Inferring Transferable Rewards via Active Inverse Reinforcement Learning

Victor Villin[1], **Till Freihaut**[2,†], **Andreas Schlaginhaufen**[3,†], **Giorgia Ramponi**[2], **Christos Dimitrakakis**[1], **Maryam Kamgarpour**[3]
{victor.villin,christos.dimitrakakis}@unine.ch,
{freihaut,ramponi}@ifi.uzh.ch,
{andreas.schlaginhaufen,maryam.kamgarpour}@epfl.ch

[1]**Computer Science Department, University of Neuchâtel, Switzerland**
[2]**Department of Informatics, University of Zurich, Switzerland**
[3]**SYCAMORE, EPFL, Switzerland**

[†] Equal contribution

Abstract

Inverse Reinforcement Learning (IRL) aims to recover expert rewards that transfer across environments with differing dynamics. Prior work has shown that such transfer is possible when gathering expert demonstrations across sufficiently different dynamics. In practice, however, expert demonstrations are costly, making it critical to select the right dynamics to reduce the quantity of expert data required for learning transferable rewards. We propose a two-phase approach for doing so: we cycle through the available dynamics to obtain reliable initial reward estimates, before we actively choose them by minimizing uncertainty about the expert reward. Crucially, this leads to new sample complexity guarantees in terms of transferability. Empirically, our method significantly reduces the amount of expert data needed to learn a transferable reward.

1 Introduction

Learning a *transferable* representation of expert behavior is valuable: it lets us predict how the expert would act in previously unseen circumstances. Inverse Reinforcement Learning (IRL) (Ng et al., 2000) formalizes this by inferring a reward function from expert demonstrations. Unlike imitation learning methods such as Behavior Cloning (BC) (Pomerleau, 1988), which directly replicate observed behavior, IRL aims to recover the underlying objective the expert is optimizing. This distinction matters: a cloned policy is tied to the training environment dynamics, whereas a learned reward can, in principle, transfer to new dynamics.

Despite its appeal, IRL is fundamentally ill-posed. Given expert demonstrations in a fixed environment specified by a given Markov decision process (MDP), the reward is identifiable only up to a subspace specific to that MDP (Ng et al., 1999). Consequently, such rewards transfer only to dynamics whose identifiable subspace sufficiently overlaps with that of the original MDP (Schlaginhaufen & Kamgarpour, 2024). Recent work shows that observing expert behavior across sufficiently diverse dynamics can resolve this ambiguity (Cao et al., 2021; Rolland et al., 2022; Schlaginhaufen & Kamgarpour, 2024). However, these works do not address the problem of actively selecting the dynamics in which expert demonstrations should be collected.

In practice, expert demonstrations are costly, and only a fixed set of dynamics is available. We must therefore choose carefully which dynamics to query the expert in. This raises the central question:

How should we select dynamics to learn a transferable reward from a minimal number of expert demonstrations?

To address this question, we propose an *active* IRL setting, where the learner actively selects the dynamics under which expert demonstrations are collected. This perspective is related to curriculum learning in RL (Narvekar et al., 2020) and environment design (Bruce et al., 2024; Buening et al., 2024), where dynamics are typically altered through initial states, terrain, or environmental conditions such as weather. Note that our objective differs: we choose dynamics to *learn rewards* rather than policies. To the best of our knowledge, in the setting where the learner selects dynamics, this is the first work to provide both sample complexity guarantees for reward inference in terms of expert demonstrations and a tractable method for choosing them.

Contributions. Our main contributions are as follows:

1. We formalize the active learning of transferable rewards as selecting the dynamics under which expert demonstrations are collected, so as to minimize the worst-case suboptimality of the recovered reward across all available dynamics (Equation (1)).
2. To this end, we derive novel ellipsoidal confidence sets that capture how uncertain we are about the expert reward given the demonstrations (Lemma 3.2). Building on them, we show that, after a short warm-up, how well the recovered reward transfers across all dynamics is governed by a single worst-case uncertainty measure (Lemma 3.3).
3. This points directly to which dynamics to select and motivates a simple two-phase strategy: first cycle through the dynamics to obtain a reliable reward estimate (Lemma 3.4), then actively choose dynamics to reduce the remaining uncertainty. After n rounds of dynamics selection and expert demonstrations, this yields a transfer guarantee of $\tilde{\mathcal{O}}(1/n)$, independent of the number of available dynamics (Theorem 3.6).
4. We empirically validate our approach against standard imitation learning and IRL baselines, as well as the environment design method of Buening et al. (2024) (Section 4).

2 Problem formulation

Notations. Let I_d denote the identity matrix, and $\|\cdot\|_2$ the Euclidean norm in $\mathbb{R}^d$. For a symmetric positive semidefinite matrix $X \in \mathbb{R}^{d\times d}$, we write $\|X\|_2$ for its spectral norm, $\mathrm{tr}(X)$ for its trace, $\det(X)$ for its determinant, and $\|v\|_X := \sqrt{\langle v, Xv\rangle}$ for the associated Mahalanobis norm of a vector $v \in \mathbb{R}^d$. For a finite set $\mathcal{Z}$, we write $\Delta^{\mathcal{Z}}$ for the probability simplex over $\mathcal{Z}$.

Regularized Markov Decisions Processes (MDPs). A regularized MDP is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, \nu_1, r, H, \eta)$. Here, $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, $T : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$ is the transition function, $\nu_1 \in \Delta^{\mathcal{S}}$ denotes the initial state distribution, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the reward, $H \in \mathbb{N}^+$ is the horizon. We consider entropy-regularized MDPs, with $\eta > 0$ denoting the inverse temperature that controls the strength of the regularization. Starting from state $s_1 \sim \nu_1$, an agent interacting with the MDP takes at each timestep h an action $a_h \in \mathcal{A}$, receives the reward $r(s_h, a_h)$ and the environment transitions to a new state $s_{h+1} \sim T(\cdot \mid s_h, a_h)$. Denoting the dynamics by $P = (\nu_1, T)$, the optimal solution of $\mathcal{M}$ is a policy $\pi = (\pi_h)_{h=1}^H$, with $\pi_h : \mathcal{S} \to \Delta^{\mathcal{A}}$, maximizing the regularized value $V^{\pi}_{r,P} := \mathbb{E}^{\pi}_P \left[\sum_{h=1}^H r(s_h, a_h) - \eta^{-1} \log\left(\pi_h(a_h \mid s_h)\right)\right]$. A state s is unreachable if for any sequence of H actions and any initial state $s_1 \sim \nu_1$, $s_h \neq s$ for all $h \in \{1, \dots, H\}$. We assume that policies follow an arbitrarily fixed action distribution (e.g., uniform) on such states. Under this convention, $\mathcal{M}$ admits a unique optimal policy (Geist et al., 2019). For a reward r and dynamics P, we write $V^{\star}_{r,P}$ and π^r_P for the corresponding optimal value and policy.

Setting. Given an unknown reward r^{E}, an expert acts optimally under the policy $\pi^{\mathrm{E}}_P := \pi^{r^{\mathrm{E}}}_P$ in any chosen dynamics. We consider a fixed, finite family of MDPs $\{\mathcal{M}_i\}_i$ that share r^{E} but differ in their

dynamics $\mathcal{P} := \{P_i\}_i$, and a budget $n \in \mathbb{N}^+$ for collecting expert demonstrations across them. A demonstration in $\mathcal{M}_i$ is a trajectory $\tau = (s_1, a_1, \ldots, s_H, a_H)$ generated by $\pi^{\mathrm{E}}_{P_i}$.

We consider an *active* setting in which a learner selects dynamics sequentially, and an expert is prompted after each selection. At each round $k \leq n$, the learner selects dynamics $P_k \in \mathcal{P}$, from a finite set $\mathcal{P}$, and observes m expert trajectories generated under P_k, gathering them in a dataset:

$$\mathcal{D}^{\mathrm{E}} = \bigcup_{k=1}^{n} \{\tau_{k,j}\}_{j=1}^{m}.$$

Objective. Given a fixed budget n, our goal is to select dynamics $(P_k)_{k=1}^n$ from $\mathcal{P}$ and estimate a reward r from $\mathcal{D}^{\mathrm{E}}$ that is *transferable* across all of $\mathcal{P}$. Formally, we quantify transferability by the worst-case suboptimality,

$$\max_{P \in \mathcal{P}} \mathrm{SubOpt}(r, P), \quad \text{where} \quad \mathrm{SubOpt}(r, P) := V^{\star}_{r^{\mathrm{E}},P} - V^{\pi^r_P}_{r^{\mathrm{E}},P}, \tag{1}$$

and we seek a reward r minimizing (1). This objective captures the *ground-truth* performance of the recovered optimal policy π^r_P across $\mathcal{P}$ and coincides with the notion of transferability studied by Schlaginhaufen & Kamgarpour (2024) for fixed dynamics. Furthermore, it can be viewed as a regularized analogue of the imitation gap studied in imitation learning (Swamy et al., 2021).

Before presenting the concrete algorithm for selecting the dynamics P_k and estimating the reward, we introduce our two main structural assumptions.

Structural assumptions. The first assumption enables efficient estimation of the expert reward r^{E} from the expert demonstrations $\mathcal{D}^{\mathrm{E}}$.

Assumption 2.1 (Expert bounded linear rewards)**.** Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a known feature map with $\|\phi(s,a)\|_2 \leq B_\phi$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and let $\Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B_\theta\}$, where $B_\phi, B_\theta > 0$ are known constants. We assume the expert reward is linear in *unknown* parameters $\theta^{\mathrm{E}} \in \Theta$, i.e. $r^{\mathrm{E}}(s,a) = \langle \theta^{\mathrm{E}}, \phi(s,a) \rangle$.

Linear reward parameterizations are standard in both RL (Jin et al., 2020) and inverse RL (Ng et al., 2000), and the feature map ϕ may be any expressive representation, from one-hot encodings ($d = |\mathcal{S}||\mathcal{A}|$) to learned dense embeddings.[1] For notational simplicity, we henceforth index linear rewards $r_\theta = \langle \theta, \phi \rangle$ by their parameters θ, and use the superscript E for expert quantities.

The second assumption restricts our attention to a set of deterministic dynamics.

Assumption 2.2 (Deterministic transitions)**.** $\mathcal{P}$ only contains deterministic dynamics.

As discussed in Appendix C.4, this simplifies both the derivation of confidence sets for the unknown reward parameter θ^{E} and the design of active selection strategies. The assumption is often realistic: after fixing initial states, many common RL environments are deterministic, including robotics benchmarks such as MuJoCo (Todorov et al., 2012). Moreover, we show in Section 4.1 that rewards learned under Assumption 2.2 transfer successfully to stochastic dynamics. Finally, in contrast to prior IRL work (Amin et al., 2017; Schlaginhaufen & Kamgarpour, 2024), we do not require the dynamics to be ergodic.

3 Actively learning transferable rewards

We now present our approach to actively learning transferable rewards minimizing (1), summarized in Algorithm 1. At a high level, the algorithm proceeds in two phases. In an initial *warm-up* phase, the learner has too little data to tell which dynamics are informative, so it queries the expert

[1] The assumption holds exactly in linear MDPs. With one-hot encodings, the bounded linear rewards span the ball $\left\{ r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}} : \|r\|_2 \leq \sqrt{|\mathcal{S}|\,|\mathcal{A}|}\, B_\phi B_\theta \right\}$.

Algorithm 1 Active IRL

1: **input:** set of dynamics $\mathcal{P}$, warm-up length n_δ
2: `// Warm-up phase`
3: **for** round $k = 1, \ldots, n_\delta$ **do**
4: Select P_k round-robin over $\mathcal{P}$; observe m trajectories in P_k
5: Estimate θ_{n_δ} from warm-up data by minimizing (2) (Section 3.1)
6: `// Active phase`
7: **for** round $k = n_\delta + 1, \ldots, n$ **do**
8: Actively pick P_k via (8) (Section 3.2)
9: Observe m trajectories in P_k; estimate θ_k by minimizing (2) (Section 3.1)

under each dynamics in a round-robin fashion to reduce its uncertainty about the reward. Once this uncertainty is small enough, the algorithm switches to an *active* phase: it uses its current reward estimate to select the dynamics that most reduce the remaining uncertainty.

In the following, we make this precise in two steps. In Section 3.1, we discuss how to estimate the reward by minimizing an empirical IRL loss (Algorithm 1, lines 5 and 9), and quantify the resulting uncertainty after n rounds of dynamics selection through a single scalar, the worst-case uncertainty g_n. In Section 3.2, we show that the worst-case suboptimality (1) is governed entirely by g_n, reducing the problem to actively selecting dynamics so as to minimize it (Algorithm 1, line 8). We then show that this yields a transfer guarantee of $\tilde{\mathcal{O}}\left(d^2/(nm\eta)\right)$ (Theorem 3.6).

3.1 Reward estimation

To understand how to actively select dynamics, we first need to quantify our uncertainty about the expert reward. We do so through ellipsoidal confidence sets (Lemma 3.2) centered at an empirical reward estimator (2). These sets track how uncertainty about θ^{E} decreases across rounds in the parameter directions that would affect the optimality of the expert under dynamics in $\mathcal{P}$. Complete proofs of this section's results are deferred to Appendix C.

We start by estimating θ^{E} from expert trajectories. Let θ_n be a minimizer in Θ of the empirical loss

$$L_n(\theta) \coloneqq \sum_{k=1}^{n} l_k(\theta), \qquad l_k(\theta) \coloneqq V^\star_{\theta,P_k} - \langle \theta, \hat{\phi}_k \rangle, \tag{2}$$

where P_k are the dynamics selected at round k, and $\hat{\phi}_k = \frac{1}{m}\sum_{j=1}^{m}\sum_{h=1}^{H}\phi(s_h, a_h)$ represents the empirical trajectory features of the expert, and $\theta_n \in \arg\min_{\theta\in\Theta} L_n(\theta)$. Each loss l_k matches the min-max formulation of IRL (Syed & Schapire, 2007) under dynamics P_k.

Dynamics information. Characterizing confidence around θ_n amounts to understanding how much *information* minimizing each l_k provides about θ^{E}. This information is captured by the loss curvature. Denote the Hessian of l_k at parameters θ by

$$\mathcal{H}_{P_k}(\theta) \coloneqq \nabla^2_\theta l_k(\theta) = \nabla^2_\theta V^\star_{\theta,P_k}.$$

In particular, the curvature at the expert parameters, $\mathcal{H}_{P_k}(\theta^{\mathrm{E}})$, quantifies how informative dynamics P_k are about θ^{E}. Intuitively, moving expert reward parameters in directions induces non-zero curvature if and only if it changes the optimal policy under P_k. These are *identifiable* directions.

Accumulating curvatures across rounds, the Hessian of the full loss (2) at expert parameters θ^{E}, $\mathcal{H}^{\mathrm{E}}_n \coloneqq \sum_{k=1}^{n}\mathcal{H}_{P_k}(\theta^{\mathrm{E}})$, summarizes the information we gathered about the expert in the previously selected dynamics $(P_k)_{k=1}^{n}$. For our confidence bound to be useful uniformly over $\mathcal{P}$, the accumulated information $\mathcal{H}^{\mathrm{E}}_n$ must cover the information $\mathcal{H}_P(\theta^{\mathrm{E}})$ relevant under every $P \in \mathcal{P}$. Fixing some $\lambda_0 > 0$ for invertibility, we measure this coverage via the transfer operator

$$\mathcal{T}^{\mathrm{E}}_n(P) \coloneqq \mathcal{H}_P(\theta^{\mathrm{E}})\left(\mathcal{H}^{\mathrm{E}}_n + \lambda_0 I_d\right)^{-1}, \tag{3}$$

whose spectral norm is small when $\mathcal{H}_n^{\mathrm{E}}$ dominates $\mathcal{H}_P(\theta^{\mathrm{E}})$ on the identifiable directions of P. Taking the worst case over $\mathcal{P}$ yields the *worst-case uncertainty*

$$g_n \;\coloneqq\; \max_{P\in\mathcal{P}} \|\mathcal{T}_n(P)\|_2 \,. \tag{4}$$

To provide intuition, g_n is large when there exist dynamics $P \in \mathcal{P}$ and an identifiable direction under P that remains poorly covered by $\mathcal{H}_n^{\mathrm{E}}$. When the selected dynamics $(P_k)_{k=1}^n$ collectively cover all identifiable directions, the following lemma ensures that g_n shrinks.

Lemma 3.1. *For any dynamics $P \in \mathcal{P}$, the Hessian $\mathcal{H}_P(\theta^E)$ admits a strictly positive spectral gap: its smallest non-zero eigenvalue is lower bounded by some constant $\kappa > 0$.*

Proof sketch. Under Assumption 2.2, $\mathcal{H}_P(\theta^{\mathrm{E}})$ is a covariance matrix of trajectory features under the expert trajectory law (see Appendix C.2). Any reward perturbation that changes optimality induces non-zero variability in these features. Combining lower bounds on expert trajectory probabilities and feature variability yields a lower bound $\kappa > 0$ on the smallest non-zero eigenvalue of $\mathcal{H}_P(\theta^{\mathrm{E}})$. □

Ellipsoidal confidence. Crucially, we find that reducing g_n below a threshold leads to ellipsoidal confidence sets aligned with the geometry of $\mathcal{H}_n^{\mathrm{E}}$.

