TOWARDS RECOVERABLE REWARD TRANSFER GEOMETRY FROM IRL GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse reinforcement learning (IRL) aims to recover reward functions from expert demonstrations, but reward ambiguity fundamentally limits their transferability across environments. Recent theory characterizes transfer sensitivity through principal-angle geometry between Potential Shaping Spaces, yet computing this latent geometry requires known or estimated transition laws. We show that this model-level route can be bypassed: we establish an observable route through IRL gradients, linking the transfer-relevant reward structure to gradient-space geometry and enabling direct estimation of this geometry. In practice, however, noisy or approximately evaluated gradients can lead to rank inflation and degeneracy of empirical principal angles, while complementary representations introduce structural angles unrelated to transfer. To enable estimation, we develop **Dimension-Calibrated Truncated SVD (DC-TSVD)**, which uses intrinsic dimensions determined by IRL identifiability to suppress perturbation-induced directions and remove structural angles. Theoretical analysis establishes principal-angle estimation error bounds, and experiments on synthetic and finite-demonstration IRL environments demonstrate reliable estimation of the latent transfer geometry.

1 INTRODUCTION

Reinforcement learning (RL) has achieved impressive empirical successes (Sutton & Barto, 2018; Mnih et al., 2015; Silver et al., 2016; Chu et al., 2026), yet specifying a reward function that faithfully captures the desired behavior remains a fundamental challenge. Inverse reinforcement learning (IRL) addresses this challenge by inferring reward functions from expert demonstrations (Ng & Russell, 2000; Abbeel & Ng, 2004). Compared with directly imitating expert policies, IRL aims to recover task objectives that are potentially transferable across environments.

A fundamental difficulty of IRL is that the recovered reward is generally not unique: multiple reward functions can induce identical optimal behaviors, with potential-based shaping transformations representing an important class of such ambiguities (Ng et al., 1999). To characterize this ambiguity, a prominent line of work adopts maximum-entropy (soft) optimality as a model of expert behavior (Ziebart et al., 2008). Under this setting, recent theoretical studies show that expert rewards can be identified up to potential-based shaping transformations (Cao et al., 2021; Skalse et al., 2023). Building on these results, Schlaginhaufen & Kamgarpour (2023) further showed that guaranteeing transferability across arbitrary transition laws requires identifying the reward up to a constant. Such identification can be achieved either by restricting the reward class, for example to state-only rewards (Amin et al., 2017; Fu et al., 2018; Zhang et al., 2026; Li et al., 2026; Gangwani & Peng, 2020; Zhou et al., 2024), or by leveraging demonstrations from multiple experts with the same reward but different transition laws under suitable rank conditions (Cao et al., 2021; Rolland et al., 2022; Likmeta et al., 2021; Tateo et al., 2017; Giwa & Lee, 2021). More broadly, recent studies have explored reward identifiability through feasible reward sets, algebraic characterizations, and partial identifiability (Metelli et al., 2021; 2023; Lazzati et al., 2024; Poiani et al., 2024; Shehab et al., 2024; Skalse & Abate, 2026).

Beyond determining whether a reward is identifiable, recent work has further investigated how reward ambiguity affects transfer across environments. For two source environments, Schlaginhaufen & Kamgarpour (2024) established a theoretical connection between reward transfer error and the principal angles between their Potential Shaping Spaces (PS Spaces): when the two spaces share a

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

shaping direction (e.g., the constant-shift ambiguity), transfer error can be substantially amplified as the second principal angle approaches zero. An illustrative example is provided in Appendix A. Nevertheless, when the transition laws are unknown, Schlaginhaufen & Kamgarpour (2024) assumes empirical transition-law estimates are available for constructing the PS Spaces, without specifying how these estimates are obtained. Moreover, estimating the transition model itself requires recovering a model-level object before the target principal-angle geometry can be computed. We ask whether this model-level route can be bypassed: *Can the transfer-relevant latent geometry be estimated directly from the learning signals generated by IRL?*

Our key insight is that latent transfer geometry can be represented through observable IRL gradients. The principal-angle geometry of the PS Spaces is encoded in their complementary Identifiable Reward Spaces and is approximately preserved under the reward-Jacobian mapping to the theoretical gradient spaces. This establishes an observable route from IRL gradients to latent transfer geometry without explicitly reconstructing transition dynamics.

In practice, estimating this geometry from gradient observations faces two challenges. Noisy or approximately evaluated gradients can expand the observed gradient geometry beyond its intrinsic structure, leading to rank inflation and even degeneracy of empirical principal angles. Moreover, complementary representations introduce structural angles unrelated to the original transfer geometry. To address these challenges, we develop **Dimension-Calibrated Truncated SVD (DC-TSVD)**. Rather than relying on the empirical rank (subject to rank inflation), DC-TSVD exploits the identifiability-implied gradient-space dimension to identify the intrinsic gradient geometry and then remove complementary structural angles. This makes the observable gradient route operational: latent transfer geometry can be estimated from noisy gradient observations, including the second principal angle associated with transfer sensitivity. Our contributions are summarized as follows:

- We establish an observable gradient-based representation of reward-transfer geometry, bypassing the model-level route based on estimated transition dynamics. By relating PS-space principal angles to Identifiable Reward Spaces and Jacobian-induced gradient representations, we characterize the conditions under which this geometry can be recovered from IRL gradients.
- We develop DC-TSVD to estimate transfer-relevant reward geometry from perturbed gradient observations. By exploiting the identifiability-implied gradient-space dimension, DC-TSVD suppresses perturbation-induced rank inflation and removes the structural angles introduced by orthogonal complementation.
- We establish perturbation bounds that separate gradient-observation error from Jacobian-induced geometric distortion. We further evaluate how gradient coverage, sampling accuracy, and reward parameterization affect geometric recovery.

2 Problem Formulation and Transfer Geometry

Regularized Inverse Reinforcement Learning. We consider the framework of regularized inverse reinforcement learning (IRL). Let $(\mathcal{S}, \mathcal{A}, P, \nu_0, r, \gamma, h)$ be a regularized Markov decision process (Sutton & Barto, 2018; Puterman, 1994), where $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, $P \in \Delta_{|\mathcal{S}|}^{|\mathcal{S}|\times|\mathcal{A}|}$ denotes the transition law, ν_0 is the initial state distribution, $r \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the reward function, $\gamma \in (0,1)$ is the discount factor, and h is a strictly convex regularizer. The optimal policy is obtained by solving $\max_\pi \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - h(\pi(\cdot|s_t)))\right]$. Given expert demonstrations $\mu_{P^k}^{\mathcal{B}}$ collected under transition law P^k, inverse reinforcement learning estimates a reward function by optimizing a general objective

$$\ell_{P^k}(r_\theta, \mu_{P^k}^{\mathcal{B}}),$$

where $r_\theta : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is a parameterized reward function with parameter $\theta \in \mathbb{R}^d$, and ℓ_{P^k} measures the discrepancy between the induced behavior and the expert demonstrations. Here, d denotes the reward-parameter dimension.

Potential Shaping Space. The Potential Shaping Space associated with the transition operator P is defined by

$$\mathcal{U}_P = \mathrm{im}(E - \gamma P) \subseteq \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$

where $E : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the state-action expansion operator and $P : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the transition operator. Rewards differing by an element of $\mathcal{U}_P$ induce the same optimal behavior under regularized IRL. Therefore, the reward ambiguity is characterized by the shaping space $\mathcal{U}_P$, while the **Identifiable Reward Space** is represented by its orthogonal complement $\mathcal{U}_P^{\perp}$. For subsequent dimension accounting, let

$$D := |\mathcal{S}||\mathcal{A}|$$

denote the dimension of the state-action reward space. Note that D is distinct from the reward-parameter dimension d. Under the standard rank condition $\mathrm{rank}(E - \gamma P) = |\mathcal{S}|$, the dimensions of the Potential Shaping Space and the Identifiable Reward Space are

$$r := \dim(\mathcal{U}_P) = |\mathcal{S}|, \qquad m := \dim(\mathcal{U}_P^{\perp}) = D - r = |\mathcal{S}|(|\mathcal{A}| - 1),$$

Principal Angles. (Galántai, 2003) Let $\mathcal{V}, \mathcal{W} \subseteq \mathbb{R}^n$ be two subspaces of dimension p and q respectively, with $r = \min(p, q)$. The principal angles $0 \le \theta_1 \le \theta_2 \le \cdots \le \theta_r \le \pi/2$ between $\mathcal{V}$ and $\mathcal{W}$ are defined recursively by:

$$\cos\theta_i = \max_{\substack{\mathbf{v}\in\mathcal{V}_i \\ \mathbf{w}\in\mathcal{W}_i}} \langle \mathbf{v}, \mathbf{w} \rangle,$$

where:

$$\mathcal{V}_i = \{\mathbf{v} \in \mathcal{V} \mid \|\mathbf{v}\| = 1, \langle \mathbf{v}, \mathbf{v_j} \rangle = 0 \text{ for } j = 1, \ldots, i-1\}$$
$$\mathcal{W}_i = \{\mathbf{w} \in \mathcal{W} \mid \|\mathbf{w}\| = 1, \langle \mathbf{w}, \mathbf{w_j} \rangle = 0 \text{ for } j = 1, \ldots, i-1\}.$$

The maximizing vectors from previous steps are denoted by $\mathbf{v_j}, \mathbf{w_j}$. Throughout the paper, principal angles and their cosines are indexed in ascending and descending order, respectively.

Transfer Geometry under Environment Shift. Recent theoretical analysis (Schlaginhaufen & Kamgarpour, 2024) characterizes reward transferability through the geometry of Potential Shaping Spaces. In particular, for two experts under transition laws P_0 and P_1 whose Potential Shaping Spaces admit a shared shaping direction (i.e., $\theta_1(P^0, P^1) = 0$, e.g., the constant-shift ambiguity), Theorem 3.10 provides a transferability guarantee controlled by their second principal angle:

$$\epsilon = \frac{\hat{\epsilon}}{\eta\sigma_R \sin^2(\theta_2(P^0, P^1)/2)}, \tag{1}$$

where $\hat{\epsilon}$ denotes the reward recovery error in the source environments, η and σ_R are problem-dependent regularity constants, and $\theta_2(P^0, P^1) := \theta_2(\mathcal{U}_P^0, \mathcal{U}_P^1)$ is the second principal angle between the two Potential Shaping Spaces. Therefore, a near-zero second principal angle indicates transfer regimes that are highly sensitive to reward estimation errors. We next investigate whether transfer-relevant geometry can be recovered directly from observable IRL gradients.

3 Observable Gradient Geometry

This section connects latent transfer geometry to observable IRL gradients. The proofs are deferred to Appendix C, and Figure 1 provides an overview of the geometric relationships developed below.

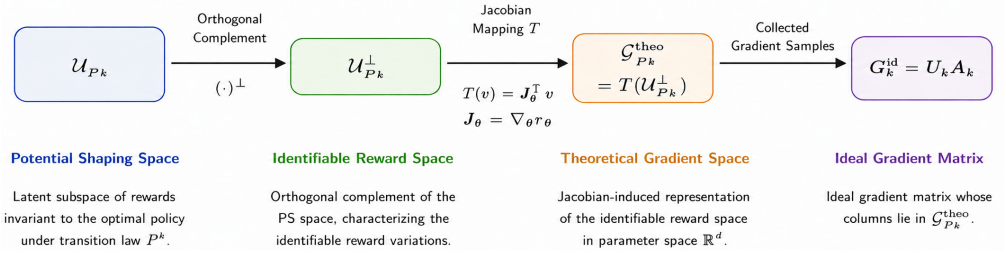


Figure 1: From latent reward geometry to observable gradient representations.

3.1 Observable Gradient Representation of Transfer Geometry

For transition law P^k, consider the IRL objective $\ell_{P^k}(r_\theta, \mu^{\mathcal{B}}_{P^k})$, where $r_\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the reward vector induced by the parameterized reward model with parameter $\theta \in \mathbb{R}^d$, and let

$$J_\theta := \nabla_\theta r_\theta \tag{2}$$

denote the Jacobian of reward parameterization.

Although the transfer geometry is originally defined through the PS Spaces $\mathcal{U}_{P^k}$, these subspaces are latent because they depend on the transition dynamics. We therefore turn to their orthogonal complements $\mathcal{U}^\perp_{P^k}$, which admit observable gradient representations. Define the excess complement dimension

$$q := m - r = D - 2r = |\mathcal{S}|(|\mathcal{A}| - 2).$$

The following Proposition 1 establishes the precise geometric relation between the two representations; the proof is provided in Appendix C.1.

