action choice, this lemma shows that if our policy has a significant probability of high regret, there must be at least one state where it has a significant probability of making a poor action choice.

Now we will build on this lemma to formulate our first theorem to take a step further: if we apply this lemma to a policy that has a significant probability of being approximately correct in each state, but the lemma also gives us a state where it has a significant probability of making a poor choice, we have two contradictory hypotheses that an expert demonstration can help us resolve – we formalize this as a lower bound on the information gain from observing the expert action in such a state.

**Theorem 1.** *For $\epsilon > 0$ and $\delta \in (0, \frac{1}{2}]$, assume that no policy $\pi$ is $(\epsilon, \delta)$-probably-approximately-correct, i.e., $\mathbb{P}[R_r^{\pi} \geq \epsilon] > \delta,\ \forall \pi$. Then, there exists a state $s \in \mathcal{S}$ such that observing a new expert demonstration at $s$ has an expected information gain with respect to the variable $E^{\pi^A}$ (Eq. 5) of at least*

$$EIG_{min}(\epsilon, \delta) = \frac{\delta e^{-\beta(1-\gamma)\epsilon}(1 - e^{-\beta(1-\gamma)\epsilon/2})^2}{4|\mathcal{A}|^5|\mathcal{S}|}. \tag{10}$$

*Proof.* Let $\pi^{\mathrm{A}}$ be an apprentice policy that in each state maximizes the probability of being approximately optimal in the immediate regret sense, i.e., $\pi^{\mathrm{A}}(s) \in \operatorname{argmax}_a \mathbb{P}[Q^*(s, \pi^*(s)) - Q^*(s, a) < \epsilon(1-\gamma)/2|\mathcal{D}_n],\ \forall s \in \mathcal{S}$. To informally outline our proof strategy: we will prove the theorem by showing that under its assumptions, there exists a state $s$ and an alternative action $a' \neq \pi^{\mathrm{A}}(s)$ that has a chance of being significantly better than the apprentice action. If it is significantly better, it is significantly more likely to get selected by the expert than if it is inferior to $\pi^{\mathrm{A}}(s)$. Since the observation distributions in the two cases are different, this allows us to put a lower bound on the expected information gained by observing the expert at $s$. Now let us turn to the proof in full detail.

Under the assumptions of this theorem and using Lemma 2, there exists a state $s$ such that

$$\mathbb{P}_{r|\mathcal{D}_n}\left[Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi^{\mathrm{A}}(s)) \geq (1-\gamma)\epsilon\right] \geq \frac{\delta}{|\mathcal{S}|}.$$

Conditioning on the event $\left\{Q_r^*(s, \pi_r^*(s)) - Q_r^*(s, \pi^{\mathrm{A}}(s)) \geq (1-\gamma)\epsilon\right\}$, let $a'$ be the action that has the highest posterior predictive probability of being optimal. By the pigeonhole principle, this conditional probability of being optimal is at least $\frac{1}{|\mathcal{A}|-1}$. Furthermore, if $a'$ is optimal, then it also has a higher probability of being selected by the Boltzmann-rational expert than any other action, so conditional on being optimal, it must have a probability of at least $1/|\mathcal{A}|$ of being selected. Using just the first fact we get

$$\mathbb{P}\left[\, Q_r^*(s, a') - Q_r^*(s, \pi^{\mathrm{A}}(s)) \geq (1-\gamma)\epsilon\right] \geq \frac{\delta}{|\mathcal{S}|(|\mathcal{A}|-1)}. \tag{22}$$

Using both facts, conditional on $\{Q_r^*(s,\pi_r^*(s)) - Q_r^*(s,\pi^{\mathrm{A}}(s)) \geq (1-\gamma)\epsilon\}$, we have

$$\mathbb{P}\big[A = a' \text{ and } a' \text{ optimal } |Q_r^*(s,\pi_r^*(s)) - Q_r^*(s,\pi^{\mathrm{A}}(s)) \geq (1-\gamma)\big] \geq \frac{1}{|A|(|A|-1)} > \frac{1}{|A|^2} \tag{23}$$

where $A$ is a random variable representing the action that the expert would choose at $s$. This implies also

$$\mathbb{P}\left[A = a' | Q_r^*(s,a') - Q_r^*(s,\pi^{\mathrm{A}}(s)) \geq (1-\gamma)\epsilon\right] > \frac{1}{|A|^2}\,, \tag{24}$$

since the event of $a'$ being optimal is a subset of the event $\{Q_r^*(s,a') - Q_r^*(s,\pi^{\mathrm{A}}(s)) \geq (1-\gamma)\}$ conditional on $\{Q_r^*(s,a') - Q_r^*(s,\pi^{\mathrm{A}}(s)) \geq (1-\gamma)\epsilon\}$ and $\{A = a'\}$ is a superset of $\{A = a' \text{ and } a' \text{ optimal}\}$.