Lemma 3.2. *Let $\delta \in (0,1)$ and the warm-up length be*

$$n_\delta \coloneqq \min\left\{ k : g_k \;\le\; \frac{(1-e^{-1})^2 \kappa m}{32 d\eta H^2 B_\phi^2 \log(d/\delta)} \right\}. \tag{5}$$

Then, with probability at least $1-\delta$, it holds for all $n \ge n_\delta$ that

$$\|\theta_n - \theta^{\mathrm{E}}\|^2_{\mathcal{H}_n^E} \;\lesssim\; \frac{d\,\log(d/\delta)}{\eta m} \,.$$

Proof sketch. The proof leverages the pseudo self-concordance of L_n (2) (in the sense of Bach (2010)), combined with the warm-up condition (5). This can be viewed as an extension of the analysis from Schlaginhaufen & Kamgarpour (2026) (Theorem 4.3) to multiple dynamics. □

The warm-up length (5) is governed by the rate at which information accumulates across identifiable directions, as captured by κ (Lemma 3.1), and scales quadratically with the horizon H and feature scale B_ϕ. Once this warm-up phase is completed, sufficient information has been gathered in all identifiable directions for local curvature arguments to apply, as is standard in analyses of self-concordant objectives (Chaudhuri et al., 2015; Lee & Oh, 2025).

A key consequence is that the resulting confidence bound is independent of both κ and the number of rounds n. Moreover, increasing the per-round sample size m simultaneously reduces estimation variance and relaxes the warm-up condition 5. Thus, when collecting additional trajectories is cheaper than switching dynamics, larger values of m are preferable.

3.2 Dynamics selection

In this section, we introduce new transfer guarantees (Lemma 3.3) that hold for any sequence $(P_k)_{k=1}^n$, unveiling that we must actively minimize g_n (4). We propose a two-phase selection rule for doing so: we first obtain reliable estimates by picking dynamics passively for a number of warm-up rounds, scaling linearly in $|\mathcal{P}|$ (Lemma 3.4). We then actively choose dynamics approximately minimizing worst-case uncertainty g_n. We show this retrieves $\tilde{\mathcal{O}}(d^2/(nm\eta))$-optimal policies across $\mathcal{P}$ (Theorem 3.6). Complete proofs for the results of this section can be found in Appendix D.

General transfer guarantees. To design a provably efficient selection strategy, we must understand how well the minimizer θ_n of L_n (2) is guaranteed to transfer to all dynamics in $\mathcal{P}$, given the dynamics selected for demonstrations. After the warm-up condition (5) is verified, we find that transferability is governed by our worst-case uncertainty g_n.

Lemma 3.3. *Fix $\delta \in (0,1)$ and $n \geq n_\delta$. For any $P \in \mathcal{P}$, it holds with probability at least $1-\delta$ that*

$$\mathrm{SubOpt}(\theta_n, P) \lesssim g_n \left(\frac{d \log(d/\delta)}{\eta m} + \lambda_0 B_\theta \right). \tag{6}$$

Proof sketch. We first remark that $\mathrm{SubOpt}(\theta_n, P) = \int_0^1 (1-t)\mathcal{H}_P(t\theta^{\mathrm{E}} + (1-t)\theta_n)\mathrm{d}t$. Pseudo self-concordance of $V^\star_{\theta,P}$ yields quadratic control over suboptimality: $\mathrm{SubOpt}(\theta_n, P) \leq \psi_n \left\|\theta_n - \theta^{\mathrm{E}}\right\|^2_{\mathcal{H}_P(\theta^{\mathrm{E}})} \leq \psi_n \left\|\mathcal{T}^{\mathrm{E}}_n(P)\right\|_2 \left\|\theta_n - \theta^{\mathrm{E}}\right\|^2_{\mathcal{H}^{\mathrm{E}}_n + \lambda_0 I_d}$, for some factor ψ_n. Using that $\psi_n \leq e-2$ for all $n \geq n_\delta$, together with the definition of g_n and Lemma 3.2, gives the result. □

This result provides a strong directive to pick dynamics minimizing g_n. The smaller g_n is, the sooner the warm-up condition (5) is met and the tighter the resulting transfer bound (6) becomes. However, g_n depends on the unknown θ^{E}. In practice, one may instead sequentially select each P_k using estimates θ_k. An obstacle, however, is that these estimates can be highly inaccurate during the warm-up phase. This reflects a classical challenge in active learning (Chaudhuri et al., 2015; Rittler & Chaudhuri, 2023; Xie et al., 2024), where initial uncertainty prevents reliable adaptive selection.

A two-phase strategy. To overcome this challenge, we propose a two-phase approach. The inital phase uses a *passive* strategy that avoids reliance on early estimates. In the second phase, we *actively* pick the dynamics to approximately minimize worse-case uncertainty g_n.

Let $k(P)$ denote how many times P was selected after k rounds. In the passive phase, we simply cycle through $\mathcal{P}$ by picking the least chosen dynamics so far:

$$P_{k+1} \in \arg\min_{P \in \mathcal{P}} k(P). \tag{7}$$

We show this gives a finite warm-up length that grows linearly with the number of dynamics $|\mathcal{P}|$.

Lemma 3.4. *Let us follow the passive rule (7) until Condition (5) is verified. Then, it holds that*

$$n_\delta = |\mathcal{P}| \left\lceil \frac{32 d \eta H^2 B_\phi^2 \log(d/\delta)}{(1-e^{-1})^2 \kappa m} \right\rceil.$$

Proof sketch. Remark that for any $k > 0$ and $P \in \mathcal{P}$, following the rule (7) ensures $\mathcal{H}^{\mathrm{E}}_n(\theta^{\mathrm{E}}) \succeq \lfloor n/|\mathcal{P}| \rfloor \mathcal{H}_P(\theta^{\mathrm{E}})$. By the definition of g_k, for any $k \geq |\mathcal{P}|$, this leads to $g_k \leq \lfloor k/|\mathcal{P}| \rfloor^{-1}$. Plugging in the definition of n_δ (Equation (5)) recovers the result. □

Once the condition (5) is met ($k \geq n_\delta$), θ_k becomes accurate enough to support reliable active selection. The approximate transfer operator $\mathcal{T}_k(P) = \mathcal{H}_P(\theta_k) \left(\lambda_0 I_d + \sum_{i=1}^k \mathcal{H}_{P_i}(\theta_k)\right)^{-1}$ is a reliable proxy for minimizing g_k. For any $k \geq n_\delta$ and $J \in \{\|\cdot\|_2, \mathrm{tr}, \log\det(I_d + \cdot)\}$, we greedily select the dynamics whose identifiable directions remain the most uncertain:

$$P_{k+1} = \max_{P \in \mathcal{P}} J\left(\mathcal{T}_k(P)\right). \tag{8}$$

Remark 3.5. Minimizing worst-case uncertainty g_n (4) relates to G-optimal designs (Kiefer & Wolfowitz, 1959; Kiefer, 1974). Since G-optimal designs are hard to compute, it is common to rely on approximations (Krause & Guestrin, 2007). In the case where Hessians are rank-1 and parameter-independent, the choices for J in Equation (8) all correspond to greedy D-optimal design.

Theorem 3.6. *Follow the active rule (8) for all $k \in \{n_\delta, \dots, n\}$. It holds for every $P \in \mathcal{P}$ that*

$$\mathrm{SubOpt}(\theta_n, P) \lesssim \frac{d}{n - n_\delta + 1} \log\left(\frac{nm}{\log(d/\delta)}\right)\left(\frac{d\log(d/\delta)}{\eta m} + \lambda_0 B_\phi\right). \tag{9}$$

Proof sketch. For any $k \in \{n_\delta, \dots, n\}$, we first show that the Hessians $\mathcal{H}_P(\theta_k)$ and $\mathcal{H}_P(\theta^{\mathrm{E}})$ coincide up to a e-factor. We then upper bound g_n using an elliptical potential argument relating the spectral norm to the log-determinant. Combining this bound with Lemma 3.3 yields the result. □

Theorem 3.6 shows that, after satisfying the warm-up condition (5), the active selection rule (8) identifies a reward whose induced policies are $\tilde{\mathcal{O}}(d^2/(nm\eta))$-optimal over $\mathcal{P}$. The dependence on the inverse temperature η reflects the reduced information available from more strongly regularized expert policies, while n and m contribute similarly, indicating that switching dynamics less frequently does not worsen the guarantees. The result also holds with the effective dimension $d_{\mathrm{eff}} \leq d$ of the identifiable parameter subspace (see Appendix B). Finally, the bound in (9) is independent of $|\mathcal{P}|$: rather than exploring $\mathcal{P}$ uniformly, the selection rule concentrates sampling effort on the most informative dynamics.

To summarize, Algorithm 1 follows the passive rule (7) until the warm-up condition (5) is satisfied, which is guaranteed to occur after finitely many rounds (Lemma 3.4). The algorithm then switches to the active rule (8), yielding the transfer guarantees of Theorem 3.6.

Remark 3.7 (Offline active learning). When sequential interaction is impractical, one can select all dynamics in advance using $\theta_0 = 0$ instead of θ_n. This offline variant is analyzed in Appendix E.

Computational complexity. Under Assumption 2.2, Hessians coincide with feature covariance matrices (Lemma C.1). In contrast to the approach of Buening et al. (2024), which requires $\mathcal{O}(k|\mathcal{P}|)$ RL solves at round k, our selection rule (8) only requires estimating Hessians from sampled trajectories. Full algorithmic details are provided in Appendix F.

4 Experiments

Our experiments serve two purposes: *(i)* to evaluate when active learning improves sample efficiency in terms of expert demonstrations, and *(ii)* to identify which dynamics are most useful for efficiently learning transferable rewards.

We instantiate our two-phase strategy (Section 3.2) using the log-determinant criterion, with the warm-up phase occupying one third of the total demonstration budget.[2] We compare against several baselines to isolate the benefits of: (a) the passive warm-up phase, (b) leveraging multiple dynamics, and (c) learning transferable rewards rather than directly imitating behavior.

To evaluate (a), we include an *Active only* ablation that follows the active rule (8) from the start. To evaluate (b), we compare against the environment design (ED) approach of Buening et al. (2024), a passive baseline following (7), and single-dynamics methods that learn from fixed dynamics sampled uniformly from $\mathcal{P}$: fixed-dynamics IRL and GAIL (Ho & Ermon, 2016). All IRL methods except GAIL optimize the min-max IRL objective (2). Finally, to evaluate (c), we compare against imitation methods: Behavioral Cloning (BC) and robust imitation via RIME (Chae et al., 2022).

All methods follow the same evaluation protocol. At each round k, the learner selects a dynamics $P_k \in \mathcal{P}$, observes $m = 1$ expert trajectories optimal with respect to unknown parameters θ^{E}, and updates its reward estimate. For single-dynamics methods, the environment is uniformly sampled once from $\mathcal{P}$ and then fixed. Evaluation proceeds by learning a reward from collected data and deploying a policy optimized for each tested dynamics; imitation methods (BC, RIME) instead deploy a single policy across all dynamics. In all experiments, state–action features are indicator features over states and actions.

[2] We report log-det results for clarity, as it performed best empirically; other criteria are reported in the appendix.

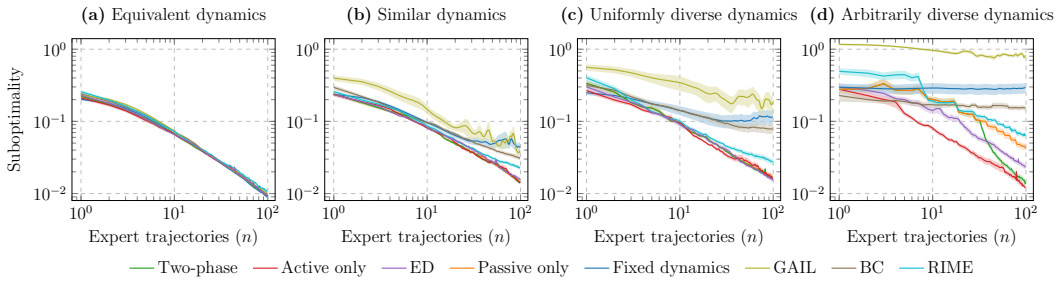

Figure 1: Worst-case suboptimality over the training set $\mathcal{P}$ as a function of the number of observed expert trajectories on randomly generated MDPs. Each plot corresponds to a different structure of $\mathcal{P}$.

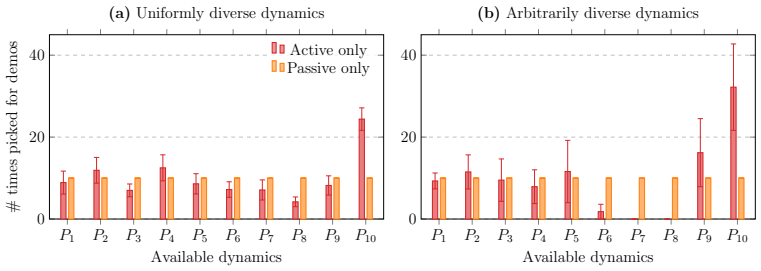

Figure 2: Number of times each dynamics in the training set $\mathcal{P}$ is selected after 100 rounds, for two configuration of $\mathcal{P}$. Dynamics are ordered by increasing average dissimilarity to the rest of $\mathcal{P}$, so that P_{10} corresponds to the most isolated dynamics.

Performance is measured by suboptimality (1) and by reward estimation error (up to constants), defined as $\min_{c\in\mathbb{R}} \left\|\theta^{\mathrm{E}} - \theta_n + c\right\|_2$. Results are averaged over 20 randomly seeded runs with standard errors. We do not address scalability here and focus on tractable environments without function approximation, leaving such extensions to future work. Further details and results are in Appendix G.

4.1 Controlled sets of dynamics

In this experiment, we investigate when active dynamics selection improves sample efficiency over passive selection. To this end, we construct training sets of dynamics $\mathcal{P}$ with controlled transfer structure, i.e., controlling how well rewards learned in one dynamics transfer to the others in $\mathcal{P}$. We quantify transferability through principal angles between dynamics (Galántai, 2004; Schlaginhaufen & Kamgarpour, 2024).

We consider four types of dynamics sets: (1) Equivalent dynamics, where rewards transfer perfectly across $\mathcal{P}$; (2) Similar dynamics, where rewards transfer consistently well; (3) Uniformly diverse dynamics, where transfer is uniformly weak across pairs; (4) Arbitrarily diverse dynamics, where transferability is heterogeneous across the set. For each random seed, we generate a new set of $|\mathcal{P}| = 10$ dynamics. After $n = 100$ demonstrations, methods are additionally evaluated on held-out dynamics. All environments use $|\mathcal{S}| = 20$ states and $|\mathcal{A}| = 5$ actions. Training dynamics are deterministic, while held-out evaluation dynamics are stochastic.

Results. Figure 1 shows that single-dynamics methods (Fixed dynamics, GAIL, and BC) fail when the dynamics set is too diverse (Figures 1(c)–(d)). In contrast, active selection becomes particularly beneficial in arbitrarily diverse sets (Figure 1(d)). Specifically, the Active only variant performs the best, empirically beating the theoretically motivated two-phase strategy, suggesting that the warm-up phase may be overly conservative in practice. Table 1 further confirms that it yields the most transferable rewards under limited demonstration budgets, widely outperforming ED.

Table 1: Worst-case suboptimality on sets of *held-out* stochastic dynamics after learning from $n = 100$ expert trajectories collected on either uniformly or arbitrarily diverse sets of dynamics.

	Worst-case suboptimality ($\times 100$) on held-out dynamics	
Algorithm	**Uniformly diverse training set**	**Arbitrarily diverse training set**
Active only	$\mathbf{02.86 \pm 0.02}$	$\mathbf{12.63 \pm 0.14}$
Two-phase	02.93 ± 0.02	13.63 ± 0.15
ED	03.32 ± 0.02	16.52 ± 0.22
Passive only	03.15 ± 0.02	19.59 ± 0.19
Fixed dynamics	18.18 ± 0.16	24.36 ± 0.09
GAIL	26.62 ± 0.37	96.06 ± 0.42
BC	15.05 ± 0.18	20.01 ± 0.10
RIME	04.91 ± 0.03	24.47 ± 0.34

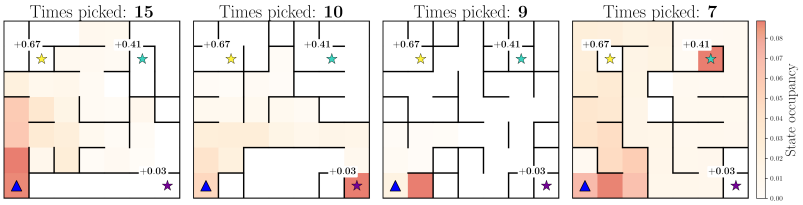

Figure 3: Mazes selected most frequently after following the active rule (8) for 100 rounds, on a representative random seed. The blue triangle corresponds to the initial state, stars are goal states, the displayed rewards are the expert rewards, and the heatmap represents the state occupancy of policies optimized under the inferred rewards, i.e., the proportion of time spent in each state.