Proposition 1 (Transfer Geometry in Orthogonal Complements) *Let* $\{\theta^{\mathrm{PS}}_i\}_{i=1}^r$, $\{\theta^{\mathrm{comp}}_j\}_{j=1}^m$ *denote the principal angles between* $(\mathcal{U}_{P^0}, \mathcal{U}_{P^1})$, $(\mathcal{U}^\perp_{P^0}, \mathcal{U}^\perp_{P^1})$, *respectively. Then*

$$\theta^{\mathrm{comp}}_j = 0, \quad j = 1, \dots, q, \qquad \theta^{\mathrm{comp}}_{q+i} = \theta^{\mathrm{PS}}_i, \quad i = 1, \dots, r.$$

Equivalently, for $\rho^{\mathrm{PS}}_i := \cos\theta^{\mathrm{PS}}_i$ *and* $\rho^{\mathrm{comp}}_j := \cos\theta^{\mathrm{comp}}_j$,

$$\rho^{\mathrm{comp}}_j = 1, \quad j = 1, \dots, q, \qquad \rho^{\mathrm{comp}}_{q+i} = \rho^{\mathrm{PS}}_i, \quad i = 1, \dots, r.$$

Remark 1 (Structural Zero Angles Induced by Orthogonal Complementation) *Proposition 1 shows that orthogonal complementation introduces* $q = |\mathcal{S}|(|\mathcal{A}| - 2)$ *additional zero principal angles before the original PS-space principal-angle spectrum. These zero angles arise solely from the dimensional structure of the orthogonal complements and are therefore not part of the original PS-space transfer geometry. Importantly, they are distinct from any intrinsic zero principal angles already present between the PS spaces, which are contained in* $\{\theta^{\mathrm{comp}}_{q+i}\}_{i=1}^r$.

To obtain an observable gradient representation, we require the following local expressivity condition:

Assumption 1 (Local Reward Representation) *The reward Jacobian is locally expressive over the joint Identifiable Reward Space:*

$$\mathcal{U}^\perp_{P^0} + \mathcal{U}^\perp_{P^1} \subseteq \mathrm{im}(J_\theta).$$

This assumption requires the Jacobian to have sufficient local rank to represent the joint identifiable reward space: $d \geq \dim(\mathcal{U}^\perp_{P^0} + \mathcal{U}^\perp_{P^1}) = 2m - (q + z) = D - z$, where $z = \dim(\mathcal{U}_{P^0} \cap \mathcal{U}_{P^1})$ and $q + z = \dim(\mathcal{U}^\perp_{P^0} \cap \mathcal{U}^\perp_{P^1})$. Define the Jacobian-induced mapping $T : \mathbb{R}^D \to \mathbb{R}^d$, $T(v) = J_\theta^\top v$. The **Theoretical Gradient Space** associated with P^k is

$$\mathcal{G}^{\mathrm{theo}}_{P^k} := T(\mathcal{U}^\perp_{P^k}) = \{J_\theta^\top v : v \in \mathcal{U}^\perp_{P^k}\} \subseteq \mathbb{R}^d.$$

The Theoretical Gradient Space is a Jacobian-induced representation of the Identifiable Reward Space in the parameter space. Under Assumption 1, T is injective on $\mathcal{U}^\perp_{P^0} + \mathcal{U}^\perp_{P^1}$, and thus preserves both the dimension and intersection structure of the Identifiable Reward Spaces:

$$\dim(\mathcal{G}^{\mathrm{theo}}_{P^k}) = m, \quad k \in \{0, 1\}, \tag{3}$$

$$T(\mathcal{U}^\perp_{P^0}) \cap T(\mathcal{U}^\perp_{P^1}) = T(\mathcal{U}^\perp_{P^0} \cap \mathcal{U}^\perp_{P^1}). \tag{4}$$

These properties are proved in Appendix C.2. We next connect it to observable IRL gradients.

For transition law P^k, define the **gradient random vector**

$$\mathbf{g}^k := \nabla_\theta \ell_{P^k}(r_\theta, \mu^{\mathcal{B}}_{P^k}),$$

where the randomness is induced by demonstration sampling.

By the chain rule, we have $\mathbf{g}^k = J_\theta^\top \nabla_r \ell_{P^k}(r_\theta, \mu^{\mathcal{B}}_{P^k})$. Since the IRL objective is invariant along potential-shaping directions, for any $u \in \mathcal{U}_{P^k}$, we have $\ell_{P^k}(r + tu, \mu^{\mathcal{B}}_{P^k}) = \ell_{P^k}(r, \mu^{\mathcal{B}}_{P^k})$. Differentiating with respect to t at $t = 0$ gives

$$\langle \nabla_r \ell_{P^k}, u \rangle = 0.$$

Hence, $\nabla_r \ell_{P^k}(r_\theta, \mu^{\mathcal{B}}_{P^k}) \in \mathcal{U}^\perp_{P^k}$, and consequently,

$$\mathbf{g}^k \in T(\mathcal{U}^\perp_{P^k}) = \mathcal{G}^{\text{theo}}_{P^k}. \tag{5}$$

Therefore, IRL gradients provide samples from $\mathcal{G}^{\text{theo}}_{P^k}$, the Jacobian-induced representation of $\mathcal{U}^\perp_{P^k}$. At a fixed reward parameter θ, these samples necessarily share the same theoretical gradient space; more generally, this space is retained across varying sampled parameters whenever the Jacobian-induced identifiable image remains invariant; parameter-dependent spaces are left for future work.

To estimate the principal-angle geometry of the original PS Spaces, two questions remain: i. whether it is approximately preserved under the Jacobian-induced mapping into $\mathcal{G}^{\text{theo}}_{P^k}$; and ii. whether $\mathcal{G}^{\text{theo}}_{P^k}$ can be recovered from gradient samples $\{g^k_i\}_{i=1}^n$. We address both below.

3.2 Jacobian-Induced Representation

We first study i. In the identity-parameterization case $J_\theta = I$, $\mathcal{G}^{\text{theo}}_{P^k}$ coincides exactly with $\mathcal{U}^\perp_{P^k}$: $\mathcal{G}^{\text{theo}}_{P^k} = J_\theta^\top \mathcal{U}^\perp_{P^k} = \mathcal{U}^\perp_{P^k},\ k \in \{0, 1\}$. Their principal-angle geometry is therefore preserved exactly. For nonlinear reward parameterizations, however, J_θ generally varies with θ and may induce anisotropic deformation. To identify which part of this deformation affects the subspace geometry, consider the polar decomposition

$$J_\theta^\top = U_\theta T_\theta, \qquad T_\theta = (J_\theta J_\theta^\top)^{1/2},$$

where U_θ is a partial isometry on $\text{im}(T_\theta)$ and thus preserves the relevant principal angles. Hence, any Jacobian-induced change in the principal-angle geometry is determined by the symmetric stretching component T_θ. Let P_k denote the orthogonal projection onto $\mathcal{U}^\perp_{P^k}$. Since $P_k v = v$ for $v \in \mathcal{U}^\perp_{P^k}$, we decompose

$$T_\theta v = \underbrace{P_k T_\theta P_k v}_{\text{within-space component}} + \underbrace{(I - P_k) T_\theta P_k v}_{\text{off-subspace component}}.$$

The first component acts entirely within $\mathcal{U}^\perp_{P^k}$, whereas the second introduces components outside this space. Define the within-space transformation

$$\mathcal{A}_k : \mathcal{U}^\perp_{P^k} \to \mathcal{U}^\perp_{P^k}, \qquad \mathcal{A}_k v = P_k T_\theta P_k v. \tag{6}$$

To quantify the off-subspace component relative to the within-space component, we define

$$\varepsilon_k = \left\| (I - P_k) T_\theta P_k \mathcal{A}_k^{-1} \right\|_2. \tag{7}$$

We assume that this relative off-subspace contribution remains small.

Assumption 2 (Local Identifiable Space Stability) *There exists a sufficiently small constant $c > 0$ such that*

$$\varepsilon_k \le c, \qquad k \in \{0, 1\}.$$

Remark 2 (Interpretation of Identifiable Space Stability) *Assumption 2 provides a condition for preserving the geometry of identifiable reward components under the Jacobian mapping. This condition is motivated by the fact that identifiable reward components determine preference-relevant variations in IRL: substantial Jacobian-induced off-subspace displacement would alter the associated gradient geometry and weaken the correspondence between reward and gradient spaces (see Appendix B for more details).*

Importantly, this condition does not restrict anisotropic deformation within the Identifiable Reward Space. Since $\mathcal{A}_k$ is an invertible map from $\mathcal{U}^\perp_{P^k}$ to itself (see Appendix C.3), for every $u \in \mathcal{U}^\perp_{P^k}$, there exists $v = \mathcal{A}_k^{-1} u \in \mathcal{U}^\perp_{P^k}$ such that $u = \mathcal{A}_k v$. Hence,

$$\mathcal{A}_k \mathcal{U}^\perp_{P^k} = \mathcal{U}^\perp_{P^k},$$

regardless of its anisotropy. Thus, even strong anisotropic within-space deformation still preserves the subspace itself; the assumption only restricts the relative off-subspace displacement that affects the principal-angle geometry.

The following theorem quantifies the deviation in principal-angle geometry induced by relative off-subspace displacement (see Appendix C.3 for the proof).

Theorem 1 (Approximate Angle Preservation under Subspace Leakage) *Let ρ_j^{comp} and ρ_j^{grad} denote the ordered principal-angle cosines between $(\mathcal{U}_{P^0}^{\perp}, \mathcal{U}_{P^1}^{\perp})$ and $(\mathcal{G}_{P^0}^{\mathrm{theo}}, \mathcal{G}_{P^1}^{\mathrm{theo}})$, respectively. Under Assumption 1,*

$$\left|\rho_j^{\mathrm{grad}} - \rho_j^{\mathrm{comp}}\right| \le \frac{\varepsilon_0}{\sqrt{1+\varepsilon_0^2}} + \frac{\varepsilon_1}{\sqrt{1+\varepsilon_1^2}}, \qquad j = 1, \dots, m.$$

Under Assumption 2, this further gives

$$\left|\rho_j^{\mathrm{grad}} - \rho_j^{\mathrm{comp}}\right| \le 2c.$$

Combining Theorem 1 with Proposition 1 gives, for $i = 1, \dots, r$,

$$\left|\rho_{q+i}^{\mathrm{grad}} - \rho_i^{\mathrm{PS}}\right| \le \frac{\varepsilon_0}{\sqrt{1+\varepsilon_0^2}} + \frac{\varepsilon_1}{\sqrt{1+\varepsilon_1^2}} \le 2c. \tag{8}$$

When $J_\theta = I$, $\varepsilon_k = 0$, and (8) yields exact recovery of the PS-space principal-angle geometry.

3.3 Ideal Gradient Observations

We next address ii. Let $g_1^k, \dots, g_n^k$ denote collected gradient samples generated from the gradient random vector $\mathbf{g}^k$.

The **Ideal Gradient Observation Model** (abbreviated as **Ideal Model**) idealizes the collected gradient samples as exactly lying in the intrinsic theoretical gradient space. By (5), we have

$$g_i^k \in \mathcal{G}_{P^k}^{\mathrm{theo}}. \tag{9}$$

Therefore, there exists a coefficient vector $\boldsymbol{\alpha}_i^k \in \mathbb{R}^m$ such that $g_i^k = U_k \boldsymbol{\alpha}_i^k$. Collecting the ideal gradient samples gives the **Ideal Gradient Matrix** $G_k^{\mathrm{id}} = [g_1^k, \dots, g_n^k]$. Collecting the coefficient vectors gives $A_k = [\boldsymbol{\alpha}_1^k, \boldsymbol{\alpha}_2^k, \dots, \boldsymbol{\alpha}_n^k] \in \mathbb{R}^{m \times n}$, where each coefficient vector $\boldsymbol{\alpha}_i^k \in \mathbb{R}^m$ represents the coordinates of the gradient sample g_i^k with respect to the basis U_k. Hence, the Ideal Model admits the factorization

$$G_k^{\mathrm{id}} = U_k A_k. \tag{10}$$

The following theorem establishes exact recovery of the Theoretical Gradient Space under sufficient *Gradient Coverage*, i.e., $\mathrm{rank}(A_k) = m$, so that the sampled gradients span all identifiable directions; see Appendix C.4 for the proof.

Theorem 2 (Gradient Subspace Recovery under the Ideal Model) *Under the Ideal Model, if*

$$\mathrm{rank}(A_k) = m, \qquad k \in \{0, 1\},$$

then the compact SVD

$$G_k^{\mathrm{id}} = \hat{U}_k \hat{\Sigma}_k \hat{V}_k^{\top}, \qquad \hat{U}_k \in \mathbb{R}^{d \times m},$$

satisfies $\mathrm{span}(\hat{U}_k) = \mathcal{G}_{P^k}^{\mathrm{theo}},\ k \in \{0, 1\}$. *Consequently,*

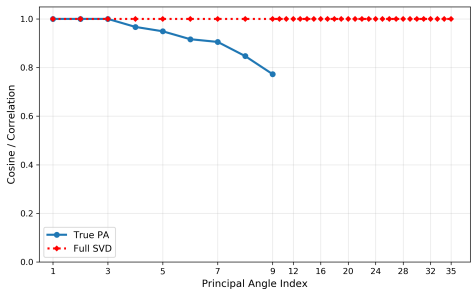
$$\sigma_j(\hat{U}_0^{\top} \hat{U}_1) = \rho_j^{\mathrm{grad}}, \qquad j = 1, \dots, m.$$

Simulation results in Appendix D confirm this exact principal-angle equivalence.

