INVERSE REINFORCEMENT LEARNING WITHOUT AN OPTIMAL DEMONSTRATOR: A FEASIBLE REWARD SET APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse reinforcement learning (IRL) typically assumes demonstrations from a single optimal demonstrator, but in many applications data come from multiple imperfect demonstrators with heterogeneous suboptimality levels. We study reward learning in this setting through a feasible-reward-set framework: for each demonstrator, we encode its declared suboptimality level as a linear constraint and intersect the resulting feasible sets across demonstrators. Our theoretical analysis shows that the joint feasible set shrinks monotonically as data are added and provides a sufficient condition for strict shrinkage. We further establish optimality guarantees for the unobserved optimal demonstrator under every reward in the resulting feasible set. On the practical side, we introduce strategies to address the inherent reward ambiguity in the obtained reward set and provide an offline algorithm with function approximation for high-dimensional environments. Experiments in tabular grid-world and large language model (LLM) fine-tuning settings support the theoretical predictions and demonstrate improved performance over baselines.

1 INTRODUCTION

IRL aims to infer rewards that explain observed behavior from demonstrations, typically under the assumption that the demonstrator policy is optimal under the true reward (Ng & Russell, 2000; Adams et al., 2022). In many practical applications, however, demonstration data is not collected from a single optimal demonstrator. Instead, we often observe demonstrations from multiple imperfect demonstrators, each reflecting limited skill, resource constraints, or data corruption. For example, in LLMs, supervision commonly comes from imperfect and heterogeneous human raters rather than from a single optimal expert (Bukharin et al., 2024; Park et al., 2024; Zhang et al., 2025). In robotics, demonstrations are often collected from operators with heterogeneous proficiency (Chen et al., 2020; 2021), or in constrained domains (Jayanthi et al., 2023). Such scenarios underscore the need for IRL algorithms that can recover optimal behavior from suboptimal demonstrations generated by multiple demonstrators with heterogeneous optimality levels.

Conventional IRL algorithms, such as maximum entropy (MaxEnt) IRL (Ziebart et al., 2008) and Bayesian IRL (Ramachandran & Amir, 2007), typically assume a probabilistic behavioral model and recover a single reward consistent with that model. Building on these algorithms, Beliaev et al. (2022); Beliaev & Pedarsani (2025) account for varying demonstrator quality by estimating demonstrator expertise or demonstrator-specific noise parameters. However, this approach leaves a fundamental question unresolved: can suboptimal demonstrations alone identify rewards under which an unobserved optimal demonstrator is nearly optimal? The difficulty is that maximum likelihood estimation (MLE) tends to assign higher reward to frequently visited state-action pairs, even when they do not reflect optimal behavior. Ranked-demonstration methods address this question more directly, giving a sufficient condition for extrapolation beyond suboptimal demonstrations (Brown et al., 2019; 2020). However, their guarantee of performance improvement does not ensure that the ground-truth optimal demonstrator is nearly optimal under the learned reward. Moreover, their ranking-based algorithm does not apply without strict preferences, such as when demonstrators have the same optimality level but specialize in different tasks.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this question, we adopt the *feasible reward set* approach in this paper. Rather than selecting a single reward under an assumed model, this approach characterizes the set of rewards consistent with demonstrated behavior through linear constraints (Metelli et al., 2021; 2023). Specifically, for each demonstrator, we construct the feasible set of rewards under which its behavior is at most ϵ-suboptimal, where ϵ is a given upper bound on that demonstrator's suboptimality. We then intersect these sets across demonstrators to obtain a joint feasible set. This approach provides a model-free way to encode the suboptimality levels of the demonstrators, rather than representing them indirectly through a probabilistic model assumption.

This offers several advantages over previous work. On the theory side, we show that the joint feasible set shrinks monotonically as demonstrations are added, and provide a sufficient condition for strict shrinkage. We then establish optimality guarantees for the unobserved optimal demonstrator under every reward in the resulting feasible set. The bound depends on how closely a mixture of demonstrator occupancies approximates the optimal occupancy. We also provide additional sufficient conditions showing that demonstrators with relatively large suboptimality bounds can still help reduce the gap. We also present instances demonstrating an advantage of our method over MaxEnt and ranking-based baselines. On the practical side, we introduce practical strategies to address reward ambiguity issues, and extend the proposed framework to high-dimensional environments with function approximation. Taken together, these results yield a principled, model-free framework for IRL from multiple suboptimal demonstrations, providing an alternative to model-dependent likelihood-based methods. Experiments in both tabular grid-world and high-dimensional LLM fine-tuning settings support the theoretical predictions and demonstrate the effectiveness of the proposed framework over baselines.

Related work. Our work builds on optimization-based views of IRL, including the original constraint formulation (Ng & Russell, 2000) and recent feasible-reward-set analyses (Metelli et al., 2021; 2023; Kim et al., 2024). Poiani et al. (2024) show that suboptimal-demonstrator data shrink the feasible reward set, but assume access to an optimal demonstrator, whereas our focus is on reward-set recovery without one. Lazzati & Metelli (2026) study principled representative selection from the feasible reward set via closed-form reward centroids; we adopt a related idea to address reward ambiguity issues. Learning from imperfect or negatively labeled behavior has also been studied in IRL and imitation learning (IL): failed demonstrations can reduce reward ambiguity (Shiarlis et al., 2016), ranked-demonstration methods learn from ordered imperfect trajectories (Brown et al., 2019; 2020), margin-based methods use separation constraints (Ratliff et al., 2006; 2009), and other work learns policies directly from imperfect demonstrations (Gao et al., 2018; Zhang et al., 2021). In contrast to these works, our framework does not rely on probabilistic model assumptions, ordered demonstrations, or access to an optimal demonstrator. See Appendix A for further discussion.

Organization of the paper. Section 2 introduces notation and background on the LP view of IRL. Section 3.1 presents the feasible reward set formulation for multiple suboptimal demonstrations, establishes monotonic shrinkage and optimality guarantees, and compares with existing baselines. Section 3.2 introduces strategies for addressing the reward ambiguity issue in IRL, and Section 3.3 proposes an offline algorithm with function approximation for high-dimensional environments. Finally, Section 4 provides empirical results on both a grid-world environment and LLM fine-tuning.

2 PRELIMINARIES

Let $[K]$ denote the set $\{1, 2, \dots, K\}$ and let $\Delta(S)$ denote the probability simplex over S. $\|\cdot\|_p$ denotes the ℓ_p norm. $\mathbf{1}\{e\} \in \{0, 1\}$ denotes the indicator of event e, and $\mathbf{1}$ denotes the all-ones vector of appropriate dimension. $\mathrm{supp}(\cdot)$ denotes the support of a function (or distribution), while $\mathrm{conv}(\cdot)$ and $\mathrm{relint}(\cdot)$ denote the convex hull and relative interior of a set, respectively. For ease of notation, we use both function and vector notation for rewards and occupancy measures.

2.1 MARKOV DECISION PROCESSES (MDPS) AND OCCUPANCY MEASURES

We consider an infinite-horizon discounted MDP $\mathcal{M}_r := (S, A, P, r, \mu_0, \gamma)$, where S and A are finite state and action spaces, $P(\cdot|s, a) \in \Delta(S)$ is the transition kernel, $r : S \times A \mapsto [0, 1]$ is the reward function, $\mu_0 \in \Delta(S)$ is the initial-state distribution, and $\gamma \in (0, 1)$ is the discount factor. When the

reward is not specified, we write $\mathcal{M} := (S, A, P, \mu_0, \gamma)$ for the corresponding reward-free MDP. For a policy $\pi(\cdot|s) \in \Delta(A)$, the value function $v_r^\pi(s)$ is defined as the expected total discounted reward received when initiating from state s and following π, such that

$$v_r^\pi(s) := \mathbb{E}^\pi \left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \mid s_0 = s \right]. \tag{1}$$

Let $v_r^\star$ denote the optimal value function that maximizes the value function, which satisfies the Bellman equations. Then, it is well known that $v_r^\star$ is the optimal solution to the following linear program (Puterman, 1994):

$$\min_{v \in \mathbb{R}^{|S|}} (1-\gamma)\mu_0^\top v \quad \text{s.t.} \quad M^\top v \geq r, \tag{2}$$

where the dynamics operator $M \in \mathbb{R}^{|S| \times (|S||A|)}$ is defined by $M(s', (s, a)) := \mathbf{1}\{s' = s\} - \gamma P(s'|s, a)$. The dual program can be expressed as

$$\max_{d \in \mathbb{R}^{|S||A|}} r^\top d \quad \text{s.t.} \quad Md = (1-\gamma)\mu_0, \quad d \geq 0. \tag{3}$$

The dual variable $d \in \mathbb{R}^{|S||A|}$ is a *discounted occupancy measure*, i.e., the normalized discounted frequency with which a policy visits each state–action pair. For a policy π and initial distribution μ_0, d^π is defined by

$$d^\pi(s, a) := (1-\gamma)\, \mathbb{E}^\pi \left[\sum_{h=0}^{\infty} \gamma^h \, \mathbf{1}\{(s_h, a_h) = (s, a)\} \,\middle|\, s_0 \sim \mu_0 \right] \tag{4}$$

which is feasible for the dual LP equation 3.

Throughout the paper, we fix the reward-free MDP $\mathcal{M}$ and denote the admissible occupancy polytope under $\mathcal{M}$ by

$$\Phi := \{d \in \mathbb{R}_+^{|S||A|} \mid Md = (1-\gamma)\mu_0\}. \tag{5}$$

For a reward r, define the suboptimality gap of an occupancy measure $d \in \Phi$ as

$$\mathrm{SubOpt}(r, d) := \max_{\tilde{d} \in \Phi} r^\top \tilde{d} - r^\top d. \tag{6}$$

Then, $\mathrm{SubOpt}(r, d) = 0$ if and only if d maximizes the reward over Φ. We denote the set of optimal occupancy measures under reward r by $\Phi^\star(r) \subseteq \Phi$:

$$\Phi^\star(r) := \{d \in \Phi \mid \mathrm{SubOpt}(r, d) = 0\} = \arg\max_{d \in \Phi} r^\top d. \tag{7}$$

2.2 IRL FROM A SINGLE OPTIMAL DEMONSTRATOR

Feasible reward set. Given an optimal demonstrator policy π_e, let $d_e := d^{\pi_e} \in \Phi$ denote its occupancy measure, where the subscript e denotes expert. Since multiple rewards can make the demonstrator occupancy measure optimal, we consider the set of rewards for which d_e is optimal, referred to as the feasible reward set in the IRL literature (Metelli et al., 2021; 2023):

$$\mathcal{R}(d_e) := \{r \in \Delta(S \times A) \mid d_e \in \Phi^\star(r)\}. \tag{8}$$

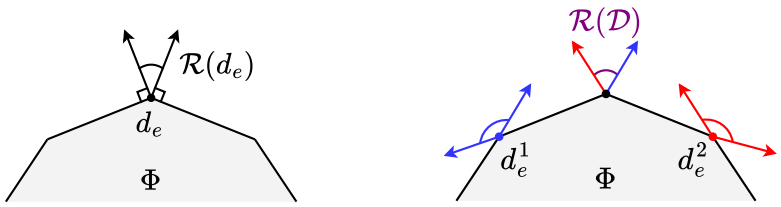
The simplex constraint $r \in \Delta(S \times A)$ is introduced to remove scaling ambiguity in the suboptimality measure, and all feasible reward sets in the theory below use this normalization. For other reward scales, the suboptimality levels must be rescaled by the corresponding constant. With this normalization, the feasible reward set has a simple geometric interpretation: it is the intersection of the normal cone to Φ at d_e and the simplex $\Delta(S \times A)$, i.e., $\mathcal{R}(d_e) = N(d_e; \Phi) \cap \Delta(S \times A)$, where $N(d_e; \Phi) := \{r \mid r^\top (d_e - d) \geq 0 \quad \forall d \in \Phi\}$.

LP characterization. By strong duality, d_e is optimal for the dual LP equation 3 under reward r if and only if there exists a primal variable v such that Karush–Kuhn–Tucker (KKT) conditions hold:

$$\begin{aligned} (1-\gamma)\mu_0^\top v - r^\top d_e = 0 \quad &\text{(Zero duality gap)}, \\ M^\top v \geq r \quad &\text{(Primal feasibility)}. \end{aligned} \tag{9}$$

Note that the dual feasibility condition is unnecessary since it is automatically satisfied by the assumption $d_e \in \Phi$. Then, the feasible reward set can be expressed as the following polytope (Kim et al., 2024)

$$\mathcal{R}(d_e) = \{\, r \in \Delta(S \times A) \mid \exists v \in \mathbb{R}^{|S|} \quad \text{s.t.} \quad (1-\gamma)\mu_0^\top v - r^\top d_e = 0,\ M^\top v \geq r \,\}. \tag{10}$$



(a) (b)

Figure 1: Illustration of feasible reward sets for (a) one optimal demonstrator (black cone) and (b) two suboptimal demonstrators (intersection of blue and red cones).

3 IRL FROM MULTIPLE SUBOPTIMAL DEMONSTRATORS

3.1 PROPOSED IRL FRAMEWORK WITH THEORETICAL GUARANTEES

Feasible reward set. We start by extending the feasible reward set of a single optimal demonstrator to the multiple suboptimal demonstrations setting. Consider K demonstrators with policies $\{\pi_e^k\}_{k=1}^K$ and corresponding discounted occupancy measures $d_e^k := d^{\pi_e^k}$. Suppose that each demonstrator policy is known to be at most ϵ^k suboptimal under true rewards, where $\epsilon^k \in [0, 1]$. Let $\mathcal{D} := \{(d_e^k, \epsilon^k)\}_{k=1}^K$ denote the set of occupancy-suboptimality pairs.

Unless we have $\epsilon^k = 0$ for some k, we do not have access to an optimal demonstrator. Therefore, we first recover the feasible reward set that makes d_e^k ϵ^k-suboptimal, and find the intersection over all k. Analogously to the optimal occupancy set $\Phi^\star(r)$ equation 7 in IRL from a single optimal demonstrator, we define the ϵ-suboptimal occupancy set for any reward r as

$$\Phi^\star(r; \epsilon) := \{d \in \Phi \mid \mathrm{SubOpt}(r, d) \leq \epsilon\}. \tag{11}$$

We then define the feasible reward set for which d_e^k is at most ϵ^k suboptimal, and take the intersection over all demonstrators:

Definition 3.1 (Feasible reward set for multiple suboptimal demonstrations)**.**

$$\mathcal{R}(\mathcal{D}) := \bigcap_{k=1}^{K} \mathcal{R}(d_e^k; \epsilon^k), \quad \text{where} \quad \mathcal{R}(d_e^k; \epsilon^k) := \{r \in \Delta(S \times A) \mid d_e^k \in \Phi^\star(r; \epsilon^k)\}. \tag{12}$$

Proposition 3.2. *The feasible reward set $\mathcal{R}(\mathcal{D})$ can be expressed as:*

$$\mathcal{R}(\mathcal{D}) := \{r \in \Delta(S \times A) \mid \exists v \in \mathbb{R}^{|S|} \text{ s.t. } (1-\gamma)\mu_0^\top v - r^\top d_e^k \leq \epsilon^k \quad \forall k \in [K],\ M^\top v \geq r\}. \tag{13}$$

The proof is provided in Appendix B. The 2D illustration of the feasible reward sets for a single optimal demonstrator and multiple suboptimal demonstrators is shown in Figure 1(a) and (b), respectively.