Figure 2 provides insight into the dynamics selected by our active strategy. In uniformly sets (Figure 2(a)), selection closely matches the passive strategy. In contrast, for arbitrarily sets (Figure 2(b)), the algorithm concentrates demonstrations on dynamics that either transfer well on average across $\mathcal{P}$ ($P_1, \ldots, P_5$) or are highly isolated from the rest of the set (P_{10}).

4.2 Random mazes

We consider a maze navigation task as another convenient benchmark, which provides intuitive visuals on the dynamics selected. The goal of the maze task is to reach one of three goal states in the least amount of steps (Buening et al., 2024) (See Figure 2).

We generate sets of $|\mathcal{P}| = 100$ mazes by randomly placing walls, producing a wide range of difficulty levels, from nearly empty layouts to highly constrained or even unsolvable configurations (e.g., blocked goals). All mazes within a set share the same reward function and initial state, with rewards for goal states sampled uniformly in $[0, 1]$ at generation time. We reward an agent for waiting in a goal state. Thus, within each set, only the maze layout, and therefore the transition function, varies. Here, $|\mathcal{S}| = 49$, $|\mathcal{A}| = 5$ (four movement actions and a waiting action).

Results. On Figure 4, we see that our active learning method without warm-up achieves neatly superior sample efficiency: its suboptimality after 20 expert trajectories is only attained at 100 trajectories under passive learning. It also quickly attains low reward error, indicating that it may also identify the ground truth expert reward more efficiently.

Figure 3 illustrates the maze selection behavior of our active rule (8). The method often selects unsolvable mazes, which, counterintuitively, help reduce uncertainty in the rewards. In such dynamics, where the goal is unreachable but the layout remains open, the entropy regularized expert acts uniformly, revealing that rewards over reachable states are effectively constant. This highlights that informative dynamics do not need to be solvable, but should reveal reward structure relevant for performance for the set of available dynamics.

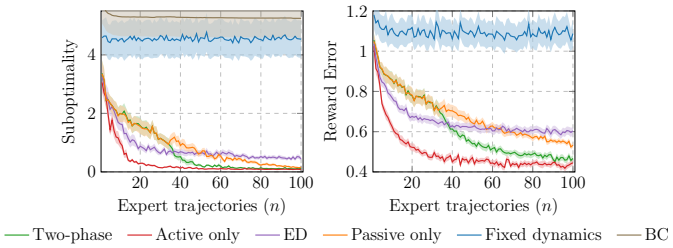

Figure 4: Worst-case suboptimality over the training set $\mathcal{P}$ of 100 randomly generated mazes, and the corresponding reward error, as a function of the number of observed expert trajectories.

5 Related Work

Active IRL aims to minimize expert interactions, the main practical bottleneck, by intelligently querying the expert. Prior work has mostly focused on the single-dynamics setting. For instance, Lopes et al. (2009) and Cui & Niekum (2018) adopt a Bayesian IRL perspective to actively select states or trajectories, respectively, maximizing information about the reward function. Lindner et al. (2023) instead target the feasible set of rewards consistent with expert behavior, using confidence sets over plausible rewards to guide exploration toward the most informative regions of the Markov Decision Process (MDP). In contrast, our approach actively selects the dynamics under which the expert acts, rather than querying arbitrary states or trajectories.

Identifying rewards and ensuring their **transferability** across dynamics are central challenges in IRL. Cao et al. (2021) and Rolland et al. (2022) show that sufficiently diverse dynamics are necessary to recover rewards up to constants (in the entropy-regularized setting), assuming access to the expert policy. Removing this assumption, Schlaginhaufen & Kamgarpour (2024) provide sufficient conditions for transferring learned rewards to new dynamics. All these works rely on an ergodicity assumption on the dynamics. In contrast, our analysis avoids this requirement. Importantly, none of these works address active selection of dynamics.

Learning from **multiple dynamics** has been studied in both imitation learning and IRL. Chae et al. (2022) learn policies that are robust to dynamics perturbations by matching averaged expert occupancy measures across environments, but without sample complexity guarantees or adaptive selection of dynamics. In IRL, Amin et al. (2017) consider a framework where the learner selects policies while an adversary selects dynamics, and the expert provides feedback only when the learner is suboptimal. This leads to guarantees on reward identification in terms of the number of suboptimal actions. More closely related to our setting, where the learner actively chooses the dynamics for expert demonstrations, Nitschke et al. (2024) and Bajgar et al. (2024) propose information-directed strategies, such as maximizing behavior entropy or expected information gain, while Buening et al. (2024) select dynamics under which the current reward estimate is believed to perform worst. However, these methods either focus on state-only rewards, deterministic experts, rely on computationally expensive selection rules, or lack formal sample complexity guarantees for active dynamics selection. In contrast, our work provides a principled and provably efficient framework for actively choosing dynamics to learn transferable rewards.

6 Conclusion

We introduced a novel active learning method for inferring transferable rewards, by adaptively selecting the dynamics under which expert demonstrations are collected. Our approach first follows a passive exploration phase to obtain a sufficiently reliable estimate of the expert reward, before actively selecting dynamics that minimize uncertainty. We showed that this strategy benefits from transfer guarantees, ensuring that policies optimized under the inferred rewards remain near-optimal across the set of dynamics available during training. Empirically, our method learns transferable

rewards using substantially fewer expert demonstrations than passive dynamics selection and outperforms another relevant active learning approach.

Limitations and future work. Our analysis currently relies on a warm-up phase, although our empirical results suggest that this requirement may be unnecessary in practice. Removing this requirement is therefore an important direction for future work. More importantly, our study is restricted to discrete environments and linear rewards. Extending the framework to continuous settings and expressive reward parametrizations therefore remains an open challenge. One promising direction is to model rewards with neural networks while treating intermediate representations as learned feature embeddings for Hessian estimation (Ash et al., 2021). Advancing along these directions should help bridge our theoretical results with the practical challenges of actively learning transferable rewards.

Appendix

A Notations and Preliminaries

We enumerate below the most important notations used in our appendix.

Table 2: Symbols

$\mathcal{P}$	$:=$	finite set of dynamics that is available for expert trajectories
$\mathcal{H}_P(\theta)$	$:=$	$\nabla^2_\theta V^\star_{\theta,P}$
$\mathcal{H}^{\mathrm{E}}_n$	$:=$	$\sum_{k=1}^n \mathcal{H}_{P_k}(\theta^{\mathrm{E}})$
$G_P(\theta_1,\theta_2)$	$:=$	$\int_0^1 (1-t)\mathcal{H}_P\left(\theta_2 t + (1-t)\theta_1\right) dt$
π^θ_P	$:=$	$\max_\pi V^\pi_{\theta,P}$
$\nu^\theta_P(s)$	$:=$	$\mathbb{E}^{\pi^\theta_P}_P\left[\sum_{h=1}^H \mathbb{1}\left[s_h = s\right]\right]$
$\mu^\theta_P(s,a)$	$:=$	$\nu^\theta_P(s)\pi^\theta_P(a \mid s)$
$\mathrm{H}(\mu^\theta_P)$	$:=$	$-\sum_{s,a}\mu^\theta_P(s,a)\log\left(\pi^\theta_P(a \mid s)\right)$
Θ	$:=$	$\left\{\theta \in \mathbb{R}^d : \left\|\theta\right\|_2 \le B_\theta\right\}$
$\Theta^\perp_P$	$:=$	$\left\{\theta \in \Theta : \mu^{\mathrm{E}}_P = \mu^{\theta+\theta^{\mathrm{E}}}_P\right\}$
$\Theta^\perp_{\mathcal{P}}$	$:=$	$\bigcap_{P\in\mathcal{P}}\Theta^\perp_P$
Π_P	$:=$	orthogonal projector onto Θ_P
$\Pi_{\mathcal{P}}$	$:=$	orthogonal projector onto $\Theta_{\mathcal{P}}$
$k(P)$	$:=$	$\sum_{i=1}^k \mathbb{1}\left[P_i = P\right]$
R	$:=$	$2\eta H B_\phi$
$\phi(\tau)$	$:=$	$\sum_{h=1}^H \phi(s_h,a_h)$
ϕ^θ_P	$:=$	$\mathbb{E}^{\pi^\theta_P}_P\left[\phi(\tau)\right]$
$\lambda^{\mathcal{X}}_{\min}(X)$	$:=$	$\min_{v\in\mathcal{X},\left\|v\right\|_2=1}\left\|v\right\|^2_X$
$A \oplus B$	$:=$	$\{a+b \mid a\in A,\, b\in B\}$
x	$:=$	desc

A.1 Preliminaries

MDPs. Throughout this appendix, we work under the same regularized MDP setup introduced in the main text in Section 2. Introducing state occupancies as $\nu^\pi_P(s) = \mathbb{E}^\pi_P\left[\sum_{h=1}^H \mathbb{1}\left[s_h = s\right]\right]$ and state-action occupancies as $\mu^\pi_P(s,a) = \nu^\pi_P(s)\pi(a \mid s)$, we can conveniently express the value objective as:

$$V^\pi_{r,P} := \langle r, \mu^\pi_P\rangle + \frac{1}{\eta}h(\mu^\pi_P),$$

where $h(\mu^\pi_P) = -\sum_{s,a}\mu^\pi_P(s,a)\log\left(\pi(a \mid s)\right)$ is the expected Shannon entropy of policy π in dynamics P. The optimal policy in regularized MDPs has a softmax characterization (Geist et al., 2019):

$$\pi^\theta_{h,P}(a \mid s) = \frac{\exp\left(\eta Q^\theta_{h,P}(s,a)\right)}{\sum_{a'\in\mathcal{A}}\exp\left(\eta Q^\theta_{h,P}(s,a)\right)}, \tag{10}$$

where Q^θ_P is the optimal regularized Q-function:

$$Q^\theta_{h,P}(s,a) := \left\{\langle\theta,\phi(s,a)\rangle - \frac{1}{\eta}\log\left(\pi^\theta_{h,P}(a \mid s)\right) + \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} T(s' \mid s,a)\pi^\theta_{h+1,P}(a' \mid s')Q^\theta_{h+1,P}(s',a')\right\}.$$

Eigenvalues. The smallest eigenvalue of X is denoted as $\lambda_{\min}(X) := \min_{\left\|v\right\|_2=1}\left\|v\right\|^2_X$ and the smallest eigenvalue restricted to a subspace $\mathcal{X}\subset\mathbb{R}^d$ as $\lambda^{\mathcal{X}}_{\min}(X) := \min_{\left\|v\right\|_2=1,\, v\in\mathcal{X}}\left\|v\right\|^2_X$.

Features. For convenience, we use the $\phi(\tau) = \sum_{h=1}^{H} \phi(s_h, a_h)$ to denote trajectory features, and $\phi_P^{\Theta} = \mathbb{E}_P^{\pi_P^{\theta}}[\phi(\tau)]$ to denote feature occupancies, with $\phi_P^{\mathrm{E}} = \phi_P^{\theta^{\mathrm{E}}}$.

Fréchet derivatives. For a function $f : \mathbb{R}^d \to \mathbb{R}$ that is three times differentiable at $x \in \mathbb{R}^d$, we denote the first three directional derivatives along $u, v, w \in \mathbb{R}^d$ by

$$Df(x)[u] := \langle \nabla f(x), u \rangle, \quad D^2 f(x)[u, v] := \langle u, \nabla^2 f(x) v \rangle,$$
$$D^3 f(x)[u, v, w] := D\left(D^2 f(x)[u, v]\right)[w].$$

where $D^n f(x)$ denotes the n-th Fréchet derivative of f.

B Identifiably under multiple dynamics

Ideally, we seek reward parameters that induce expert-optimal behavior across all dynamics in $\mathcal{P}$. This defines the *feasible set*

$$\Theta_{\mathcal{P}}^{\star} := \bigcap_{P \in \mathcal{P}} \Theta_P^{\star}, \qquad \Theta_P^{\star} := \{\theta \in \Theta : \mathrm{SubOpt}(\theta, P) = 0\},$$

namely the intersection of the per-dynamics optimality sets.

A fundamental difficulty in IRL is that reward parameters are not uniquely identifiable from behavior alone. Indeed, for a fixed dynamics P, there exist transformations of the reward that leave the optimal policies unchanged. Consequently, the expert parameter θ^{E} can only be identified up to the subspace of such invariant transformations.

We now characterize these unidentifiable directions. They arise from two distinct sources. First, potential-based shaping transformations preserve optimal policies (Ng et al., 1999). Define

$$\Theta_P^{\mathrm{shaping}} := \{\theta \in \Theta : \langle \theta, \phi \rangle \in \mathrm{im}(E - T)\},$$

where $E, T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ are the linear operators

$$(Ev)(s, a) = v(s), \qquad (Tv)(s, a) = \sum_{s'} T(s' \mid s, a) v(s').$$

Second, rewards assigned to unreachable state-action pairs are unconstrained, since they never affect the induced behavior. We define

$$\Theta_P^{\mathrm{unreachable}} := \left\{\theta \in \Theta : \langle \theta, \phi(s, a) \rangle = 0 \quad \forall (s, a) \text{ such that } \max_{\pi} \nu_P^{\pi}(s, a) > 0\right\}.$$

Together, these directions span the full unidentifiable subspace of P. Denoting

$$\Theta_P^{\perp} := \Theta_P^{\mathrm{shaping}} + \Theta_P^{\mathrm{unreachable}},$$

the feasible set admits therefore the decomposition

$$\Theta_P^{\star} = \{\theta^{\mathrm{E}}\} \oplus \Theta_P^{\perp}.$$

Taking the intersection, the feasible set of our problem reduces to:

$$\mathrm{IRL}(\mathcal{P}) \subset \{\theta^{\mathrm{E}}\} \oplus \Theta_{\mathcal{P}}^{\perp}, \qquad \text{where} \qquad \Theta_{\mathcal{P}}^{\perp} = \bigcap_{P \in \mathcal{P}} \Theta_P^{\perp}.$$

In particular, Rolland et al. (2022) and Schlaginhaufen & Kamgarpour (2024) show that $\Theta_{\mathcal{P}}^{\perp}$ reduces to constant reward shifts when two dynamics in $\mathcal{P}$ are ergodic and their shaping subspaces are sufficiently misaligned. Naturally, $\Theta_{\mathcal{P}}$ represents the subspace of *identifiable* parameter transformations.

Other types of characterizations of feasible sets in IRL have also been established, for instance in the fixed-dynamics and unregularized setting (Metelli et al., 2021).

C Proofs of Section 3.1

C.1 Reward estimation preliminaries

In this paper, we consider minimizing the following aggregated empirical IRL loss to estimate rewards:

$$L_n(\theta) := \sum_{k=1}^{n} l_k(\theta_n), \qquad l_k(\theta) = V^\star_{\theta,P_k} - \langle \theta, \hat{\phi}_k \rangle.$$

where $\hat{\phi}_k = \frac{1}{m}\sum_{j=1}^{m} \phi(\tau_{k,j})$, and $\tau_{k,j}$ are the j-th trajectories generated by the expert at round k. We denote its minimizer over Θ by $\theta_n := \arg\min_{\theta\in\Theta} L_n(\theta)$. We furthermore define the population loss

$$L_n^{\text{pop}}(\theta) = \sum_{k=1}^{n} \left(V^\star_{\theta,P_k} - \langle \theta, \phi^{\text{E}}_P \rangle \right), \tag{11}$$

minimized at the expert parameters ($L_n^{\text{pop}}(\theta^{\text{E}}) = 0$). Defining $S_n := \sum_{k=1}^{n}(\hat{\phi}_k - \phi^{\text{E}}_{P_k})$, the aggregated empirical loss decomposes as

$$L_n(\theta) = L_n^{\text{pop}}(\theta) \; - \; \langle \theta, \, S_n \rangle.$$

Additionally, we introduce the following key quantities of our analysis:

$$\mathcal{H}_P(\theta) := \nabla^2_\theta V^\star_{\theta,P}, \qquad \mathcal{H}^E_n := \nabla^2 L_n(\theta^E) = \nabla^2 L_{\text{pop}}(\theta^E) = \sum_{k=1}^{n} \mathcal{H}_{P_k}(\theta^E),$$

$$\lambda_n := \lambda_{\min}^{\Theta_{\mathcal{P}}}(\mathcal{H}^E_n), \qquad R := 2\eta H B_\phi, \qquad R_n := \frac{R}{\sqrt{\lambda_n}},$$

$$\rho_n := \|\theta_n - \theta^E\|_{\mathcal{H}^E_n}, \qquad \omega_n := \|S_n\|_{(\mathcal{H}^E_n)^\dagger}.$$

Here, $(\mathcal{H}^E_n)^\dagger$ denotes the Moore-Penrose inverse of $\mathcal{H}^{\text{E}}_n$.