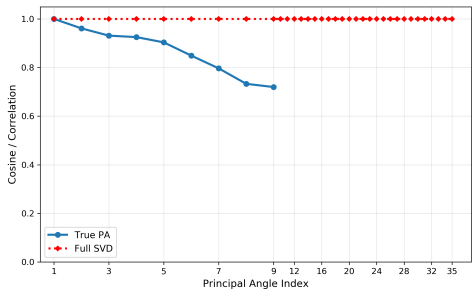
Combining Theorem 2 with (8) yields

$$\left|\cos\theta_{q+i}(\hat{U}_0, \hat{U}_1) - \cos\theta_i(\mathcal{U}_{P^0}, \mathcal{U}_{P^1})\right| \le \frac{\varepsilon_0}{\sqrt{1+\varepsilon_0^2}} + \frac{\varepsilon_1}{\sqrt{1+\varepsilon_1^2}} \le 2c, \qquad i = 1, \dots, r.$$

Thus, under the Ideal Model, the principal-angle geometry of the theoretical gradient spaces provides a direct approximation to the latent transfer geometry.



3×3, windless vs. windy



3×3, horizontal vs. vertical wind

Figure 2: Noise-induced collapse of Full-SVD principal-angle estimation.

4 Subspace Recovery Failure beyond the Ideal Model

Section 3 shows that gradient observations recover the Theoretical Gradient Spaces under an Ideal Model. In practice, off-subspace perturbations can inflate the observed gradient dimension, causing subspace recovery failure and full-SVD degeneration. This section characterizes this failure.

4.1 Empirical Failure of Naive Spectral Subspace Estimation

We first evaluate whether full-SVD estimates recover the latent transfer geometry in realistic environments. We use WindyGridWorld (Sutton & Barto, 2018) with a 3×3 grid, yielding $|\mathcal{S}| = 9$ and $|\mathcal{A}| = 4$ (up, down, left, and right), and set the discount factor to $\gamma = 0.99$. In these environments, the agent follows the intended transition with probability $1-\beta$. With probability β, the wind perturbs the transition by displacing the agent one step along the wind direction, where $\beta \in [0, 1]$ controls the wind strength. For each environment, we collect sampled gradient matrices $G_k = [g_1^k, \ldots, g_n^k]$ and sample size $n = 10 \times |\mathcal{S}| \cdot |\mathcal{A}| = 360$. Each gradient column corresponds to a sampled reward parameter under the standard tabular parameterization $r_\theta = \theta$, for which $J_\theta = I$, and is estimated from 100 Monte Carlo trajectories with horizon 100, without explicit transition-model or occupancy estimation. Two types of environment are considered: windless (G_0) vs. windy (G_1) environments, and horizontal (G_0) vs. vertical (G_1) wind, with $\beta = 0.5$.

Figure 2 compares the ground-truth principal-angle cosines between $\mathcal{U}_{P^0}$ and $\mathcal{U}_{P^1}$ (blue) with the **full-SVD** estimates (red), using all $d - 1 = 35$ left singular vectors retained for singular values exceeding the threshold $10^{-10}\sigma_1(G_k)$. For the windless vs. windy and horizontal vs. vertical pairs, the $(\sigma_{35}, \sigma_{36})$ are $(2.44 \times 10^{-3}, 9.55 \times 10^{-15})$ and $(3.51 \times 10^{-3}, 4.37 \times 10^{-14})$ for G_0, respectively, and $(3.56 \times 10^{-3}, 1.45 \times 10^{-15})$ for G_1 in both cases; all 35 estimated cosines collapse to one.

4.2 Noisy Gradient Model

Under the Ideal Model, the gradients satisfying the shaping invariance property are given by $G_k^{\mathrm{id}} = U_k A_k$, which preserves the intrinsic dimension of $\mathcal{G}_{P^k}^{\mathrm{theo}}$: $\mathrm{rank}(G_k^{\mathrm{id}}) = m$, $\mathrm{span}(G_k^{\mathrm{id}}) = \mathcal{G}_{P^k}^{\mathrm{theo}}$.

In practical IRL training, however, the actually collected gradient matrix $G_k \in \mathbb{R}^{d \times n}$ may deviate from this ideal low-dimensional structure. Such deviations can arise from approximate gradient estimation or numerical errors that do not exactly preserve the shaping invariance property. For instance, Monte Carlo approximations used to estimate the objective gradient may introduce off-subspace errors and inflate the observed rank beyond the intrinsic dimension: $\mathrm{rank}(G_k) > m$.

Such rank inflation can introduce spurious alignments and distort the estimated principal-angle geometry. The following theorem quantifies this effect; its proof is deferred to Appendix E.

Theorem 3 (Degeneration under Rank Inflation) *Let $\mathcal{H} \subseteq \mathbb{R}^d$ be an h-dimensional ambient subspace, and let $G_0, G_1 \in \mathbb{R}^{d \times n}$ have ranks ℓ_0, ℓ_1, respectively, with $\mathrm{span}(G_k) \subseteq \mathcal{H}$ for $k \in \{0, 1\}$. Denote $s := \max\{0, \ell_0 + \ell_1 - h\}$. Then the first s principal-angle cosines between $\mathrm{span}(G_0)$ and $\mathrm{span}(G_1)$ satisfy*

$$\rho_i^{\mathrm{obs}} = 1, \qquad i = 1, \ldots, s.$$

To characterize how such rank inflation arises, we consider the following signal–deviation model:

$$G_k = U_k \Sigma_k A_k + \epsilon_k, \qquad k \in \{0, 1\}, \tag{11}$$

where $U_k \in \mathbb{R}^{d\times m}$ spans $\mathcal{G}_{P^k}^{\text{theo}}$, $\Sigma_k \in \mathbb{R}^{m\times m}$ captures anisotropic scaling, $A_k \in \mathbb{R}^{m\times n}$ contains the coefficients, and $\epsilon_k \in \mathbb{R}^{d\times n}$ collects observation perturbations. The signal $U_k\Sigma_k A_k$ remains confined to the intrinsic m-dimensional gradient space and reduces to the Ideal Model $G_k^{\text{id}} = U_k A_k$ when $\Sigma_k = I_m$, whereas off-subspace components of ϵ_k can inflate the observed rank beyond m. For continuously distributed ϵ_k with $n \geq d$, G_k can become full row rank almost surely. Taking $\mathcal{H} = \mathbb{R}^d$ and $h = d$, Theorem 3 gives $\ell_0 = \ell_1 = d$, so all d principal-angle cosines equal one.

The WindyGridWorld experiments exhibit the same mechanism. Each observed gradient is the difference of two finite-horizon discounted occupancy estimates, whose total discounted mass satisfies

$$\mathbf{1}^\top \widehat{\mu} = (1-\gamma)\sum_{t=0}^{H-1}\gamma^t = 1 - \gamma^H.$$

Hence $\mathbf{1}^\top g = 0$, so the gradients lie in $\mathcal{H} = \mathbf{1}^\perp$, with $h = d-1$, implying $\operatorname{rank}(G_k) \leq d-1$. In our experiments, both G_0 and G_1 have numerical rank $d-1$, so the gradient spaces saturate $\mathbf{1}^\perp$, and all $d-1$ principal-angle cosines collapse to one, consistent with Theorem 3.

Overall, rank inflation can make direct SVD-based principal-angle estimation unreliable.

5 Dimension-Calibrated Recovery of Transfer Geometry

The preceding analysis reveals two dimension mismatches in recovering the latent transfer geometry. First, the r-dimensional PS Spaces correspond to m-dimensional Identifiable Reward Spaces, introducing q structural zero angles (Proposition 1). Second, observation deviations can inflate the intrinsic gradient dimension m to empirical ranks $\ell_k > m$, forcing at least $s = \max\{0, \ell_0 + \ell_1 - h\}$ zero-angle alignments (Theorem 3).

DC-TSVD corrects these mismatches in two stages: it first truncates each observed gradient matrix to the theoretically prescribed dimension m, suppressing excess directions induced by rank inflation, and then discards the first $q = m - r$ complement-induced structural zero angles by index to recover the r-dimensional latent transfer geometry. Both calibrations are determined by the theoretical dimensions, requiring neither empirical rank selection nor angle thresholding.

We analyze perturbed-observation subspace recovery in the first stage. For each environment $k \in \{0, 1\}$, let $G_k^\star := U_k\Sigma_k A_k$ denote the rank-m latent component. Assume $\operatorname{rank}(G_k^\star) = m$ and define

$$\tau_k := \sigma_m(G_k^\star) > 0 \tag{12}$$

as its smallest nonzero singular value.

Assumption 3 (Signal–Perturbation Separation) *For each $k \in \{0, 1\}$, assume that there exists a constant $c < 1$ such that*

$$\|\epsilon_k\|_2/\tau_k \leq c.$$

Define the observation-induced perturbation factor for m-dimensional subspace recovery as

$$\xi_k := \frac{\|\epsilon_k\|_2}{\tau_k - \|\epsilon_k\|_2} \leq \frac{c}{1-c}. \tag{13}$$

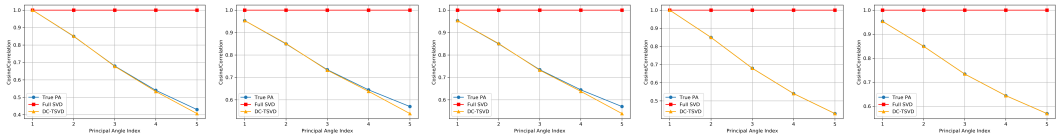
Thus, smaller c yields smaller ξ_k. Let $C_{\text{theo}} := U_0^\top U_1$, so that $\sigma_j(C_{\text{theo}}) = \rho_j^{\text{grad}}$, $j = 1, \ldots, m$. The following result characterizes Stage-1 subspace recovery; the proof is provided in Appendix F.

Theorem 4 (Principal-Angle Recovery under Rank-m Truncated SVD) *For each $k \in \{0, 1\}$, let $G_k = \hat{U}_k\hat{\Sigma}_k\hat{V}_k^\top$ be the full SVD of G_k, and let $\hat{U}_{k,m} \in \mathbb{R}^{d\times m}$ contain only the leading m left singular vectors, i.e., the rank-m truncated-SVD basis. Define the truncated-SVD principal-angle cosines by*

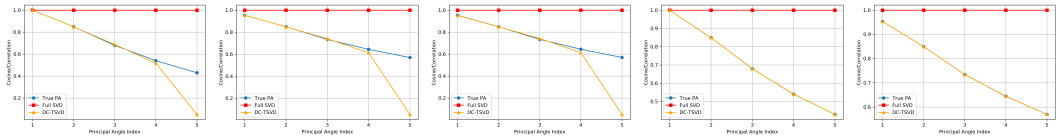
$$\hat{\rho}_j = \sigma_j(\hat{U}_{0,m}^\top \hat{U}_{1,m}), \qquad j = 1, \ldots, m.$$

Under Assumption 3,

$$|\hat{\rho}_j - \rho_j^{\text{grad}}| \leq \sqrt{2}(\xi_0 + \xi_1) + 2\xi_0\xi_1, \qquad j = 1, \ldots, m. \tag{14}$$



(a) $\Sigma_1,\ \boldsymbol{\rho}_1$ (b) $\Sigma_1,\ \boldsymbol{\rho}_2$ (c) $\Sigma_1',\ \boldsymbol{\rho}_2$ (d) $\Sigma_2,\ \boldsymbol{\rho}_1$ (e) $\Sigma_2,\ \boldsymbol{\rho}_2$



(f) $\Sigma_1,\ \boldsymbol{\rho}_1$ (g) $\Sigma_1,\ \boldsymbol{\rho}_2$ (h) $\Sigma_1',\ \boldsymbol{\rho}_2$ (i) $\Sigma_2,\ \boldsymbol{\rho}_1$ (j) $\Sigma_2,\ \boldsymbol{\rho}_2$

Figure 3: Principal-angle recovery under different spectral configurations and noise levels. The top row corresponds to $\epsilon_k \sim \mathcal{N}(0, 0.01^2)$, while the bottom row corresponds to $\epsilon_k \sim \mathcal{N}(0, 0.5^2)$.

The bound in (14) for full m-dimensional recovery is governed by the weakest signal component through τ_k. This suggests that weak signal directions are susceptible to perturbations, while stronger directions may exhibit stable recovery; a partial-subspace extension would instead depend on their spectral separation. Theorem 4 provides a deterministic perturbation guarantee, with no asymptotic rate in n; empirically, the subspace discrepancy decreases as n increases (Table 1 in Appendix H).