On the other hand, since $\pi^{\mathrm{A}}$ was chosen as the policy maximizing the probability of being "strongly approximately correct", we must have

$$\mathbb{P}\left[Q_r^*(s,\pi_r^*(s)) - Q_r^*(s,\pi^{\mathrm{A}}(s)) < (1-\gamma)\epsilon/2\right] > \frac{1}{|\mathcal{A}|}.$$

(Since in each state, under each reward function, one of the $|\mathcal{A}|$ actions must be optimal, at least one action must have a probability of at least $\frac{1}{|\mathcal{A}|}$ of being optimal and thus also strongly approximately correct. Since $\pi^{\mathrm{A}}$ is maximizing this probability, it must also have a probability of at least $\frac{1}{|\mathcal{A}|}$ of being strongly approximately correct).

Since $Q_r^*(s,a') \leq Q_r^*(s,\pi_r^*(s))$, it is that case that $\{Q_r^*(s,\pi_r^*(s)) - Q_r^*(s,\pi^{\mathrm{A}}(s)) < (1-\gamma)\epsilon/2\} \implies \{Q_r^*(s,a') - Q_r^*(s,\pi^{\mathrm{A}}(s)) < (1-\gamma)\epsilon/2\}$, so the probability of the latter event must be at least as high as that of the former event and we have

$$\mathbb{P}\left[Q_r^*(s,a') - Q_r^*(s,\pi^{\mathrm{A}}(s)) < (1-\gamma)\epsilon/2\right] > \frac{1}{|\mathcal{A}|}. \tag{25}$$

Now, if we denote by $A$ the action taken by the expert in state $s$ seen as a random variable, we can decompose the mutual information between $A$ and $E_{s,a'}$ as

$$I(A; E_{s,a'}) = \sum_{E_{s,a'} \in \mathcal{E}} \mathbb{P}[E_{s,a'}]\; D_{\mathrm{KL}}(p(A|E_{s,a'}) \| p(A)).$$

where $\mathcal{E} :=$ {"strongly approximately correct", "weakly approximately correct", "not approximately correct"}. Going forward, let us shorten the three options to $e_{\mathrm{s}}$,$e_w$, and $e_{\mathrm{n}}$ respectively.

To finish the proof, we need to put a lower bound on this mutual information. We have already given lower bounds on $\mathbb{P}[e_{\mathrm{n}}]$ (Eq. 22) and $\mathbb{P}[e_{\mathrm{s}}]$ (Eq. 25). We will now provide a lower bound on the corresponding KL terms by first lower-bounding the total variation distance between $p(A|e_{\mathrm{n}})$ and $p(A|e_{\mathrm{s}})$.

In the event $e_{\mathrm{s}}$, we have $Q_r^*(s,\pi^{\mathrm{A}}(s)) \geq Q_r^*(s,a') - \epsilon(1-\gamma)/2$, so under the Boltzmann-rational policy, for any given reward $r_{\mathrm{s}}$ compatible with $e_{\mathrm{s}}$, we have

$$\begin{aligned}
\mathbb{P}\left[A = \pi^{\mathrm{A}}(s) | s; r_{\mathrm{s}}\right] &= \frac{1}{Z} e^{\beta Q_r^*(s,\pi^{\mathrm{A}}(s))} \\
&\geq \frac{1}{Z} e^{\beta(Q_r^*(s,a') - (1-\gamma)\epsilon/2)} \\
&= e^{-\beta(1-\gamma)\epsilon/2} \mathbb{P}\left[A = a' | s; r_{\mathrm{s}}\right] \\
&= \mathbb{P}\left[A = a' | s; r_{\mathrm{s}}\right] - (1 - e^{-\beta(1-\gamma)\epsilon/2}) \mathbb{P}\left[A = a' | s; r_{\mathrm{s}}\right],
\end{aligned}$$

where $Z$ is a normalizing constant, so

$$\mathbb{P}\left[A = a' | s; r_{\mathrm{s}}\right] - \mathbb{P}\left[A = \pi^{\mathrm{A}}(s) | s; r_{\mathrm{s}}\right] \leq (1 - e^{-\beta(1-\gamma)\epsilon/2}) \mathbb{P}\left[A = a' | s; r_{\mathrm{s}}\right]$$