C.2 Supporting lemmas

We present several technical lemmas laying out various properties of the optimal regularized value function and population loss 11. To make the dependence of the optimal value function on reward parameters explicit, we sometimes use the notation $V^\star_{\theta,P} = V^\star_P(\theta)$.

Lemma C.1 (Value derivatives)**.** *Let Assumption 2.1 hold. Then, for any $\theta, v, u \in \mathbb{R}^d$ and deterministic P, it holds that*

$$\begin{aligned} DV^\star_P(\theta)[v] &= \langle v, \phi^\theta_P \rangle, \\ D^2 V^\star_P(\theta)[v,v] &= \|v\|^2_{\mathcal{H}_P(\theta)} = \eta \mathbb{E}_P^{\pi^\theta_P}\left[\langle v, \phi(\tau) - \phi^\theta_P \rangle^2 \right]. \\ D^3 V^\star_P(\theta)[v,v,u] &= \eta^2 \mathbb{E}_P^{\pi^\theta_P}\left[\langle v, \phi(\tau) - \phi^\theta_P \rangle^2 \langle u, \phi(\tau) - \phi^\theta_P \rangle \right]. \end{aligned}$$

Proof. Fix θ and define

$$A^\theta_P(\tau) := \sum_{h=1}^{H} A^\theta_{h,P}(s_h, a_h),$$

with

$$A^\theta_{h,P}(s,a) = \mathbb{E}_P^{\pi^\theta}\left[\sum_{t=h}^{H} \phi(s_t,a_t) \middle| s_t = s, a_t = a \right] - \mathbb{E}_P^{\pi^\theta}\left[\sum_{t=h}^{H} \phi(s_t,a_t) \middle| s_t = s \right].$$

By Lemma A.13 in (Schlaginhaufen & Kamgarpour, 2026), using the assumed linear structure of rewards, we have for any $\theta, u, v \in \mathbb{R}^d$

$$\begin{aligned} DV_P^\star(\theta)[v] &= \langle v, \phi_P^\theta \rangle, \\ D^2 V_P^\star(\theta)[v,v] &= \eta \mathbb{E}_P^{\pi_P^\theta}\left[\langle v, A_P^\theta \rangle^2\right], \\ D^3 V_P^\star(\theta)[v,v,u] &= \eta^2 \mathbb{E}_P^{\pi_P^\theta}\left[\langle v, A_P^\theta \rangle^2 \langle u, A_P^\theta \rangle\right]. \end{aligned}$$

We assumed dynamics are deterministic. We therefore remark that for any parameters $\theta \in \Theta$, dynamics P and $t \in [1, \dots, H-1]$, the feature advantage cancels out with

$$\mathbb{E}_P^{\pi^\theta}\left[\sum_{h=t}^{H} \phi(s_h, a_h) \middle| s_h = s_t, a_h = a_t\right] - \mathbb{E}_P^{\pi^\theta}\left[\sum_{h=t+1}^{H} \phi(s_h, a_h) \middle| s_h = s_{t+1}\right] = \phi(s_t, a_t),$$

since conditioning on (s_t, a_t) is equivalent to conditioning on s_{t+1} when dynamics are deterministic. Hence, the trajectory feature advantage accepts the following form via telescoping:

$$\begin{aligned} A_P^\theta(\tau) &= \left(\sum_{h=1}^{H} \phi(s_h, a_h)\right) - \mathbb{E}_P^{\pi^\theta}\left[\sum_{h=1}^{H} \phi(s_h, a_h) \middle| s_h = s_1\right] \\ &= \phi(\tau) - \mathbb{E}_P^{\pi^\theta}\left[\phi(\tau)\right] \\ &= \phi(\tau) - \phi_P^\theta, \end{aligned}$$

where the conditioning on the initial state s_1 is discarded since it is deterministic. This concludes the proof. □

Proposition C.2 (Hessian upper bound)**.** *Let $\theta \in \mathbb{R}^d$, P be deterministic dynamics, and Assumption 2.1 hold. Then, $\|\mathcal{H}_P(\theta)\|_2 \leq R^2/\eta$.*

To bound the spectral norm of the Hessian, we bound for any unit vector $v \in \mathbb{R}^d$ using Lemma C.1:

$$\begin{aligned} \|v\|^2_{\mathcal{H}_P(\theta)} &= \eta \mathbb{E}_P^{\pi_P^\theta}\left[\langle v, \phi(\tau) - \phi_P^\theta \rangle^2\right] \\ &\leq \eta \max_\tau \langle v, \phi(\tau) - \phi_P^\theta \rangle^2 \\ &\leq \eta \max_\tau \left(2\langle v, \phi(\tau)\rangle^2 + 2\langle v, \phi_P^\theta \rangle^2\right) \\ &\leq 4\eta B_\phi^2 H^2 = R^2/\eta. \end{aligned}$$

Proposition C.3 (Hessian image)**.** *Fix $\theta \in \mathbb{R}^d$ and let P be arbitrary deterministic dynamics. The image of the value Hessian at arbitrary parameters θ corresponds to the subspace of identifiable parameter directions under P:*

$$\Theta_P = \operatorname{im}\left(\mathcal{H}_P(\theta)\right).$$

Proof. Let $v \in \ker\left(\mathcal{H}_P(\theta)\right)$ for arbitrary dynamics P and parameters θ. We therefore have:

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}_P^\theta}\left[\langle v, \phi(\tau) - \phi_P^\theta \rangle^2\right] = 0 \\ &\iff \langle v, \phi(\tau) - \phi_P^\theta \rangle = 0 \ \text{ for all feasible trajectories } \tau \\ &\iff \langle v, \phi(\tau) \rangle = \text{constant for all feasible trajectories } \tau. \end{aligned}$$

In other words, if a reward parameter vector lies in the nullspace of $\mathcal{H}_P(\theta)$, every trajectory that is feasible following policy π_P^θ in dynamics P has no advantage, or in other word, produce the same rewards. Consequently, under such a reward vector, no trajectory generated from π_P^θ is preferred over another. This is can be explained by two cases:

1. The rewards are not reachable under the policy π_P^θ. However, since the policy π_P^θ is optimal with respect to an entropy regularized objective, it must visit every reachable state with some probability. Hence

$$v \in \left\{\theta \in \mathbb{R}^d : \phi(s,a)^\top \theta = 0 \quad \forall (s,a) \text{ s.t. } \nu_P^\theta(s) = 0\right\} = \Theta_P^{\text{unreachable}}$$

2. Or, the rewards are in the subspace of potential shaping transformation inducing parameters: $v \in \Theta_P^{\text{shaping}}$.

This naturally implies:

$$\Theta_P^{\text{unreachable}} \cup \Theta_P^{\text{shaping}} = \Theta_P^\perp = \ker(\mathcal{H}_P(\theta)).$$

□

Lemma C.4 (Pseudo Self-Concordance)**.** *Under Assumptions 2.1 and 2.2, for any $\theta, v, u \in \mathbb{R}^d$ and dynamics P, the optimal regularized value function satisfies*

$$\left|D^3 V_P^\star(\theta)[v,v,u]\right| \leq R \left\|u\right\|_{\Pi_P} D^2 V_P^\star(\theta)[v,v].$$

Proof. Fix $\theta, v, u \in \mathbb{R}^d$ and let P be arbitrary deterministic dynamics. By Lemma C.1 and using that for any vector $u \notin \Theta_P$, $\langle u, \phi(\tau)\rangle = 0$ (Lemma C.3).

$$\begin{aligned}
\left|\nabla_\theta^3 V_P^\star(\theta)[v,v,u]\right| &= \eta^2 \left|\mathbb{E}_P^{\pi_P^\theta}\left[\langle v, \phi(\tau) - \phi_P^\theta\rangle^2 \langle u, \phi(\tau) - \phi_P^\theta\rangle\right]\right| \\
&\leq \eta^2 \max_\tau \left|\langle u, \phi(\tau) - \phi_P^\theta\rangle\right| \mathbb{E}_P^{\pi_P^\theta}\left[\langle v, \phi(\tau) - \phi_P^\theta\rangle^2\right] \\
&= \left\|\Pi_P(u)\right\|_2 (2\eta H B_\phi) \nabla_\theta^2 V_P^\star(\theta)[v,v] \\
&= R \left\|u\right\|_{\Pi_P} \nabla_\theta^2 V_P^\star(\theta)[v,v].
\end{aligned}$$

□

Lemma C.5. *Let P be arbitrary deterministic dynamics and Assumption 2.1 hold. Fix $\theta_1, \theta_2 \in \mathbb{R}^d$, $u = \theta_2 - \theta_1$. Then,*

1.

$$\exp\left(-R\left\|u\right\|_{\Pi_P}\right) \mathcal{H}_P(\theta_2) \preceq \mathcal{H}_P(\theta_1) \preceq \exp\left(R\left\|u\right\|_{\Pi_P}\right) \mathcal{H}_P(\theta_2),$$

2. Let $\psi(x) = \frac{e^x - x - 1}{x^2}$ and $G_P(\theta_1, \theta_2) = \int_0^1 (1-t)\mathcal{H}_P(\theta_2 t + (1-t)\theta_1) dt$. Then, for $\theta \in \{\theta_1, \theta_2\}$,

$$\psi\left(-R\left\|u\right\|_{\Pi_P}\right) \mathcal{H}_P(\theta) \preceq G_P(\theta_1, \theta_2) \preceq \psi\left(R\left\|u\right\|_{\Pi_P}\right) \mathcal{H}_P(\theta).$$

Proof. Part 1. Fix $\theta_1, \theta_2 \in R^d$, $v \in \Theta_P$ and let P be arbitrary deterministic dynamics. Furthermore, define $u = \theta_2 - \theta_1$ and $\theta_t = \theta_2 t + (1-t)\theta_1$. The pseudo-self concordance inequality from Lemma C.4 can be rewritten with

$$\left|\frac{d \log D^2 V_P^\star(\theta_t)[v,v]}{dt}\right| \leq R\left\|u\right\|_{\Pi_P}. \tag{12}$$

Now, using $v \in \Theta_P$ ($D^2 V_P^\star(\theta_2)[v,v] > 0$), by the fundamental theorem of calculus,

$$\begin{aligned}
\left|\log D^2 V_P^\star(\theta_1)[v,v] - \log D^2 V_P^\star(\theta_2)[v,v]\right| &= \int_0^1 \left|\frac{d \log D^2 V_P^\star(\theta_t)[v,v]}{dt}\right| dt \\
&\leq R\left\|u\right\|_{\Pi_P}.
\end{aligned}$$

Exponentiating both sides, we obtain the first inequality:

$$e^{-R\|u\|_{\Pi_P}} \leq \frac{D^2V_P^\star(\theta_1)[v,v]}{D^2V_P^\star(\theta_2)[v,v]} \leq e^{R\|u\|_{\Pi_P}}.$$

Part 2. To obtain the second inequality, we first integrate Equation (12) over t, in order to obtain, for $\theta \in \{\theta_1, \theta_2\}$:

$$\exp\left(-tR\|u\|_{\Pi_P}\right) D^2V_P^\star(\theta)[v,v] \leq D^2V_P^\star(\theta_t)[v,v] \leq \exp\left(tR\|u\|_{\Pi_P}\right) D^2V_P^\star(\theta)[v,v]. \quad (13)$$

Multiplying by $(1-t)$ and integrating t over $[0,1]$ yields the second desired result.

$$\psi\left(-R\|u\|_{\Pi_P}\right) D^2V_P^\star(\theta)[v,v] \leq G_P(\theta_1,\theta_2) \leq \psi\left(R\|u\|_{\Pi_P}\right) D^2V_P^\star(\theta)[v,v],$$

Conclusion. Now, let $v \notin \Theta_P$ and notice the three desired inequalities still hold, since $D^2V_P^\star(\theta)[v,v] = 0$. This concludes the proof. □

Lemma C.6. *Let Assumption 2.1, and define $\chi(x) = \frac{e^x - x}{x}$. Then, for any $n > 0$,*

$$\chi\left(-R_n\rho_n\right)\rho_n^2 \leq \langle \theta_n - \theta^E, \nabla L_n^{pop}(\theta_n) - \nabla L_n^{pop}(\theta^E)\rangle \leq \chi\left(R_n\rho_n\right)\rho_n^2.$$

Proof. First, by the definition of ρ_n and R_n, setting $u = \theta_n - \theta^{\mathrm{E}}$, we have

$$R\|u\|_{\sum_{k=1}^n \Pi_{P_k}} \leq R\frac{\|u\|_{\mathcal{H}_n^{\mathrm{E}}}}{\sqrt{\lambda_n}} = R_n\rho_n.$$

Following a similar logic to how we derived (13), using $\theta_1 = \theta^{\mathrm{E}}$ and $\theta_n = \theta_2$, we get for any $v \in \mathbb{R}^d$

$$\exp\left(-tR_n\rho_n\right)\|v\|_{\mathcal{H}_n^{\mathrm{E}}}^2 \leq D^2L_n^{\mathrm{pop}}(\theta_t)[v,v] \leq \exp\left(tR_n\rho_n\right)\|v\|_{\mathcal{H}_n^{\mathrm{E}}}^2. \quad (14)$$

By the fundamental theorem of calculus,

$$\langle u, \nabla L_n^{\mathrm{pop}}(\theta^{\mathrm{E}}) - \nabla L_n^{\mathrm{pop}}(\theta_n)\rangle = \int_0^1 \|u\|_{\sum_{k=1}^n \mathcal{H}_{P_k(\theta_t)}} \mathrm{d}t.$$

Sandwiching the above using (14), choosing $v = u$, and using that $\int_0^1 e^{xt}dt = \chi(x)$ gives the result. □

Lemma 3.1. *For any dynamics $P \in \mathcal{P}$, the Hessian $\mathcal{H}_P(\theta^E)$ admits a strictly positive spectral gap: its smallest non-zero eigenvalue is lower bounded by some constant $\kappa > 0$.*

Proof. Let us fix arbitrary dynamics $P \in \mathcal{P}$ and denote $\mathbb{P}_P^{\mathrm{E}}$ the trajectory law of the expert. A feasible trajectory τ under P verifies:

$$T(s' \mid s,a) > 0, \forall (s,a,s') \in \tau,$$

We first consider the case where there exists at least two trajectories τ and τ' feasible under P, such that $\phi(\tau) \neq \phi(\tau')$. In our problem setting, we have $-HB_\phi B_\theta \leq Q_P^\theta(s,a) \leq HB_\phi B_\theta + \eta^{-1}\log|A|$ for all (s,a) (See (10)). Since dynamics in $\mathcal{P}$ are deterministic under Assumption 2.2, we can lower bound expert trajectory probabilities for any feasible trajectory τ, with

$$\begin{aligned}
\mathbb{P}_P^{\mathrm{E}}(\tau) &= \nu_1(s_1)\prod_{h=1}^H T(s_{h+1} \mid s_h, a_h)\pi^{\mathrm{E}}(a_h \mid s_h) \\
&\geq \min_{s,a}\left(\pi^{\mathrm{E}}(a \mid s)\right)^H \\
&\geq \left(\frac{\exp\left(-\eta HB_\phi B_\theta\right)}{\sum_{a\in\mathcal{A}}\exp\left(\eta HB_\phi B_\theta + \log|A|\right)}\right)^H \\
&= \frac{\exp\left(-RHB_\theta\right)}{|\mathcal{A}|^{H+1}}.
\end{aligned}$$

Secondly, let v be a unit vector lying in $\mathrm{im}(\mathcal{H}_P(\theta^{\mathrm{E}}))$. Then, using the covariance form of Hessians (Lemma C.1), we have

$$\|v\|^2_{\mathcal{H}_P(\theta^{\mathrm{E}})} = \frac{\eta}{2}\sum_{\tau,\tau'} \mathbb{P}^{\mathrm{E}}_P(\tau)\mathbb{P}^{\mathrm{E}}_P(\tau')\langle v, \phi(\tau) - \phi(\tau')\rangle^2 > 0.$$

This implies that there must exist at least some pair τ, τ' such that $\langle v, \phi(\tau) - \phi(\tau')\rangle^2 > 0$. This is valid since we previously assumed that there are at least two feasible trajectories $\tau \neq \tau'$ under dynamics P such that $\phi(\tau) \neq \phi(\tau')$. Define the minimal identifiable feature gap:

$$\begin{aligned}\Delta^2 = &\min_{\substack{v\in \mathrm{im}(\mathcal{H}_P(\theta^{\mathrm{E}})), \|v\|_2=1 \\ \tau,\tau' \in \mathrm{supp}(\mathbb{P}^{\mathrm{E}}_{P_k})}} \langle v, \phi(\tau) - \phi(\tau')\rangle^2, \\ &\text{s.t.} \quad \langle v, \phi(\tau) - \phi(\tau')\rangle^2 > 0.\end{aligned}$$

Then, using the definition of Δ and the lower bound on trajectory probabilities, we finally obtain for any unit vector $v \in \mathrm{im}(\mathcal{H}_P(\theta^{\mathrm{E}}))$:

$$\begin{aligned}\|v\|^2_{\mathcal{H}_P(\theta^{\mathrm{E}})} &= \frac{\eta}{2}\sum_{\tau,\tau'} \mathbb{P}^{\mathrm{E}}_P(\tau)\mathbb{P}^{\mathrm{E}}_P(\tau')\langle v, \phi(\tau) - \phi(\tau')\rangle^2 \\ &\geq \frac{\eta}{2}\frac{\exp\left(-RHB_\theta\right)}{|\mathcal{A}|^{H+1}}\Delta^2 =: \kappa.\end{aligned}$$

By Lemma C.3, we then can write that

$$\lambda^{\Theta_P}_{\min}(\mathcal{H}_p(\theta^{\mathrm{E}})) \coloneqq \min_{v\in\Theta_P, \|v\|_2=1} \|v\|^2_{\mathcal{H}_P(\theta^{\mathrm{E}})} \geq \kappa.$$

and $\lambda^{\Theta_P^\perp}_{\min}(\mathcal{H}_p(\theta^{\mathrm{E}})) = 0$. This proves that $\mathcal{H}_P(\theta^{\mathrm{E}})$ has a spectral gap of at least κ.