The intersection-preservation property in (4) preserves the multiplicity of unit principal-angle cosines, ensuring that the q dimension-mismatch-induced structural unit cosines remain identifiable separately from the transfer-relevant angles. Combining Theorem 4 with (8) yields

$$\left|\hat{\rho}_{q+j} - \rho_j^{\mathrm{PS}}\right| \le \sqrt{2}(\xi_0 + \xi_1) + 2\xi_0\xi_1 + \frac{\varepsilon_0}{\sqrt{1+\varepsilon_0^2}} + \frac{\varepsilon_1}{\sqrt{1+\varepsilon_1^2}}, \quad j = 1, \dots, r. \tag{15}$$

The bound separates two sources of estimation error: observation-induced perturbation and Jacobian subspace leakage. The complete DC-TSVD procedure is summarized in Algorithm 1 (Appendix G).

6 Simulations and Experiments

In this section, we evaluate the accuracy of DC-TSVD in recovering the latent transfer geometry under the noisy gradient model (11), predicted by (15). We set $d = 200$, $m = 5$, and $n = 6000$. The additive noise is generated by $(\epsilon_k)_{ij} \sim \mathcal{N}(0, 0.01^2)$ and $\mathcal{N}(0, 0.5^2)$, while the coefficient matrices are sampled as $(A_k)_{ij} \sim \mathcal{N}(0, 1)$. We consider three typical singular-value spectra:

$$\Sigma_1 = \mathrm{diag}(25, 10, 2, 0.5, 0.01),\ \Sigma_1' = \mathrm{diag}(2, 10, 25, 0.5, 0.01),\ \Sigma_2 = \mathrm{diag}(10, 10, 10, 10, 10).$$

Here, Σ_1 represents an anisotropic signal spectrum, Σ_1' redistributes signal spectral strengths, and Σ_2 represents an isotropic spectrum. We consider two ground-truth principal-angle cosine spectra

$$\boldsymbol{\rho}_1 = [1.00, 0.85, 0.70, 0.55, 0.40], \quad \boldsymbol{\rho}_2 = [0.95, 0.85, 0.75, 0.65, 0.55]. \tag{16}$$

For each configuration, we generate subspaces U_0 and U_1 with the prescribed principal-angle structure (16) and construct G_0 and G_1. As shown in Figure 3, where the horizontal axis indexes the principal angles, Full-SVD (red) collapses all estimated principal-angle cosines to one across all cases, reproducing the same degeneration observed in WindyGridWorld. In contrast, DC-TSVD (yellow) closely recovers the prescribed ground-truth principal-angle cosines (blue) across different spectral configurations; under $\epsilon_k \sim \mathcal{N}(0, 0.5^2)$, only the weakest signal component (0.01 in Σ_1) is overwhelmed by noise, while the remaining signal directions are accurately recovered.

Note that this simulation does not consider structural dimension mismatch, i.e., we set $m = r$, thereby isolating Stage 1 recovery of the PS-space principal angles from noisy gradient observations.

The complete recovery pipeline is further evaluated on WindyGridWorld (Sutton & Barto, 2018), which involves a substantial structural dimension mismatch between m and r. We consider isotropic and anisotropic Jacobian transformations, multiple wind strengths, and two distinct transfer geometries: windless vs. windy pairs with multiple zero-angle alignments, and horizontal vs. vertical pairs with only one. Gradient samples are obtained across varying reward parameters while the Jacobian-induced identifiable image remains invariant. We further examine the importance of the structurally determined dimension m for geometric recovery. Experimental details are provided in Appendix H.

AI USE STATEMENT

In this work, we used generative AI tools primarily to improve the readability and presentation of the manuscript, including language polishing, rephrasing, and suggestions on the organization and structure of the paper. We also used generative AI tools in a feedback capacity, including providing critical comments on research methodology and experimental presentation, as well as reviewing author-developed mathematical derivations and proofs for potential errors or inconsistencies.

Generative AI tools were not used to originate the core research idea or develop the central theoretical framework. The theoretical results and proof arguments were developed by the authors, while generative AI was used only in the feedback and verification capacity described above. The final experimental design decisions and scientific conclusions were also determined by the authors.

All AI-assisted suggestions were critically reviewed by the authors. We take full responsibility for the final content of this work, including text, claims, proofs, experiments, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide complete proofs of the theoretical results, detailed experimental settings, implementation details, and additional experimental results in the appendix.

REFERENCES

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, pp. 1, 2004.

Kareem Amin, Nan Jiang, and Satinder Singh. Repeated inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Haoyang Cao, Samuel N. Cohen, and Łukasz Szpruch. Identifiability in inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 12362–12373, 2021.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. In *International Conference on Learning Representations (ICLR)*, 2026.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2018.

Aurél Galántai. *Projectors and Projection Methods*. Springer Science & Business Media, 2003.

Tanmay Gangwani and Jian Peng. State-only imitation with transition dynamics mismatch. In *International Conference on Learning Representations (ICLR)*, 2020.

Babatunde H. Giwa and Chi-Guhn Lee. A marginal log-likelihood approach for the estimation of discount factors of multiple experts in inverse reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7786–7791. IEEE, 2021.

Filippo Lazzati, Mirco Mutti, and Alberto Maria Metelli. Offline inverse RL: New solution concepts and provably efficient algorithms. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 26085–26151. PMLR, 2024.

Xinhu Li, Ayush Jain, Zhaojing Yang, Yigit Korkmaz, and Erdem Bıyık. When a robot is more capable than a human: Learning from constrained demonstrators. In *International Conference on Learning Representations (ICLR)*, 2026.

Amarildo Likmeta, Alberto Maria Metelli, Giorgia Ramponi, Andrea Tirinzoni, Matteo Giuliani, and Marcello Restelli. Dealing with multiple experts and non-stationarity in inverse reinforcement learning: an application to real-life problems. *Machine Learning*, 110:2541–2576, 2021.

Alberto Maria Metelli, Giorgia Ramponi, Alessandro Concetti, and Marcello Restelli. Provably efficient learning of transferable rewards. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 7665–7676. PMLR, 2021.

Alberto Maria Metelli, Filippo Lazzati, and Marcello Restelli. Towards theoretical understanding of inverse reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 24555–24591. PMLR, 2023.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pp. 663–670, 2000.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pp. 278–287, 1999.

Riccardo Poiani, Gabriele Curti, Alberto Maria Metelli, and Marcello Restelli. Sub-optimal experts mitigate ambiguity in inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 85778–85823, 2024.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

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

Andreas Schlaginhaufen and Maryam Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 30224–30251. PMLR, 2023.

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

Mohamad Louai Shehab, Antoine Aspeel, Nikos Arechiga, Andrew Best, and Necmiye Ozay. Learning true objectives: Linear algebraic characterizations of identifiability in inverse reinforcement learning. In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference (L4DC)*, volume 242, pp. 1266–1277. PMLR, 2024.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

Joar Skalse and Alessandro Abate. Partial identifiability and misspecification in inverse reinforcement learning. *Artificial Intelligence*, 356:104525, 2026.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 32033–32058, 2023.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Davide Tateo, Matteo Pirotta, Marcello Restelli, and Andrea Bonarini. Gradient-based minimization for multi-expert inverse reinforcement learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8. IEEE, 2017.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

Huaipin Zhang, Yingying Yang, Wei Zhao, and Dong Yue. Distributed data-driven inverse reinforcement learning for multi-agent systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 73:2892–2899, 2026.

Yirui Zhou, Lei Wang, Mengxiao Lu, Zhiyuan Xu, Jian Tang, Yangchun Zhang, and Yaxin Peng. Generalization error for portable rewards in transfer imitation learning. *Knowledge-Based Systems*, 300:112230, 2024.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of 23rd AAAI Conference on Artificial Intelligence*, pp. 1433–1438, 2008.

APPENDIX

A Illustration of the Risk of a Near-Zero Second Principal Angle **14**

B Behavioral Structure of Identifiable Reward Spaces **14**

B.1 Behavioral Asymmetry of Reward Components 15

B.2 Discussion 15

C Proofs of Theoretical Results in Section 3 **16**

C.1 Proof of Proposition 1 16

C.2 Injectivity and Structural Preservation of the Jacobian Mapping 17

C.3 Proof of Theorem 1 17

C.4 Proof of Theorem 2 20

D Validation of Gradient Subspace Recovery under the Ideal Model **20**

E Proof of Theorem 3 **21**

F Proof of Theorem 4 **21**

G Dimension-Calibrated Truncated SVD Algorithm **24**

H Real-Environment Validation **24**

I Limitations and Future Work **30**

A Illustration of the Risk of a Near-Zero Second Principal Angle

We provide an illustrative example showing how a near-zero second principal angle can arise and lead to unreliable reward transfer.

Consider a rescue robot task where a reward is learned from multiple source environments and then transferred to new environments. Suppose that the source environments share a behavior-preserving ambiguity, yielding a zero first principal angle between their Potential Shaping Spaces. This setting may arise, for example, from a common constant-shift ambiguity. The critical question is then whether an additional shared shaping direction exists, corresponding to a zero second principal angle. Excessive similarity between the source environments may cause this second principal angle to approach zero. For example, two flat terrains exhibiting similar dynamics may make different energy preferences behaviorally indistinguishable, resulting in an energy-related ambiguity direction shared by the Potential Shaping Spaces of both environments. A reward estimate perturbed along this ambiguous direction,

$$\hat{r} = r^\star + \epsilon r_{\text{energy}},$$

may remain behaviorally indistinguishable in the source environments, but its effect can be substantially amplified after transfer to an energy-intensive environment, causing excessive energy consumption and eventual failure.

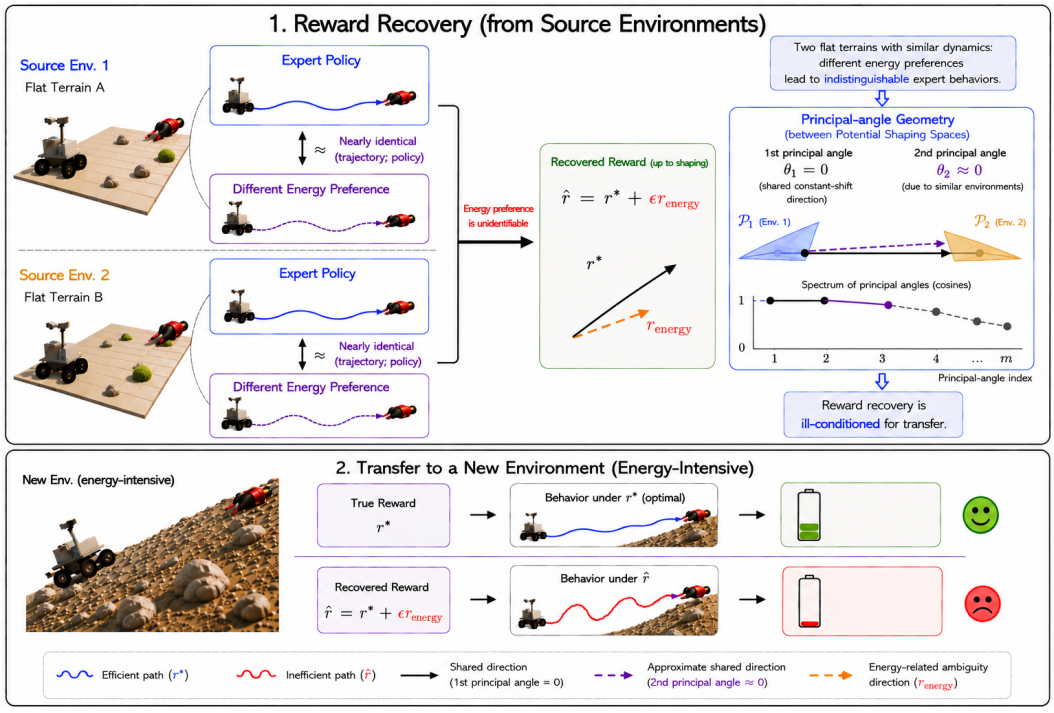


Figure 4: figure
Risk of a Near-Zero Second Principal Angle

B Behavioral Structure of Identifiable Reward Spaces

This section provides geometric intuition for Assumption 2. It does not claim that Jacobian-induced leakage is universally small; rather, it explains why the behavioral asymmetry between identifiable reward components and shaping ambiguities motivates local stability of the identifiable reward spaces.

B.1 Behavioral Asymmetry of Reward Components

The key observation is that Identifiable Reward and Potential Shaping components have fundamentally different behavioral effects in IRL.

Identifiable Reward Space. Consider an identifiable reward direction $v \in \mathcal{U}_P^{\perp}$. Perturbations along v can change relative reward preferences, for example through differences of the form

$$r(s, a_1) - r(s, a_2),$$

and can therefore alter action preferences and the induced policy. Identifiable reward components are thus directly exposed to the behavioral learning signal and constrained by the IRL objective.

Potential Shaping Space. In contrast, a shaping direction $u \in \mathcal{U}_P$ represents a behavior-preserving reward ambiguity. In particular, potential-based shaping transformations of the form

$$r' = r + \Phi - \gamma P \Phi$$

preserve the induced optimal policy,

$$\pi^{(r')} = \pi^{(r)}.$$

Variations along shaping components are therefore indistinguishable from behavior alone and are not subject to the same behavioral constraint.