Along similar lines, in case of $e_{\mathrm{n}}$ and a reward $r_{\mathrm{n}}$ compatible with it, we have

$$\mathbb{P}\left[A = a' | s; r_{\mathrm{n}}\right] - \mathbb{P}\left[A = \pi^{\mathrm{A}}(s) | s; r_{\mathrm{n}}\right] > (1 - e^{-\beta(1-\gamma)\epsilon}) \mathbb{P}\left[A = a' | s; r_{\mathrm{n}}\right].$$

Marginalizing over the reward and using bound from Equation 24

$$\begin{aligned}
\mathbb{P}[A = a' | s; e_{\mathrm{n}}] - \mathbb{P}\big[A = \pi^{\mathrm{A}}(s) | s; e_{\mathrm{n}}\big] &= \int_{r_{\mathrm{n}}} \left(\mathbb{P}[A = a' | s; r_{\mathrm{n}}] - \mathbb{P}\big[A = \pi^{\mathrm{A}}(s) | s; r_{\mathrm{n}}\big]\right) p(r_{\mathrm{n}} | e_{\mathrm{n}}) \mathrm{d}r_{\mathrm{n}} \\
&\geq \int_{r_{\mathrm{n}}} (1 - e^{-\beta(1-\gamma)\epsilon}) \mathbb{P}\left[A = a' | s; r_{\mathrm{n}}\right] p(r_{\mathrm{n}} | e_{\mathrm{n}}) \mathrm{d}r_{\mathrm{n}} \\
&= (1 - e^{-\beta(1-\gamma)\epsilon}) \mathbb{P}\left[A = a' | s; e_{\mathrm{n}}\right] \\
&> (1 - e^{-\beta(1-\gamma)\epsilon}) \frac{1}{|\mathcal{A}|^2}.
\end{aligned}$$

and similarly for the case of $e_{\mathrm{s}}$. We can use this to put a lower bound on the total variation distance between the posterior predictive distributions under the events $e_{\mathrm{n}}$ and $e_{\mathrm{s}}$:

$$\begin{aligned}
&D_{\mathrm{TV}}(p(A|e_{\mathrm{s}}), p(A)) + D_{\mathrm{TV}}(p(A|e_{\mathrm{n}}), p(A)) \\
&\qquad \geq D_{\mathrm{TV}}(p(A|e_{\mathrm{s}}), p(A|e_{\mathrm{n}})) \\
&\qquad \geq \frac{1}{2}\big(\mathbb{P}\left[A = \pi^{\mathrm{A}}(s) | e_{\mathrm{s}}\right] - \mathbb{P}\left[A = \pi^{\mathrm{A}}(s) | e_{\mathrm{n}}\right] + \mathbb{P}\left[A = a' | e_{\mathrm{n}}\right] - \mathbb{P}\left[A = a' | e_{\mathrm{s}}\right]\big) \\
&\qquad = \frac{1}{2}\big(\mathbb{P}\left[A = a' | e_{\mathrm{n}}\right] - \mathbb{P}\left[A = \pi^{\mathrm{A}}(s) | e_{\mathrm{n}}\right]\big) - \frac{1}{2}\big(\mathbb{P}\left[A = a' | e_{\mathrm{s}}\right] - \mathbb{P}\left[A = \pi^{\mathrm{A}}(s) | e_{\mathrm{s}}\right]\big) \\
&\qquad > \frac{1}{2}(1 - e^{-\beta(1-\gamma)\epsilon}) \frac{1}{|\mathcal{A}|} - \frac{1}{2}(1 - e^{-\beta(1-\gamma)\epsilon/2}) \frac{1}{|\mathcal{A}|} \\
&\qquad = \frac{1}{2} e^{-\beta(1-\gamma)\epsilon/2} (1 - e^{-\beta(1-\gamma)\epsilon/2}) \frac{1}{|\mathcal{A}|^2}
\end{aligned}$$

where we used the triangle inequality in the first step, the definition of total variation distance in the second step (omitting non-negative terms corresponding to actions other than $\pi^{\mathrm{A}}(s)$ and $a'$), the two inequalities we just derived for $e_{\mathrm{n}}$ and $e_{\mathrm{s}}$ in the fourth step.

Applying Pinsker's inequality to both TV terms gives us

$$\begin{aligned}
D_{\mathrm{KL}}(p(A|e_{\mathrm{s}}) \| p(A)) + D_{\mathrm{KL}}(p(A|e_{\mathrm{n}}) \| p(A)) &\geq 2(D_{\mathrm{TV}}(p(A|e_{\mathrm{s}}), p(A))^2 + D_{\mathrm{TV}}(p(A|e_{\mathrm{n}}), p(A))^2) \\
&\geq (D_{\mathrm{TV}}(p(A|e_{\mathrm{s}}), p(A)) + D_{\mathrm{TV}}(p(A|e_{\mathrm{n}}), p(A)))^2 \\
&> \frac{1}{4|\mathcal{A}|^4} e^{-\beta(1-\gamma)\epsilon} (1 - e^{-\beta(1-\gamma)\epsilon/2})^2.
\end{aligned}$$

where we applied the inequality $(a+b)^2 \leq 2(a^2 + b^2)$ in the second step and the previous inequality on the sum of TV distances in the third step.