Now suppose that all feasible trajectories under P share the same feature vector. Then, for any $\theta \in \mathbb{R}^d$, the trajectory reward $\langle \theta, \phi(\tau)\rangle$ is identical across all trajectories τ. As a result, every reward parameter induces the same optimal policy. Consequently, there are no identifiable directions under P, i.e., $\Theta_P = \{0\}$, and therefore $\mathcal{H}_P(\theta^{\mathrm{E}}) = 0$, with no non-zero eigenvalues. This concludes the proof. □

Lemma C.7 (Hessian growth)**.** *It holds for all $n > 0$ that*

$$\lambda_n \geq \frac{\kappa}{g_n} - \lambda_0.$$

Proof. Let v_n be the the eigenvector of λ_n:

$$\lambda_n \coloneqq \lambda^{\Theta_{\mathcal{P}}}_{\min}(\mathcal{H}^{\mathrm{E}}_n) = \|v_n\|^2_{\mathcal{H}^{\mathrm{E}}_n}.$$

Specifically, this implies that $\left(\mathcal{H}^{\mathrm{E}}_n + \lambda_0 I_d\right)^{-1} v_n = \frac{1}{\lambda_0+\lambda_n} v_n$.

Furthermore, since $v_n \in \Theta_{\mathcal{P}}$, by definition there must exist $P^\star \in \mathcal{P}$ such that we also have $v_n \in \Theta_{P^\star}$. By Lemma 3.1, we know

$$\|v_n\|^2_{\mathcal{H}_{P^\star}(\theta^{\mathrm{E}})} \geq \lambda^{\Theta_{P^\star}}_{\min}\mathcal{H}_{P^\star}(\theta^{\mathrm{E}}) \geq \kappa.$$

We then have

$$\begin{aligned}\left\|\mathcal{T}^{\mathrm{E}}_n(P^\star)\right\|_2 &\geq \|v_n\|^2_{\mathcal{T}^{\mathrm{E}}_n(P^\star)} \\ &= v_n^\top \mathcal{H}_{P^\star}(\theta^{\mathrm{E}})\left(\mathcal{H}^{\mathrm{E}}_n + \lambda_0 I_d\right)^{-1} v_n \\ &= \frac{\|v_n\|^2_{\mathcal{H}_{P^\star}(\theta^{\mathrm{E}})}}{\lambda_0 + \lambda_n} \\ &\geq \frac{\kappa}{\lambda_0 + \lambda_n}.\end{aligned}$$

Now using the definition of g_n, we recover:

$$\begin{aligned} g_n &= \max_{P \in \mathcal{P}} \left\| \mathcal{T}_n^{\mathrm{E}}(P) \right\|_2 \\ &\geq \left\| \mathcal{T}_n^{\mathrm{E}}(P^\star) \right\|_2 \\ &\geq \frac{\kappa}{\lambda_0 + \lambda_n}. \end{aligned}$$

Rearranging, we retrieve the desired result.

□

C.3 Ellipsoidal confidence sets

Remark C.8. We note that for any $u \in \mathbb{R}^d$,

$$\langle S_n, u \rangle = \langle ((\mathcal{H}_n^{\mathrm{E}})^\dagger)^{1/2} S_n, (\mathcal{H}_n^{\mathrm{E}})^{1/2} u \rangle.$$

Indeed, let $v \in \ker(\mathcal{H}_n^{\mathrm{E}})$. Since

$$\mathcal{H}_n^{\mathrm{E}} = \sum_{k=1}^{n} \mathcal{H}_{P_k}(\theta^{\mathrm{E}}),$$

and each $\mathcal{H}_{P_k}(\tau^{\mathrm{E}})$ is positive semidefinite, it follows that $v \in \ker(\mathcal{H}_{P_k}(\theta^{\mathrm{E}}))$ for every k. By Lemma C.1,

$$\mathbb{E}_{P_k}^{\pi_{P_k}^{\mathrm{E}}} \left[\langle v, \phi(\tau) - \phi_{P_k}^{\mathrm{E}} \rangle^2 \right] = 0.$$

Since the integrand is nonnegative, we obtain

$$\langle v, \phi(\tau) - \phi_{P_k}^{\mathrm{E}} \rangle = 0 \qquad \text{a.s.}$$

for every k. Therefore, by definition of S_n,

$$\langle v, S_n \rangle = 0,$$

showing that $S_n \in \mathrm{im}(\mathcal{H}_n^{\mathrm{E}})$. The identity then follows from the Moore–Penrose pseudoinverse properties.

To proceed, we use the first order optimality of θ_n for L_n over Θ and of θ^E for L_n^{pop} to provide a deterministic localization bound.

Lemma C.9 (Deterministic localization bound)**.** *For all $n > 0$, it holds that*

$$\chi(-R_n \, \rho_n) \, \rho_n \;\leq\; \omega_n. \tag{15}$$

Additionally, on the localization event

$$\mathcal{E} := \left\{ \omega_n \leq \frac{1 - e^{-1}}{R_n} \right\}, \tag{16}$$

we have $R_n \rho_n \leq 1$ and

$$\rho_n \leq \frac{1}{1 - e^{-1}} \omega_n. \tag{17}$$

Proof. Since θ_n minimizes L_n over Θ, the first-order optimality condition gives

$$\langle \nabla L_n(\theta_n), \theta - \theta_n \rangle \geq 0 \quad \forall \, \theta \in \Theta.$$

Since θ^E minimizes L_n^{pop} (each L_k is minimized at θ^E), we have $\nabla L_n^{\mathrm{pop}}(\theta^E) = 0$. Using the decomposition $\nabla L_n(\theta_n) = \nabla L_n^{\mathrm{pop}}(\theta_n) - S_n$, we have:

$$\langle \nabla L_n^{\mathrm{pop}}(\theta_n) - S_n, \; \theta^E - \theta_n \rangle \geq 0.$$

Rearranging,

$$\langle \nabla L_n^{\text{pop}}(\theta_n) - \nabla L_n^{\text{pop}}(\theta^E),\ \theta_n - \theta^E \rangle \leq \langle S_n,\ \theta_n - \theta^E \rangle. \tag{18}$$

Next, let us apply Cauchy Schwarz in the H_n^E-norm on the right side of (18). This gives

$$\langle S_n, \theta_n - \theta^E \rangle = \langle ((\mathcal{H}_n^{\text{E}})^{\dagger})^{1/2} S_n, (\mathcal{H}_n^{\text{E}})^{1/2}(\theta_n - \theta^E) \rangle \leq \omega_n\, \rho_n. \tag{19}$$

Combining the self-concordance property of L_{pop} (Lemma C.6) with (18) and (19):

$$\chi(-R_n \rho_n) \leq \omega_n\, \rho_n.$$

Dividing by ρ_n (if $\rho_n > 0$; the bound is trivial otherwise) yields (15).

Last, note that on the event $\mathcal{E}$, we have

$$\chi(-R_n\, \rho_n)\, \rho_n \leq \omega_n \leq \frac{1 - e^{-1}}{R_n} = \frac{\chi(-1)}{R_n}.$$

Since $x \mapsto \chi(-x)\, x$ is increasing on $[0, \infty)$ and equals $\chi(-1)/R_n$ at $x = 1/R_n$, we conclude that $\rho_n \leq 1/R_n$, i.e., $R_n \rho_n \leq 1$. Since χ is decreasing on $(-\infty, 0)$ with $\chi(-1) = 1 - e^{-1}$, we get $\chi(-R_n \rho_n) \geq 1 - e^{-1}$, and therefore

$$(1 - e^{-1})\, \rho_n \leq \chi(-R_n \rho_n)\, \rho_n \leq \omega_n,$$

which gives (17). □

In the next step, we are bounding the excess risk on the event $\mathcal{E}$.

Lemma C.10 (Excess risk bound on $\mathcal{E}$)**.** *On the event $\mathcal{E}$ defined in* (16)*,*

$$L_n^{pop}(\theta_n) - L_n^{pop}(\theta^E) \leq \frac{1}{1 - e^{-1}} \omega_n^2. \tag{20}$$

Proof. By the optimality of θ_n for L_n over Θ and $\theta^E \in \Theta$, we have $L_n(\theta_n) \leq L_n(\theta^E)$. Using $L_n(\theta) = L_n^{\text{pop}}(\theta) - \langle \theta, S_n \rangle$:

$$\begin{aligned} L_n^{\text{pop}}(\theta_n) - L_n^{\text{pop}}(\theta^E) &= \langle \theta_n - \theta^E, S_n \rangle + \big(L_n(\theta_n) - L_n(\theta^E)\big) \\ &\leq \langle \theta_n - \theta^E, S_n \rangle. \end{aligned}$$

On $\mathcal{E}$, applying Cauchy–Schwarz and then Lemma C.9 gives

$$\langle \theta_n - \theta^E, S_n \rangle \leq \omega_n\, \rho_n \leq \frac{1}{1 - e^{-1}}\, \omega_n^2.$$

□

In a last step, we need to derive the concentration of ω_n. We can conclude this by noting that S_n forms a martingale difference sequence.

Lemma C.11 (Concentration of ω_n via matrix Freedman)**.** *Let $\delta \in (0, 1)$. Under Assumptions 2.1 and 2.2, the noise term $S_n = \sum_{k=1}^n \sum_{j=1}^m X_{k,j}$ with $X_{k,j} := \frac{1}{m}(\phi(\tau_{k,j}) - \phi_{P_k}^E)$ forms a matrix martingale difference sequence with respect to $\mathcal{G}_{k,j} := \sigma(P_1, \tau_{1,1}, \ldots, \tau_{1,m}, \ldots; P_k, \tau_{k,1}, \ldots, \tau_{k,j})$. Consequently, with probability at least $1 - \delta$,*

$$\omega_n \leq \sqrt{\frac{2d \log(d/\delta)}{\eta m}} + \frac{4B_\phi H\ \log(d/\delta)}{3m\sqrt{\lambda_n}}. \tag{21}$$

Proof. First, we note that S_n is indeed a martingale difference sequence. For this remark that P_k is $\mathcal{F}_{k-1}$-measurable and at iteration k the m trajectories are chosen i.i.d. given P_k. Therefore, each $X_{k,j}$ is $\mathcal{G}_{k,j}$- measurable with

$$\mathbb{E}[X_{k,j} \mid \mathcal{G}_{k,j-1}] = \frac{1}{m}\left(\mathbb{E}[\phi(\tau_{k,j}) \mid P_k] - \phi^{\mathrm{E}}_{P_k}\right) = 0.$$

Additionally, note that S_n is almost surely bounded as $\|\phi(\tau)\|_2 \leq B_\phi H$. Additionally, since the m trajectories are i.i.d. given P_k, it also holds that (Lemma C.1)

$$\mathbb{E}[X_{k,j}X_{k,j}^\top \mid \mathcal{G}_{k,j-1}] = \frac{1}{m^2}\mathrm{Cov}^{\pi^{\mathrm{E}}_{P_k}}_{P_k}[\phi(\tau)] = \frac{1}{m^2\eta}\mathcal{H}_{P_k}(\theta^{\mathrm{E}}).$$

Now, summing over $j = 1, \ldots, m$ and $k = 1, \ldots, n$, we get that

$$V_n = \sum_{k=1}^{n} \frac{m}{\eta m^2}\mathcal{H}_{P_k}(\theta^{\mathrm{E}}) = \frac{1}{\eta m}\mathcal{H}^{\mathrm{E}}_n.$$

We can conclude that S_n has 0 conditional mean, is almost surely bounded, and has predictable quadratic variation. Therefore, to conclude the proof, we can apply the matrix Freedman inequality (Tropp, 2011). For this, we first compute the normalized variance and truncation parameters:

$$\sigma_n^2 := \left\|\left((H_n^E)^\dagger\right)^{1/2} V_n \left((H_n^E)^\dagger\right)^{1/2}\right\|_2 = \frac{1}{\eta m}, \qquad b_n := \frac{2B_\phi H}{m\sqrt{\lambda_n}}.$$

Next, applying the matrix Freedman inequality, we get with probability at least $1-\delta$ that

$$\begin{aligned}\omega_n = \|S_n\|_{(H_n^E)^\dagger} &\leq \sqrt{2\sigma_n^2\, d\log(d/\delta)} + \frac{2}{3}\, b_n \log(d/\delta) \\ &= \sqrt{\frac{2d\log(d/\delta)}{\eta m}} + \frac{4B_\phi H\, \log(d/\delta)}{3m\sqrt{\lambda_n}}.\end{aligned}$$

□

Additionally note that we can replace the dimension d with the effective dimension d_{eff}.

Remark C.12 (Effective dimension). The dimension factor d in (21) can be replaced by the effective dimension

$$d_{\mathrm{eff}} := \dim(\Theta_{\mathcal{P}}) = \dim\Big(\mathrm{span}\bigcup_{P\in\mathcal{P}} \mathrm{im}(\mathcal{H}_P(\theta^{\mathrm{E}}))\Big) \leq d.$$

Indeed, $S_n \in \mathrm{im}(H_n^E) = \Theta_{\mathcal{P}}$ almost surely, and the norm $\|\cdot\|_{(H_n^E)^\dagger}$ acts on this same subspace via the pseudoinverse. The matrix Freedman inequality applied to the projected sum $\Pi_{\mathcal{P}} S_n$ then yields

$$\omega_n \leq \sqrt{\frac{2d_{\mathrm{eff}}\log(d_{\mathrm{eff}}/\delta)}{\eta m}} + \frac{4B_\phi H\log(d_{\mathrm{eff}}/\delta)}{3m\sqrt{\lambda_n}},$$

since the union bound over eigendirections only needs to range over $\mathrm{im}(H_n^E)$, whose dimension is at most d_{eff}. The same substitution applies for the proof of Lemma 3.2.

We can now conclude with our main confidence result by combining the previously stated lemmas.

Lemma 3.2. *Let $\delta \in (0,1)$ and the warm-up length be*

$$n_\delta := \min\left\{k : g_k \;\leq\; \frac{(1-e^{-1})^2\kappa m}{32 d\eta H^2 B_\phi^2 \log(d/\delta)}\right\}. \tag{5}$$

Then, with probability at least $1-\delta$, it holds for all $n \geq n_\delta$ that

$$\|\theta_n - \theta^{\mathrm{E}}\|^2_{\mathcal{H}_n^E} \;\lesssim\; \frac{d\,\log(d/\delta)}{\eta m}\,.$$

Proof. First, note that by Lemma C.11, with probability at least $1-\delta$,

$$\omega_n \leq \sqrt{\frac{2d\log(d/\delta)}{\eta m}} + \frac{4B_\phi H\ \log(d/\delta)}{3m\sqrt{\lambda_n}}.$$

The localization event $\mathcal{E}$ (Definition (16)) requires $\omega_n \leq (1-e^{-1})/R_n$.

Therefore, we need that both terms in the concentration bound to be at most $(1-e^{-1})/(2R_n)$. Indeed it holds

$$\sqrt{\frac{2d\log(d/\delta)}{\eta m}} \leq \frac{(1-e^{-1})\sqrt{\lambda_n}}{4\eta H B_\phi},$$
$$\frac{4B_\phi H\ \log(d/\delta)}{3m\sqrt{\lambda_n}} \leq \frac{(1-e^{-1})\sqrt{\lambda_n}}{4\eta H B_\phi}.$$

Squaring the first and rearranging the second leads to

$$\lambda_n \geq \max\left\{\frac{32d\eta H^2 B_\phi^2 \log(d/\delta)}{m(1-e^{-1})^2},\ \frac{16\eta H^2 B_\phi^2 \log(d/\delta)}{3m(1-e^{-1})}\right\}. \tag{22}$$

Since $d \geq 1$, the first term always dominates. By Lemma C.7, we recover the warm-up condition (5).