Next, we define the ground-truth optimal policy and the target reward set we aim to recover.

Definition 3.3 (Ground-truth solution)**.** Let $\pi_e^\star$ denote the ground-truth optimal policy and $d_e^\star := d^{\pi_e^\star} \in \Phi$ denote its occupancy measure. Let $\mathcal{R}^\star := \mathcal{R}(d_e^\star)$ denote the feasible reward set of $\pi_e^\star$.

Note that neither $d_e^\star$ nor $\mathcal{R}^\star$ can be directly estimated because demonstrations from the optimal demonstrator $\pi_e^\star$ are unavailable. Instead, we observe the dataset $\mathcal{D} := \{(d_e^k, \epsilon^k)\}_{k=1}^K$ and analyze how the feasible reward set $\mathcal{R}(\mathcal{D})$ shrinks toward $\mathcal{R}^\star$ or a subset of it.

Monotonic shrinkage property. First, we investigate the monotonic shrinkage property of the feasible reward set as new demonstrations are added. Specifically, we show that the feasible reward set becomes strictly tighter when a new demonstrator visits previously unseen state-action pairs and its declared suboptimality level is sufficiently small relative to the newly covered mass. This follows from the fact that our formulation is intersection-based: each additional demonstration contributes another linear compatibility constraint. The following theorem gives a coverage-based sufficient condition under which adding a new demonstrator strictly shrinks the feasible reward set.

Theorem 3.4. *Let $\mathcal{D}^+ := \mathcal{D} \cup \{(d'_e, \epsilon')\}$ denote the augmented dataset with the additional datapoint. Let $\mathcal{Z}_\mathcal{D} := \bigcup_{k=1}^{K} \mathrm{supp}(d_e^k)$ be the set of state-action pairs visited in $\mathcal{D}$, and let $\alpha(d'_e, \mathcal{D}) := \sum_{(s,a)\notin\mathcal{Z}_\mathcal{D}} d'_e(s,a)$ be the mass that the new demonstrator places outside the previously visited set. If*

$$\epsilon' < \frac{\alpha(d'_e, \mathcal{D})}{|\mathcal{Z}_\mathcal{D}|}, \tag{14}$$

then $\mathcal{R}(\mathcal{D}^+) \subsetneq \mathcal{R}(\mathcal{D})$.

The proof is provided in Appendix C. We make a few remarks about this theorem. First, the theorem provides only a sufficient condition on the new data for strictly shrinking the set $\mathcal{R}(\mathcal{D})$, and the necessary and sufficient condition is given in Lemma C.2. Second, the denominator $|\mathcal{Z}_\mathcal{D}|$ may appear large because it scales with the number of visited state-action pairs, but this dependence is a consequence of the simplex normalization in our feasible reward set definition: under this normalization, the average reward over the full state-action space is only $1/(|S||A|)$. If a different constant reward scale is used, the condition rescales accordingly.

Optimality guarantees for the recovered rewards. Next, we bound the mismatch between $\mathcal{R}(\mathcal{D})$ and the feasible reward set of the ground-truth optimal demonstrator, $\mathcal{R}^\star$. To quantify this mismatch, we use the worst-case suboptimality of $d_e^\star$ over rewards in $\mathcal{R}(\mathcal{D})$:

$$\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) := \max_{r\in\mathcal{R}(\mathcal{D})} \mathrm{SubOpt}(r, d_e^\star). \tag{15}$$

A small gap ensures that every retained reward makes $d_e^\star$ nearly optimal. Specifically, if $\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) \to 0$, then $\mathcal{R}(\mathcal{D})$ converges to $\mathcal{R}^\star$ in the one-sided Hausdorff distance induced by the ℓ_∞ norm (see Appendix D for the proof). If $\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) = 0$, then $\mathcal{R}(\mathcal{D}) \subseteq \mathcal{R}^\star$, and any selected reward in $\mathcal{R}(\mathcal{D})$ exposes an optimal face of Φ containing $d_e^\star$. If the selected reward additionally lies in $\mathrm{relint}(\mathcal{R}^\star)$, it exposes exactly the minimal face containing $d_e^\star$ (see Appendix E for detailed geometric characterization).

The following theorem gives an explicit bound in terms of how closely a mixture of demonstrator occupancies approximates the optimal occupancy and the corresponding mixture of suboptimality bounds.

Theorem 3.5. *The following inequality holds:*

$$\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) \le \min_{(d,\epsilon)\in\mathrm{conv}(\mathcal{D})} \{\|d - d_e^\star\|_1 + \epsilon\}. \tag{16}$$

The proof is provided in Appendix F. Intuitively, the theorem guarantees a small gap when a mixture of demonstrator occupancies is close to the optimal occupancy and the corresponding weighted average of declared suboptimality bounds is small.

We provide two additional sufficient conditions for a small gap beyond the occupancy-proximity bound in Theorem 3.5. Appendix G gives a local separation condition: each reward that makes $d_e^\star$ at least δ-suboptimal must admit a reference reward and a demonstrator satisfying directional separation and slack conditions that force a feasibility violation. A complementary coverage-based bound is given in Theorem H.1 in Appendix H. If a nonnegative weighted sum of demonstrator occupancies dominates $d_e^\star$ coordinatewise, then the gap is bounded by the same weighted sum of their declared suboptimality bounds. The weights need not sum to one, so this condition does not require a convex mixture to be close to $d_e^\star$. A key implication of these conditions is that $\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star)$ can be effectively reduced even with demonstrators whose declared suboptimality bounds ϵ^k are relatively large. We illustrate this point with a simple bandit example in Appendix I.

Comparison with baselines. Next, we compare our approach with two standard baselines for multiple suboptimal demonstrations: a MaxEnt estimator with demonstrator-specific inverse temperatures (Beliaev et al., 2022; Beliaev & Pedarsani, 2025), where we determine β_k as a function of the declared suboptimality levels, and a ranking-based estimator based on Bradley–Terry–Luce (BTL) pairwise likelihoods (Bradley & Terry, 1952; Luce, 1959), as in T-REX/D-REX-style methods (Brown et al., 2019; 2020). Both baselines optimize a single global objective over all demonstrations, unlike our intersection-based feasible-set formulation, and therefore need not track the shrinking feasible reward set. The next proposition shows that each can retain a constant mismatch with $\mathcal{R}^\star$ even when the bound in Theorem 3.5 is arbitrarily small; formal definitions and proofs are in Appendices J and K.

Proposition 3.6. *For each of the MaxEnt and ranking-based estimators described in Appendix J and Appendix K, there exists a constant $C > 0$ such that, for every $\eta > 0$, one can construct a finite dataset $\mathcal{D}$ for which the upper bound in Theorem 3.5 is at most η, but every optimizer $\hat{r}$ satisfies* $\mathrm{SubOpt}(\hat{r}, d_e^\star) \geq C$.

3.2 Addressing reward ambiguity issue in IRL

Reward ambiguity issue. The reward set $\mathcal{R}(\mathcal{D})$ suffers from the reward ambiguity issue as in IRL from a single optimal demonstrator (Ng & Russell, 2000; Metelli et al., 2021; 2023; Kim et al., 2024). The main issue is the existence of *degenerate or sparse rewards* in $\mathcal{R}(\mathcal{D})$, such as $r = \frac{1}{|S||A|}\mathbf{1}$, that cannot separate the demonstrators' behaviors from other suboptimal behaviors. An LP formulation that only imposes constraints in equation 13 does not rule out such degenerate solutions, and a solver may return one depending on the objective or tie-breaking rule. We consider two complementary ways to reduce this ambiguity: imposing additional constraints and setting the objective function.

Imposing additional constraints. We consider two ways of imposing additional constraints beyond the constraints in equation 13 to reduce the feasible reward set. First, a coverage-based penalty can discourage rewards that make unobserved state–action pairs appear optimal by enforcing positive Bellman inequality slack on pairs with zero aggregate demonstrator occupancy. Geometrically, this is closely related to selecting rewards from the interior of the feasible set (Lazzati & Metelli, 2026). Second, pairwise performance-gap information, when available, can be encoded as additional linear constraints. These constraints further shrink $\mathcal{R}(\mathcal{D})$ and rule out rewards that cannot separate better demonstrators from worse ones. Formal definitions and their advantages are discussed in detail in Appendices L and M.

Reward selection within the feasible set. Even after these constraints are imposed, the feasible set may contain many rewards. We select a representative reward using a baseline policy known, or assumed, to be substantially worse than the demonstrators, by maximizing the return gap between a convex combination of demonstrator occupancies and the baseline occupancy:

$$\hat{r} \in \operatorname*{arg\,max}_{r \in \mathcal{R}(\mathcal{D})} \; r^\top (\bar{d}_e - d_{\mathrm{base}}). \tag{17}$$

Here, $\bar{d}_e := \sum_{k=1}^K w_k d_e^k$, with $w_k \geq 0$ and $\sum_{k=1}^K w_k = 1$, is a convex combination representing the aggregate behavior of the demonstrators. For example, one may choose $w_k \propto 1 - \epsilon^k$ to assign greater weight to demonstrators with tighter suboptimality bounds. The baseline occupancy d_{base} corresponds to the baseline policy and is assumed to be inexpensive to estimate (when no natural baseline is available, one can use the uniform policy). Although the maximizer need not be unique, this objective excludes rewards that assign the same value to the demonstrators and the baseline. This strategy is particularly well-suited to LLM fine-tuning, where the pretrained base model provides a baseline and demonstrators improve over it only on the regions where they are competent.

3.3 Offline algorithm with function approximation

Datasets and function approximators. Next, we describe an offline function-approximation implementation of the tabular objective equation 17 for high-dimensional MDPs. The goal is to preserve the structure of the LP: maximize the empirical return gap between demonstrators and a baseline, while enforcing the demonstrator-suboptimality and Bellman-feasibility constraints in equation 13 through penalties. For each demonstrator k, we observe an offline dataset of N_k trajectories of horizon H, sampled i.i.d. from the initial state distribution μ_0, the unknown transition P, and π_e^k:

$$\mathcal{D}^k = \left\{ \tau^{k,n} = (s_0^{k,n}, a_0^{k,n}, \ldots, s_H^{k,n}) \right\}_{n=1}^{N_k}. \tag{18}$$

Then, the empirical estimate of d_e^k is computed from $\mathcal{D}^k$ by the standard truncated discounted occupancy estimate:

$$\hat{d}_e^k(s,a) := (1-\gamma)\frac{1}{N_k}\sum_{n=1}^{N_k}\sum_{h=0}^{H-1}\gamma^h \mathbf{1}\{(s_h^{k,n}, a_h^{k,n}) = (s,a)\} \quad \forall (s,a) \in S \times A. \tag{19}$$

We parameterize the reward function and the value function as $r_\theta : S \times A \mapsto [0,1]$ and $v_\phi : S \mapsto \mathbb{R}$, respectively. Given an offline dataset $\mathcal{D} = \{(\mathcal{D}^k, \epsilon^k)\}_{k=1}^K$, we train (θ, ϕ) by optimizing the following empirical penalized objective.

Loss function. The first term below is the empirical version of the reward-selection objective equation 17; the remaining terms penalize violations of the two LP constraints in equation 13.

$$L(r_\theta, v_\phi) := -r_\theta^\top \left(\sum_{k=1}^K w_k \hat{d}_e^k - \hat{d}_{\text{base}}\right) + \lambda_{\text{sub}} \frac{1}{K}\sum_{k=1}^K \rho_{\leq}\Big((1-\gamma)\mu_0^\top v_\phi - r_\theta^\top \hat{d}_e^k - \epsilon^k\Big) \tag{20}$$

$$+ \lambda_{\text{bell}} \mathbb{E}_{(s,a)\sim\bar{\mathcal{D}}}\big[\rho_{\leq}\big(r_\theta(s,a) + \gamma \mathbb{E}_{s'\sim\bar{\mathcal{D}}(\cdot|s,a)}[v_\phi(s')] - v_\phi(s)\big)\big], \tag{21}$$

where $\lambda_{\text{sub}}, \lambda_{\text{bell}} > 0$ weight the demonstrator suboptimality penalty and the Bellman feasibility penalty, respectively, and $\rho_{\leq}$ denotes a convex non-decreasing penalty function, such as squared hinge $\rho_{\leq}(x) = \max\{0, x\}^2$. Here $(s,a) \sim \bar{\mathcal{D}}$ denotes the empirical state-action distribution of the full offline dataset, and $\bar{\mathcal{D}}(\cdot \mid s,a)$ denotes its empirical next-state distribution. This Bellman term is therefore a support-limited relaxation of the pointwise constraint $M^\top v \geq r$ in equation 13. (θ, ϕ) can be updated by stochastic gradient methods on $L(r_\theta, v_\phi)$, enforcing bounded rewards by parameterization (e.g., sigmoid layer).

Complexity and convergence. In the tabular setting, equation 17 is an LP with $|S||A| + |S|$ variables, K demonstrator-gap constraints, Bellman inequalities over $S \times A$, and the simplex constraints, so generic LP methods run in polynomial time in these dimensions. Under function approximation, the cost scales with the number of gradient updates, the number of sampled transitions and demonstrators included in each batch, and the forward/backward cost of the reward and value networks. When an explicit reward is not needed, this cost can be reduced by directly parameterizing the advantage function instead of separately learning r_θ and v_ϕ; the resulting advantage-parameterized objective is given in Appendix N. While convergence of specific optimization algorithms is beyond the scope of this paper, the empirical objective can be viewed as a standard sample-average approximation of its population analogue (Shapiro et al., 2021). Sample-complexity bounds for estimating $\hat{d}_e^k$ are available in (Kim et al., 2024).

Sensitivity and calibration of suboptimality bounds. The proposed algorithm requires specifying the suboptimality bounds ϵ^k, which generally must be estimated in practice. The effect of misspecification depends on whether these bounds are overestimated or underestimated. Increasing valid bounds from ϵ^k to $\epsilon^k + \Delta^k$, with $\Delta^k \geq 0$, loosens the upper bound in Theorem 3.5 by at most $\max_k \Delta^k$. Conversely, bounds that are too small may exclude rewards in $\mathcal{R}^\star$ and lead to degenerate solutions.

This trade-off motivates a practical tuning heuristic when reliable bounds are unavailable. We start with loose bounds and gradually tighten them while monitoring constraint violations, reward degeneracy, and an external quality signal, such as human feedback or a task-specific rubric, to evaluate the induced policies. When finite-sample confidence bounds are available, we can additionally account for occupancy estimation error by introducing confidence-based margins. Specifically, if $\|\hat{d}_e^k - d_e^k\|_1 \leq \eta_k$ holds with high probability, bounded rewards imply that replacing ϵ^k with $\epsilon^k + \eta_k$ in the empirical constraint preserves compatibility with the population constraint on the same event, provided the population-level bound is valid. This heuristic offers a practical approach to bound calibration, while more systematic calibration methods remain a direction for future work. We use this heuristic in our numerical experiments and evaluate sensitivity to misspecified bounds.