This finally allows us to establish that

$$\begin{aligned}
I(A; E_{s,a'}) &= \sum_{E_{s,a'} \in \{e_{\mathrm{n}}, e_{\mathrm{s}}, e_{\mathrm{w}}\}} \mathbb{P}[E_{s,a'}] D_{\mathrm{KL}}(p(A|E_{s,a'}) \| p(A)) \\
&\geq \min\{\mathbb{P}[e_{\mathrm{s}}], \mathbb{P}[e_{\mathrm{n}}]\} \left(D_{\mathrm{KL}}(p(A|e_{\mathrm{s}}) \| p(A)) + D_{\mathrm{KL}}(p(A|e_{\mathrm{n}}) \| p(A))\right) \\
&> \min\left\{\frac{1}{|\mathcal{A}|}, \frac{\delta}{(|\mathcal{A}| - 1)|\mathcal{S}|}\right\} \frac{1}{4|\mathcal{A}|^2} e^{-\beta(1-\gamma)\epsilon} (1 - e^{-\beta(1-\gamma)\epsilon/2})^2 \\
&= \frac{\delta}{4|\mathcal{A}|^5 |\mathcal{S}|} e^{-\beta(1-\gamma)\epsilon} (1 - e^{-\beta(1-\gamma)\epsilon/2})^2.
\end{aligned}$$

The first inequality follows from the fact that all three terms in the sum are non-negative. The second inequality just plugs in results previously derived in this proof. The final step resolves the minimum as its second term using the assumption that $\delta \leq \frac{1}{2}$ and $|\mathcal{A}| \geq 2$ and uses the fact that $\frac{1}{|\mathcal{A}|} < \frac{1}{|\mathcal{A}|-1}$.

Since the random variable $E_{s,a'}$ is coarser than the variable $E$, which is the collection of the variables $E_{s,a}$ across all state-action pairs, we have $I(A;E) \geq I(A;E_{s,a'})$, which completes the proof. □

Now that we have a lower bound on the information that we gain in each step, we can use it to bound the number of steps needed to reach the PAC condition.

**Theorem 2.** *Let $h_{max}$ be an upper bound on the entropy of the prior distribution over $E$, the PAC-EIG discretized regret values $E^{\pi^A}$ aggregated across all apprentice policies $\pi^A$. Then, the expected number of steps needed to reach the PAC condition is upper bounded by*

$$\frac{h_{max}}{EIG_{min}(\epsilon,\delta)} = \frac{4h_{max}|\mathcal{A}|^5|\mathcal{S}|}{\delta e^{-\beta(1-\gamma)\epsilon}(1-e^{-\beta(1-\gamma)\epsilon/2})^2}. \tag{11}$$

*Proof.* The minimal expected information gain guaranteed by Theorem 1 is fully derived from components of the random variable $E$. Thus, at every step of active learning where we have not yet achieved the PAC criterion, we can gain at least $\text{EIG}_{\text{min}}(\epsilon,\delta)$ information about $E$.

Let $H_n$ denote the expected entropy of $E$ after $n$ steps of active learning. By the properties of entropy and information gain:

1. $H_n \geq 0$ for all $n$ (non-negativity of entropy)
2. $H_0 = h_{\text{prior}}$ (initial entropy)
3. $H_{n+1} \leq H_n - \text{EIG}_{\text{min}}(\epsilon,\delta)$ for all $n$ where the PAC criterion is not met (guaranteed information gain)

Let $N$ be the number of steps needed to reach the PAC criterion. Then:

$$\begin{aligned} 0 &\leq H_N \\ &\leq H_0 - N \cdot \text{EIG}_{\text{min}}(\epsilon,\delta) \\ &= h_{\text{prior}} - N \cdot \text{EIG}_{\text{min}}(\epsilon,\delta) \end{aligned}$$

Solving for $N$:

$$N \leq \frac{h_{\text{prior}}}{\text{EIG}_{\text{min}}(\epsilon,\delta)} \tag{26}$$

The result follows by substituting the expression for $\text{EIG}_{\text{min}}(\epsilon,\delta)$ from Theorem 1. □