Next, on the event $\mathcal{E}$, Lemma C.9 gives $\rho_n \leq \frac{1}{1-e^{-1}}\omega_n$. Therefore, it holds that

$$\rho_n^2 \leq \frac{1}{(1-e^{-1})^2}\omega_n^2.$$

Now, using Lemma C.11, we get

$$\omega_n^2 \leq \frac{4d\log(d/\delta)}{\eta m} + \frac{32B_\phi^2 H^2 \log^2(d/\delta)}{9m^2\lambda_n} \leq \frac{4d\log(d/\delta)}{\eta m} + \frac{\log(d/\delta)(1-e^{-1})^2}{9\eta m}.$$

where we additionally used the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and employed the previous lower bound on λ_n. Therefore, we conclude with

$$\|\theta_n - \theta^E\|^2_{\mathcal{H}_n^{\mathrm{E}}} \lesssim \frac{d\log(d/\delta)}{\eta m}.$$

□

C.4 Discussion of Assumption 2.2

Restricting to deterministic dynamics is a genuine limitation of our analysis, and we discuss here where it enters and what would be needed to relax it. The assumption plays two distinct roles.

Simpler Hessians. As shown in Schlaginhaufen & Kamgarpour (2026) and used in the proof of Lemma C.1, under deterministic dynamics the Hessian of the optimal value reduces to a trajectory-feature covariance,

$$\mathcal{H}_P(\theta) = \mathbb{E}_P^{\pi_\theta}\left[\left(\phi(\tau) - \phi_P^\theta\right)\left(\phi(\tau) - \phi_P^\theta\right)^\top\right],$$

rather than the advantage-based form

$$\mathcal{H}_P(\theta) = \mathbb{E}_P^{\pi_\theta}\left[\left(A_P^\theta(\tau) - \mathbb{E}_P^{\pi_\theta} A_P^\theta(\tau)\right)\left(A_P^\theta(\tau) - \mathbb{E}_P^{\pi_\theta} A_P^\theta(\tau)\right)^\top\right]$$

required in the stochastic case. This simplifies the active selection rule (8), since the learner only has to estimate the covariance of trajectory features under π_θ rather than of the advantage $A_P^\theta(\tau)$.

Confidence-set construction. The proof of Lemma 3.2 additionally relies on two properties that hold under Assumption 2.2 but may fail otherwise: *(i)* the feature deviations lie in the right subspace, $\phi(\tau) - \phi_P^{\mathrm{E}} \in \operatorname{im} \mathcal{H}_P(\theta^{\mathrm{E}})$ (Remark C.8), and *(ii)* the effective dimension satisfies $d_{\text{eff}} \leq d$, which can fail in the stochastic case (Schlaginhaufen & Kamgarpour, 2026, Corollary A.17). While *(i)* can in principle be resolved by projecting onto $\operatorname{im} \mathcal{H}_P(\theta^{\mathrm{E}})$, an increased effective dimension seems unavoidable for the empirical IRL loss (2) in the stochastic case. We leave the details of this extension to the stochastic setting to future work.

D Proofs of Section 3.2

We now lay out the results employed to prove Lemmas 3.3, 3.4 and our Theorem 3.6.

D.1 Supporting lemmas

We first present useful identities of the suboptimality (1).

Proposition D.1 (Reminder form)**.** *Let Assumption 2.1 hold, P be arbitrary dynamics and $\theta_1, \theta_2 \in \Theta$ be any pair of parameters. Then,*

$$V^\star_{\theta_2,P} - V^{\pi^{\theta_1}_P}_{\theta_2,P} = \|\theta_1 - \theta_2\|^2_{G_P(\theta_1,\theta_2)} \,.$$

Proof. Let P be any arbitrary dynamics and $\theta_1, \theta_2 \in \Theta$ any pair of parameters. Using $u = \theta_2 - \theta_1$, we begin first by writing a Taylor expansion of the expert's optimal value function, together with Lemma C.1:

$$\begin{aligned} V^\star_{\theta_2,P} &= V^\star_{\theta_1,P} + \nabla V^\star_{\theta_1,P}[u] + G_P(\theta_1,\theta_2)[u,u] \\ &= V^\star_{\theta_1,P} + \langle u, \phi^{\theta_1}_P \rangle + \|u\|^2_{G_P(\theta_1,\theta_2)} \,. \end{aligned}$$

Expressing now the value gap provides the desired result:

$$\begin{aligned} V^\star_{\theta_2,P} - V^{\pi^{\theta_1}_P}_{\theta_2,P} &= \left(V^\star_{\theta_1,P} + \langle u, \phi^{\theta_1}_P \rangle + \|u\|^2_{G_P(\theta_1,\theta_2)} \right) - \left(V^\star_{\theta_1,P} + \langle u, \phi^{\theta_1}_P \rangle \right) \\ &= \|u\|^2_{G_P(\theta_1,\theta_2)} \,. \end{aligned}$$

□

Proposition D.2 (Pseudo self-concordant suboptimality bound)**.** *Let Assumption 2.1 hold, P be arbitrary dynamics and $\theta_1, \theta_2 \in \Theta$ be any pair of parameters. Then, for any symmetric positive definite matrix X of dimension $d \times d$ and $\theta \in \{\theta_1, \theta_2\}$,*

$$V^\star_{\theta_2,P} - V^{\pi^{\theta_1}_P}_{\theta_2,P} \leq \|\theta_1 - \theta_2\|^2_X \left\| \mathcal{H}_P(\theta) X^{-1} \right\|_2 \psi\left(R \|\theta_1 - \theta_2\|_{\Pi_P} \right) .$$

Proof. Let P be arbitrary dynamics and fix $\theta_1, \theta_2 \in \Theta$ any pair of parameters. Define $u = \theta_2 - \theta_1$. Then, employing Lemma D.1 together with Lemma C.5 and denoting

$$\psi_u = \psi\left(R \|u\|_{\Pi_P} \right) ,$$

we have for $\theta \in \{\theta_1, \theta_2\}$ and for any symmetric positive definite matrix X of dimension $d \times d$:

$$\begin{aligned} V^\star_{\theta_2,P} - V^{\pi^{\theta_1}_P}_{\theta_2,P} &= \|\theta_1 - \theta_2\|^2_{G_P(\theta_1,\theta_2)} \\ &\leq \psi_u \|\theta_1 - \theta_2\|^2_{\mathcal{H}_P(\theta)} \\ &= \psi_u (\theta_1 - \theta_2)^\top \mathcal{H}_P(\theta)(\theta_1 - \theta_2) \\ &= \psi_u (\theta_1 - \theta_2)^\top X^{1/2} \left(X^{-1/2} \mathcal{H}_P(\theta) X^{-1/2} \right) X^{1/2} (\theta_1 - \theta_2) \\ &\leq \psi_u \|\theta_1 - \theta_2\|^2_X \left\| \mathcal{H}_P(\theta) X^{-1} \right\|_2 . \end{aligned}$$

□

We now show that once the warm-up phase has terminated, i.e, Condition (5), pseudo self-concordance factors (Lemma C.5) are upper bounded by their localized constants.

Lemma D.3 (Localized self-concordance constants)**.** *Let Assumptions 2.1 and 2.2 hold. Furthermore, let us assume fix $n \geq n_\delta$ where n_δ verifies the warm-up condition (5). Then, for any $P \in \mathcal{P}$,*

$$\psi\left(R\left\|\theta^E - \theta_n\right\|_{\Pi_P}\right) \leq e-2, \quad \exp\left(R\left\|\theta^E - \theta_n\right\|_{\Pi_P}\right) \leq e.$$

Proof. Fix $P \in \mathcal{P}$ and $n \geq n_\delta$. We know by the definition of n_δ that $\lambda_n > 0$ (See Equation (22)). We therefore always have:

$$\begin{aligned} R\left\|\theta^{\mathrm{E}} - \theta_n\right\|_{\Pi_P} &\leq R\left\|\theta^{\mathrm{E}} - \theta_n\right\|_{\Pi_{\mathcal{P}}} \\ &\leq \rho_n R_n. \end{aligned}$$

The last inequality holds since $\lambda_n > 0$ implies that $\mathrm{im}(\mathcal{H}_n^{\mathrm{E}}) = \Theta_{\mathcal{P}}$ (Lemma C.3). Using the fact that we are on the localization event $\mathcal{E}$ (16) for any $n \geq n_\delta$, together with Lemma C.10 yields:

$$\rho_n \leq \frac{1}{1-e^{-1}}\omega_n \leq \frac{1}{R_n}.$$

This ultimately implies that,

$$\psi\left(R\left\|\theta^{\mathrm{E}} - \theta_n\right\|_{\Pi_P}\right) \leq \psi(1) = e-2,$$

and

$$\exp\left(R\left\|\theta^{\mathrm{E}} - \theta_n\right\|_{\Pi_P}\right) \leq \exp(1) = e.$$

□

D.2 Transfer guarantees

Having stated the required lemmas, we subsequently prove our main transfer results. We first present the general transfer guarantee.

Lemma 3.3. *Fix $\delta \in (0,1)$ and $n \geq n_\delta$. For any $P \in \mathcal{P}$, it holds with probability at least $1-\delta$ that*

$$\mathrm{SubOpt}(\theta_n, P) \;\lesssim\; g_n\left(\frac{d\,\log(d/\delta)}{\eta m} + \lambda_0 B_\theta\right). \tag{6}$$

Proof. The proof starts from Lemma D.2, with $X = \mathcal{H}_n^{\mathrm{E}}$. For each $P \in \mathcal{P}$, we write

$$\begin{aligned} \mathrm{SubOpt}(\theta_n, P) &\leq \psi\left(R\left\|\theta^{\mathrm{E}} - \theta_n\right\|_{\Pi_P}\right)\left\|\theta^{\mathrm{E}} - \theta_n\right\|_{\mathcal{H}_n^{\mathrm{E}}+\lambda_0 I_d}\left\|\mathcal{T}_n^{\mathrm{E}}(P)\right\|_2 \\ &\leq \psi\left(R\left\|\theta^{\mathrm{E}} - \theta_n\right\|_{\Pi_P}\right)\left\|\theta^{\mathrm{E}} - \theta_n\right\|_{\mathcal{H}_n^{\mathrm{E}}+\lambda_0 I_d}\; g_n. \end{aligned}$$

To apply the confidence results of Theorem 3.2, we fix $n \geq n_\delta$ so that g_n verifies the warm-up condition (5). Hence, via Lemma D.3 and Lemma 3.2, it holds with probability at least $1-\delta$ ($\delta \in (0,1)$) that

$$\begin{aligned} \mathrm{SubOpt}(\theta_n, P) &\leq (e-2)\left(\left\|\theta^{\mathrm{E}} - \theta_n\right\|_{\mathcal{H}_n^{\mathrm{E}}} + 4B_\theta\lambda_0\right) g_n \\ &\lesssim \left(\frac{d\log(d/\delta)}{\eta m} + B_\theta\lambda_0\right) g_n. \end{aligned}$$

□

We then show how the passive rule (7) achieves finite warm-up time.

Lemma 3.4. *Let us follow the passive rule (7) until Condition (5) is verified. Then, it holds that*

$$n_\delta = |\mathcal{P}| \left\lceil \frac{32 d\eta H^2 B_\phi^2 \log(d/\delta)}{(1-e^{-1})^2 \kappa m} \right\rceil .$$

Proof. Let us follow the 7 rule every round $n > 0$. Then, we can decompose and lower bound the expert accumulated Hessian at round k as follows, for any $P \in \mathcal{P}$:

$$\begin{aligned}
\mathcal{H}_k^{\mathrm{E}} &= \sum_{P' \in \mathcal{P}} k(P') \mathcal{H}_{P'}(\theta^{\mathrm{E}}) \\
&\succeq \left\lfloor \frac{k}{|\mathcal{P}|} \right\rfloor \sum_{P' \in \mathcal{P}} \mathcal{H}_{P'}(\theta^{\mathrm{E}}) \\
&\succeq \left\lfloor \frac{k}{|\mathcal{P}|} \right\rfloor \mathcal{H}_P(\theta^{\mathrm{E}}).
\end{aligned}$$

Hence, we can write for any $n \geq |\mathcal{P}|$

$$\begin{aligned}
\mathcal{T}_n^{\mathrm{E}}(P) &\preceq \mathcal{H}_P(\theta^{\mathrm{E}}) \left(\lambda_0 I_d + \left\lfloor \frac{n}{|\mathcal{P}|} \right\rfloor \mathcal{H}_P(\theta^{\mathrm{E}}) \right)^{-1} \\
&\preceq \left\lfloor \frac{n}{|\mathcal{P}|} \right\rfloor^{-1} I_d.
\end{aligned}$$

We simply plug in the definition of g_n to obtain the bound $g_n \leq \left\lfloor \frac{n}{|\mathcal{P}|} \right\rfloor^{-1}$. To get the required warm-up time given some $\delta \in (0,1)$, we combine the above's bound with the warm-up condition (5) to get:

$$n_\delta = \min \left\{ n : \left\lfloor \frac{n}{|\mathcal{P}|} \right\rfloor \geq \frac{32 d R^2 \log(d/\delta)}{(1-e^{-1})^2 \eta \kappa m} \right\} .$$

This thus provides the following warm-up time:

$$n_\delta = |\mathcal{P}| \left\lceil \frac{32 d R^2 \log(d/\delta)}{(1-e^{-1})^2 \eta \kappa m} \right\rceil ,$$

where the condition $n_\delta \geq |\mathcal{P}|$ is always satisfied. □

Theorem 3.6. *Follow the active rule (8) for all $k \in \{n_\delta, \dots, n\}$. It holds for every $P \in \mathcal{P}$ that*

$$\mathrm{SubOpt}(\theta_n, P) \lesssim \frac{d}{n - n_\delta + 1} \log \left(\frac{nm}{\log(d/\delta)} \right) \left(\frac{d \log(d/\delta)}{\eta m} + \lambda_0 B_\phi \right) . \tag{9}$$

Proof. Let $w(x) = \frac{x}{\log(1+x)}$. We recall the definition of the online transfer operator:

$$\mathcal{T}_n(P) = \mathcal{H}_P(\theta_n) \left(\sum_{k=1}^{n} \mathcal{H}_P(\theta_k) + \lambda_0 I_d \right)^{-1} .$$

Employing the 8 rule with $J = \log\det(I_d + \cdot)$, together with the pseudo self-concordance property (Lemma C.5) to move from $\mathcal{T}_k$ to $\mathcal{T}_k^{\mathrm{E}}$ combined to Lemma D.3, we get for any $k > 0$ and $P \in \mathcal{P}$

$$\begin{aligned}
\|\mathcal{T}_k(P)\|_2 &\leq w\left(\|\mathcal{T}_k(P)\|_2\right) \log\det\left(I_d + \mathcal{T}_k(P)\right) \\
&\leq w\left(\|\mathcal{T}_k(P)\|_2\right) \log\det\left(I_d + \mathcal{T}_k(P_{k+1})\right) \\
&\leq e^2 w\left(\|\mathcal{T}_k(P)\|_2\right) \log\det\left(I_d + \mathcal{T}_k^{\mathrm{E}}(P_{k+1})\right) .
\end{aligned}$$

The above derivation can also equivalently be obtained using $J = \|\cdot\|_2$ or $J = \mathrm{tr}$. For $\delta \in (0,1)$ and n_δ such that it verifies the warm-up condition (5). Using the following determinant identity:

$$\det(\mathcal{H}^{\mathrm{E}}_{k+1}) = \det\left(\mathcal{H}^{\mathrm{E}}_k\right) \det\left(I_d + \mathcal{T}^{\mathrm{E}}_k(P_{k+1})\right)$$

and a telescoping sum from n_δ to n, we retrieve for any $P \in \mathcal{P}$:

$$\begin{aligned}
\|\mathcal{T}_n(P)\|_2 &= \frac{1}{n-n_\delta+1}\sum_{k=n_\delta}^{n} \|\mathcal{T}_n(P)\|_2 \\
&\leq \frac{1}{n-n_\delta+1}\sum_{k=n_\delta}^{n} \|\mathcal{T}_k(P)\|_2 && (\mathcal{T}_n \preceq \cdots \preceq \mathcal{T}_1) \\
&\leq w\left(\|\mathcal{T}_n(P)\|_2\right) \frac{e^2}{n-n_\delta+1}\sum_{k=n_\delta}^{n} \log\det\left(I_d + \mathcal{T}^{\mathrm{E}}_k(P_{k+1})\right) \\
&\leq w\left(\|\mathcal{T}_n(P)\|_2\right) \frac{e^2}{n-n_\delta+1}\sum_{k=n_\delta}^{n} \log\det \frac{\mathcal{H}^{\mathrm{E}}_{k+1}}{\mathcal{H}^{\mathrm{E}}_k} \\
&\leq w\left(\|\mathcal{T}_n(P)\|_2\right) \frac{e^2}{n-n_\delta+1} \log\det \frac{\mathcal{H}^{\mathrm{E}}_{n+1}}{\mathcal{H}^{\mathrm{E}}_{n_\delta}}.
\end{aligned}$$

Using $\left\|\mathcal{H}^{\mathrm{E}}_n\right\|_2 \leq nR^2/\eta + \lambda_0$ (Proposition C.2), and that the warm-up condition (5) is verified, we use the lower bound on λ_n (22), we get

$$\begin{aligned}
\log\det \frac{\mathcal{H}^{\mathrm{E}}_{n+1}}{\mathcal{H}^{\mathrm{E}}_{n_\delta}} &\leq d\left(\log\left((n+1)R^2/\eta\right) - \log\left[\frac{32R^2 d\log(d/\delta)}{\eta m(1-e^{-1})^2}\right]\right) \\
&= d\log\left(\frac{(n+1)\eta m(1-e^{-1})^2}{32 d\log(d/\delta)}\right)
\end{aligned}$$

Denoting $x_n = \frac{e^2}{n-n_\delta+1} d\log\left(\frac{(n+1)m(1-e^{-1})^2}{d\log(d/\delta)}\right)$, we obtain via the definition of w:

$$\begin{aligned}
\|\mathcal{T}_n(P)\|_2 &\lesssim w\left(\|\mathcal{T}_n(P)\|_2\right) x_n \\
&\leq e^{x_n} - 1 \\
&\lesssim \frac{e^2}{n-n_\delta+1} d\log\left(\frac{(n+1)m(1-e^{-1})^2}{d\log(d/\delta)}\right).
\end{aligned}$$

Applying one more time Lemma C.5 together with Lemma D.3, we conclude for all $P \in \mathcal{P}$:

$$\left\|\mathcal{T}^{\mathrm{E}}_n(P)\right\|_2 \lesssim \frac{e^4 d}{n-n_\delta+1} \log\left(\frac{(n+1)m(1-e^{-1})^2}{d\log(d/\delta)}\right).$$

By definition of g_n (4), and leveraging Lemma 6, we get the result.