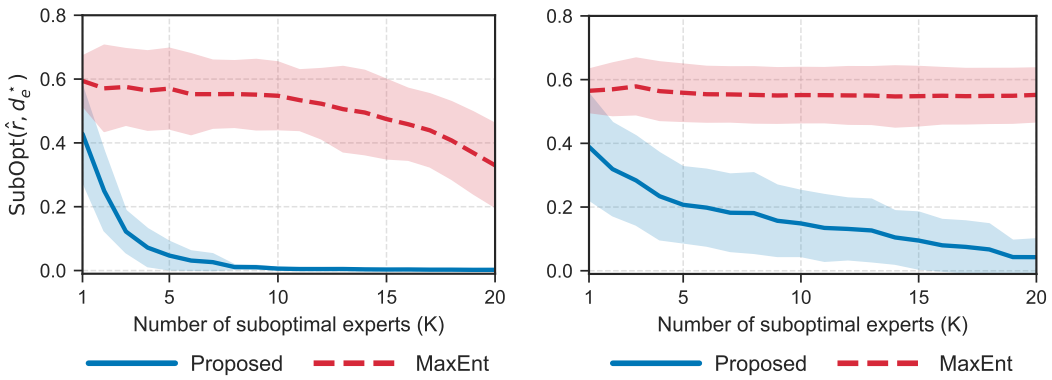


Figure 2: Suboptimality of the recovered reward for the proposed method and the MaxEnt baseline. Curves denote means and shaded bands denote one standard deviation over 20 random maps. Left: Case 1 (increasing quality and coverage). Right: Case 2 (fixed quality with increasing coverage).

4 Numerical Experiments

4.1 Grid-world with detour portals

MDP setting. We use a tabular deterministic 20×20 Grid-world with terminal state $(20, 20)$, uniform initial distribution, and 20 detour portals $\{(i_p, o_p)\}_{p=1}^{20}$ with distinct entrances. The action space is $A = \{\mathsf{U}, \mathsf{D}, \mathsf{L}, \mathsf{R}, \mathsf{IN}\}$: movement actions are standard grid moves, while IN teleports from i_p to o_p at portal entrances and otherwise stays in place.

Datasets. Let $\pi_e^\star$ denote the optimal policy that knows the correct action at all portal entrances. We consider two cases for suboptimal demonstration data. In Case 1, each demonstrator π_e^k has knowledge about k portals, i.e., $\{(i_p, o_p)\}_{p=1}^{k}$. This reflects the condition in Theorem 3.5 that the quality of the added demonstrator increases with K. When $K = 20$, we have an optimal demonstration in the dataset. In Case 2, each demonstrator π_e^k has knowledge about only one portal $p = k$, i.e., (i_k, o_k). This setting examines whether highly suboptimal demonstrators can collectively recover useful rewards through complementary coverage, as suggested by the occupancy-coverage analysis in Appendix H. For portals unknown to a demonstrator, we assume that the demonstrator follows the optimal policy of the grid-world in which detour portals are ignored. We illustrate the environment and the two cases of suboptimal demonstrators in Figure 5 in Appendix O.1.

For each demonstrator π_e^k, we sample $N = 100$ trajectories of horizon $H = 40$ and compare the solution of equation 17 with a solution of the MaxEnt baseline described in Appendix J. We omit the ranking-based baseline because it performs worse than both methods in Case 1 and is not applicable in Case 2, where the strict ordering between demonstrators is unavailable. The performance of the learned reward is evaluated by the suboptimality of the optimal occupancy, i.e., $\mathrm{SubOpt}(\widehat{r}, d_e^\star)$, and the results are reported as the mean and standard deviation over 20 random maps for each value of K. Other hyperparameters are $\gamma = 0.95, \epsilon^k = 0.1, \beta_k = 1.0, w_k = 1/K$ for all k.

Results. The results are shown in Figure 2. In Case 1, both methods benefit from clearer quality differences, but the proposed method improves faster: its mean suboptimality decreases from 0.428 at $K = 1$ to 0.002 at $K = 20$, whereas MaxEnt decreases from 0.594 to 0.330. In Case 2, the gap is more pronounced: the MaxEnt baseline remains nearly flat, with mean suboptimality 0.564 at $K = 1$ and 0.550 at $K = 20$, as the reward is fitted to the aggregated data with only highly suboptimal demonstrators. By contrast, the proposed feasible-set method decreases from 0.389 at $K = 1$ to 0.043 at $K = 20$, consistent with our theoretical analysis and the intuition behind the intersection-based formulation. Overall, these results confirm that our framework recovers substantially better rewards from suboptimal demonstrations than the MaxEnt baseline. Additionally, we conduct a sensitivity analysis in the Case-2 setting by systematically overestimating and underestimating the declared suboptimality bounds; the results and their implications for bound calibration are reported in Appendix O.2.

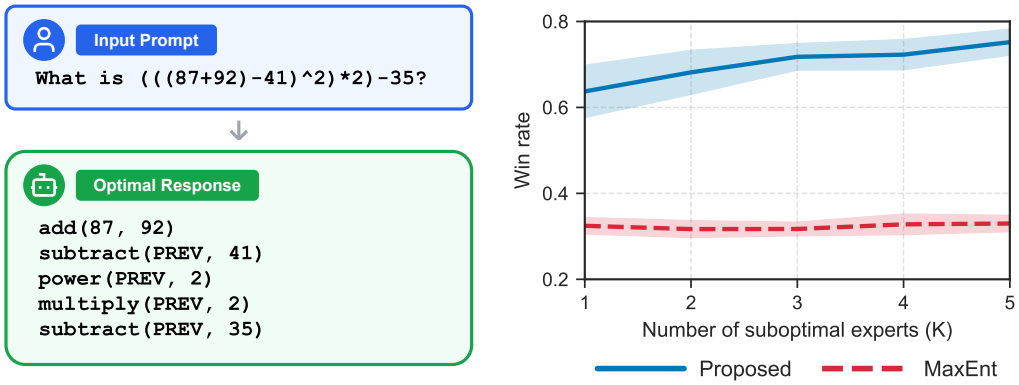


Figure 3: LLM experiment on arithmetic tool-use tasks. Left: Example of prompt and the optimal response. Right: Win rate of optimal policy under the trained advantage model.

4.2 LLM FINE-TUNING ON ARITHMETIC TOOL-USE TASKS

Arithmetic tool-use tasks. We evaluate the function approximation algorithm from Section 3.3 in the LLM domain with arithmetic tool-use tasks. Each task consists of a single-turn arithmetic prompt whose answer is obtained through a chain of 1–5 calls to five arithmetic tools (`add`, `subtract`, `multiply`, `divide`, `power`). An optimal response is a newline-separated trace with one call per line, using a dedicated `PREV` token when an operand is the output of the previous line (Figure 3 left).

Datasets and evaluation. Each of the 5 suboptimal demonstrators knows how to use one tool and generates an incorrect (dummy) tool call for all other tools; increasing K from 1 to 5 expands coverage of the tool set while keeping per-demonstrator quality low. We used Qwen2.5-1.5B (Yang et al., 2024) with an additional advantage head and trained it using the proposed method with the advantage-parameterized loss equation 163 and the MaxEnt baseline (Appendix J). Since the direct evaluation of the suboptimality measure can be blurred by an extra policy-optimization stage (for computing an optimal policy) and is sensitive to the reward (or advantage) scale, instead we report the win-rate of the optimal response with respect to the near-optimal response (which has at most one error in tool calling) as a proxy. Additional details are provided in Appendix O.3.

Results. Figure 3 shows the mean and standard deviation of the win-rate depending on the number of demonstrators K, over 10 runs. The MaxEnt baseline achieves an average win-rate of only 0.323, below the 0.5 win-rate of a constant reward, because it fits the reward to aggregated data from only suboptimal demonstrators. The proposed method improves the win-rate (average win-rate 0.702) over MaxEnt, increasing from 0.637 at $K = 1$ to 0.752 at $K = 5$, consistent with the benefit of combining complementary demonstrations observed in the tabular experiments above. These results suggest that our framework remains effective beyond tabular domains when implemented with the proposed function approximation algorithm.

5 CONCLUSION

This work develops a feasible-set-based IRL framework for learning from multiple imperfect demonstrators without assuming access to an optimal demonstrator. By encoding demonstrator suboptimality as linear constraints and intersecting them, the approach progressively refines the feasible reward set and establishes conditions under which every retained reward makes the unobserved optimal demonstrator nearly optimal. Experiments in both tabular and high-dimensional settings support these predictions. Overall, the framework opens a new avenue for principled IRL beyond likelihood-based methods, particularly in realistic scenarios with imperfect and heterogeneous demonstrations. Finally, we note that this paper focuses primarily on the theoretical advantages of the proposed framework. Improving implementation efficiency, evaluating optimization methods, and developing online extensions remain directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to improve readability, identify relevant literature, assist with experimental code, and construct numerical examples. Code assistance included writing programs to generate synthetic datasets, which were generated programmatically rather than directly by an LLM. We have not used generative AI tools to develop theoretical models, formulate mathematical claims, assist with proofs, design research methodology or experiments, or interpret results. We have reviewed all AI-assisted work, including the text, identified relevant literature, code, and numerical examples. We take responsibility for the final content of this work and for all AI-assisted contributions.

ETHICS STATEMENT

This work is primarily a theoretical contribution to IRL from multiple imperfect demonstrators. Potential positive impacts include reducing reliance on a single near-optimal demonstrator and enabling reward learning from more realistic heterogeneous data, which may be useful in applications such as RLHF, robotics, and assistive decision-making where demonstrations from optimal demonstrators are costly or unavailable. At the same time, reward-learning methods can inherit biases or systematic errors present in the demonstrations, so care is needed when applying them in safety-critical or high-stakes domains.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the problem setting, assumptions, proposed framework, and complete proofs of all theoretical results. The experimental settings and implementation details are provided in Appendix O, including the construction of the environments and synthetic datasets, model architecture, training procedure, evaluation protocol, and hyperparameters.

REFERENCES

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, 2004.

Stephen Adams, Tyler Cody, and Peter A Beling. A survey of inverse reinforcement learning. *Artificial Intelligence Review*, 55(6):4307–4346, 2022.

Monica Babes, Vukosi Marivate, Kaushik Subramanian, and Michael L Littman. Apprenticeship learning about multiple intentions. In *International Conference on Machine Learning*, pp. 897–904, 2011.

Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine intelligence 15*, pp. 103–129, 1995.

Mark Beliaev and Ramtin Pedarsani. Inverse reinforcement learning by estimating expertise of demonstrators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15532–15540, 2025.

Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, and Ramtin Pedarsani. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*, pp. 1732–1748. PMLR, 2022.

Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 182–189. JMLR Workshop and Conference Proceedings, 2011.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International Conference on Machine Learning*, pp. 783–792. PMLR, 2019.

Daniel S Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on Robot Learning*, pp. 330–359. PMLR, 2020.

Alexander Bukharin, Ilgee Hong, Haoming Jiang, Zichong Li, Qingru Zhang, Zixuan Zhang, and Tuo Zhao. Robust reinforcement learning from corrupted human feedback. *Advances in Neural Information Processing Systems*, 37:124093–124113, 2024.

Letian Chen, Rohan Paleja, Muyleng Ghuy, and Matthew Gombolay. Joint goal and strategy inference across heterogeneous demonstrators via reward network distillation. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pp. 659–668, 2020.

Letian Chen, Rohan Paleja, and Matthew Gombolay. Learning from suboptimal demonstration via self-supervised reward regression. In *Conference on robot learning*, pp. 1262–1277. PMLR, 2021.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Christos Dimitrakakis and Constantin A Rothkopf. Bayesian multitask inverse reinforcement learning. In *European Workshop on Reinforcement Learning*, pp. 273–284, 2011.

Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pp. 49–58. PMLR, 2016.

Till Freihaut and Giorgia Ramponi. On feasible rewards in multi-agent inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 38:114230–114284, 2025.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Yang Gao, Huazhe Xu, Ji Lin, Fisher Yu, Sergey Levine, and Trevor Darrell. Reinforcement learning from imperfect demonstrations. *arXiv preprint arXiv:1802.05313*, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Alan J Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4), 1952.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

Sravan Jayanthi, Letian Chen, Nadya Balabanska, Van Duong, Erik Scarlatescu, Ezra Ameperosa, Zulfiqar Haider Zaidi, Daniel Martin, Taylor Keith Del Matto, Masahiro Ono, et al. Droid: Learning from offline heterogeneous demonstrations via reward-policy distillation. In *Conference on Robot Learning*, pp. 1547–1571. PMLR, 2023.

Kihyun Kim, Jiawei Zhang, Asuman E Ozdaglar, and Pablo Parrilo. A unified linear programming framework for offline reward learning from human demonstrations and feedback. In *International Conference on Machine Learning*, pp. 24694–24712. PMLR, 2024.

Filippo Lazzati and Alberto Maria Metelli. Generalizing behavior via inverse reinforcement learning with closed-form reward centroids. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, pp. 685–693. PMLR, 2026.

David Lindner, Andreas Krause, and Giorgia Ramponi. Active exploration for inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5843–5853, 2022.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

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

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

Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 663–670, 2000.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. Rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

Javier Pena, Juan C Vera, and Luis F Zuluaga. New characterizations of hoffman constants for systems of linear constraints. *Mathematical Programming*, 187(1):79–109, 2021.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Riccardo Poiani, Gabriele Curti, Alberto M Metelli, and Marcello Restelli. Sub-optimal experts mitigate ambiguity in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 37:85778–85823, 2024.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

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

Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *International Conference on Machine Learning*, pp. 729–736, 2006.

Nathan D Ratliff, David Silver, and J Andrew Bagnell. Learning to search: Functional gradient techniques for imitation learning. *Autonomous Robots*, 27(1):25–53, 2009.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Kyriacos Shiarlis, Joao Messias, and Shimon Whiteson. Inverse reinforcement learning from failure. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 1060–1068, 2016.

Harshit Sikchi, Akanksha Saran, Wonjoon Goo, and Scott Niekum. A ranking game for imitation learning. *Transactions on machine learning research*, 2023.

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

Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pp. 1032–1039, 2008.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Michael Jq Zhang, Zhilin Wang, Jena D Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree and do models know? In *International Conference on Machine Learning*, pp. 76193–76212. PMLR, 2025.

Songyuan Zhang, Zhangjie Cao, Dorsa Sadigh, and Yanan Sui. Confidence-aware imitation learning from demonstrations with varying optimality. *Advances in Neural Information Processing Systems*, 34:12340–12350, 2021.