**Corollary 1.** *For any prior distribution over rewards, the expected number of steps to reach the PAC condition is at most*

$$\log(3)|\mathcal{S}||\mathcal{A}|^2/EIG_{min}(\epsilon,\delta) = \frac{4\log(3)|\mathcal{A}|^7|\mathcal{S}|^2}{\delta e^{-\beta(1-\gamma)\epsilon}(1-e^{-\beta(1-\gamma)\epsilon/2})^2}. \tag{27}$$

*Proof.* The random variable $E$ aggregates $|\mathcal{S}||\mathcal{A}|^2$ ternary random variables (in each state, apprentice policies can take $|\mathcal{A}|$ actions, and we are considering each action's immediate regret relative to $|\mathcal{A}|-1 < |\mathcal{A}|$ alternative actions), so $E$ can take at most $3^{|\mathcal{S}||\mathcal{A}|}$ values. Thus, its maximum entropy is $-\log(1/3^{|\mathcal{S}||\mathcal{A}|^2}) = |\mathcal{S}||\mathcal{A}|^2\log(3)$. The result follows by plugging this maximum entropy into Theorem 2. □

While we do not claim this bound is tight, it provides a useful characterization of how the sample complexity scales with the problem parameters. In particular, it shows polynomial dependence on the size of the state and action spaces, and inverse dependence on both the allowed suboptimality $\epsilon$ and failure probability $\delta$. More importantly, on a qualitative level, the result shows that the learning process does continue as long as the PAC condition is not satisfied, so it does not get stuck forever querying an uninformative state as is the case with e.g. the acquisition function of Kweon et al. (2023).

### B.1 Notes on possible improvements

#### B.1.1 Tighter bound for large state spaces

Note that the bound from Lemma 2 can be tightened if the state space is large and only a subset is reachable within an effective horizon. In that case $|S|$ can be replaced by the number of states reachable from the initial states within $1/(1-\gamma)$ steps. Also, if there is a limited horizon, or the apprentice policy always reaches a terminal state in a certain number of steps, we can also have fewer steps than that (and can further reduce the $(1-\gamma)$ factor in the exponent of our theoretical results.

#### B.1.2 Static policy

The bound includes the entropy of each action in each state. In fact, it may be enough to focus on a single action in each state, since we want to identify only a particular PAC policy, rather than reducing entropy of all of the components of $E(s)$ in every state. This should allow us to exclude a factor of $|A|$ from the bound on the expected number of steps to reach the PAC condition.

## C Accounting for visitation frequencies

A practically useful acquisition function should account for one more important point: in order to get a policy with low expected regret, we do not need to reduce the expected immediate regret of all points, but just those that are likely to get visited by the apprentice policy $\pi^{\mathrm{A}}$. Let $\nu_A(s) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}[s_t = s]]$ be the discounted expected occupancy of state $s$. Then, we can set

$$\tilde{\alpha}(s) := \nu_A(s) I(A_s, E_s). \tag{28}$$

to be the acquisition function for querying single states.

However, we also wish to have an acquisition function for collecting full trajectories. A naive approach employed by some prior work (Kweon et al., 2023) would be summing the individual

$$\alpha'_n(s_0) := \mathbb{E}_{r|\mathcal{D}_n}\left[\mathbb{E}_{\tau|r}\left[\sum_{s\in\tau} \tilde{\alpha}(s)\right]\right] \tag{29}$$

However, the sum in Eq. (29) neglects correlation between the regrets in different states (and, worse, autocorrelation if a state is visited multiple times). We can instead estimate the full expected information gain about our variable $E$ from the new expert trajectory. Then, dropping the weighing for the moment,

$$\mathrm{EIG}_E(s_0) = \mathbb{E}_{\tau,E}\big[\log p(\tau|E) - \log p(\tau)\big]. \tag{30}$$

We assume that the Bayesian IRL method we use to estimate this is able to give us samples from the current posterior over Q-values. Given a Q-value and an initial state, we can sample the corresponding hypothetical expert trajectories $\tau|Q$. Also, $E$ is a cheaply-computable function of Q, so we can also easily convert the samples of Q into samples of $E$. Then, the only remaining challenge in computing $\mathrm{EIG}_E$ is estimating $p(\tau|E)$.

If the state space is small and we have a lot of Q samples, we can estimate $p(\tau|E) = \frac{1}{|\mathcal{Q}_E|}\sum_{Q\in\mathcal{Q}_E} p(\tau|Q)$. However, note that for a given policy, there are $3^{|S|}$ possible values of $E^\pi$, so even for a moderate size of the state space, the number of Q corresponding to each $E$ could be small. At the same time, the components of $E^\pi$ corresponding to states far away from the trajectory are unlikely to share much mutual information with the trajectory. Thus we suggest using the bound

$$I(\tau, E) \geq I(\tau, E^\pi_\tau) \tag{31}$$

where $E^\pi_\tau := \big(E_s\big)_{s\in\mathcal{S}_\tau}$ for $\mathcal{S}_\tau$ being some neighbourhood of $\tau$ in the state space, including all states on the trajectory $\tau$ plus states that can be quickly reached from the trajectory.