This asymmetry can also be characterized locally through the curvature of the behavioral IRL objective. Let

$$H_r = \nabla_r^2 \mathcal{L}(r^\star),$$

where $r^\star$ is a stationary point. For a shaping direction $u \in \mathcal{U}_P$, behavioral invariance implies

$$\mathcal{L}(r^\star + tu) = \mathcal{L}(r^\star),$$

and hence

$$H_r u = 0, \qquad u \in \mathcal{U}_P. \tag{17}$$

By contrast, under nondegenerate local curvature on the Identifiable Reward Space,

$$v^\top H_r v \geq \mu \|v\|^2, \qquad v \in \mathcal{U}_P^{\perp}, \tag{18}$$

for some $\mu > 0$. Consequently, an identifiable perturbation $\Delta r = \epsilon v$ produces

$$\nabla_r \mathcal{L}(r^\star + \epsilon v) = \epsilon H_r v + o(\epsilon),$$

whereas a shaping perturbation produces no corresponding first-order behavioral response.

As shown in (17) and (18), the IRL objective itself distinguishes between $\mathcal{U}_P^{\perp}$ and $\mathcal{U}_P$ through its local curvature, showing that this separation is intrinsic to the IRL objective rather than an artificial decomposition of the reward space.

Since the IRL objective sharply distinguishes identifiable from shaping variations, this behavioral separation motivates the expectation of limited Jacobian-induced displacement within the identifiable reward spaces under suitable parameterizations, thereby supporting the plausibility of a small c.

B.2 Discussion

While the IRL objective sharply distinguishes identifiable from shaping variations, formally establishing that this distinction controls Jacobian-induced displacement would require a deeper characterization of the interaction between the reward parameterization and the IRL learning dynamics. Nevertheless, this intuition is consistent with our empirical observations: the parameterization experiments in Appendix H (especially in Table 1) exhibit small principal-angle estimation errors even under strongly anisotropic reward parameterizations.

Understanding when and how IRL learning naturally induces such local identifiable-space stability is an interesting theoretical question that we leave for future work.

C Proofs of Theoretical Results in Section 3

C.1 Proof of Proposition 1

Let $X, Y \in \mathbb{R}^{D\times r}$ be orthonormal basis matrices for $\mathcal{U}_{P^0}$ and $\mathcal{U}_{P^1}$, and let $X_\perp, Y_\perp \in \mathbb{R}^{D\times m}$ be orthonormal basis matrices for $\mathcal{U}_{P^0}^\perp$ and $\mathcal{U}_{P^1}^\perp$, respectively. Define the cross-Gram matrix

$$A := X^\top Y.$$

By the definition of principal angles,

$$\sigma_i(A) = \rho_i^{\mathrm{PS}} = \cos\theta_i^{\mathrm{PS}}, \qquad i = 1, \dots, r.$$

Further define

$$B := X_\perp^\top Y, \qquad H := X_\perp^\top Y_\perp.$$

Since $[X, X_\perp] \in \mathbb{R}^{D\times D}$ is orthogonal, $XX^\top$ and $X_\perp X_\perp^\top$ are complementary orthogonal projectors, and hence

$$XX^\top + X_\perp X_\perp^\top = I_D.$$

It follows that

$$A^\top A + B^\top B = Y^\top (XX^\top + X_\perp X_\perp^\top) Y = Y^\top Y = I_r. \tag{19}$$

By (19), we have

$$B^\top B = I_r - A^\top A.$$

Recall that the eigenvalues of $A^\top A$ are $(\rho_i^{\mathrm{PS}})^2$, $i = 1, \dots, r$. For an eigenvector v_i associated with $(\rho_i^{\mathrm{PS}})^2$, we have

$$B^\top B v_i = (I_r - A^\top A) v_i = \big(1 - (\rho_i^{\mathrm{PS}})^2\big) v_i.$$

Hence, the eigenvalues of $B^\top B$ are

$$1 - (\rho_i^{\mathrm{PS}})^2, \qquad i = 1, \dots, r.$$

Moreover, $B^\top B$ and $BB^\top$ have the same nonzero eigenvalues with the same multiplicities. Since $B \in \mathbb{R}^{m\times r}$ and $m = r + q$, passing from the $r \times r$ matrix $B^\top B$ to the $m \times m$ matrix $BB^\top$ introduces exactly q additional zero eigenvalues. Hence, the spectrum of $BB^\top$ consists of

$$1 - (\rho_i^{\mathrm{PS}})^2, \qquad i = 1, \dots, r,$$

together with q additional zeros.

Similarly, since $[Y, Y_\perp] \in \mathbb{R}^{D\times D}$ is orthogonal,

$$YY^\top + Y_\perp Y_\perp^\top = I_D.$$

Therefore,

$$BB^\top + HH^\top = X_\perp^\top (YY^\top + Y_\perp Y_\perp^\top) X_\perp = X_\perp^\top X_\perp = I_m. \tag{20}$$

Combining (19) and (20), the eigenvalues of $HH^\top$ consist of

$$(\rho_i^{\mathrm{PS}})^2, \qquad i = 1, \dots, r,$$

together with q additional ones. Hence, the singular values of $H = X_\perp^\top Y_\perp$ are

$$\underbrace{1, \dots, 1}_{q \text{ additional}}, \rho_1^{\mathrm{PS}}, \dots, \rho_r^{\mathrm{PS}}.$$

Since these are precisely the principal-angle cosines between $\mathcal{U}_{P^0}^\perp$ and $\mathcal{U}_{P^1}^\perp$, we obtain

$$\rho_j^{\mathrm{comp}} = 1, \qquad j = 1, \dots, q, \qquad \rho_{q+i}^{\mathrm{comp}} = \rho_i^{\mathrm{PS}}, \qquad i = 1, \dots, r.$$

Equivalently,

$$\theta_j^{\mathrm{comp}} = 0, \qquad j = 1, \dots, q, \qquad \theta_{q+i}^{\mathrm{comp}} = \theta_i^{\mathrm{PS}}, \qquad i = 1, \dots, r,$$

which completes the proof.

C.2 Injectivity and Structural Preservation of the Jacobian Mapping

For notational convenience, let

$$\mathcal{V}_k := \mathcal{U}_{P^k}^{\perp}, \qquad k \in \{0, 1\}.$$

We first prove (3). By Assumption 1,

$$\mathcal{V}_0 + \mathcal{V}_1 \subseteq \mathrm{im}(J_\theta). \tag{21}$$

Using the standard identity

$$\ker(J_\theta^\top) = \mathrm{im}(J_\theta)^\perp, \tag{22}$$

together with (21), this gives

$$(\mathcal{V}_0 + \mathcal{V}_1) \cap \ker(J_\theta^\top) = \{0\}. \tag{23}$$

Thus, no nonzero vector in $\mathcal{V}_0 + \mathcal{V}_1$ is mapped to zero by $T(v) = J_\theta^\top v$, so T is injective on $\mathcal{V}_0 + \mathcal{V}_1$, and therefore on each $\mathcal{V}_k$. Since

$$\mathcal{G}_{P^k}^{\mathrm{theo}} = T(\mathcal{V}_k)$$

and $\dim(\mathcal{V}_k) = m$, it follows that

$$\dim(\mathcal{G}_{P^k}^{\mathrm{theo}}) = \dim T(\mathcal{V}_k) = \dim(\mathcal{V}_k) = m, \qquad k \in \{0, 1\},$$

which proves (3).

Next, we prove (4). Since T is injective on $\mathcal{V}_0 + \mathcal{V}_1$, it preserves the intersection structure of $\mathcal{V}_0$ and $\mathcal{V}_1$. Specifically, for any $y \in T(\mathcal{V}_0) \cap T(\mathcal{V}_1)$, there exist $v_0 \in \mathcal{V}_0$ and $v_1 \in \mathcal{V}_1$ such that

$$y = T(v_0) = T(v_1).$$

Therefore,

$$T(v_0 - v_1) = 0.$$

Because $v_0 - v_1 \in \mathcal{V}_0 + \mathcal{V}_1$ and T is injective on this space, we have

$$v_0 = v_1 \in \mathcal{V}_0 \cap \mathcal{V}_1.$$

Hence,

$$T(\mathcal{V}_0) \cap T(\mathcal{V}_1) \subseteq T(\mathcal{V}_0 \cap \mathcal{V}_1).$$

The reverse inclusion follows directly from the definition of intersection, which proves

$$T(\mathcal{V}_0) \cap T(\mathcal{V}_1) = T(\mathcal{V}_0 \cap \mathcal{V}_1).$$

C.3 Proof of Theorem 1

We first show that the within-space transformation $\mathcal{A}_k$ defined in (6) is invertible on $\mathcal{V}_k$. Recall the polar decomposition

$$J_\theta^\top = U_\theta T_\theta, \qquad T_\theta = (J_\theta J_\theta^\top)^{1/2}.$$

For any $v \in \mathcal{V}_k$, since $P_k v = v$,

$$v^\top \mathcal{A}_k v = v^\top P_k T_\theta P_k v = v^\top T_\theta v. \tag{24}$$

Because $T_\theta \succeq 0$,

$$v^\top T_\theta v = \|T_\theta^{1/2} v\|_2^2,$$

so

$$v^\top T_\theta v = 0 \quad \Longrightarrow \quad T_\theta v = 0. \tag{25}$$

Moreover, since $T_\theta = (J_\theta J_\theta^\top)^{1/2}$,

$$\ker(T_\theta) = \ker(J_\theta^\top). \tag{26}$$

By (23), (26), and $\mathcal{V}_k \subseteq \mathcal{V}_0 + \mathcal{V}_1$, we derive

$$\mathcal{V}_k \cap \ker(T_\theta) = \{0\}.$$

Therefore, by (24) and (25),

$$v^\top \mathcal{A}_k v = 0 \quad \Longrightarrow \quad v = 0, \qquad v \in \mathcal{V}_k.$$

Since $T_\theta \succeq 0$, (24) also implies

$$v^\top \mathcal{A}_k v \geq 0.$$

Hence,

$$v^\top \mathcal{A}_k v > 0, \qquad \forall\, v \in \mathcal{V}_k \setminus \{0\}.$$

Thus, $\mathcal{A}_k$ is positive definite on $\mathcal{V}_k$, and therefore invertible on $\mathcal{V}_k$.

We next characterize the displacement of $\mathcal{V}_k$ under T_θ. For $v \in \mathcal{V}_k$,

$$T_\theta v = P_k T_\theta P_k v + (I - P_k) T_\theta P_k v = \mathcal{A}_k v + (I - P_k) T_\theta P_k v.$$

Since $\mathcal{A}_k$ is invertible, let

$$u = \mathcal{A}_k v \in \mathcal{V}_k, \qquad v = \mathcal{A}_k^{-1} u.$$

Then we have

$$T_\theta v = u + (I - P_k) T_\theta P_k \mathcal{A}_k^{-1} u.$$

Define

$$L_k = (I - P_k) T_\theta P_k \mathcal{A}_k^{-1}.$$

Since $I - P_k$ is the orthogonal projector onto $\mathcal{V}_k^\perp$, for any $u \in \mathcal{V}_k$,

$$L_k u = (I - P_k) T_\theta P_k \mathcal{A}_k^{-1} u \in \mathrm{im}(I - P_k) = \mathcal{V}_k^\perp.$$

Moreover, by (7),

$$\|L_k\|_2 = \left\|(I - P_k) T_\theta P_k \mathcal{A}_k^{-1}\right\|_2 = \varepsilon_k.$$

Therefore,

$$T_\theta \mathcal{V}_k = \{u + L_k u : u \in \mathcal{V}_k\}. \tag{27}$$

Let $\widehat{P}_k$ denote the orthogonal projector onto $T_\theta \mathcal{V}_k$. From (27) and $L_k u \in \mathcal{V}_k^\perp$, we have

$$P_k(u + L_k u) = u, \tag{28}$$

$$(I - P_k)(u + L_k u) = L_k u. \tag{29}$$

Thus, for $x = u + L_k u \in T_\theta \mathcal{V}_k$, $P_k x$ is its orthogonal projection onto $\mathcal{V}_k$, while $(I - P_k)x$ is the corresponding orthogonal residual. The relative magnitude of this residual therefore gives the sine of the angular deviation of x from $\mathcal{V}_k$. Consequently, the largest principal angle between $T_\theta \mathcal{V}_k$ and $\mathcal{V}_k$ satisfies

$$\begin{aligned} \sin \Theta_{\max}(T_\theta \mathcal{V}_k, \mathcal{V}_k) &= \sup_{x \in T_\theta \mathcal{V}_k \setminus \{0\}} \frac{\|(I - P_k)x\|_2}{\|x\|_2} \\ &= \sup_{u \in \mathcal{V}_k \setminus \{0\}} \frac{\|(I - P_k)(u + L_k u)\|_2}{\|u + L_k u\|_2}. \end{aligned} \tag{30}$$

Substituting (29) into (30) and using $u \perp L_k u$, we obtain

$$\sin \Theta_{\max}(T_\theta \mathcal{V}_k, \mathcal{V}_k) = \sup_{u \in \mathcal{V}_k \setminus \{0\}} \frac{\|L_k u\|_2}{\sqrt{\|u\|_2^2 + \|L_k u\|_2^2}}.$$