Lei Zhao, Mengdi Wang, and Yu Bai. Is inverse reinforcement learning harder than standard reinforcement learning? a theoretical perspective. In *International Conference on Machine Learning*, pp. 60957–61020. PMLR, 2024.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, pp. 1433–1438, 2008.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A ADDITIONAL RELATED WORK

IRL from a single optimal demonstrator and apprenticeship learning. Classical IRL studies the recovery of a reward function that rationalizes a single optimal demonstrator's behavior. The foundational formulation of Ng & Russell (2000) characterizes rewards under which the demonstrated policy is optimal, thereby making explicit the non-identifiability of IRL and motivating solution concepts based on constraints. Abbeel & Ng (2004) shifted the emphasis toward apprenticeship learning through feature expectation matching, while Ratliff et al. (2006) introduced max-margin formulations that select rewards or cost functions separating expert behavior from alternatives. The game-theoretic and LP apprenticeship-learning formulations (Syed & Schapire, 2007; Syed et al., 2008) are also relevant to our occupancy-measure viewpoint, as they cast learning from demonstrations as an optimization problem over policies and rewards. Our contribution differs from this lineage in that we do not assume access to an optimal demonstrator to match or compete with; instead, we build a feasible reward set by intersecting slack-bounded constraints from multiple imperfect demonstrators. Probabilistic approaches model expert stochasticity more directly: Bayesian IRL (Ramachandran & Amir, 2007) places a posterior over rewards, and MaxEnt IRL (Ziebart et al., 2008) resolves ambiguity by choosing the distribution over trajectories with highest entropy subject to matching demonstration statistics. These likelihood-based ideas have since become a dominant template for scalable IRL, including scalable variants such as Relative Entropy IRL (Boularias et al., 2011), Guided Cost Learning (Finn et al., 2016), and adversarial IRL (AIRL) (Fu et al., 2017).

Feasible-set approaches in IRL. Our work is most closely related to the feasible-reward-set view of IRL, which frames reward learning not as selecting a single maximum-likelihood reward, but as characterizing the full set of rewards consistent with the demonstrations through optimality constraints. This perspective is implicit in the original IRL formulation (Ng & Russell, 2000) and is developed explicitly in Metelli et al. (2021), which studies the estimation and transfer of compatible rewards, and Metelli et al. (2023), which formalizes feasible-reward-set estimation in finite-horizon MDPs and establishes sample-complexity bounds. Lindner et al. (2022) further study an efficient trajectory-sampling strategy that relaxes the need for a state-action generative model when estimating the feasible reward set. In the offline setting, Zhao et al. (2024) and Kim et al. (2024) analyze sample complexity using pessimism-based and LP-based approaches, respectively. Freihaut & Ramponi (2025) study the feasible reward set for multi-agent IRL in entropy-regularized Markov games. Poiani et al. (2024) show that incorporating additional suboptimal demonstrations can shrink the feasible set and reduce ambiguity. More recently, Lazzati & Metelli (2026) study how to select principled representatives, such as reward centroids, from the feasible region to improve generalization. Our method fits squarely within this feasible-set tradition, but differs in two important respects: we consider the setting with multiple suboptimal demonstrations only and no optimal demonstrator, and we aim to recover the ground-truth reward set rather than merely shrink the feasible set.

IRL from multiple demonstrations via MLE-based approaches. A natural extension of single-optimal-demonstrator MaxEnt IRL to heterogeneous demonstrators is to assume a shared latent reward while introducing demonstrator-specific noise, confidence, or inverse-temperature parameters, then fit the reward by maximizing an aggregate likelihood over all demonstrations. This line includes Bayesian multi-task IRL formulations that cluster demonstrations and infer task-specific rewards (Dimitrakakis & Rothkopf, 2011; Babes et al., 2011), as well as more recent methods that explicitly estimate demonstrator expertise so that more reliable demonstrators exert more influence on the learned reward (Beliaev et al., 2022; Beliaev & Pedarsani, 2025). Relatedly, IRL from failure shows that successful and failed demonstrations can reduce reward ambiguity (Shiarlis et al., 2016). The main strength of this family is statistical efficiency and compatibility with standard probabilistic modeling, since heterogeneity is absorbed into soft-optimality parameters or latent clusters rather than hard constraints. However, these methods still commit to a single point estimate chosen to explain the observed visitation frequencies, which can be problematic when all demonstrators are suboptimal and no optimal demonstrator is available: the failure mode emphasized in our main text.

Ranking and preference-based reward learning. Another line of work leverages ranking or margin information rather than assuming a probabilistic model for likelihood maximization. Classical choice models such as Bradley–Terry (Bradley & Terry, 1952), Luce choice (Luce, 1959), and Plackett–Luce ranking models (Plackett, 1975) provide statistical foundations for pairwise and

listwise preference likelihoods. The idea of enforcing a margin constraint appears in early imitation-learning work such as maximum margin planning (MMP) (Ratliff et al., 2006; 2009), which imposes a margin in the reward-optimization problem so that the demonstrated policy attains higher expected reward than alternative policies. In the suboptimal-data regime, this idea reappears in trajectory-ranking and preference-based formulations such as T-REX (Brown et al., 2019), which learns a reward from approximately ranked demonstrations and can extrapolate beyond imperfect behavior, and D-REX (Brown et al., 2020), which automatically generates rankings by perturbing a cloned policy. More recently, the ranking-game framework of Sikchi et al. (2023) proposes a two-player formulation that jointly learns a reward and a policy from expert demonstrations and offline preferences. Although these methods are closely related in spirit to our objective because they use comparative structure to rule out degenerate rewards, they typically rely on explicit preference or ranking information rather than the maximum suboptimality level considered here. In practical settings, explicit preference or ranking information is often difficult to obtain because demonstrators may have different areas of expertise.

Imitation learning (IL). IL learns policies directly from demonstrations rather than first recovering an explicit reward function. The simplest formulation is behavior cloning, which casts IL as supervised learning over expert state-action pairs but is prone to distribution shift in sequential decision-making (Bain & Sammut, 1995). Reduction-based and interactive methods such as SMILe and DAgger address the resulting compounding-error and covariate-shift issues by analyzing sequential prediction and querying the expert on learner-induced states (Ross & Bagnell, 2010; Ross et al., 2011). Value-aware extensions such as AggreVaTe incorporate expert cost-to-go information so that imitation reflects long-horizon decision quality (Ross & Bagnell, 2014), while GAIL reframes imitation as adversarial occupancy-measure matching (Ho & Ermon, 2016). More directly related to our motivation, Zhang et al. (2021) learn confidence scores and policies from demonstrations with varying optimality, and Gao et al. (2018) study reinforcement learning from imperfect demonstrations using an actor-critic objective designed to surpass demonstrator performance. These works address the practical problem of learning good policies from imperfect data; our focus is instead reward recovery and feasible-set ambiguity under declared suboptimality bounds.

Reinforcement learning from human feedback (RLHF). Our motivation from LLM raters connects the proposed setting to RLHF. Christiano et al. (2017) learn rewards from non-expert human pairwise preferences, Ibarz et al. (2018) combine demonstrations and trajectory preferences in Atari, Ziegler et al. (2019) fine-tune language models from human preferences, and Ouyang et al. (2022) establish the InstructGPT pipeline using demonstrations, rankings, reward modeling, and policy optimization. These works use human comparisons to train reward models and policies at scale. In our framework, such comparisons can impose additional linear constraints on the reward set, thereby reducing ambiguity, as discussed in Appendix M.

B Proof of Proposition 3.2

By Definition 3.1,

$$\mathcal{R}(\mathcal{D}) = \left\{ r \in \Delta(S \times A) \mid \mathrm{SubOpt}(r, d_e^k) \le \epsilon^k \quad \forall k \in [K] \right\}. \tag{22}$$

Thus, for a fixed reward vector r, it suffices to prove

$$\mathrm{SubOpt}(r, d_e^k) \le \epsilon^k \quad \forall k \in [K] \tag{23}$$

$$\iff \quad \exists v \in \mathbb{R}^{|S|} \quad \text{s.t.} \quad M^\top v \ge r,\ (1-\gamma)\mu_0^\top v - r^\top d_e^k \le \epsilon^k \quad \forall k \in [K]. \tag{24}$$

Consider the LP formulation of an MDP

$$\max_{d \in \Phi} r^\top d, \tag{25}$$

whose dual is

$$\min_{v \in \mathbb{R}^{|S|}} (1-\gamma)\mu_0^\top v \quad \text{s.t.} \quad M^\top v \ge r. \tag{26}$$

By strong duality, these two problems have the same optimal value, and the dual optimum is attained. Let

$$J^\star(r) := \min_{v:\, M^\top v \ge r} (1-\gamma)\mu_0^\top v = \max_{d \in \Phi} r^\top d. \tag{27}$$

Then, for every $k \in [K]$,

$$\mathrm{SubOpt}(r, d_e^k) \le \epsilon^k \quad \iff \quad J^\star(r) \le \epsilon^k + r^\top d_e^k. \tag{28}$$

Therefore,

$$\mathrm{SubOpt}(r, d_e^k) \le \epsilon^k \quad \forall k \in [K] \quad \iff \quad J^\star(r) \le \epsilon^k + r^\top d_e^k \quad \forall k \in [K]. \tag{29}$$

Since the dual optimum is attained, there exists $v^\star \in \mathbb{R}^{|S|}$ such that

$$M^\top v^\star \ge r, \qquad (1-\gamma)\mu_0^\top v^\star = J^\star(r). \tag{30}$$

This $v^\star$ depends only on r, not on k. Substituting $J^\star(r) = (1-\gamma)\mu_0^\top v^\star$ into the previous display yields

$$\mathrm{SubOpt}(r, d_e^k) \le \epsilon^k \quad \forall k \in [K] \tag{31}$$

$$\iff \quad \exists v^\star \in \mathbb{R}^{|S|} \quad \text{s.t.} \quad M^\top v^\star \ge r,\ (1-\gamma)\mu_0^\top v^\star - r^\top d_e^k \le \epsilon^k \quad \forall k \in [K], \tag{32}$$

which proves the forward implication. Conversely, suppose there exists some $v \in \mathbb{R}^{|S|}$ satisfying $M^\top v \ge r$ and

$$(1-\gamma)\mu_0^\top v - r^\top d_e^k \le \epsilon^k \quad \forall k \in [K]. \tag{33}$$

Because $J^\star(r)$ is the minimum dual value over all such feasible v,

$$J^\star(r) - r^\top d_e^k \le (1-\gamma)\mu_0^\top v - r^\top d_e^k \le \epsilon^k \quad \forall k \in [K]. \tag{34}$$

Thus $\mathrm{SubOpt}(r, d_e^k) \le \epsilon^k$ for every k, completing the proof.

C Proof of Theorem 3.4

Lemma C.1 (Coverage gap from newly visited mass)**.** *Let* $\mathcal{Z}_\mathcal{D} := \bigcup_{k=1}^K \mathrm{supp}(d_e^k)$ *and*

$$\alpha(d'_e, \mathcal{D}) := \sum_{(s,a) \notin \mathcal{Z}_\mathcal{D}} d'_e(s,a). \tag{35}$$

There exists a reward $r_\mathcal{Z} \in \mathcal{R}(\mathcal{D})$ *such that*

$$\mathrm{SubOpt}(r_\mathcal{Z}, d'_e) = \frac{\alpha(d'_e, \mathcal{D})}{|\mathcal{Z}_\mathcal{D}|}. \tag{36}$$

Proof. The idea is to build a reward that makes all previously observed state-action pairs equally valuable and assigns zero reward elsewhere. Define

$$r_\mathcal{Z}(s,a) := \begin{cases} 1/|\mathcal{Z}_\mathcal{D}|, & (s,a) \in \mathcal{Z}_\mathcal{D}, \\ 0, & (s,a) \notin \mathcal{Z}_\mathcal{D}. \end{cases} \tag{37}$$

Then $r_\mathcal{Z} \in \Delta(S \times A)$. Since every previous demonstrator has support contained in $\mathcal{Z}_\mathcal{D}$ and every occupancy measure has unit mass,

$$r_\mathcal{Z}^\top d_e^k = \frac{1}{|\mathcal{Z}_\mathcal{D}|} \quad \forall k \in [K]. \tag{38}$$

For any $d \in \Phi$, we also have $r_\mathcal{Z}^\top d \le 1/|\mathcal{Z}_\mathcal{D}|$. Thus, $\mathrm{SubOpt}(r_\mathcal{Z}, d_e^k) = 0 \le \epsilon^k$ for every k, and hence $r_\mathcal{Z} \in \mathcal{R}(\mathcal{D})$.

The only way the new demonstrator loses value under this reward is by placing mass outside the old support. Therefore,

$$r_\mathcal{Z}^\top d'_e = \frac{1 - \alpha(d'_e, \mathcal{D})}{|\mathcal{Z}_\mathcal{D}|}. \tag{39}$$

Since the previous demonstrators attain value $1/|\mathcal{Z}_\mathcal{D}|$ under $r_\mathcal{Z}$,

$$\mathrm{SubOpt}(r_\mathcal{Z}, d'_e) = \max_{d \in \Phi} r_\mathcal{Z}^\top d - r_\mathcal{Z}^\top d'_e = \frac{1}{|\mathcal{Z}_\mathcal{D}|} - \frac{1 - \alpha(d'_e, \mathcal{D})}{|\mathcal{Z}_\mathcal{D}|} = \frac{\alpha(d'_e, \mathcal{D})}{|\mathcal{Z}_\mathcal{D}|}. \tag{40}$$

□

Proof of Theorem 3.4. By Definition 3.1,

$$\mathcal{R}(\mathcal{D}^+) = \mathcal{R}(\mathcal{D}) \cap \mathcal{R}(d'_e; \epsilon'). \tag{41}$$

The augmented feasible set is therefore obtained by adding one more constraint to the old feasible set, so $\mathcal{R}(\mathcal{D}^+) \subseteq \mathcal{R}(\mathcal{D})$. By Lemma C.1, there exists $r_{\mathcal{Z}} \in \mathcal{R}(\mathcal{D})$ such that

$$\mathrm{SubOpt}(r_{\mathcal{Z}}, d'_e) = \frac{\alpha(d'_e, \mathcal{D})}{|\mathcal{Z}_{\mathcal{D}}|} > \epsilon'. \tag{42}$$

Therefore, $r_{\mathcal{Z}} \notin \mathcal{R}(d'_e; \epsilon')$. Thus the added constraint removes at least one reward that was previously feasible, namely $r_{\mathcal{Z}}$. Hence $\mathcal{R}(\mathcal{D}^+) \subsetneq \mathcal{R}(\mathcal{D})$. □

Lemma C.2 (Exact characterization of strict shrinkage)**.** $\mathcal{R}(\mathcal{D}^+) \subsetneq \mathcal{R}(\mathcal{D})$ *if and only if*

$$\epsilon' < \max_{r \in \mathcal{R}(\mathcal{D})} \mathrm{SubOpt}(r, d'_e). \tag{43}$$

Proof. The identity

$$\mathcal{R}(\mathcal{D}^+) = \mathcal{R}(\mathcal{D}) \cap \mathcal{R}(d'_e; \epsilon') \tag{44}$$

implies $\mathcal{R}(\mathcal{D}^+) \subseteq \mathcal{R}(\mathcal{D})$. Therefore, it remains to identify exactly when the last inclusion is strict. First, assume

$$\epsilon' < \max_{r \in \mathcal{R}(\mathcal{D})} \mathrm{SubOpt}(r, d'_e). \tag{45}$$

Then there exists some $\tilde{r} \in \mathcal{R}(\mathcal{D})$ such that

$$\mathrm{SubOpt}(\tilde{r}, d'_e) > \epsilon'. \tag{46}$$

Therefore, $\tilde{r} \notin \mathcal{R}(d'_e; \epsilon')$, and hence

$$\tilde{r} \notin \mathcal{R}(\mathcal{D}) \cap \mathcal{R}(d'_e; \epsilon') = \mathcal{R}(\mathcal{D}^+), \tag{47}$$

which yields the desired strict inclusion $\mathcal{R}(\mathcal{D}^+) \subsetneq \mathcal{R}(\mathcal{D})$.