To again add the weighing, we can note that since the transition probabilities are the same for both components,

$$\log p(\tau|E) - \log p(\tau) = \sum_{s,a\in\tau} \log p(a|s;E) - \log p(\tau|s) \tag{32}$$

which naturally allows us to re-introduce the weights as

$$\sum_{s,a\in\tau} \nu_\pi(s,a)(\log p(a|s;E) - \log p(\tau|s)) \tag{33}$$

resulting in an acquisition function

$$\text{EIG}_{\nu\text{-PAC}}(s_0) = \mathbb{E}_{\tau,E}\big[\sum_{s,a\in\tau} \nu_\pi(s,a)(\log p(a|s;E_\tau) - \log p(\tau|s))\big]. \tag{34}$$

# D Alternative Acquisition Functions

In this appendix, we discuss alternative formulations of acquisition functions for active IRL, which could help to see the reasoning process that led to $PAC-EIG$, and explain why some, possibly more, obvious alternatives were not chosen.

## D.1 From Apprentice Return to Regret

When the goal is to produce a well-performing apprentice policy (as opposed to learning the reward for its own sake), a natural starting point is to directly optimize the apprentice's expected return. This suggests minimizing the loss:

$$\mathcal{L}_{\text{ret}}(\xi_1,\dots,\xi_N) = -\mathbb{E}_{s_0\sim\rho}\mathbb{E}_{\tau|s_0,\pi_N^{\text{A}}} G_r(\tau) \tag{35}$$

where $\pi_N^{\text{A}}$ is the apprentice policy after observing $N$ expert trajectories from initial states $\xi_1,\dots,\xi_N$.

Since the optimal value $V^*(s_0)$ is independent of our choice of queries, minimizing $\mathcal{L}_{\text{ret}}$ is equivalent to minimizing the regret loss:

$$\mathcal{L}_{\text{reg}}(\xi_1,\dots,\xi_N) = R_r^{\pi_N^{\text{A}}} = \mathbb{E}_{s_0\sim\rho}\left[V_r^*(s_0) - V_r^{\pi_N^{\text{A}}}(s_0)\right] \tag{36}$$

## D.2 The Challenge of Direct Regret Optimization

Directly optimizing this regret loss is computationally intractable even in the greedy case. The one-step acquisition function would be:

$$\alpha_n^{\text{reg}}(\xi) = -\mathbb{E}_{r|\mathcal{D}_n}\mathbb{E}_{\tau_{n+1}|\xi,\pi_r^{\text{E}}} R_r^{\pi_{n+1}^{\text{A}}} \tag{37}$$

Computing this requires: 1. For each possible reward $r$ in our posterior 2. For each possible expert trajectory $\tau$ from initial state $\xi$ 3. Computing the updated posterior $p(r|\mathcal{D}_n\cup\{\tau\})$ 4. Finding the optimal apprentice policy for this updated posterior 5. Evaluating its regret

This nested optimization involving repeated Bayesian IRL updates is prohibitively expensive.

## D.3 Information Gain About Regret

Following the approach in Bayesian optimization (Wang & Jegelka, 2017), rather than directly optimizing the hard-to-compute expected improvement, we can instead maximize information gain about the quantity of interest. This suggests the acquisition function:

$$\alpha_n^{\text{Regret-EIG}}(s_0) = I(\tau; R_r^{\pi^A}|s_0,\mathcal{D}_n) \tag{38}$$

However, this formulation has a critical flaw. Consider this example:

- In state $s$, the apprentice can take action $a_0$ yielding return 0
- Actions $a_1$ and $a_2$ yield returns of +50 and -100, but we don't know which is which
- With equal probability on both orderings, the apprentice chooses $a_0$ (expected return 0 vs -25)
- The regret is known with certainty to be 50
- Since there's no uncertainty about regret, Regret-EIG assigns zero value to querying this state
- Yet the apprentice is definitely choosing suboptimally!

### D.4 Immediate Regret EIG

The solution is to decompose regret more finely. The total regret can be written as:

$$R_r^{\pi^A} = \mathbb{E}_{\tau\sim\rho,\pi^A} \sum_{s_t,a_t\in\tau} \gamma^t \underbrace{\left[V_r^*(s_t) - Q_r^*(s_t,\pi^A(s_t))\right]}_{R^*_{\pi^A,r}(s_t)} \tag{39}$$

where $R^*_{\pi^A,r}(s)$ is the *immediate regret* – the value lost by following the apprentice policy in state $s$ without considering future consequences.