□

Remark D.4 (Effective dimension). In the above proof, using $\left\|\mathcal{H}^{\mathrm{E}}_n\right\|_2 \leq nR^2 + \lambda_0$, and that for all $k > 0$, $\dim\left(\mathrm{im}\,\mathcal{H}_{P_k}(\theta_k)\right) = d_{\text{eff}} \leq d$, we can conclude that $\mathcal{H}^{\mathrm{E}}_n$ has at least $d - d_{\text{eff}}$ eigenvalues equal to λ_0, and at most d_{eff} eigenvalues bounded above by $nR^2 + \lambda_0$. The result of Theorem 3.6 still holds by replacing d by d_{eff}.

E Offline Active Learning

In some cases, sequential interaction with the expert may be impractical, for instance due to limited expert availability or high environment setup costs. We therefore also consider an offline variant,

where the entire sequence of dynamics is selected in advance. Setting $\theta_0 = 0$, we construct the offline information proxy $\mathcal{H}_n^0 := \mathcal{H}_0 + \sum_{k=1}^n \mathcal{H}_{P_k}(\theta_0)$, its offline transfer operator $\mathcal{T}_n^0$ accordingly, and select dynamics with

$$P_{k+1} = \max_{P \in \mathcal{P}} J\left(\mathcal{T}_n^0(P)\right), \quad \text{with} \quad J \in \{\|\cdot\|_2, \mathrm{tr}, \log\det(I_d + \cdot)\}. \tag{O}$$

This rule enables constructing the full sequence $(P_k)_{k=1}^n$ without any interaction with the expert.

Theorem E.1. *Let us follow* (O) *every round. Then, for any* $n > 0$ *and* $P \in \mathcal{P}$,

$$g_n \lesssim de^{2R} \cdot \frac{\log(nR)}{n}, \quad \textit{and} \quad n_\delta \eqsim \left\lceil \frac{d^2R^3\log(d/\delta)e^{2R}}{\eta\kappa m} \log\left[\frac{dR\log(d/\delta)}{\eta\kappa m}\right]\right\rceil.$$

Proof. To begin, notice the following identity between the spectral norm, the trace and log determinant for symmetric, PSD matrices $\|X\|_2 \le B$ with $B > 0$:

$$\|X\|_2 \le \mathrm{tr}(X) \le w(B)\log\det\left(I_d + X\right),$$

where $w(x) = \frac{x}{\log(1+x)}$.

Employing the O rule with $J = \log\det\left(I_d + \cdot\right)$, we get for any $k > 0$ and $P \in \mathcal{P}$

$$\begin{aligned}\left\|\mathcal{T}_k^0(P)\right\|_2 &\le w\left(\left\|\mathcal{T}_k^0(P)\right\|_2\right)\log\det\left(I_d + \mathcal{T}_k^0(P)\right)\\ &\le w\left(\left\|\mathcal{T}_k^0(P)\right\|_2\right)\log\det\left(I_d + \mathcal{T}_k^0(P_{k+1})\right).\end{aligned}$$

The above derivation can also equivalently be obtained using $J = \|\cdot\|_2$ or $J = \mathrm{tr}$. Via the following determinant identity:

$$\det(\mathcal{H}_{k+1}^0) = \det\left(\mathcal{H}_k^0\right)\det\left(I_d + \mathcal{T}_k^0(P_{k+1})\right)$$

and a telescoping sum, we retrieve:

$$\begin{aligned}\left\|\mathcal{T}_n^0(P)\right\|_2 &= \frac{1}{n}\sum_{k=0}^{n-1}\left\|\mathcal{T}_n^0(P)\right\|_2\\ &\le \frac{1}{n}\sum_{k=0}^{n-1}\left\|\mathcal{T}_k^0(P)\right\|_2 && (\mathcal{T}_n^0 \preceq \cdots \preceq \mathcal{T}_1^0)\\ &\le w\left(|\mathcal{P}|\left\|\mathcal{T}_n^0(P)\right\|_2\right)\frac{1}{n}\sum_{k=0}^{n-1}\log\det\left(I_d + \mathcal{T}_k^0(P_{k+1})\right)\\ &\le w\left(\left\|\mathcal{T}_n^0(P)\right\|_2\right)\frac{1}{n}\sum_{k=0}^{n-1}\log\det\frac{\mathcal{H}_{k+1}^0}{\mathcal{H}_k^0}\\ &\le w\left(\left\|\mathcal{T}_n^0(P)\right\|_2\right)\frac{1}{n}\log\det\frac{\mathcal{H}_n^0}{\lambda_0 I_d}.\end{aligned}$$

Using $\left\|\mathcal{H}_n^0\right\|_2 \le nR^2/\eta + \lambda_0$ (Proposition C.2),

$$\det\mathcal{H}_n^0 \le \left(nR^2/\eta + \lambda_0\right)^d.$$

Hence,

$$\left\|\mathcal{T}_n^0(P)\right\|_2 \le dw\left(\left\|\mathcal{T}_n^0(P)\right\|_2\right)\frac{\log\left(nR^2/\eta\right)}{n}.$$

Using the definition of w, we further simply to:

$$\left\|\mathcal{T}_n^0(P)\right\|_2 \le \left(nR^2/\eta\right)^{d/n} - 1 \lesssim \frac{d\log(nR^2/\eta)}{n}.$$

We now use the pseudo self-concordance property (Lemma C.5) to move from $\mathcal{T}_k^0$ to $\mathcal{T}_k^{\mathrm{E}}$. Since we pick $\theta_0 = 0$ for the offline rule, we pay the squared worst-case pseudo self-concordance constant

$$M = \exp(2R) \geq \left(\max_{P \in \mathcal{P}} \exp\left(R \left\| \theta^{\mathrm{E}} - \theta_0 \right\|_{\Pi_P} \right) \right)^2,$$

to obtain for all $P \in \mathcal{P}$:

$$\left\| \mathcal{T}_n^{\mathrm{E}}(P) \right\|_2 \leq M \left\| \mathcal{T}_n^0(P) \right\|_2 \lesssim d\,|\mathcal{P}|\, \frac{\log\left(nR^2/\eta\right)}{n}.$$

By definition of g_n:

$$g_n \lesssim d e^{2R} \cdot \frac{\log(nR/\eta)}{n}.$$

We must now find the warm-up time n_δ (5). Setting $a = \frac{d^2 R^2 \log(d/\delta) e^{2R}}{\eta \kappa m}$ and $b = \frac{d^2 R^2 \log(d/\delta) e^{2R} \log(R/\eta)}{\eta \kappa m}$, the condition is satisfied for all

$$n \gtrsim -a\, W_{-1}\left(-\frac{1}{a} e^{-b/a} \right),$$

where W_{-1} is the negative branch of the Lambert W function. Using standard bounds on W_{-1} yields the explicit estimate

$$\begin{aligned} n &\gtrsim a \log a + b \\ &\gtrsim \frac{d^2 R^3 \log(d/\delta) e^{2R}}{\eta \kappa m} \log\left[\frac{dR \log(d/\delta)}{\eta \kappa m} \right]. \end{aligned}$$

Taking the smallest n verifying the above inequality retrieves the warm-up time n_δ.

□

While looser due to the less precise offline information $\mathcal{H}_n^0$, this strategy retains the key benefit of entirely removing the dependence on $|\mathcal{P}|$, and provides a practical alternative when interaction is limited. The resulting offline algorithm is given in Appendix F.

F Algorithms

Two phase active IRL. A detailed procedure for our two-phase active learning approach is provided in Algorithm 2. The algorithm operates as follows. First, for a chosen number of warm-up rounds n_δ, we pick dynamics using the passive rule (7), observe m expert trajectories on each chosen dynamics, and make initial reward estimates θ_{n_δ} by minimizing the minmax IRL loss (2). In practice this is done by following Algorithm 4. Subsequently, every next round k we:

1. Estimate Hessians $\mathcal{H}_P(\theta_k)$ by sampling trajectories on each $P \in \mathcal{P}$ with policies $\pi_P^{\theta_k}$ previously recovered via Algorithm 4. By Lemma C.1, Hessians are straightforwardly estimated via $\mathcal{H}_P(\theta_k) = \mathrm{Cov}_P^{\pi_P^{\theta_k}}[\phi(\tau)]$.
2. Observe m new expert trajectories on the selected dynamics.
3. Learn new estimates θ_{k+1} incorporating the newest expert demonstrations into the dataset.

Offline active IRL. We further provide a procedure for actively constructing a sequence of dynamics without observing the expert in Algorithm 3. The main difference with the online Algorithm 2 is that all dynamics are chosen using $\theta_0 = 0$, and does not have a warm-up phase. Then, expert trajectories are observed over the chosen dynamics $(P_k)_{k=1}^n$ all at once. Finally, rewards are inferred from expert trajectories with Algorithm 4.

Algorithm 2 Active IRL

Input: $\mathcal{P}$, $n, n_\delta > 0$, $\lambda_0 > 0$, criterion J
Initialize: $\theta_0 \leftarrow 0$, $\pi_P^{\theta_0}(a) = \mathrm{Unif}(\mathcal{A})$ for all $P \in \mathcal{P}$; $\mathcal{D}_0^{\mathrm{E}} \leftarrow \emptyset$, $k(P) \leftarrow 0 \;\; \forall P \in \mathcal{P}$
for $k = 1, \ldots, n$ **do**
  if $k < n_\delta$ **then** (Warm-up)
    $P_k \in \arg\min_{P \in \mathcal{P}} k(P)$
    $k(P_k) \leftarrow 1 + k(P_k)$
  else (Active learning)
    Estimate $\mathcal{H}_P(\theta_{k-1})$ by sampling trajectories with $\pi_P^{\theta_{k-1}}$ on each $P \in \mathcal{P}$
    $\mathcal{H}_{k-1} \leftarrow \sum_{j=1}^{k} \mathcal{H}_{P_j}(\theta_{k-1})$
    $P_k \leftarrow \arg\max_{P \in \mathcal{P}} \; J\left(\mathcal{H}_P(\theta_{k-1})(\mathcal{H}_{k-1} + \lambda_0 I_d)^{-1}\right)$
  $\mathcal{D}_k^{\mathrm{E}} \leftarrow \mathcal{D}_{k-1}^{\mathrm{E}} \cup \left\{\tau_{k,j} \sim \mathbb{P}_{P_k}^{\theta^{\mathrm{E}}}\right\}_{j=1}^{m}$ (Observe expert)
  if $k \geq n_\delta$ **then**
    $\theta_k, \left\{\pi_P^{\theta_k}\right\}_{P \in \mathcal{P}} \leftarrow IRL(\mathcal{D}_k)$ (Algorithm 4)
return: $\theta_n, \left\{\pi_P^{\theta_n}\right\}_{P \in \mathcal{P}}$

Algorithm 3 Offline Active IRL

Input: $\mathcal{P}$, $n \in \mathbb{N}^+$, $\delta \in (0,1)$, $\lambda > 0$
Initialize: $\theta_0 \leftarrow 0$, $\pi_P^{\theta_0}(a) = \mathrm{Unif}(\mathcal{A})$ for all $P \in \mathcal{P}$; $\mathcal{H}_0^0 = 0$.
for $P \in \mathcal{P}$ **do**
  Estimate $\mathcal{H}_P(\theta_0)$ by sampling trajectories with $\pi_P^{\theta_0}$ on each $P \in \mathcal{P}$
for $k = 1, \ldots, n$ **do**
  $P_k \leftarrow \arg\max_{P \in \mathcal{P}} \; J\left(\mathcal{H}_P(\theta_0)(\mathcal{H}_k^0 + \lambda_0 I_d)^{-1}\right)$ (Offline Active Learning)
  $\mathcal{H}_{k+1}^0 \leftarrow \mathcal{H}_k^0 + \mathcal{H}_{P_k}(\theta_0)$
$\mathcal{D} \leftarrow \bigcup_{k=1}^{n} \left\{\tau_{k,j} \sim \mathbb{P}_{P_k}^{\theta^{\mathrm{E}}}\right\}_{j=1}^{m}$ (Observe expert)
$\theta_n, \left\{\pi_P^{\theta_n}\right\}_{P \in \mathcal{P}} \leftarrow \mathrm{IRL}(\mathcal{D})$ (Algorithm 4)
return: $\theta_n, \left\{\pi_P^{\theta_n}\right\}_{P \in \mathcal{P}}$

G Additional experimental details and results

G.1 Common Experimental Details

All expert trajectories have length $H = 100$. We set $\eta = 20$ and use $B_\phi = B_\theta = 1$ throughout all experiments. Policies corresponding to every combination of dynamics and reward functions are computed using soft policy iteration.

In all experiments, the feature map ϕ is chosen as the indicator function over state–action pairs. Consequently, the reward function takes the form

$$r_\theta(s,a) = \sum_{s',a'} \mathbb{1}\{s' = s,\ a' = a\}\, \theta_{s',a'}.$$

Reported suboptimality is evaluated as the worst-case value over the considered set of dynamics $\mathcal{P}$:

$$\max_{P \in \mathcal{P}} \mathrm{SubOpt}(\theta_n, P).$$

Implementation details. The implementation is publicly available at https://anonymous.4open.science/r/transferable-active-irl.

Algorithm 4 IRL with multiple dynamics

Input: α, $\mathcal{D}^{\mathrm{E}}$, previously picked dynamics $(P_i)_{i=1}^k$, $\mathcal{P}$
Initialize: θ_0 arbitrarily and pre-compute $\pi_P^{\theta_0}$ for each $P \in \mathcal{P}$, $t = 0$
Compute $\hat{\phi}^{\mathrm{E}} \leftarrow \frac{1}{m} \sum_{\tau \in \mathcal{D}^{\mathrm{E}}} \phi(\tau)$
while not converged **do**
  for $i = 1, \ldots, k$ **do**
    Compute $\hat{\phi}_{P_i}^{\theta_t} = \mathbb{E}_{P_i}^{\pi_{P_i}^{\theta}} [\phi(\tau)]$ sampling trajectories on P_i with $\pi_{P_i}^{\theta_t}$
  $G_t \leftarrow \hat{\phi}^{\mathrm{E}} - \sum_{i=1}^k \phi_{P_i}^{\theta_t}$
  $\theta_{t+1} \leftarrow \Pi_\Theta (\theta_t + \alpha G_t)$ (Project onto Θ)
  Compute policies $\pi_P^{\theta_{t+1}}$ for each $P \in \mathcal{P}$ (RL)
  $t \leftarrow t + 1$
return: $\theta_t, \left\{\pi_P^{\theta_t}\right\}_{P \in \mathcal{P}}$

Compute resources. All experiments were conducted on a single `AMD EPYC 7302P` machine.

G.2 Controlled sets of dynamics

Principal angles. Here we provide the details on principal angles, used for the first experimental setup. Intuitively, principal angles (Galántai, 2004) can be seen as a natural generalization of angles between two lines or planes to higher dimensional subspaces. Formally we define them as follows.