Conversely, assume $\mathcal{R}(\mathcal{D}^+) \subsetneq \mathcal{R}(\mathcal{D})$. Since this inclusion is strict, there exists some $\hat{r} \in \mathcal{R}(\mathcal{D})$ such that $\hat{r} \notin \mathcal{R}(\mathcal{D}^+)$. Using $\mathcal{R}(\mathcal{D}^+) = \mathcal{R}(\mathcal{D}) \cap \mathcal{R}(d'_e; \epsilon')$ and $\hat{r} \in \mathcal{R}(\mathcal{D})$, we must have $\hat{r} \notin \mathcal{R}(d'_e; \epsilon')$. By Definition 3.1, this means

$$\mathrm{SubOpt}(\hat{r}, d'_e) > \epsilon'. \tag{48}$$

Since $\hat{r} \in \mathcal{R}(\mathcal{D})$, it follows that

$$\max_{r \in \mathcal{R}(\mathcal{D})} \mathrm{SubOpt}(r, d'_e) \geq \mathrm{SubOpt}(\hat{r}, d'_e) > \epsilon', \tag{49}$$

which proves the reverse implication. □

D One-sided Hausdorff distance and maximum suboptimality

Definition D.1 (One-sided Hausdorff distance)**.** For two nonempty sets $A, B \subset \mathbb{R}^m$, the one-sided Hausdorff distance from A to B with respect to the ℓ_∞ norm is defined as

$$d_H(A \mid B) := \sup_{x \in A} \inf_{y \in B} \|x - y\|_\infty. \tag{50}$$

Proposition D.2. *There exists a constant $C > 0$ depending on the reward-free MDP $\mathcal{M}$ and the fixed optimal occupancy $d_e^\star$ such that*

$$d_H\big(\mathcal{R}(\mathcal{D}) \mid \mathcal{R}^\star\big) \leq C \ \mathrm{Gap}\big(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star\big). \tag{51}$$

In particular, $\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) \to 0$ *implies that* $\mathcal{R}(\mathcal{D}) \to \mathcal{R}^\star$ *in the one-sided Hausdorff distance induced by the ℓ_∞ norm.*

Proof. Since Φ is a polytope, $\mathcal{R}^\star$ can equivalently be written as the simplex intersected with the finitely many inequalities $r^\top(d - d_e^\star) \leq 0$ for the vertices d of Φ. The Hoffman error bound states that, for a finite feasible system of linear inequalities, the distance to the feasible set is bounded by a constant times the maximum constraint violation (Hoffman, 1952; Pena et al., 2021). Therefore, there exists $C > 0$ such that for all $r \in \Delta(S \times A)$,

$$\inf_{r' \in \mathcal{R}^\star} \|r - r'\|_\infty \leq C \max_{d \in \Phi} \big(r^\top(d - d_e^\star)\big)_+ = C \ \mathrm{SubOpt}(r, d_e^\star). \tag{52}$$

Taking the supremum over $r \in \mathcal{R}(\mathcal{D})$ gives

$$d_H\big(\mathcal{R}(\mathcal{D}) \mid \mathcal{R}^\star\big) \leq C \ \mathrm{Gap}\big(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star\big). \tag{53}$$

Therefore, if $\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) \to 0$, then $d_H\big(\mathcal{R}(\mathcal{D}) \mid \mathcal{R}^\star\big) \to 0$. □

E Optimal-face recovery from relative-interior rewards

We state the standard geometric fact underlying the face-recovery statement in Section 3. For a face F of Φ, define its normal cone by

$$N_\Phi(F) := \left\{ r \mid r^\top d = r^\top d' \ \ \forall d, d' \in F, \quad r^\top y \le r^\top d \ \ \forall y \in \Phi, \ \forall d \in F \right\}. \tag{54}$$

Let $F_e^\star$ denote the minimal face of Φ containing $d_e^\star$. Then, we prove that relative interior points of the feasible reward set expose the expert face $F_e^\star$.

Lemma E.1. *For any $r \in \mathcal{R}^\star$, the optimal occupancy set $\Phi^\star(r)$ is an exposed face of Φ containing $d_e^\star$. Moreover, if $r \in \mathrm{relint}(\mathcal{R}^\star)$, then $\Phi^\star(r) = F_e^\star$.*

Proof. If $r \in \mathcal{R}^\star$, then $d_e^\star \in \Phi^\star(r)$ by definition. Since $\Phi^\star(r) = \arg\max_{d \in \Phi} r^\top d$ is the maximizer set of a linear functional over the polytope Φ, it is an exposed face of Φ. Thus $\Phi^\star(r)$ is an exposed face containing $d_e^\star$.

It remains to prove the second claim. Let $\Delta := \Delta(S \times A)$. Since $F_e^\star$ is the minimal face containing $d_e^\star$, we have

$$\mathcal{R}^\star = N_\Phi(F_e^\star) \cap \Delta. \tag{55}$$

To apply the relative-interior intersection rule, we first verify that $\mathrm{relint}(N_\Phi(F_e^\star))$ intersects $\mathrm{relint}(\Delta)$. Choose any $u \in \mathrm{relint}(N_\Phi(F_e^\star))$. Every occupancy measure in Φ has total mass one, so adding a constant vector to a reward does not change its maximizers over Φ. Thus $u + \alpha \mathbf{1}$ remains in $\mathrm{relint}(N_\Phi(F_e^\star))$ for any α, and choosing $\alpha > 0$ sufficiently large makes all coordinates positive. After normalizing, we obtain

$$\frac{u + \alpha \mathbf{1}}{\mathbf{1}^\top (u + \alpha \mathbf{1})} \in \mathrm{relint}(N_\Phi(F_e^\star)) \cap \mathrm{relint}(\Delta). \tag{56}$$

Therefore $\mathrm{relint}(N_\Phi(F_e^\star)) \cap \mathrm{relint}(\Delta) \neq \emptyset$, and the standard relative-interior intersection rule for polyhedra yields

$$\mathrm{relint}(\mathcal{R}^\star) = \mathrm{relint}(N_\Phi(F_e^\star)) \cap \mathrm{relint}(\Delta). \tag{57}$$

Thus $r \in \mathrm{relint}(\mathcal{R}^\star)$ implies $r \in \mathrm{relint}(N_\Phi(F_e^\star))$. By the definition of the normal cone, every vector in $N_\Phi(F_e^\star)$ is constant on $F_e^\star$ and satisfies the weak inequalities

$$r^\top y \le r^\top d \qquad \forall d \in F_e^\star, \ \forall y \in \Phi. \tag{58}$$

We now show that these inequalities are strict for points outside $F_e^\star$. Fix $d \in F_e^\star$ and $y \in \Phi \setminus F_e^\star$. Since $F_e^\star$ is a face of the polytope Φ, there exists some $\bar{r} \in N_\Phi(F_e^\star)$ that exposes $F_e^\star$ and therefore satisfies

$$\bar{r}^\top y < \bar{r}^\top d. \tag{59}$$

Suppose, toward a contradiction, that our relative-interior point $r \in \mathrm{relint}(N_\Phi(F_e^\star))$ satisfies $r^\top y = r^\top d$. Because r lies in the relative interior of the cone, the point

$$r + \varepsilon(r - \bar{r}) \tag{60}$$

also lies in $N_\Phi(F_e^\star)$ for all sufficiently small $\varepsilon > 0$. However,

$$\big(r + \varepsilon(r - \bar{r})\big)^\top y > \big(r + \varepsilon(r - \bar{r})\big)^\top d, \tag{61}$$

contradicting the defining inequality of the normal cone. Therefore,

$$r^\top y < r^\top d \qquad \forall d \in F_e^\star, \ \forall y \in \Phi \setminus F_e^\star. \tag{62}$$

Therefore the maximizers of $r^\top d$ over Φ are precisely the points in $F_e^\star$, i.e., $\Phi^\star(r) = F_e^\star$. □

Consequently, if $\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) = 0$, then every selected reward in $\mathcal{R}(\mathcal{D})$ exposes an optimal face containing $d_e^\star$. If the selected reward additionally lies in $\mathrm{relint}(\mathcal{R}^\star)$, then policy optimization recovers exactly the minimal face.

F Proof of Theorem 3.5

Fix any $\tilde{r} \in \mathcal{R}(\mathcal{D})$ and any $(d, \epsilon) \in \mathrm{conv}(\mathcal{D})$. By definition of the convex hull, there exist coefficients $\{\lambda_k\}_{k=1}^K$ such that

$$\lambda_k \geq 0, \qquad \sum_{k=1}^K \lambda_k = 1, \qquad d = \sum_{k=1}^K \lambda_k d_e^k, \qquad \epsilon = \sum_{k=1}^K \lambda_k \epsilon^k. \tag{63}$$

Since $\tilde{r} \in \mathcal{R}(\mathcal{D})$, we have $\mathrm{SubOpt}(\tilde{r}, d_e^k) \leq \epsilon^k$ for every $k \in [K]$. Using the definition of SubOpt and the linearity in the second argument,

$$\begin{aligned}
\mathrm{SubOpt}(\tilde{r}, d) &= \max_{\tilde{d} \in \Phi} \tilde{r}^\top (\tilde{d} - d) && (64)\\
&= \max_{\tilde{d} \in \Phi} \tilde{r}^\top \tilde{d} - \tilde{r}^\top \sum_{k=1}^K \lambda_k d_e^k && (65)\\
&= \sum_{k=1}^K \lambda_k \left(\max_{\tilde{d} \in \Phi} \tilde{r}^\top \tilde{d} - \tilde{r}^\top d_e^k \right) && (66)\\
&= \sum_{k=1}^K \lambda_k \, \mathrm{SubOpt}(\tilde{r}, d_e^k) \leq \sum_{k=1}^K \lambda_k \epsilon^k = \epsilon. && (67)
\end{aligned}$$

Next, we compare the suboptimality of $d_e^\star$ with that of the convex-combined demonstrator occupancy d:

$$\begin{aligned}
\mathrm{SubOpt}(\tilde{r}, d_e^\star) &= \max_{\tilde{d} \in \Phi} \tilde{r}^\top (\tilde{d} - d_e^\star) && (68)\\
&= \max_{\tilde{d} \in \Phi} \tilde{r}^\top (\tilde{d} - d) + \tilde{r}^\top (d - d_e^\star) && (69)\\
&\leq \mathrm{SubOpt}(\tilde{r}, d) + \left| \tilde{r}^\top (d - d_e^\star) \right| && (70)\\
&\leq \epsilon + \|\tilde{r}\|_\infty \|d - d_e^\star\|_1 && (71)\\
&\leq \epsilon + \|d - d_e^\star\|_1, && (72)
\end{aligned}$$

where the last inequality uses $\tilde{r} \in [0,1]^{|S||A|}$. Since $(d, \epsilon) \in \mathrm{conv}(\mathcal{D})$ was arbitrary, we obtain

$$\mathrm{SubOpt}(\tilde{r}, d_e^\star) \leq \min_{(d,\epsilon) \in \mathrm{conv}(\mathcal{D})} \left\{ \|d - d_e^\star\|_1 + \epsilon \right\}. \tag{73}$$

Finally, taking the maximum over $\tilde{r} \in \mathcal{R}(\mathcal{D})$ yields

$$\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) = \max_{\tilde{r} \in \mathcal{R}(\mathcal{D})} \mathrm{SubOpt}(\tilde{r}, d_e^\star) \leq \min_{(d,\epsilon) \in \mathrm{conv}(\mathcal{D})} \left\{ \|d - d_e^\star\|_1 + \epsilon \right\}, \tag{74}$$

which proves the theorem.

G A gap bound from a local separation condition

The following theorem gives a sufficient condition under which each reward with a large mismatch is excluded by a demonstrator constraint. The directional separation mechanism is illustrated in Figure 4.

Theorem G.1. *Fix $\delta > 0$. Suppose that every reward $\tilde{r} \in \Delta(S \times A)$ satisfying $\mathrm{SubOpt}(\tilde{r}, d_e^\star) \geq \delta$ admits a reference reward $r' \in \mathcal{R}^\star$ and a demonstrator pair $(d_e^k, \epsilon^k) \in \mathcal{D}$ such that*

$$(\tilde{r} - r')^\top (d_e^\star - d_e^k) > 0, \qquad \mathrm{SubOpt}(r', d_e^k) \in [\epsilon^k - \delta, \epsilon^k]. \tag{75}$$

Then $\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) < \delta$. Moreover, while these conditions are imposed pointwise over infinitely many rewards, finitely many demonstrators suffice to satisfy them simultaneously.

Proof. Fix any $\tilde{r} \in \Delta(S \times A)$ with $\mathrm{SubOpt}(\tilde{r}, d_e^\star) \geq \delta$, and choose a reference reward $r' \in \mathcal{R}^\star$ and a demonstrator pair $(d_e^k, \epsilon^k) \in \mathcal{D}$ satisfying equation 75. Since $d_e^\star$ is optimal under r', we have

$$r'^\top (d_e^\star - d_e^k) = \mathrm{SubOpt}(r', d_e^k). \tag{76}$$

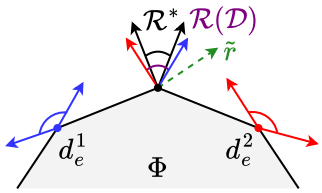


Figure 4: Illustration of the local separation condition in Theorem G.1: the candidate reward $\tilde{r}$ (green arrow) is ruled out by demonstrator d_e^1 (blue region).