Dividing the numerator and denominator by $\|u\|_2$ gives

$$\sin \Theta_{\max}(T_\theta \mathcal{V}_k, \mathcal{V}_k) = \sup_{u \in \mathcal{V}_k \setminus \{0\}} \frac{\|L_k u\|_2 / \|u\|_2}{\sqrt{1 + (\|L_k u\|_2 / \|u\|_2)^2}}.$$

Since

$$\sup_{u \in \mathcal{V}_k \setminus \{0\}} \frac{\|L_k u\|_2}{\|u\|_2} = \|L_k\|_2$$

and $x/\sqrt{1+x^2}$ is increasing for $x \geq 0$, it follows from (7) that

$$\sin\Theta_{\max}\left(T_\theta \mathcal{V}_k, \mathcal{V}_k\right) = \frac{\|L_k\|_2}{\sqrt{1+\|L_k\|_2^2}} = \frac{\varepsilon_k}{\sqrt{1+\varepsilon_k^2}}. \tag{31}$$

Since $\mathcal{A}_k$ is invertible, $T_\theta \mathcal{V}_k$ and $\mathcal{V}_k$ have the same dimension. For two subspaces of equal dimension, the spectral norm of the difference between their orthogonal projectors equals the sine of their largest principal angle. Therefore, using (31),

$$\|\widehat{P}_k - P_k\|_2 = \sin\Theta_{\max}\left(T_\theta \mathcal{V}_k, \mathcal{V}_k\right) = \frac{\varepsilon_k}{\sqrt{1+\varepsilon_k^2}}.$$

We next relate the transformed subspaces to the Theoretical Gradient Spaces. From the polar decomposition,

$$\mathcal{G}_{P^k}^{\text{theo}} = J_\theta^\top \mathcal{V}_k = U_\theta T_\theta \mathcal{V}_k.$$

Since U_θ is isometric on $\text{im}(T_\theta)$, for any $x, y \in \text{im}(T_\theta)$,

$$\langle U_\theta x, U_\theta y\rangle = \langle x, y\rangle, \qquad \|U_\theta x\|_2 = \|x\|_2.$$

Moreover,

$$T_\theta \mathcal{V}_0,\, T_\theta \mathcal{V}_1 \subseteq \text{im}(T_\theta).$$

Hence, U_θ preserves all normalized inner products between vectors in the transformed subspaces, and therefore preserves their principal angles. Consequently,

$$\theta_i\left(\mathcal{G}_{P^0}^{\text{theo}}, \mathcal{G}_{P^1}^{\text{theo}}\right) = \theta_i\left(T_\theta \mathcal{V}_0, T_\theta \mathcal{V}_1\right), \qquad i = 1, \ldots, m.$$

It remains to compare the principal-angle cosines between the original and transformed subspaces.

Let V_k and $\widehat{V}_k$ be orthonormal bases of $\mathcal{V}_k$ and $T_\theta \mathcal{V}_k$, respectively. The corresponding principal-angle cosine spectra are given by

$$\rho_i^{\text{comp}} = \cos\theta_i^{\text{comp}} = \sigma_i(V_0^\top V_1),$$

and

$$\rho_i^{\text{grad}} = \cos\theta_i^{\text{grad}} = \sigma_i(\widehat{V}_0^\top \widehat{V}_1), \qquad i = 1, \ldots, m.$$

The orthogonal projection matrices onto the two pairs of subspaces are

$$P_k = V_k V_k^\top, \qquad \widehat{P}_k = \widehat{V}_k \widehat{V}_k^\top.$$

Therefore,

$$\begin{aligned} P_0 P_1 &= V_0 V_0^\top V_1 V_1^\top = V_0 (V_0^\top V_1) V_1^\top, \\ \widehat{P}_0 \widehat{P}_1 &= \widehat{V}_0 \widehat{V}_0^\top \widehat{V}_1 \widehat{V}_1^\top = \widehat{V}_0 (\widehat{V}_0^\top \widehat{V}_1) \widehat{V}_1^\top. \end{aligned}$$

Since V_k and $\widehat{V}_k$ have orthonormal columns, left and right multiplication by these matrices preserves the nonzero singular values. Hence,

$$\rho_i^{\text{comp}} = \sigma_i(P_0 P_1), \qquad \rho_i^{\text{grad}} = \sigma_i(\widehat{P}_0 \widehat{P}_1), \qquad i = 1, \ldots, m.$$

Applying Weyl's perturbation inequality for singular values yields

$$\left|\rho_i^{\text{grad}} - \rho_i^{\text{comp}}\right| \leq \left\|\widehat{P}_0 \widehat{P}_1 - P_0 P_1\right\|_2.$$

Furthermore,

$$\begin{aligned} \left\|\widehat{P}_0 \widehat{P}_1 - P_0 P_1\right\|_2 &= \left\|(\widehat{P}_0 - P_0)\widehat{P}_1 + P_0(\widehat{P}_1 - P_1)\right\|_2 \\ &\leq \|\widehat{P}_0 - P_0\|_2 + \|\widehat{P}_1 - P_1\|_2 \\ &= \frac{\varepsilon_0}{\sqrt{1+\varepsilon_0^2}} + \frac{\varepsilon_1}{\sqrt{1+\varepsilon_1^2}}. \end{aligned}$$

Therefore,

$$\left|\rho_i^{\text{grad}} - \rho_i^{\text{comp}}\right| \le \frac{\varepsilon_0}{\sqrt{1+\varepsilon_0^2}} + \frac{\varepsilon_1}{\sqrt{1+\varepsilon_1^2}}, \qquad i = 1, \dots, m.$$

Under Assumption 2, we have $\varepsilon_k \le c$, and therefore

$$\left|\rho_i^{\text{grad}} - \rho_i^{\text{comp}}\right| \le \varepsilon_0 + \varepsilon_1 \le 2c, \qquad i = 1, \dots, m,$$

which proves the result.

C.4 Proof of Theorem 2

Consider the singular value decomposition of A_k, $A_k = R_k \Lambda_k Q_k^\top$, where $R_k \in \mathbb{R}^{m\times m}$ and $Q_k \in \mathbb{R}^{n\times m}$ are orthonormal matrices, and $\Lambda_k \in \mathbb{R}^{m\times m}$ is diagonal with positive singular values due to $\operatorname{rank}(A_k) = m$.

Substituting into $G_k^{\text{id}} = U_k A_k$ yields

$$G_k^{\text{id}} = U_k R_k \Lambda_k Q_k^\top.$$

Since

$$U_k^\top U_k = I_m, \qquad R_k^\top R_k = I_m,$$

we have

$$(U_k R_k)^\top (U_k R_k) = R_k^\top U_k^\top U_k R_k = I_m.$$

Therefore, $G_k^{\text{id}} = (U_k R_k)\Lambda_k Q_k^\top$ is a compact SVD of G_k^{id}, where $U_k R_k$ provides an orthonormal basis for the left singular subspace of G_k^{id}.

Let $\hat{U}_k = U_k R_k$. Because R_k is an orthogonal matrix, it only performs a rotation within the subspace spanned by U_k. Therefore,

$$\operatorname{span}(\hat{U}_k) = \operatorname{span}(U_k).$$

This equality holds regardless of the particular choice of SVD basis. Next,

$$\hat{U}_0^\top \hat{U}_1 = R_0^\top (U_0^\top U_1) R_1.$$

Since R_0 and R_1 are orthogonal matrices, left and right multiplication by $R_0^\top$ and R_1 does not alter singular values. Therefore,

$$\sigma_i\left(\hat{U}_0^\top \hat{U}_1\right) = \sigma_i\left(U_0^\top U_1\right), \qquad i = 1, \dots, m.$$

By the definition of principal angles,

$$\theta_i(\hat{U}_0, \hat{U}_1) = \theta_i(U_0, U_1), \qquad i = 1, \dots, m.$$

D Validation of Gradient Subspace Recovery under the Ideal Model

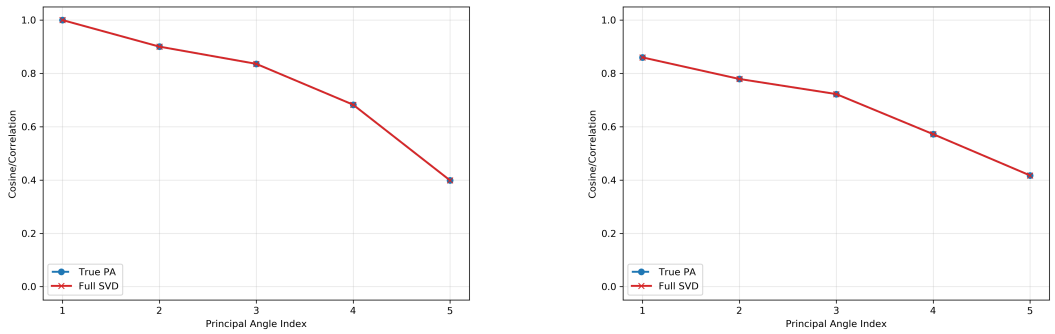
We first validate the recovery property of compact SVD under the Ideal Model. Recall the Ideal Model introduced in Section 3:

$$G_k^{\text{id}} = U_k A_k, \qquad k \in \{0, 1\}.$$

In our simulations, we set $d = 50$, $m = 5$, and the sample size to $n = 10000$. The coefficient matrices $A_0, A_1 \in \mathbb{R}^{m\times n}$ are independently generated, with their entries drawn i.i.d. from $\mathcal{N}(0, 1)$. We prescribe two principal-angle cosine spectra:

$$\boldsymbol{\rho}_1 = [1.00, 0.90, 0.80, 0.60, 0.40], \qquad \boldsymbol{\rho}_2 = [0.85, 0.75, 0.65, 0.55, 0.45].$$

We perform the same operations as in section 6 to construct G_0, G_1, and then apply compact SVD to the two gradient matrices and compute the principal-angle cosines between the recovered left singular subspaces. The blue curve denotes the prescribed ground-truth principal-angle cosines between U_0 and U_1, while the red curve represents the compact SVD estimates. As shown in Figure 5, the close agreement between the two curves verifies the theoretical equivalence under the Ideal Model.



$\boldsymbol{\rho}_1$ $\boldsymbol{\rho}_2$

Figure 5: Recovery of principal-angle geometry by compact SVD under the ideal gradient model.

E Proof of Theorem 3

Let $\mathcal{G}_k^{\text{obs}} := \text{span}(G_k)$, $k \in \{0, 1\}$. Since both spaces are contained in the same h-dimensional ambient subspace $\mathcal{H}$,

$$\dim(\mathcal{G}_0^{\text{obs}} \cap \mathcal{G}_1^{\text{obs}}) = \ell_0 + \ell_1 - \dim(\mathcal{G}_0^{\text{obs}} + \mathcal{G}_1^{\text{obs}}) \geq \ell_0 + \ell_1 - h.$$

Therefore,

$$\dim(\mathcal{G}_0^{\text{obs}} \cap \mathcal{G}_1^{\text{obs}}) \geq s := \max\{0, \ell_0 + \ell_1 - h\}.$$

Let Q_k be orthonormal bases of $\mathcal{G}_k^{\text{obs}}$. The shared s-dimensional intersection implies that $Q_0^\top Q_1$ has at least s singular values equal to one. Hence

$$\rho_i^{\text{obs}} = 1, \qquad i = 1, \ldots, s.$$

This completes the proof.

F Proof of Theorem 4

To prove Theorem 4, we first recall Wedin's classical sin-Θ theorem (Wedin, 1972), which controls the perturbation of singular subspaces under spectral separation.

Lemma 1 (Spectral-norm version of Wedin's sin-Θ theorem) *Let $A \in \mathbb{R}^{p\times q}$ have the singular value decomposition*

$$A = [U_1, U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} [V_1, V_2]^\top,$$

where $U_1 \in \mathbb{R}^{p\times m}$ spans the m-dimensional left singular subspace of interest. Let

$$\widehat{A} = A + E, \qquad E \in \mathbb{R}^{p\times q},$$

and let

$$\widehat{A} = [\widehat{U}_1, \widehat{U}_2] \begin{bmatrix} \widehat{\Sigma}_1 & 0 \\ 0 & \widehat{\Sigma}_2 \end{bmatrix} [\widehat{V}_1, \widehat{V}_2]^\top$$

be the corresponding perturbed SVD, where $\widehat{U}_1 \in \mathbb{R}^{p\times m}$.

If

$$\Delta := \sigma_{\min}(\Sigma_1) - \sigma_{\max}(\widehat{\Sigma}_2) > 0,$$

then

$$\left\|\sin\Theta(\widehat{U}_1, U_1)\right\|_2 \leq \frac{\|E\|_2}{\Delta}.$$

We now apply Lemma 1 to the noisy gradient model. For each $k \in \{0, 1\}$,

$$G_k = G_k^\star + \epsilon_k, \qquad G_k^\star = U_k \Sigma_k A_k \in \mathbb{R}^{d\times n},$$

where

$$U_k \in \mathbb{R}^{d\times m}, \qquad \Sigma_k \in \mathbb{R}^{m\times m}, \qquad A_k \in \mathbb{R}^{m\times n},$$

and

$$\operatorname{rank}(G_k^\star) = m.$$

Let

$$G_k = \widehat{U}_k \widehat{\Sigma}_k \widehat{V}_k^\top$$

be the full SVD of the observed gradient matrix, and let $\widehat{U}_{k,m} \in \mathbb{R}^{d\times m}$ contain its leading m left singular vectors.