Definition G.1 (Principal angles (Galántai, 2004; Schlaginhaufen & Kamgarpour, 2024))**.** Let $\mathcal{V}, \mathcal{W} \subseteq \mathbb{R}^d$ be two subspaces. The principal angles $0 \leq \alpha_1(\mathcal{V}, \mathcal{W}) \leq \cdots \leq \alpha_m(\mathcal{V}, \mathcal{W}) := \alpha_{\max}(\mathcal{V}, \mathcal{W}) \leq \frac{\pi}{2}$ between $\mathcal{V}$ and $\mathcal{W}$ are defined recursively by

$$\cos\big(\alpha_i(\mathcal{V}, \mathcal{W})\big) = \max_{v \in \mathcal{V},\, w \in \mathcal{W}} \langle v, w \rangle \quad \text{s.t.} \quad \|v\|_2 = \|w\|_2 = 1, \quad \langle v, v_j \rangle = 0, \quad \langle w, w_j \rangle = 0, \quad j = 1, \ldots, i-1,$$

where $v_j \in \mathcal{V}$ and $w_j \in \mathcal{W}$ are maximizers corresponding to $\alpha_j(\mathcal{V}, \mathcal{W})$.

In the context of IRL, our work builds on the geometric perspective introduced by Schlaginhaufen & Kamgarpour (2024). In particular, Theorems 3.10 and 3.11 characterize reward transferability through the relative geometry of the identifiable subspaces associated with different transition dynamics, quantified via principal angles. More precisely, Theorem 3.10 shows that when the *second* principal angle between two unidentifiable subspaces $\Theta_P^\perp$ and $\Theta_{P'}^\perp$ is sufficiently large, observing the expert on both dynamics enables the recovery of rewards that transfer to arbitrary dynamics. Conversely, Theorem 3.11 establishes a complementary regime: when the *largest* principal angle between Θ_P and $\Theta_{P'}$ is sufficiently small, a reward learned from demonstrations in one dynamics transfers reliably to the other. Taken together, these results reveal a geometric trade-off. Large diversity between environments, captured by a large second principal angle, improves global identifiability and enables broad transfer. In contrast, strong similarity between dynamics, captured by a small maximal principal angle, guarantees local transferability of learned rewards across environments.

Let $\alpha_i(P, P')$ denote the i-th principal angle between the subspaces $\Theta_P^\perp$ and $\Theta_{P'}^\perp$, and let $\bar{\alpha}_i(P) = \frac{1}{|\mathcal{P}|} \sum_{P' \in \mathcal{P}} \alpha_i(P, P')$ be the average i-th principal angle of P over the set $\mathcal{P}$.

Setup. In our experiments, we constructed four classes of dynamics sets, each characterized by the geometry of their identifiable subspaces through principal-angle statistics. For a set of dynamics

$\mathcal{P} = \{P_1, \dots, P_{|\mathcal{P}|}\}$, we define

$$\mu_i(\mathcal{P}) \coloneqq \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} \alpha_i(P), \qquad \sigma_i^2(\mathcal{P}) \coloneqq \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} \big(\alpha_i(P) - \mu_i(\mathcal{P})\big)^2,$$

Table 3 summarizes the statistics of the evaluated dynamics sets, averaged over 20 runs.

Regarding rewards, for each different seed, θ^{E} is uniformly sampled in $[0,1]^d$, and then rescaled such that $\left\|r^{\mathrm{E}}\right\|_2 \leq 1$. To estimate rewards, we run Algorithm 4 for 100 iterations, with $\alpha = 0.02$.

Table 3: Principal-angle statistics (in radian) of the controlled sets of dynamics, averaged over 20 runs.

Dynamics set	$\mu_2(\mathcal{P}) \pm \sigma_2(\mathcal{P})$	$\mu_{\max}(\mathcal{P}) \pm \sigma_{\max}(\mathcal{P})$
Equivalent dynamics	0.00 ± 0.00	0.00 ± 0.00
Similar dynamics	0.00 ± 0.00	0.98 ± 0.05
Uniformly diverse dynamics	0.03 ± 0.03	1.15 ± 0.17
Arbitrarily diverse dynamics	0.57 ± 0.11	1.44 ± 0.05

For the held-out dynamics used to evaluate the transferability of learned rewards (Tables 1 and 4), each set was generated by sampling 1000 dynamics using Dirichlet distributions. Specifically, denoting by $\mathcal{P}^{\text{test}} = \{P_i^{\text{test}}\}_{k=1}^{1000}$ the set of held-out dynamics, we used

$$P_i^{\text{test}}(s' \mid s, a) \sim \mathrm{Dir}(\underbrace{\alpha_i, \dots, \alpha_i}_{\times |S|}), \quad \forall (s', s, a), \quad \text{with} \quad \alpha_i \sim \log\mathrm{uniform}(e^{-2}, e^{0.1}).$$

Results. Figure 5 reports the suboptimality attained by our methods when using either the $\log\det(I_d + \cdot)$, tr, or $\|\cdot\|_2$ criterion for the active selection objective J defined in Equation (8). In these experiments, all methods operate without a warm-up phase, meaning that dynamics are selected actively from the very first round. We additionally compare against the offline variant introduced in Section E.

We first observe that all methods achieve comparable performance when the dynamics set is uniformly distributed with respect to principal angles. Differences only emerge in the arbitrarily distributed setting, where the spectral norm criterion and the offline trace-based variant perform slightly worse, reaching approximately twice the suboptimality of the best-performing methods. Interestingly, the offline variant remains competitive in this setting.

Moreover, Table 4 shows that the log-determinant and trace criteria provide the strongest transfer performance to previously unseen dynamics. In contrast, the offline variants generally exhibit weaker transferability overall, although the trace-based offline criterion performs best among the three offline alternatives.

G.3 Random mazes

Setup. Denote $l = 7$ the dimension of the square mazes. Each set of $|\mathcal{P}|$ mazes were randomly generated as follows:

1. The i-th maze is randomly generated with a path length to the goal constrained to be at least $\left(\frac{i \times l}{|\mathcal{P}|}\right)^2$.
2. For every odd i, the i-th maze was additionally allowed to be unsolvable.

For the rewards, each different seed, we randomly select 3 goal states s_1, s_2, s_3 different from the initial state, uniformly sample the three goal state rewards for the expert uniformly from $[0,1]$, and then rescale them such that $\left\|r^{\mathrm{E}}\right\|_2 \leq 1$. For all non goal states s, or actions different from the

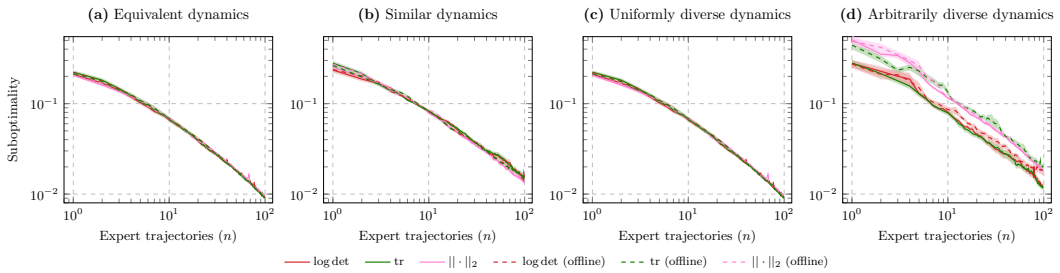

Figure 5: Worst-case suboptimality over the training set $\mathcal{P}$ as a function of the number of observed expert trajectories on randomly generated MDPs, and of the criterion J used to pick dynamics online (8) or offline (O) . Each plot corresponds to a different structure of $\mathcal{P}$.

Table 4: Worst-case suboptimality on sets of *held-out* stochastic dynamics after learning from $n = 100$ expert trajectories collected on either uniformly or arbitrarily diverse sets of dynamics.

		Worst-case suboptimality ($\times 100$) on held-out dynamics	
criterion J	**Offline**	**Uniformly diverse training set**	**Arbitrarily diverse training set**
$\log\det(I_d + \cdot)$	No	$\mathbf{02.86 \pm 0.02}$	12.63 ± 0.14
tr	No	02.96 ± 0.02	$\mathbf{12.40 \pm 0.14}$
$\|\cdot\|_2$	No	03.02 ± 0.19	15.23 ± 0.17
$\log\det(I_d + \cdot)$	Yes	03.13 ± 0.02	13.54 ± 0.15
tr	Yes	03.14 ± 0.20	12.64 ± 0.15
$\|\cdot\|_2$	Yes	03.14 ± 0.02	15.14 ± 0.17

waiting action a, $r^{\mathrm{E}}(s, a) = 0$. For reward estimation, we run Algorithm 4 for 100 iterations, with $\alpha = 0.05$ using the Adam optimizer.

Results. Beyond the results presented in the main paper, we additionally evaluate settings with dynamics sets of size $|\mathcal{P}| = 5$ and $|\mathcal{P}| = 20$ in order to study the impact of the number of candidate dynamics on the effectiveness of our active learning strategy.

Figure 6 reports both the suboptimality and reward estimation error achieved by our methods (using the log-determinant criterion), together with passive learning and ED (Buening et al., 2024), for $|\mathcal{P}| \in \{5, 20, 100\}$. Across all values of $|\mathcal{P}|$, our active selection rule (with and without warm-up) consistently outperforms the competing approaches. This suggests that actively selecting dynamics improves sample efficiency even when the set of candidate dynamics is relatively small. However, the offline variant (see Section E) performs worse than passively selecting dynamics, hinting that in general, it is hard to entirely choose dynamics offline for expert demonstrations to learn the expert reward.

We also observe that the passive warm-up phase used in our two-phase method is not beneficial. In all tested settings, removing this warm-up phase leads to improved sample efficiency, with the gains becoming even more pronounced as the number of dynamics increases.

References

Kareem Amin, Nan Jiang, and Satinder Singh. Repeated inverse reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems*, 34:8927–8939, 2021.

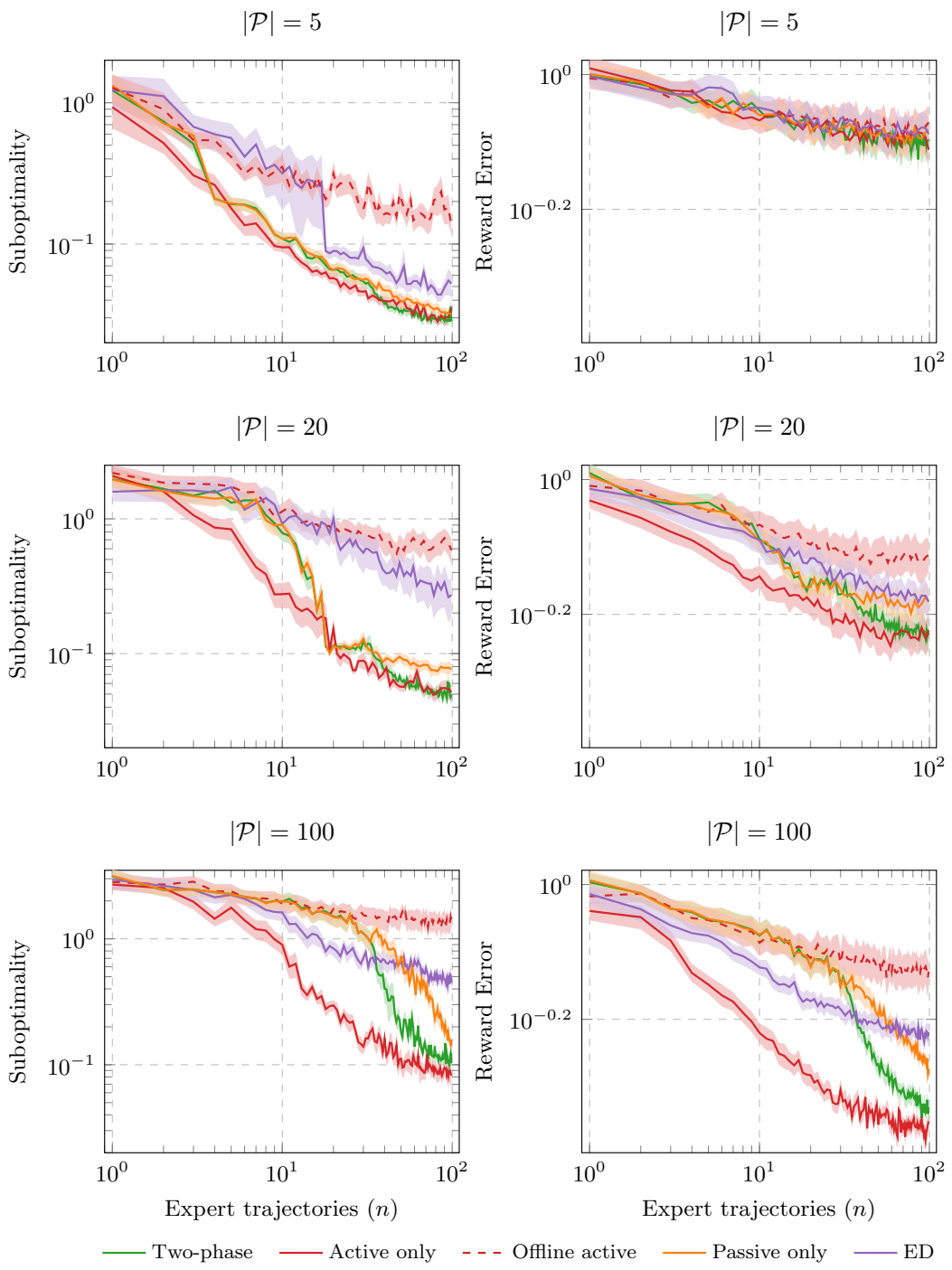

Figure 6: Worst-case suboptimality over the training set $\mathcal{P}$ of randomly generated mazes, and the corresponding reward error, as a function of the number of observed expert trajectories, and of the size $|\mathcal{P}|$ of the dynamics set.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

Ondrej Bajgar, Sid William Gould, R. Mitta, Jonathon Liu, Oliver Newcombe, and Jackson Golden. Toward information theoretic active inverse reinforcement learning. *ArXiv*, abs/2501.00381,

2024. URL https://api.semanticscholar.org/CorpusId:275212510.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391.

Thomas Kleine Buening, Victor Villin, and Christos Dimitrakakis. Environment design for inverse reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 24808–24828, 2024.

Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

Jongseong Chae, Seungyul Han, Whiyoung Jung, Myungsik Cho, Sungho Choi, and Youngchul Sung. Robust imitation learning against variations in environment dynamics. In *International Conference on Machine Learning*, pp. 2828–2852. PMLR, 2022.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. *Advances in Neural Information Processing Systems*, 28, 2015.

Yuchen Cui and Scott Niekum. Active reward learning from critiques. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6907–6914, 2018. URL https://api.semanticscholar.org/CorpusID:4490670.

Aurél Galántai. *Finite Projection Methods for Linear Systems*, pp. 83–115. Springer US, Boston, MA, 2004. DOI: 10.1007/978-1-4419-9180-5_3. URL https://doi.org/10.1007/978-1-4419-9180-5_3.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

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

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Jack Kiefer. General equivalence theory for optimum designs (approximate theory). *The annals of Statistics*, pp. 849–879, 1974.

Jack Kiefer and Jacob Wolfowitz. Optimum designs in regression problems. *The annals of mathematical statistics*, 30(2):271–294, 1959.

Andreas Krause and Carlos Guestrin. Near-optimal observation selection using submodular functions. In *AAAI*, volume 7, pp. 1650–1654, 2007.

Joongkyu Lee and Min-hwan Oh. Improved online confidence bounds for multinomial logistic bandits. *arXiv preprint arXiv:2502.10020*, 2025.

David Lindner, Andreas Krause, and Giorgia Ramponi. Active exploration for inverse reinforcement learning, 2023. URL https://arxiv.org/abs/2207.08645.

M. Lopes, Francisco S. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. In *ECML/PKDD*, 2009. URL https://doi.org/10.1007/978-3-642-04174-7_3.

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.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Paul Nitschke, Lars Lien Ankile, Eura Nofshin, Siddharth Swaroop, Finale Doshi-Velez, and Weiwei Pan. Amber: An entropy maximizing environment design algorithm for inverse reinforcement learning. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024.

Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In D. Touretzky (ed.), *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988. URL https://proceedings.neurips.cc/paper_files/paper/1988/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper.pdf.

Nicholas Rittler and Kamalika Chaudhuri. A two-stage active learning algorithm for k-nearest neighbors. In *International Conference on Machine Learning*, pp. 29103–29129. PMLR, 2023.

Paul Rolland, Luca Viano, Norman Schürhoff, Boris Nikolov, and Volkan Cevher. Identifiability and generalizability from multiple experts in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:550–564, 2022.

Andreas Schlaginhaufen and Maryam Kamgarpour. Towards the transferability of rewards recovered via regularized inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 37:21461–21501, 2024.

Andreas Schlaginhaufen and Maryam Kamgarpour. Fast rates for inverse reinforcement learning, 2026.

Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pp. 10022–10032. PMLR, 2021.

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

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. DOI: 10.1109/IROS.2012.6386109.

Joel A Tropp. Freedman's inequality for matrix martingales. *arXiv preprint arXiv:1101.3039*, 2011.

Tian Xie, Jifan Zhang, Haoyue Bai, and Robert Nowak. Deep active learning in the open world. *arXiv preprint arXiv:2411.06353*, 2024.