Therefore,

$$\begin{aligned}\mathrm{SubOpt}(\tilde{r}, d_e^k) &= \mathrm{SubOpt}(\tilde{r}, d_e^\star) + \tilde{r}^\top (d_e^\star - d_e^k) && (77)\\ &= \mathrm{SubOpt}(\tilde{r}, d_e^\star) + (\tilde{r} - r')^\top (d_e^\star - d_e^k) + \mathrm{SubOpt}(r', d_e^k) && (78)\\ &> \delta + \epsilon^k - \delta = \epsilon^k. && (79)\end{aligned}$$

The strict inequality follows from the strictly positive directional term in equation 75, together with $\mathrm{SubOpt}(\tilde{r}, d_e^\star) \geq \delta$ and $\mathrm{SubOpt}(r', d_e^k) \geq \epsilon^k - \delta$. Thus $\tilde{r}$ violates the kth demonstrator constraint and does not belong to $\mathcal{R}(\mathcal{D})$. Consequently, every retained reward satisfies $\mathrm{SubOpt}(r, d_e^\star) < \delta$. The set $\mathcal{R}(\mathcal{D})$ is compact and nonempty: it is a closed subset of the reward simplex and contains the constant reward $r = \frac{1}{|S||A|}\mathbf{1}$. Since $\mathrm{SubOpt}(\cdot, d_e^\star)$ is continuous, its maximum on this set is attained, giving

$$\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) < \delta. \tag{80}$$

Let $\mathcal{C} \subseteq \Phi \times [0, 1]$ be a nonempty family of demonstrator pairs such that the conditions in equation 75 can be satisfied for every reward in

$$\mathcal{K}_\delta := \{\tilde{r} \in \Delta(S \times A) \mid \mathrm{SubOpt}(\tilde{r}, d_e^\star) \geq \delta\}. \tag{81}$$

This set is compact because the reward simplex is compact and $\mathrm{SubOpt}(\cdot, d_e^\star)$ is continuous. If $\mathcal{K}_\delta$ is empty, any single pair from $\mathcal{C}$ suffices, since the conditions are vacuous and the preceding argument gives the bound. Otherwise, for each $\bar{r} \in \mathcal{K}_\delta$, choose a reference reward $r_{\bar{r}} \in \mathcal{R}^\star$ and a pair $(d_{\bar{r}}, \epsilon_{\bar{r}}) \in \mathcal{C}$ satisfying equation 75. The set

$$U_{\bar{r}} := \left\{\tilde{r} \in \Delta(S \times A) \mid (\tilde{r} - r_{\bar{r}})^\top (d_e^\star - d_{\bar{r}}) > 0\right\} \tag{82}$$

is open relative to the simplex and contains $\bar{r}$. The same reference reward and demonstrator pair satisfy both conditions in equation 75 for every $\tilde{r} \in U_{\bar{r}}$: the directional inequality holds by definition, and the condition on $\mathrm{SubOpt}(r_{\bar{r}}, d_{\bar{r}})$ is independent of $\tilde{r}$. Thus $\{U_{\bar{r}}\}_{\bar{r} \in \mathcal{K}_\delta}$ is an open cover of $\mathcal{K}_\delta$. By compactness, there exist $\bar{r}_1, \ldots, \bar{r}_N \in \mathcal{K}_\delta$ such that

$$\mathcal{K}_\delta \subseteq \bigcup_{i=1}^{N} U_{\bar{r}_i}. \tag{83}$$

The finite dataset

$$\mathcal{D} := \{(d_{\bar{r}_i}, \epsilon_{\bar{r}_i})\}_{i=1}^N \subseteq \mathcal{C} \tag{84}$$

therefore satisfies the conditions in equation 75 for every reward in $\mathcal{K}_\delta$. Applying the first part of the proof yields $\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) < \delta$, completing both claims. □

H A GAP BOUND FROM OCCUPANCY COVERAGE

We give a complementary bound based on coordinatewise coverage of the optimal occupancy by a nonnegative weighted sum of demonstrator occupancies. Unlike the convex-mixture bound in Theorem 3.5, the weights need not sum to one, but coverage must hold at every state-action pair.

Theorem H.1. *Suppose that $\lambda \in \mathbb{R}_+^K$ satisfies*

$$d_e^\star(s, a) \leq \sum_{k=1}^{K} \lambda_k d_e^k(s, a) \qquad \forall (s, a) \in S \times A. \tag{85}$$

Then

$$\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) \le \sum_{k=1}^{K} \lambda_k \epsilon^k. \tag{86}$$

Consequently,

$$\mathrm{Gap}(\mathcal{R}(\mathcal{D}), \mathcal{R}^\star) \le \inf_{\substack{\lambda \in \mathbb{R}_+^K \\ \sum_{k=1}^K \lambda_k d_e^k \ge d_e^\star}} \sum_{k=1}^{K} \lambda_k \epsilon^k, \tag{87}$$

where the occupancy inequality is coordinatewise and the infimum over an empty feasible set is $+\infty$.

Proof. Fix any $r \in \mathcal{R}(\mathcal{D})$ and let $v_r^\star$ be its optimal value function. The optimal Bellman slack $g_r := M^\top v_r^\star - r$ is coordinatewise nonnegative. For every $d \in \Phi$, the occupancy constraints and LP strong duality give

$$d^\top g_r = (1-\gamma)\mu_0^\top v_r^\star - r^\top d = \mathrm{SubOpt}(r, d). \tag{88}$$

Thus, by equation 85 and $g_r \ge 0$,

$$\mathrm{SubOpt}(r, d_e^\star) = (d_e^\star)^\top g_r \le \sum_{k=1}^{K} \lambda_k (d_e^k)^\top g_r \tag{89}$$

$$= \sum_{k=1}^{K} \lambda_k \,\mathrm{SubOpt}(r, d_e^k) \le \sum_{k=1}^{K} \lambda_k \epsilon^k. \tag{90}$$

The last inequality follows from feasibility of r and nonnegativity of the weights. Taking the maximum over $r \in \mathcal{R}(\mathcal{D})$ proves the first bound. Taking the infimum over all admissible weights proves equation 87. □

The bound is informative when coverage can be achieved with a small weighted sum of declared suboptimality bounds. In particular, a demonstrator with a relatively large bound can contribute useful coverage with a small coefficient. Note that coverage is impossible if every demonstrator assigns zero mass to a coordinate where $d_e^\star$ has positive mass, even if the occupancy-proximity bound is small.

I Example of gap reduction with a large suboptimality bound

We give a one-state, three-action bandit example in which adding a demonstrator with declared bound 0.4 reduces the gap from 0.1 to approximately 0.04535. For an occupancy distribution d, write

$$\mathrm{SubOpt}(r, d) = m(r) - r^\top d, \qquad m(r) := \max_{i \in \{1,2,3\}} r_i. \tag{91}$$

Let $r^\star = (0.5, 0.5, 0)$ and $d_e^\star = (0.9, 0.1, 0)$. The first two actions are optimal under $r^\star$, so their mixture $d_e^\star$ is an optimal occupancy. Consider the two demonstrator pairs

$$d_e^1 = (0.999, 0, 0.001), \quad \epsilon^1 = 0.001, \qquad d_e^2 = (0, 0.9, 0.1), \quad \epsilon^2 = 0.4. \tag{92}$$

Both demonstrators are suboptimal under the true reward:

$$\mathrm{SubOpt}(r^\star, d_e^1) = 0.0005, \qquad \mathrm{SubOpt}(r^\star, d_e^2) = 0.05. \tag{93}$$

Thus both declared bounds are valid under $r^\star$, and the second demonstrator has both higher true suboptimality and a larger declared bound. Define $\mathcal{D}_1 := \{(d_e^1, 0.001)\}$ and $\mathcal{D}_2 := \mathcal{D}_1 \cup \{(d_e^2, 0.4)\}$.

Every reward $r \in \mathcal{R}(\mathcal{D}_1)$ satisfies $m(r) - 0.999 r_1 - 0.001 r_3 \le 0.001$. Using $r_1 + r_2 + r_3 = 1$ and $r_i \ge 0$, we obtain

$$\mathrm{SubOpt}(r, d_e^\star) = m(r) - 0.9 r_1 - 0.1 r_2 \tag{94}$$

$$\le 0.099 r_1 - 0.1 r_2 + 0.001 r_3 + 0.001 \tag{95}$$

$$= 0.1 - 0.199 r_2 - 0.098 r_3 \le 0.1. \tag{96}$$

This bound is attained by $\tilde{r} = (1, 0, 0)$, since

$$\mathrm{SubOpt}(\tilde{r}, d_e^1) = 0.001, \qquad \mathrm{SubOpt}(\tilde{r}, d_e^\star) = 0.1. \tag{97}$$

Therefore,

$$\mathrm{Gap}(\mathcal{R}(\mathcal{D}_1), \mathcal{R}^\star) = 0.1. \tag{98}$$

After adding the second demonstrator, the reward $\tilde{r}$ is excluded because $\mathrm{SubOpt}(\tilde{r}, d_e^2) = 1 > 0.4$. For any retained reward, let $g_i := m(r) - r_i \geq 0$ denote the suboptimality of action i. The two demonstrator constraints imply

$$0.999g_1 + 0.001g_3 \leq 0.001, \qquad 0.9g_2 + 0.1g_3 \leq 0.4. \tag{99}$$

Since $g_3 \geq 0$, we have $g_1 \leq 1/999$ and $g_2 \leq 4/9$. Consequently,

$$\mathrm{SubOpt}(r, d_e^\star) = 0.9g_1 + 0.1g_2 \leq \frac{0.9}{999} + \frac{0.4}{9} = \frac{151}{3330} \approx 0.04535. \tag{100}$$

This bound is attained by the reward

$$\bar{r} = \frac{1}{2997}(1441, 112, 1444), \tag{101}$$

for which $g_1 = 1/999$, $g_2 = 4/9$, and $g_3 = 0$. Both demonstrator constraints hold with equality, so $\bar{r} \in \mathcal{R}(\mathcal{D}_2)$ and

$$\mathrm{Gap}(\mathcal{R}(\mathcal{D}_2), \mathcal{R}^\star) = \frac{151}{3330} \approx 0.04535 < 0.1 = \mathrm{Gap}(\mathcal{R}(\mathcal{D}_1), \mathcal{R}^\star). \tag{102}$$

Thus, adding the demonstrator with declared bound 0.4 reduces the worst-case gap by approximately 54.65%, from 0.1 to 0.04535. The first constraint tightly controls the suboptimality of the first action, whereas the added constraint also controls that of the second action, which receives positive mass under $d_e^\star$. This illustrates how a demonstrator with a relatively large declared bound can still provide useful information and substantially reduce the gap.

J Multi-demonstrator MaxEnt baseline

Algorithm. A standard maximum-causal-entropy IRL baseline assumes that all demonstrators share one latent reward r, while demonstrator k acts according to the entropy-regularized optimal policy with inverse temperature $\beta_k > 0$. In the baseline analyzed here, the inverse temperatures are fixed before reward fitting, possibly as deterministic functions $\beta_k = B(\epsilon^k)$ of the declared suboptimality levels; the likelihood is then optimized only over r. Thus, demonstrators with the same declared suboptimality level share the same inverse temperature, while demonstrators with different levels may have different temperatures. This is the usual MaxEnt / maximum-causal-entropy viewpoint (Ziebart et al., 2008; Ramachandran & Amir, 2007), with β_k controlling the strength of entropy regularization and capturing different suboptimality levels. Equivalently, for fixed r and β_k, demonstrator k solves the entropy-regularized control problem whose soft Bellman equations define the state-action and state values $Q_r^k(s,a)$ and $V_r^k(s)$:

$$\begin{aligned} V_r^k(s) &:= \frac{1}{\beta_k} \log \sum_{a \in A} \exp\big(\beta_k Q_r^k(s,a)\big), \\ Q_r^k(s,a) &:= r(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a)\, V_r^k(s'). \end{aligned} \tag{103}$$

The induced policy is the Boltzmann distribution

$$\pi_r^k(a \mid s) \propto \exp(\beta_k Q_r^k(s,a)). \tag{104}$$

Given $\mathcal{D} = \{(d_e^k, \epsilon^k)\}_{k=1}^K$, the MaxEnt estimator is

$$\hat{r} \in \operatorname*{arg\,max}_{r \in \Delta(S \times A)} \mathcal{L}(r), \qquad \mathcal{L}(r) := \sum_{k=1}^K \mathbb{E}_{(s,a) \sim d_e^k} \left[\log \pi_r^k(a \mid s)\right]. \tag{105}$$

Proof of Proposition 3.6 for MaxEnt. Consider a one-state, two-action bandit with action set $A = \{a_1, a_2\}$. Let the true optimal occupancy be $d_e^\star = e_{a_1}$, where e_a denotes the unit vector for action a, and define the reward gap

$$\Delta := r(a_1) - r(a_2) \in [-1, 1]. \tag{106}$$

Fix any $\eta > 0$ and any constant $C \in (0, 1)$. Choose $\delta := \min\{\eta/3, 1/4\}$. Let $\epsilon^g = \delta$ and $\epsilon^b = 1$, and let

$$\beta_g := B(\epsilon^g), \qquad \beta_b := B(\epsilon^b) \tag{107}$$

be the fixed inverse temperatures assigned by the baseline. Choose an integer m such that

$$m \geq \frac{\beta_g \sigma(\beta_g C)}{\beta_b \sigma(-\beta_b C)}, \qquad \sigma(x) := (1 + e^{-x})^{-1}. \tag{108}$$

Define the finite dataset

$$\mathcal{D} := \{(d^g, \epsilon^g)\} \cup \{(d^{b,i}, \epsilon^b)\}_{i=1}^m, \tag{109}$$

where

$$d^g = (1 - \delta, \delta), \tag{110}$$

and

$$d^{b,i} = (0, 1), \qquad i = 1, \ldots, m. \tag{111}$$

Since $\delta > 0$, the dataset contains no optimal demonstrator.

This dataset is consistent with $\mathcal{R}^\star$. Indeed, if $r \in \mathcal{R}^\star$, then a_1 is optimal under r, so $\Delta \in [0, 1]$. Hence

$$\mathrm{SubOpt}(r, d^g) = \delta\Delta \leq \delta = \epsilon^g, \tag{112}$$

and for each bad demonstrator,

$$\mathrm{SubOpt}(r, d^{b,i}) = \Delta \leq 1 = \epsilon^b. \tag{113}$$

Therefore $\mathcal{R}^\star \subseteq \mathcal{R}(\mathcal{D})$.

Since $(d^g, \epsilon^g) \in \mathcal{D} \subseteq \mathrm{conv}(\mathcal{D})$, Theorem 3.5 gives

$$\min_{(d,\epsilon)\in\mathrm{conv}(\mathcal{D})} \{\|d - d_e^\star\|_1 + \epsilon\} \leq \|d^g - d_e^\star\|_1 + \epsilon^g = 3\delta \leq \eta. \tag{114}$$

We next show that every MaxEnt optimum for this finite dataset gives $d_e^\star$ a constant suboptimality gap. For the dataset $\mathcal{D}$, the one-dimensional log-likelihood is

$$L_m(\Delta) := (1 - \delta) \log \sigma(\beta_g \Delta) + \delta \log \sigma(-\beta_g \Delta) + m \log \sigma(-\beta_b \Delta). \tag{115}$$

Its derivative is

$$L_m'(\Delta) = \beta_g \left((1 - \delta)\sigma(-\beta_g \Delta) - \delta\sigma(\beta_g \Delta)\right) - m\beta_b \sigma(\beta_b \Delta), \tag{116}$$

and L_m is concave in Δ. At $\Delta = -C$,

$$L_m'(-C) \leq \beta_g \sigma(\beta_g C) - m\beta_b \sigma(-\beta_b C) \leq 0 \tag{117}$$

by the choice of m. Since L_m is concave, every maximizer $\Delta^\star$ satisfies $\Delta^\star \leq -C$. Hence every MaxEnt estimator $\hat{r}$ satisfies

$$\hat{r}(a_2) - \hat{r}(a_1) \geq C. \tag{118}$$

Therefore,

$$\mathrm{SubOpt}(\hat{r}, d_e^\star) = \big(\hat{r}(a_2) - \hat{r}(a_1)\big)_+ \geq C. \tag{119}$$

Thus, for any fixed $C \in (0, 1)$, we have constructed a finite dataset $\mathcal{D}$ for which the upper bound in Theorem 3.5 is at most η, while every MaxEnt estimator satisfies $\mathrm{SubOpt}(\hat{r}, d_e^\star) \geq C$.