Recall that

$$\tau_k = \sigma_m(G_k^\star) > 0$$

is the smallest nonzero singular value of the signal component. Since $G_k^\star$ has rank m,

$$\sigma_{m+1}(G_k^\star) = 0. \tag{32}$$

By Weyl's singular-value perturbation inequality, (32) implies

$$\sigma_{m+1}(G_k) \le \sigma_{m+1}(G_k^\star) + \|\epsilon_k\|_2 = \|\epsilon_k\|_2.$$

Therefore, the separation between the smallest signal singular value of $G_k^\star$ and the complementary singular values of G_k satisfies

$$\Delta_k := \tau_k - \sigma_{m+1}(G_k) \ge \tau_k - \|\epsilon_k\|_2.$$

Under Assumption 3, $\|\epsilon_k\|_2 < \tau_k$, and hence

$$\Delta_k > 0.$$

Applying Lemma 1 with

$$A = G_k^\star, \qquad \widehat{A} = G_k, \qquad E = \epsilon_k,$$

gives the following signal-subspace recovery bound.

Lemma 2 (Recovery of the signal subspace) *Under Assumption 3,*

$$\left\|\sin\Theta(\widehat{U}_{k,m}, U_k)\right\|_2 \le \frac{\|\epsilon_k\|_2}{\Delta_k} \le \frac{\|\epsilon_k\|_2}{\tau_k - \|\epsilon_k\|_2} = \xi_k, \qquad k \in \{0,1\}.$$

The following corollary converts the subspace recovery bound into an explicit decomposition of the estimated singular-vector basis.

Corollary 1 (Aligned decomposition of $\hat{U}_{k,m}$) *Under Lemma 2, there exists an orthogonal matrix $O_k \in \mathbb{R}^{m\times m}$ such that*

$$\hat{U}_{k,m} = U_k O_k + R_k,$$

where $R_k \in \mathbb{R}^{d\times m}$ satisfies

$$\|R_k\|_2 \le \sqrt{2}\,\xi_k.$$

Proof of Corollary 1. Let $U_k^\top \hat{U}_{k,m} = Z_k \cos\Theta_k W_k^\top$ be the SVD of $U_k^\top \hat{U}_{k,m}$, where $Z_k, W_k \in \mathbb{R}^{m\times m}$ are orthogonal and

$$\cos\Theta_k = \operatorname{diag}\big(\cos\theta_{k,1}, \dots, \cos\theta_{k,m}\big),$$

with $\theta_{k,1}, \dots, \theta_{k,m}$ denoting the principal angles between $\operatorname{span}(U_k)$ and $\operatorname{span}(\hat{U}_{k,m})$.

Define

$$O_k = Z_k W_k^\top, \qquad R_k = \hat{U}_{k,m} - U_k O_k.$$

Since U_k and $\hat{U}_{k,m}$ have orthonormal columns, we have

$$\begin{aligned} R_k^\top R_k &= (\hat{U}_{k,m} - U_k O_k)^\top (\hat{U}_{k,m} - U_k O_k) \\ &= 2I_m - \hat{U}_{k,m}^\top U_k O_k - O_k^\top U_k^\top \hat{U}_{k,m}. \end{aligned}$$

Using $U_k^\top \hat{U}_{k,m} = Z_k \cos\Theta_k W_k^\top$ and $O_k = Z_k W_k^\top$, we obtain

$$R_k^\top R_k = 2W_k(I_m - \cos\Theta_k)W_k^\top.$$

Therefore,

$$\|R_k\|_2^2 = 2 \max_{1\le i\le m} (1 - \cos\theta_{k,i}).$$

Since

$$1 - \cos\theta \le \sin^2\theta, \qquad \theta \in [0, \pi/2],$$

it follows that

$$\begin{aligned}\|R_k\|_2^2 &\le 2 \max_{1\le i\le m} \sin^2\theta_{k,i} \\ &= 2\big\|\sin\Theta(\hat{U}_{k,m}, U_k)\big\|_2^2.\end{aligned}$$

By Lemma 2,

$$\big\|\sin\Theta(\hat{U}_{k,m}, U_k)\big\|_2 \le \xi_k,$$

and hence

$$\|R_k\|_2 \le \sqrt{2}\,\xi_k.$$

This completes the proof.

Applying Corollary 1 to the two environments $k = 0, 1$, we obtain the following perturbation bound for the empirical cross-Gram matrix.

Lemma 3 (Cross-Gram estimation error) *Under Corollary 1, there exist orthogonal matrices $O_0, O_1 \in \mathbb{R}^{m\times m}$ such that*

$$\left\|\hat{U}_{0,m}^\top \hat{U}_{1,m} - O_0^\top C_{\text{theo}} O_1\right\|_2 \le \sqrt{2}(\xi_0 + \xi_1) + 2\xi_0\xi_1.$$

Proof of Lemma 3. By Corollary 1, we have

$$\hat{U}_{0,m} = U_0 O_0 + R_0, \qquad \hat{U}_{1,m} = U_1 O_1 + R_1,$$

where

$$O_0, O_1 \in \mathbb{R}^{m\times m}, \qquad R_0, R_1 \in \mathbb{R}^{d\times m},$$

with

$$O_k^\top O_k = I_m, \qquad \|R_k\|_2 \le \sqrt{2}\,\xi_k.$$

Using $C_{\text{theo}} = U_0^\top U_1$, we have

$$\begin{aligned}\hat{U}_{0,m}^\top \hat{U}_{1,m} &= (U_0 O_0 + R_0)^\top (U_1 O_1 + R_1) \\ &= O_0^\top C_{\text{theo}} O_1 + O_0^\top U_0^\top R_1 + R_0^\top U_1 O_1 + R_0^\top R_1.\end{aligned}$$

Therefore,

$$\begin{aligned}&\left\|\hat{U}_{0,m}^\top \hat{U}_{1,m} - O_0^\top C_{\text{theo}} O_1\right\|_2 \\ &\quad\le \|O_0^\top U_0^\top R_1\|_2 + \|R_0^\top U_1 O_1\|_2 + \|R_0^\top R_1\|_2.\end{aligned}$$

Since U_0, U_1, O_0, O_1 have unit spectral norm,

$$\left\|\hat{U}_{0,m}^\top \hat{U}_{1,m} - O_0^\top C_{\text{theo}} O_1\right\|_2 \le \|R_0\|_2 + \|R_1\|_2 + \|R_0\|_2\|R_1\|_2.$$

Using $\|R_k\|_2 \le \sqrt{2}\,\xi_k$, we obtain

$$\left\|\hat{U}_{0,m}^\top \hat{U}_{1,m} - O_0^\top C_{\text{theo}} O_1\right\|_2 \le \sqrt{2}(\xi_0 + \xi_1) + 2\xi_0\xi_1.$$

This completes the proof.

With Lemma 3 at hand, we now prove Theorem 4.

Proof of Theorem 4. Since O_0 and O_1 are orthogonal,

$$\sigma_i(O_0^\top C_{\mathrm{theo}} O_1) = \sigma_i(C_{\mathrm{theo}}) = \rho_i^{\mathrm{grad}}.$$

Moreover, by definition,

$$\hat{\rho}_i = \sigma_i(\hat{U}_{0,m}^\top \hat{U}_{1,m}).$$

Therefore, Weyl's inequality for singular values together with Lemma 3 gives

$$\begin{aligned} |\hat{\rho}_i - \rho_i^{\mathrm{grad}}| &\le \left\| \hat{U}_{0,m}^\top \hat{U}_{1,m} - O_0^\top C_{\mathrm{theo}} O_1 \right\|_2 \\ &\le \sqrt{2}(\xi_0 + \xi_1) + 2\xi_0\xi_1, \qquad i = 1, \dots, m. \end{aligned}$$

This completes the proof.

G Dimension-Calibrated Truncated SVD Algorithm

We now summarize the resulting estimation procedure. The dimension of the Theoretical Gradient Space is structurally determined by the Identifiable Reward Space as

$$m = |\mathcal{S}|(|\mathcal{A}| - 1).$$

We therefore truncate each observed gradient matrix G_k to its leading m left singular vectors $\hat{U}_{k,m}$. We refer to this procedure as **Dimension-Calibrated Truncated SVD (DC-TSVD)**. The estimated gradient-space principal-angle cosines are

$$\hat{\rho}_i = \sigma_i\left(\hat{U}_{0,m}^\top \hat{U}_{1,m}\right), \qquad i = 1, \dots, m.$$

As characterized in Remark 1, the first $|\mathcal{S}|(|\mathcal{A}| - 2)$ cosines correspond to dimension-induced zero angles arising from the orthogonal-complement representation. Removing these terms yields

$$\hat{\rho}_{|\mathcal{S}|(|\mathcal{A}|-2)+i}, \qquad i = 1, \dots, r,$$

as estimates of the transfer-relevant principal-angle cosines between $\mathcal{U}_{P^0}$ and $\mathcal{U}_{P^1}$. Algorithm 1 summarizes DC-TSVD. The method requires no spectral tuning parameter or empirical rank selection; its recovery guarantee instead relies on the signal–perturbation separation characterized in Theorem 4.

H Real-Environment Validation

In this section, we evaluate DC-TSVD on the WindyGridWorld environments considered in Section 4, where all principal-angle cosines estimated via standard SVD degenerate to one. We consider two environment pairs: windless vs. windy environments and horizontal vs. vertical wind environments, under different wind strengths $\beta \in \{0.5,\ 1\}$. Experiments are conducted on 3×3 ($|\mathcal{S}| = 9$) and 5×5 ($|\mathcal{S}| = 25$) grids with four actions, i.e., $|\mathcal{A}| = 4$.

Following the IRL procedure in (Schlaginhaufen & Kamgarpour, 2024), we generate an expert dataset of $N = 10^4$ trajectories for each environment and use it to estimate the IRL gradients. For each environment, we collect $n = 20 \times |\mathcal{S}||\mathcal{A}|$ gradient observations to form the observed gradient matrix used for principal-angle estimation. Here, N denotes the size of the expert demonstration dataset, whereas n denotes the number of gradient observations used for subspace estimation. Each gradient observation is estimated from $M = 100$ Monte Carlo trajectories with horizon $H = 100$.

Reward Parameterization. To examine the effect of reward parameterization, we consider the standard tabular parameterization $r_\theta = \theta$, for which $J_\theta = I$ and $\frac{\varepsilon_0}{\sqrt{1+\varepsilon_0^2}} + \frac{\varepsilon_1}{\sqrt{1+\varepsilon_1^2}} = 0$, together with anisotropic linear parameterizations

$$r_\theta = B(\eta)\theta, \qquad B(\eta) = QD(\eta)Q^\top,$$

where Q is a random orthogonal matrix and $D(\eta)$ is diagonal with entries uniformly spaced over $[1 - \eta, 1 + \eta]$. We consider $\eta \in \{0,\ 0.25,\ 0.5\}$, where $\eta = 0$ corresponds to the standard tabular parameterization and larger η induces stronger anisotropy. Reward parameters vary across gradient samples, while $J_\theta = B(\eta)$ is parameter-independent and hence preserves the Jacobian-induced identifiable image. All parameterizations satisfy the Local Reward Representation Assumption 1 since their Jacobians are full rank and $d = D$. The principal-angle estimation errors therefore capture observation-induced subspace perturbation and, for $\eta > 0$, Jacobian-induced relative subspace leakage.

Algorithm 1 Dimension-Calibrated Truncated SVD (DC-TSVD)

Require: Gradient matrices G_0, G_1, $|\mathcal{S}|$, $|\mathcal{A}|$.

1: Set the structural dimensions

$$m = |\mathcal{S}|(|\mathcal{A}| - 1), \qquad r = |\mathcal{S}|.$$

2: **for** $k \in \{0, 1\}$ **do**

3: Compute the SVD

$$G_k = \hat{U}_k \hat{\Sigma}_k \hat{V}_k^\top.$$

4: Retain the leading m left singular vectors

$$\hat{U}_{k,m} = \hat{U}_k[:, 1:m].$$

5: **end for**

6: Compute the cross-Gram matrix

$$\hat{C} = \hat{U}_{0,m}^\top \hat{U}_{1,m}.$$

7: Compute the gradient-space principal-angle cosines

$$\hat{\rho}_i = \sigma_i(\hat{C}), \qquad i = 1, \ldots, m.$$

8: Discard the first $m - r$ cosines, which correspond to the dimension-induced structural zero angles

9: **return**

$$\{\hat{\rho}_{m-r+i}\}_{i=1}^{r}.$$

Ground-Truth Geometry. The two environment pairs exhibit different shared structures. For the windless vs. windy pair, the 3×3 setting contains three shared directions ($\rho_1^{\mathrm{PS}} = \rho_2^{\mathrm{PS}} = \rho_3^{\mathrm{PS}} = 1$), while the 5×5 setting contains five shared directions ($\rho_1^{\mathrm{PS}} = \cdots = \rho_5^{\mathrm{PS}} = 1$). In contrast, the horizontal vs. vertical pair contains only one shared direction ($\rho_1^{\mathrm{PS}} = 1$) for both grid sizes.