This can be further decomposed per action:

$$R^*_{\pi^A,r}(s) = \max_a R^*_{\pi^A,r}(s,a) \tag{40}$$

where $R^*_{\pi^A,r}(s,a) = \max\{0, Q_r^*(s,a) - Q_r^*(s,\pi^A(s))\}$.

The Immediate Regret EIG acquisition function is then:

$$\alpha_n^{\text{IR-EIG}}(s_0) = I(\tau; R^*|s_0, \mathcal{D}_n) \tag{41}$$

where $R^* = (R^*_{\pi^A,r}(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$.

This formulation correctly identifies informative states in our earlier example, as there is high uncertainty about which action has higher immediate regret.

### D.5 Discretization and Connection to PAC-EIG

For practical computation, the continuous immediate regret values must be discretized. Different discretization schemes lead to different acquisition functions:

1. **Multi-bucket discretization**: Using buckets of [0, $\epsilon/2$], [$\epsilon/2$, $\epsilon$], [$\epsilon$, $\infty$) like in PAC-EIG can be taken further to allow for a finer approximation of the IR-EIG. This can be viable for single state-queries, but growing the number of categories becomes untenable once we start considering the full trajectory demonstration.
2. **Two-bucket PAC discretization**: Using just two buckets – acceptable regret $[0, \epsilon]$ and unacceptable regret $(\epsilon, \infty)$ – directly captures what matters for PAC guarantees. However, the theoretical arguments as shown above do not directly apply to this case, since we loose the middle bucket that ensures separation between the two sufficiently different expert action distributions. However, we do think a weaker result could be proven even for this case.

The PAC discretization is not only computationally more tractable but also theoretically motivated: it focuses information gathering on exactly what we need to know to provide formal reliability guarantees.

### D.6 Summary

The progression from expected return optimization to PAC-EIG illustrates how principled information-theoretic thinking, combined with practical computational constraints and theoretical objectives, leads

to an effective acquisition function. While IR-EIG with fine discretization might provide marginally more information in some cases, PAC-EIG strikes the optimal balance between theoretical guarantees, computational efficiency, and practical effectiveness.

# E Experiment details

## E.1 Basic parameter values

In the three environments (structured 6x6, random 8x8, and random 10x10) we used $\beta = 4, 2, 4$ respectively, $\gamma = 0.9$, and an infinite horizon (but all environments contained terminal states). We started with an empty set of demonstrations (implemented as a single, uninformative observation of a dummy sink state) and then ran active learning for 150 steps.

For ActiveVaR, we used $\delta = 0.05$ (same as the original paper). For policy entropy, we used the entropy of the discretized distribution for each action (as proposed by the authors) with K=10 buckets. For our PAC-EIG acquisition function we used $(1 - \gamma)\epsilon = 0.01$ for the PAC condition.

## E.2 Environments

The gridworld environments have 5 actions, corresponding to staying in place and moving in the four directions. Furthermore, there is a probability of 0.1 of random action being executed instead of the intended one. If an action would result in crossing the edge, the agent instead remains in place. The gridworlds use a state-only reward (awarded upon executing any action in the given state).

The 8x8, and 10x10 fully random environments were generated as follows:

1. Each state was assigned a random reward drawn independently from $\mathcal{N}(0, 3)$ (i.e. mostly yielding rewards between -10 and 10).
2. Each state was then marked as terminal with an independent probability of 0.1.
3. The top $10\%$ of states with highest reward were further marked as terminal (producing terminal goal states, which may, however, sometimes be avoided by the optimal policy in favour of staying forever in other positive states).
4. The initial state distribution is either uniform across the whole state space, or, in the case of the 10x10 gridworld, 2 non-terminal initial states were chosen randomly uniformly.[4]

## E.3 Bayesian IRL methods

Our active learning uses a Bayesian IRL method as a key component. In our experiments, we used two methods based on Markov chain Monte Carlo (MCMC) sampling: on the structured environment, we used PolicyWalk (Ramachandran & Amir, 2007), while on the environment with a different random reward in every state, we used the faster ValueWalk (Bajgar et al., 2024), which performs the sampling primarily in the space of Q-functions before converting into rewards. We also tried a method based on variational inference (Chan & van der Schaar, 2021), but we found its uncertainty estimates unreliable for the purposes of active learning.