K RANKING-BASED BASELINE

Algorithm. A standard ranking-based baseline such as T-REX learns a reward from pairwise trajectory rankings. Given ranked trajectories $\{\tau^k\}_{k=1}^n$, T-REX assigns trajectory τ the cumulative reward

$$r(\tau) := \sum_{t=0}^{T(\tau)-1} r(s_t, a_t), \tag{120}$$

and models the preference probability by a Bradley–Terry–Luce model (Bradley & Terry, 1952; Luce, 1959) or a Plackett–Luce ranking model (Plackett, 1975):

$$\mathbb{P}_r(\tau^i \prec \tau^j) := \frac{\exp(r(\tau^j))}{\exp(r(\tau^i)) + \exp(r(\tau^j))}. \tag{121}$$

The corresponding reward estimator is

$$\hat{r} \in \operatorname*{arg\,max}_{r \in \Delta(S\times A)} \sum_{\tau^i \prec \tau^j} \log \frac{\exp(r(\tau^j))}{\exp(r(\tau^i)) + \exp(r(\tau^j))}. \tag{122}$$

D-REX differs only in how the rankings are obtained: it automatically generates ranked trajectories by injecting noise into a cloned policy and then applies the same T-REX likelihood objective (Brown et al., 2020). In the one-step bandit counterexample below, a demonstration occupancy d is scored by its expected one-step reward $r^\top d$; for a deterministic trajectory τ_a taking action a, this reduces to the action reward $r(\tau_a) = r(a)$.

Proof of Proposition 3.6 for the ranking-based estimator. Consider a one-state bandit with three actions $A = \{a_1, a_2, a_3\}$, so the occupancy set is

$$\Phi = \left\{ d \in \mathbb{R}^3_+ : \sum_{i=1}^3 d(a_i) = 1 \right\}. \tag{123}$$

Let the ground-truth optimal occupancy be $d_e^\star = e_{a_1}$, induced for example by the reward $r^\star = e_{a_1}$.

Fix any $\eta > 0$. Choose $\rho > 0$ small enough that

$$2(1-2\rho)\sigma(-(1-\rho)) > \frac{1}{2}, \qquad \sigma(x) := (1+e^{-x})^{-1}. \tag{124}$$

Such a choice is possible because the left-hand side is continuous in ρ and equals $2\sigma(-1) > 1/2$ at $\rho = 0$. Consider the finite occupancy-suboptimality dataset

$$\mathcal{D} := \{(d^g, \epsilon^g), (d^{u,1}, \epsilon^u), (d^{u,2}, \epsilon^u), (d^b, \epsilon^b)\}, \tag{125}$$

where

$$d^g = e_{a_1}, \qquad \epsilon^g = 0, \tag{126}$$

$$d^{u,1} = d^{u,2} = \rho e_{a_1} + (1-\rho) e_{a_2}, \qquad \epsilon^u = 1-\rho, \tag{127}$$

and

$$d^b = e_{a_3}, \qquad \epsilon^b = 1. \tag{128}$$

For the ranking-based estimator, score each demonstration occupancy d by $r^\top d$ and use the comparisons

$$d^{u,1} \succ d^b, \qquad d^{u,2} \succ d^b, \qquad d^g \succ d^b. \tag{129}$$

These comparisons are consistent with $r^\star$, since $(r^\star)^\top d^g = 1$, $(r^\star)^\top d^{u,i} = \rho$, and $(r^\star)^\top d^b = 0$.

We first verify consistency with $\mathcal{R}^\star$. If $r \in \mathcal{R}^\star$, then a_1 is optimal under r. Hence

$$\mathrm{SubOpt}(r, d^g) = 0 = \epsilon^g, \tag{130}$$

and for each $i \in \{1, 2\}$,

$$\mathrm{SubOpt}(r, d^{u,i}) = (1-\rho)\big(r(a_1) - r(a_2)\big) \le 1 - \rho = \epsilon^u, \tag{131}$$

while

$$\mathrm{SubOpt}(r, d^b) \le 1 = \epsilon^b. \tag{132}$$

Therefore $\mathcal{R}^\star \subseteq \mathcal{R}(\mathcal{D})$.

Since $(d^g, \epsilon^g) \in \mathcal{D} \subseteq \mathrm{conv}(\mathcal{D})$, Theorem 3.5 yields

$$\min_{(d,\epsilon)\in\mathrm{conv}(\mathcal{D})} \{\|d - d_e^\star\|_1 + \epsilon\} \le \|d^g - d_e^\star\|_1 + \epsilon^g = 0, \tag{133}$$

which is at most η.

We now analyze the actual T-REX likelihood. The T-REX objective is

$$L(r) = 2\log\sigma\big(\rho r(a_1) + (1-\rho)r(a_2) - r(a_3)\big) + \log\sigma\big(r(a_1) - r(a_3)\big). \tag{134}$$

Because $\log\sigma(\cdot)$ is increasing, any mass assigned to a_3 can be moved to a_2 while increasing both differences $\rho r(a_1) + (1-\rho)r(a_2) - r(a_3)$ and $r(a_1) - r(a_3)$. Hence every maximizer satisfies

$$r(a_3) = 0. \tag{135}$$

Therefore we may write

$$r = (t, 1-t, 0), \qquad t \in [0,1], \tag{136}$$

and the objective reduces to

$$L(t) = 2\log\sigma\big(\rho t + (1-\rho)(1-t)\big) + \log\sigma(t). \tag{137}$$

This function is concave, and

$$L'(t) = -2(1-2\rho)\sigma\big(-\rho t - (1-\rho)(1-t)\big) + \sigma(-t). \tag{138}$$

For every $t \in [0,1]$, we have $\sigma(-t) \le \frac{1}{2}$ and $\sigma\big(-\rho t - (1-\rho)(1-t)\big) \ge \sigma(-(1-\rho))$, so

$$L'(t) \le \frac{1}{2} - 2(1-2\rho)\sigma(-(1-\rho)) < 0 \tag{139}$$

by the choice of ρ. Thus L is strictly decreasing on $[0,1]$, and every maximizer is

$$\hat{r}(a_1) = 0, \qquad \hat{r}(a_2) = 1, \qquad \hat{r}(a_3) = 0. \tag{140}$$

Finally,

$$\mathrm{SubOpt}(\hat{r}, d_e^\star) = \max_{d\in\Phi} \hat{r}^\top d - \hat{r}^\top d_e^\star = 1 - 0 = 1. \tag{141}$$

Thus, with $C = 1$, we have constructed a finite dataset $\mathcal{D}$ for which the upper bound in Theorem 3.5 is at most η, while every T-REX maximizer satisfies $\mathrm{SubOpt}(\hat{r}, d_e^\star) \ge C$. Since D-REX uses the same reward-learning objective after generating automatic rankings, the same counterexample applies to that ranking-based baseline as well.

L Coverage-based penalty

In the single-optimal-demonstrator setting, reward ambiguity can be mitigated by selecting a (relative) interior point of the feasible reward set. One standard construction imposes a positive margin on the Bellman inequalities, i.e., $(M^\top v - r)(s,a) \ge \delta\mathbf{1}\{d_e(s,a) = 0\}$ for all $(s,a) \in S \times A$. The margin δ penalizes reward on state-action pairs unvisited by the expert d_e and discourages degenerate solutions on the boundary of the normal cone $N(d_e)$.

We extend this idea to multiple suboptimal demonstrations by penalizing state-action pairs that are unvisited by all demonstrators simultaneously. Concretely, we impose a positive margin on pairs that have zero mass under the demonstrators' aggregate occupancy measure. The inactive region and the feasible reward set with coverage-based penalty are defined as follows.

Definition L.1. Let $\bar{d}$ denote the aggregate demonstrator occupancy measure: $\bar{d} := \frac{1}{K}\sum_{k=1}^K d_e^k$, and $\mathcal{I}_\mathcal{D} := \{(s,a) \in S \times A \mid \bar{d}(s,a) = 0\}$ denote the inactive state-action pairs where the aggregate occupancy measure is zero.

Definition L.2 (Feasible reward set with coverage-based penalty)**.**

$$\begin{aligned}\mathcal{R}_\delta(\mathcal{D}) :=& \{r \in \Delta(S\times A) \mid \exists v \in \mathbb{R}^{|S|} \quad \text{s.t.} \quad (1-\gamma)\mu_0^\top v - r^\top d_e^k \le \epsilon^k \quad \forall k \in [K], \\ & (M^\top v - r)(s,a) \ge \delta\mathbf{1}\{(s,a) \in \mathcal{I}_\mathcal{D}\} \quad \forall (s,a) \in S\times A\}.\end{aligned} \tag{142}$$

By definition, $\mathcal{R}_0(\mathcal{D}) = \mathcal{R}(\mathcal{D})$ and $\mathcal{R}_\delta(\mathcal{D}) \subseteq \mathcal{R}(\mathcal{D})$ for any $\delta > 0$. As in the single-optimal-demonstrator case, positive Bellman inequality slack on inactive pairs discourages rewards that keep clearly under-covered behaviors competitive. The next theorem quantifies this effect: any occupancy placing mass off the demonstrators' aggregate support pays a suboptimality gap that scales linearly with the total off-support mass and the margin δ.

Theorem L.3. *If $r \in \mathcal{R}_\delta(\mathcal{D})$, then for any $d \in \Phi$, we have*

$$\mathrm{SubOpt}(r, d) \geq \delta \sum_{(s,a)\in\mathcal{I}_\mathcal{D}} d(s,a) - \min_{k\in[K]} \epsilon^k. \tag{143}$$

Proof. Fix any $r \in \mathcal{R}_\delta(\mathcal{D})$, and let $v \in \mathbb{R}^{|S|}$ be a feasible solution from equation 142. Define the slack vector

$$l := M^\top v - r. \tag{144}$$

Then $l(s,a) \geq \delta\, \mathbf{1}\{(s,a) \in \mathcal{I}_\mathcal{D}\}$ for every (s,a).

By weak duality, for every $d \in \Phi$ we have $r^\top d \leq (1-\gamma)\mu_0^\top v$. Therefore, for each demonstrator $k \in [K]$,

$$\mathrm{SubOpt}(r, d_e^k) = \max_{d\in\Phi}\ r^\top d - r^\top d_e^k \leq (1-\gamma)\mu_0^\top v - r^\top d_e^k \leq \epsilon^k, \tag{145}$$

where the last inequality is exactly the feasibility condition in equation 142. Hence d_e^k is at most ϵ^k-suboptimal for every $k \in [K]$.

Fix any $d \in \Phi$ and any demonstrator index $k \in [K]$. Since $d_e^k \in \Phi$, the definition of SubOpt gives

$$\mathrm{SubOpt}(r, d) = \max_{\tilde{d}\in\Phi} r^\top \tilde{d} - r^\top d \geq r^\top d_e^k - r^\top d. \tag{146}$$

Using $r = M^\top v - l$ and the Bellman equation constraint $Md = (1-\gamma)\mu_0$ for any $d \in \Phi$, we obtain

$$r^\top d = v^\top M d - l^\top d = (1-\gamma)\mu_0^\top v - l^\top d. \tag{147}$$

The same identity also holds with d_e^k in place of d. Hence

$$\mathrm{SubOpt}(r, d) \geq l^\top d - l^\top d_e^k. \tag{148}$$

Moreover, by the constraint of equation 142, we have

$$l^\top d_e^k = (M^\top v - r)^\top d_e^k = (1-\gamma)\mu_0^\top v - r^\top d_e^k \leq \epsilon^k. \tag{149}$$

Combining the last two inequalities yields

$$\mathrm{SubOpt}(r, d) \geq l^\top d - \epsilon^k \geq \delta \sum_{(s,a)\in\mathcal{I}_\mathcal{D}} d(s,a) - \epsilon^k. \tag{150}$$

Since this holds for every $k \in [K]$, we conclude that

$$\mathrm{SubOpt}(r, d) \geq \delta \sum_{(s,a)\in\mathcal{I}_\mathcal{D}} d(s,a) - \min_{k\in[K]} \epsilon^k, \tag{151}$$

which proves the claim. □

M Effect of additional pairwise performance-gap information

Suppose we additionally observe *pairwise performance-gap information* from comparisons: for some ordered pairs (i, j), demonstrator i is known to outperform demonstrator j by at least a margin $\epsilon_{ij} \geq 0$ under the unknown ground-truth reward. In our occupancy-measure notation, this corresponds to the linear constraints

$$r^\top (d_e^i - d_e^j) \ \geq\ \epsilon_{ij} \qquad \forall (i,j) \in \mathcal{P}, \tag{152}$$

where $\mathcal{P} \subseteq [K] \times [K]$ indexes the available comparisons.

This section shows that the constraints in equation 152 rule out degenerate rewards by forcing any recovered reward to induce a nontrivial value range between the best and worst policies. To be specific, we consider the following worst-case reward range: among all rewards in $\mathcal{R}$, it is the smallest best-vs-worst reward range achievable over Φ.

Definition M.1 (Worst-case reward range under $\mathcal{R}$)**.**

$$\mathrm{rng}(\mathcal{R}) := \min_{r\in\mathcal{R}} \left(\max_{\tilde{d}\in\Phi} r^\top \tilde{d} - \min_{d\in\Phi} r^\top d \right). \tag{153}$$

Then, we show that this measure is at least the sum of performance gaps over any path in the comparison graph.

Aggregating comparisons via paths. When many comparisons are available, one can aggregate them along a chain of pairwise comparisons to obtain a stronger lower bound than $\max_{(i,j)\in\mathcal{P}} \epsilon_{ij}$. Define the directed *comparison graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} := [K]$ and $\mathcal{E} := \mathcal{P}$, and assign each edge weight ϵ_{ij} between demonstrators i and j by the pairwise performance-gap information. For a simple directed path

$$p = (i_0 \to i_1 \to \cdots \to i_m), \tag{154}$$

in $\mathcal{G}$, define its total weight as

$$\epsilon(p) := \sum_{\ell=1}^{m} \epsilon_{i_{\ell-1} i_\ell}. \tag{155}$$

The next result shows that every such path yields a lower bound on the worst-case reward range.

Proposition M.2 (Path-based worst-case reward-range guarantee)**.** *Let*

$$\mathcal{R}_{\mathcal{P}} := \left\{ r \in \mathcal{R}(\mathcal{D}) \;\middle|\; r^\top (d_e^i - d_e^j) \geq \epsilon_{ij} \quad \forall (i,j) \in \mathcal{P} \right\}. \tag{156}$$

For any simple directed path p in $\mathcal{G}$, if $\mathcal{R}_{\mathcal{P}}$ is nonempty, then

$$\mathrm{rng}(\mathcal{R}_{\mathcal{P}}) \;\geq \epsilon(p). \tag{157}$$

Proof. Fix any simple directed path p in $\mathcal{G}$ of the form $p = (i_0 \to i_1 \to \cdots \to i_m)$, where each $(i_{\ell-1}, i_\ell) \in \mathcal{P}$. Fix any $r \in \mathcal{R}_{\mathcal{P}}$. By the definition of $\mathcal{R}_{\mathcal{P}}$, we have for each edge $r^\top (d_e^{i_{\ell-1}} - d_e^{i_\ell}) \geq \epsilon_{i_{\ell-1} i_\ell}$. Summing over $\ell = 1, \ldots, m$ telescopes to

$$r^\top (d_e^{i_0} - d_e^{i_m}) \;\geq\; \epsilon(p). \tag{158}$$

Since $d_e^{i_0}, d_e^{i_m} \in \Phi$, the reward range under r satisfies

$$\max_{\tilde{d}\in\Phi} r^\top \tilde{d} - \min_{d\in\Phi} r^\top d \;\geq\; r^\top d_e^{i_0} - r^\top d_e^{i_m} \;\geq\; \epsilon(p). \tag{159}$$

Since $r \in \mathcal{R}_{\mathcal{P}}$ was arbitrary, taking the minimum over $r \in \mathcal{R}_{\mathcal{P}}$ yields

$$\mathrm{rng}(\mathcal{R}_{\mathcal{P}}) \;=\; \min_{r\in\mathcal{R}_{\mathcal{P}}} \left[\max_{\tilde{d}\in\Phi} r^\top \tilde{d} - \min_{d\in\Phi} r^\top d \right] \;\geq\; \epsilon(p). \tag{160}$$

□

N ADVANTAGE PARAMETERIZATION OF THE OFFLINE OBJECTIVE

We derive the advantage-parameterized loss from the population optimal advantage, and then use the resulting expression with empirical occupancies.

Proposition N.1 (Advantage parameterization)**.** *For a reward r, let $v_r^\star$ be its optimal value function and define the optimal advantage*

$$A_r^\star(s,a) := r(s,a) + \gamma \sum_{s'\in S} P(s' \mid s,a) v_r^\star(s') - v_r^\star(s). \tag{161}$$

Then $A_r^\star(s,a) \leq 0$ for every (s,a), and for every occupancy $d \in \Phi$,

$$\mathrm{SubOpt}(r,d) = -\sum_{s,a} d(s,a) A_r^\star(s,a). \tag{162}$$

Consequently, the population version of the constrained reward-selection objective can be written directly in terms of $A_r^\star$. Replacing the population quantities by empirical occupancies and using a free parameterization A_ψ gives

$$\begin{aligned} L(A_\psi) := & -\sum_{s,a}\left(\sum_{k=1}^{K} w_k \hat{d}_e^k(s,a) - \hat{d}_{\text{base}}(s,a)\right) A_\psi(s,a) \\ & + \lambda_{\text{sub}} \frac{1}{K}\sum_{k=1}^{K} \rho_{\leq}\left(-\sum_{s,a} \hat{d}_e^k(s,a)\, A_\psi(s,a) - \epsilon^k\right) \\ & + \lambda_{\text{bell}}\, \mathbb{E}_{(s,a)\sim\bar{\mathcal{D}}}\left[\rho_{\leq}(A_\psi(s,a))\right]. \end{aligned} \tag{163}$$

Proof. The Bellman optimality equation gives $A_r^\star(s,a) \leq 0$ for all (s,a). Since $d \in \Phi$ satisfies $Md = (1-\gamma)\mu_0$, we have

$$\begin{aligned} \sum_{s,a} d(s,a) A_r^\star(s,a) &= r^\top d - (M^\top v_r^\star)^\top d && (164)\\ &= r^\top d - (1-\gamma)\mu_0^\top v_r^\star && (165)\\ &= r^\top d - \max_{\tilde{d}\in\Phi} r^\top \tilde{d} = -\,\text{SubOpt}(r,d), && (166) \end{aligned}$$

which proves equation 162. Thus the population suboptimality constraint $\text{SubOpt}(r, d_e^k) \leq \epsilon^k$ is equivalent to

$$-\sum_{s,a} d_e^k(s,a) A_r^\star(s,a) - \epsilon^k \leq 0. \tag{167}$$

Similarly, for any two occupancies $d, d' \in \Phi$, the flow constraints imply

$$r^\top (d - d') = \sum_{s,a} \big(d(s,a) - d'(s,a)\big) A_r^\star(s,a), \tag{168}$$

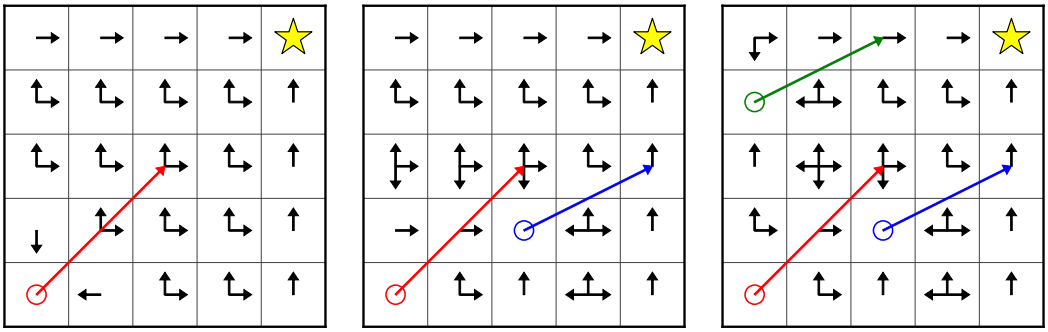
because $(M^\top v_r^\star)^\top (d - d') = v_r^{\star\top} M(d - d') = 0$. Applying this identity with $d = \sum_{k=1}^K w_k d_e^k$ and $d' = d_{\text{base}}$ rewrites the reward-selection term. Finally, Bellman feasibility is exactly the sign constraint $A_r^\star(s,a) \leq 0$, which is penalized by $\rho_{\leq}(A_\psi(s,a))$ in the direct parameterization. Replacing d_e^k and d_{base} by their empirical estimates and sampling the sign penalty over $\bar{\mathcal{D}}$ yields equation 163. □

In implementation, one may further restrict the direct parameterization to non-positive advantages, for example by setting $A_\psi(s,a) = -\text{softplus}(f_\psi(s,a))$. This enforces the Bellman-feasibility sign convention by construction while preserving the same penalty-based objective form.

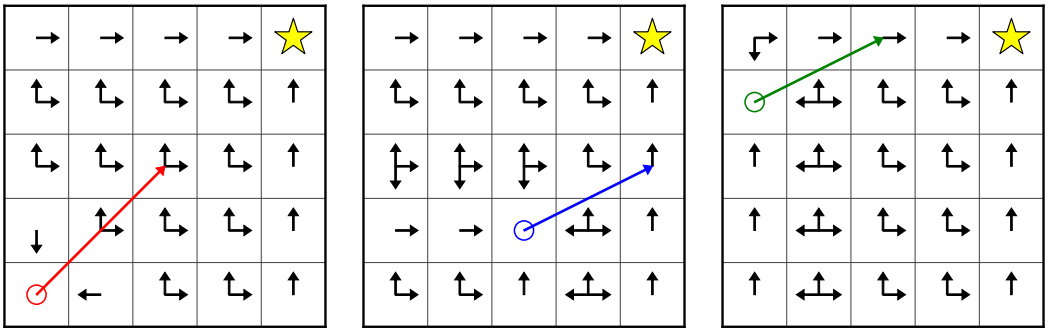
O Experimental details

O.1 Grid-world with detour portals

We illustrate the two cases of suboptimal demonstrators in Figure 5. In each figure, the yellow star marks the terminal goal state, and the red, blue, and green arrows indicate the active portal transitions from entrance to exit. The black arrows show the optimal policy. In Case 1, the demonstrator quality, or knowledge, increases with K; in Case 2, only the state-action coverage increases with K while the demonstrator quality remains limited.



(a) Case 1: Demonstrator 1 (b) Case 1: Demonstrator 2 (c) Case 1: Demonstrator 3



(d) Case 2: Demonstrator 1 (e) Case 2: Demonstrator 2 (f) Case 2: Demonstrator 3

Figure 5: Grid-world examples with detour portals for the two cases of suboptimal demonstrators.

O.2 Sensitivity to misspecified suboptimality bounds

We evaluate sensitivity to the declared suboptimality bounds in the Case-2 grid-world setting. For each demonstrator k, we compute its oracle suboptimality $\epsilon_k^\star$ under the synthetic ground-truth reward and provide the learning algorithm with

$$\hat{\epsilon}_k = \max\{\epsilon_k^\star + \Delta, 0\}. \tag{169}$$

The mean true suboptimality is 0.233 ± 0.069, and we consider perturbations

$$\Delta \in \{-0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3\}, \tag{170}$$

which are substantial relative to the scale of $\epsilon_k^\star$. All other experimental settings are unchanged: $K = 20$ demonstrators, 100 trajectories per demonstrator, and 20 random maps. For each map, we normalize the nonnegative learned reward by its maximum entry, leaving an all-zero reward unchanged, and denote the resulting reward by $\hat{r}$. We report both the suboptimality of the ground-truth optimal occupancy under this normalized reward, $\mathrm{SubOpt}(\hat{r}, d_e^\star) = \hat{r}^\top (d_{\hat{\pi}} - d_e^\star)$, where $\hat{\pi}$ is optimal under $\hat{r}$, and the population variance of $\hat{r}(s, a)$ with uniform weighting over all state-action entries. A variance of zero indicates a constant reward; a larger variance indicates greater variation across entries, but does not by itself imply more accurate reward recovery.

Table 1: Sensitivity to misspecified suboptimality bounds in the Case-2 grid-world setting. Entries are mean $\pm$ standard deviation over 20 random maps.

Δ	$\mathrm{SubOpt}(\hat{r}, d_e^\star)$	Variance of rewards
-0.3	0.000 ± 0.000	0.000 ± 0.000
-0.2	0.015 ± 0.035	0.047 ± 0.050
-0.1	0.084 ± 0.083	0.099 ± 0.044
0	0.167 ± 0.123	0.130 ± 0.014
$+0.1$	0.199 ± 0.135	0.134 ± 0.011
$+0.2$	0.201 ± 0.134	0.134 ± 0.011
$+0.3$	0.201 ± 0.134	0.134 ± 0.011

Table 1 shows that overestimating the bounds weakens the feasible-set constraints: suboptimality increases from 0.167 at the oracle declarations to approximately 0.201 and then saturates. Underestimation tightens the constraints and eventually produces reward degeneracy. In particular, $\Delta = -0.3$ yields a constant zero reward, as indicated by its zero variance. Its zero reported suboptimality is therefore vacuous rather than evidence of successful recovery, because every policy is optimal under a constant reward. Taken together, policy mismatch and the variance of normalized rewards provide complementary diagnostics of bound misspecification. The results reveal a trade-off: bounds that are too loose increase policy mismatch, while bounds that are too tight risk reward collapse. In practice, this trade-off can be controlled by calibrating the declared bounds to avoid near-zero variance of normalized rewards while maintaining a strong external trajectory-quality or task-performance proxy.

O.3 Arithmetic tool-use LLM experiment

This appendix provides additional details for the arithmetic tool-use LLM experiment in Section 4.

Task and dataset generation. We consider synthetic arithmetic tool-use tasks built from five tools: `add`, `subtract`, `multiply`, `divide`, and `power`. Each task consists of a natural-language arithmetic question whose solution requires a chain of one to five tool calls. The target completion is a multi-line tool-use trace with one call per line. When an argument is the output of the previous line, we replace it with the special token `PREV`.

For each training task, we construct one optimal completion and five demonstrator completions. The optimal completion uses the correct tool at every step. Demonstrator k is specialized to one tool: whenever the correct tool on a line matches that specialty, demonstrator k produces the correct line; otherwise, it outputs a fixed placeholder tool "unknown" with the correct arguments. This construction yields five suboptimal demonstrators with complementary coverage over the tool set. For the proposed method, we use a fixed baseline completion that replaces every tool call in the optimal trace with the same placeholder tool "unknown".

Model and training details. We initialize from `Qwen/Qwen2.5-1.5B` (Yang et al., 2024) and attach a single linear advantage head on top of the final hidden states. The base checkpoint is available at `https://huggingface.co/Qwen/Qwen2.5-1.5B` under the Apache License 2.0. The backbone is fine-tuned with Low-Rank Adaptation (LoRA).

For the proposed method, we use the advantage-parameterized objective equation 163 in Appendix N. We use the squared hinge loss for $\rho_{\leq}$. As a baseline, we compare against a MaxEnt objective that fits the demonstrator completions with demonstrator-specific inverse temperatures, as described in Appendix J. We optimize all trainable parameters with the AdamW optimizer. All LLM fine-tuning experiments were run on a single NVIDIA L40S GPU. Using the same model, dataset size, batch size, and number of epochs, the proposed advantage-parameterized objective has nearly the same wall-clock training time as the MaxEnt IRL baseline, since both optimize the same backbone with comparable loss computations. Additional optimization hyperparameters are reported in Table 2.

Evaluation. We evaluate the trained model on the validation set by scoring the optimal completion and a near-optimal completion with only a single wrong tool name, sampled uniformly at random from the available tools. For validation task i, let $y_i^\star$ denote the optimal completion, $\tilde{y}_i$ denote the

Table 2: LLM experiment hyperparameters.

Field	Value
Train / val tasks	2000 / 200
Epochs / batch size / grad. accum.	1 / 8 / 4
Learning rate	10^{-4}
Optimizer	AdamW
AdamW β_1 / β_2 / ε	0.9 / 0.999 / 10^{-8}
Weight decay / warmup ratio / grad. clip	0.01 / 0.1 / 1.0
LoRA r / α / dropout	16 / 32 / 0.05
γ (discount factor)	0.95
ϵ^k (suboptimality slack)	0.1
λ_{sub} / λ_{bell}	1.0 / 1.0
Penalty function $\rho_{\leq}(x)$	$\max\{0, x\}^2$ (squared hinge)
β (MaxEnt inverse temperature)	1.0

near-optimal completion, and $S_\theta(x_i, y)$ denote the score assigned by the trained reward or advantage model to completion y for prompt x_i. We report the win-rate

$$\text{WinRate} := \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \mathbf{1}\{S_\theta(x_i, y_i^\star) > S_\theta(x_i, \tilde{y}_i)\}\,, \tag{171}$$

with mean and standard deviation computed over 10 runs.