For MCMC sampling, we used Hamiltonian Monte Carlo (Duane et al., 1987) with the no-U-turns (NUTS) sampler (Hoffman & Gelman, 2014) and automatic step size selection during warm-up (starting with a step size of 0.1). At every step of active learning, we ran the MCMC sampling from scratch using all demonstrations available up to that point. We ran for 100 warm-up steps and then 200, 500, and 1000 on the three environments respectively. For subsequent usage, we use every other sample to reduce autocorrelation.

[4]Note that the implementation allows the two initial states to collide, producing only a single initial state in $1/81$ of the cases, but this was not the case for any of our 16 random seeds.

### E.4 Metrics

On the first two environments, we use KNN entropy estimation to calculate posterior entropy with K=5. This method is known to struggle in high dimensions, which we also observed in the case of the 10x10 gridworld (which has a 100-dimensional reward space), so there, we estimate the entropy by the entropy of a multivariate normal distribution with the mean and covariance matrix estimated from the MCMC samples.

Regret was calculated relative to the expected return of the optimal policy, calculated using value iteration with a tolerance of 1e-5. Posterior regret samples were similarly calculated relative to the optimal return with respect to each of the posterior reward samples (which were calculated using the optimal Q-value samples which get produced by the Bayesian IRL methods).

When aggregating true regret across environment instances, we also normalized the regret for each random environment instance by the average regret across all methods across the first 32 steps of active learning to account for the possibly different scales and different learning difficulties of the random environments.

### E.5 Implementation

The experiments were implemented using Python 3.10, PyTorch 2.5.1, and Pyro 1.8.6. We will publish our full code for both the experiments and the associated result analysis on Github once the anonymity requirement is lifted.

### E.6 Timing

The computational time per step of active IRL is dominated by the time necessary to collect the Bayesian IRL MCMC samples, which ranges between 5 seconds for the 100+200 samples on the structured gridworld to about 5 minutes for the 100+1000 samples on the 10x10 gridworld in a single CPU thread. The overhead of all acquisition functions on top of that is below $0.03$ and can thus be considered negligible.

Reproducing all our experiments thus takes less than a day on a CPU with 128 threads (we used AMD Ryzen Threadripper 3990X at 2.2GHz).

## F Notation overview

Table 1: Summary of notation used throughout the paper. If reward is omitted from a symbol otherwise depending on it, it means it is taken with respect to the true reward.

| Symbol | Meaning |
|---|---|
| $\mathcal{S}$ | State spstate ace of the MDP |
| $\mathcal{A}$ | Action space |
| $P(s' \mid s, a)$ | Transition kernel |
| $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ | Expected reward function |
| $\gamma \in (0, 1)$ | Discount factor |
| $t_{\max}$ | Maximum horizon (may be $\infty$) |
| $\rho$ | Initial-state distribution |
| $\pi^E$ | Expert policy (Boltzmann-rational with coefficient $\beta$) |
| $\pi^E_r$ | Hypothetical expert policy that would correspond to a reward $r$ |
| $\beta$ | Boltzmann rationality coefficient |
| $\mathcal{D}_n$ | Demonstration data after $n$ queries |
| $\pi^{\mathrm{A}},\ \pi^{\mathrm{A}}_n$ | Apprentice policy (after $n$ queries) |
| $\pi^\star_r$ | Optimal policy for reward $r$ |
| $\tau = (s_0, a_0, \ldots, s_T)$ | Trajectory |
| $\xi$ | Query (initial state for the next expert demonstration) |
| $V^\pi_r(s)$, $Q^\pi_r(s, a)$ | State- and action-value functions for policy $\pi$ and reward $r$ |
| $G_r(\tau)$ | Discounted return of trajectory $\tau$ under $r$ |
| $G^\pi_r$ | Expected discounted return of $\pi$ under $r$, $P$, and $\rho$ |
| $R^\pi_r(s_0)$ | Regret of $\pi$ from state $s_0$ |
| $R^\star_{\pi,r}(s)$, $R^\star_{\pi,r}(s, a)$ | Immediate regret of $\pi$ for state $s$ and state–action pair $s, a$ |
| $I(\tau; r \mid s_0, \mathcal{D}_n)$ | Mutual information between a trajectory and reward |
| EIG | Expected information gain |
| $\alpha^{\text{RewEIG}}_n$ | Reward-EIG acquisition function |
| $\mathrm{VaR}_\delta$ | $\delta$-value-at-risk of a loss random variable |
| $EIG_{\min}(\epsilon, \delta)$ | Per-step information-gain lower bound (Thm 3) |
| $h_{\max}$ | Upper bound on the prior entropy of $E$ |
| $\epsilon,\ \delta$ | PAC accuracy / confidence parameters |
| $\lvert\mathcal{S}\rvert,\ \lvert\mathcal{A}\rvert$ | Cardinalities of state and action spaces |