These complementary settings allow us to evaluate whether DC-TSVD accurately estimates the second principal-angle cosine both when $\rho_2^{\mathrm{PS}} = 1$ and when $\rho_2^{\mathrm{PS}} < 1$. Importantly, the former corresponds to a zero second principal angle, at which the transfer-error amplification factor in (1) diverges.

This distinction is also consistent with the environment structure: the windless vs. windy pair provides relatively redundant behavioral constraints, whereas horizontal vs. vertical winds perturb the dynamics along different directions and thus provide more complementary information for resolving reward ambiguity.

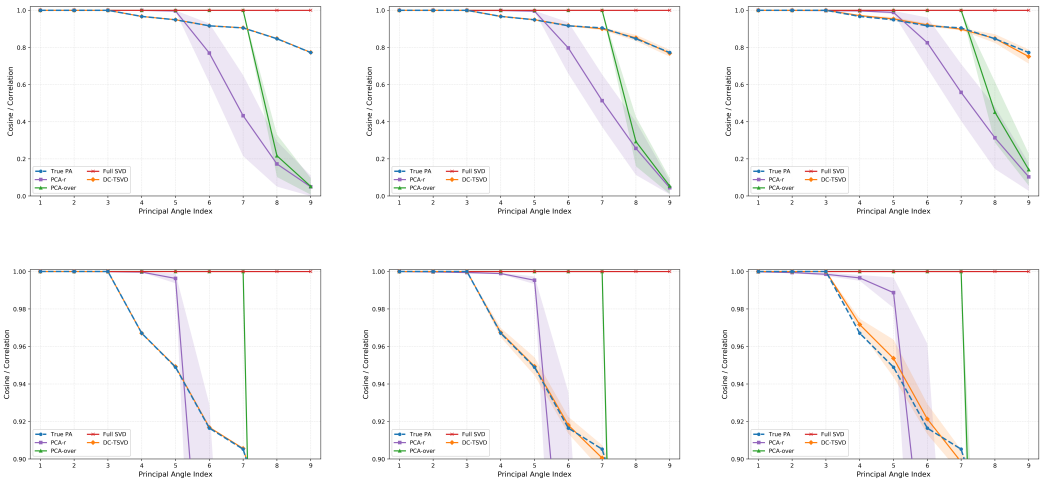
Baselines. We compare our **DC-TSVD** with three spectral baselines corresponding to different treatments of the full-rank gradient matrices.

Full SVD uses all left singular vectors associated with singular values exceeding $10^{-10}\sigma_1(G_k)$, without dimension-calibrated truncation.

PCA-over truncates each gradient matrix from dimension d to $\lceil 1.2m \rceil$, deliberately retaining more directions than the Identifiable Reward Space dimension m; principal angles are then computed between the truncated subspaces, with the last r angles retained for comparison with the latent PS-Space geometry.

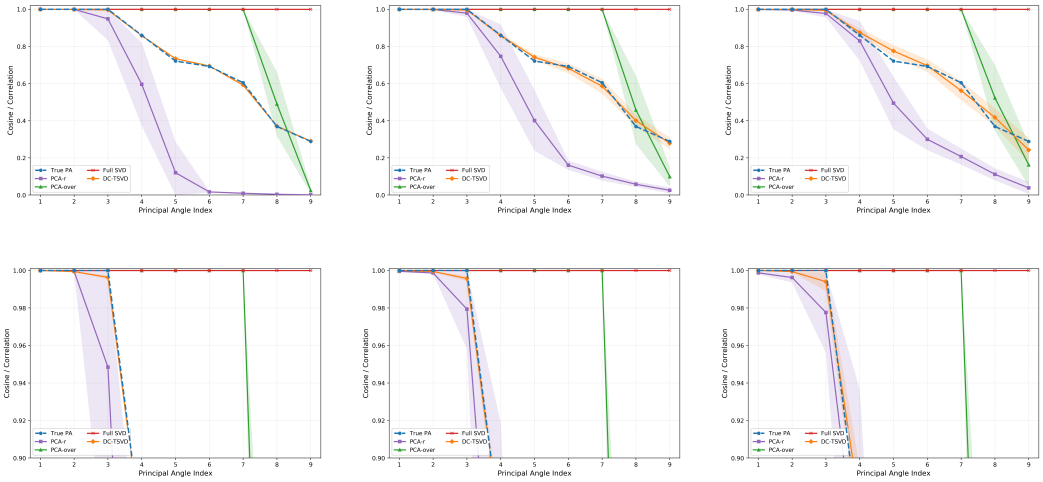
PCA-r directly truncates each gradient matrix from dimension d to r, matching the dimension of the latent PS Space.

Note that Schlaginhaufen and Kamgarpour (Schlaginhaufen & Kamgarpour, 2024) characterize principal-angle geometry from transition laws, but do not provide or experimentally evaluate a finite-sample pipeline for estimating this geometry through plug-in transition-law estimates. Moreover, such a comparison would require an additional transition-law estimator and sampling protocol, and is therefore not included as a direct empirical baseline.



(a) $\eta = 0$ (b) $\eta = 0.25$ (c) $\eta = 0.5$

Figure 6: 3×3 windless vs. windy WindyGridWorld with $\beta = 0.5$



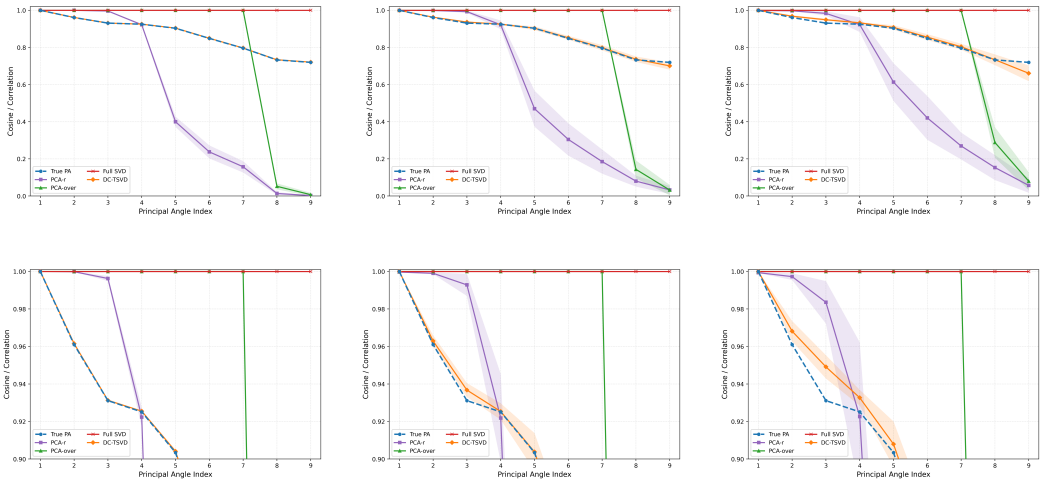
(a) $\eta = 0$ (b) $\eta = 0.25$ (c) $\eta = 0.5$

Figure 7: 3×3 windless vs. windy WindyGridWorld with $\beta = 1.0$

Evaluation. The comparison of all four methods against the ground-truth principal-angle cosines across different experimental settings is reported in Figures 6–13. For each setting, the upper panels show the complete principal-angle spectrum over $[0, 1]$, while the lower panels zoom in on $[0.9, 1]$ to highlight the differences among the principal-angle cosines close to one. Each experiment is repeated over 10 random seeds, and the curves show the mean with the corresponding standard deviation.

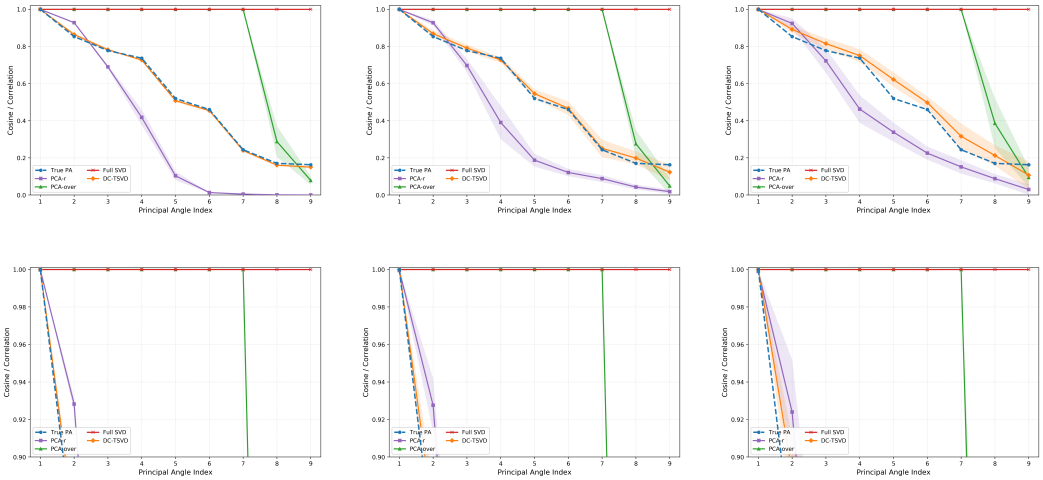
We find that across all settings, DC-TSVD accurately recovers the ground-truth principal-angle geometry, whereas the competing spectral baselines fail to reliably recover the latent geometry. Increasing reward-parameterization anisotropy leads to slightly larger estimation errors and variability for DC-TSVD, which nevertheless remain well controlled.

To further examine the two error components characterized by our theory, we report the total estimation error $|\hat{\rho}_{q+j} - \rho_j^{\mathrm{PS}}|$, the Jacobian-induced leakage error $|\rho_{q+j}^{\mathrm{grad}} - \rho_{q+j}^{\mathrm{comp}}|$, and the observation-induced error $|\hat{\rho}_{q+j} - \rho_{q+j}^{\mathrm{grad}}|$ under the 3×3 horizontal vs. vertical setting with $\beta = 0.5$ and $\eta = 0.5$. The latter two errors are controlled by the corresponding theoretical bounds: the Jacobian-induced



(a) $\eta = 0$ (b) $\eta = 0.25$ (c) $\eta = 0.5$

Figure 8: 3×3 horizontal vs. vertical wind WindyGridWorld with $\beta = 0.5$



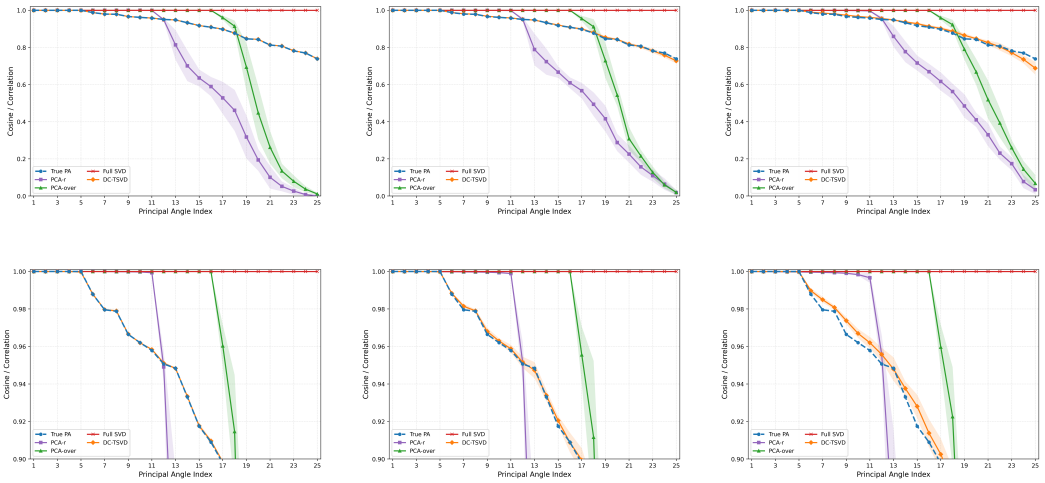
(a) $\eta = 0$ (b) $\eta = 0.25$ (c) $\eta = 0.5$

Figure 9: 3×3 horizontal vs. vertical wind WindyGridWorld with $\beta = 1.0$

bound $\frac{\varepsilon_0}{\sqrt{1+\varepsilon_0^2}} + \frac{\varepsilon_1}{\sqrt{1+\varepsilon_1^2}}$ and the observation-induced bound $\sqrt{2}(\xi_0 + \xi_1) + 2\xi_0\xi_1$, respectively. In addition to the default $n = 20 \times |\mathcal{S}||\mathcal{A}| = 720$ gradient observations, we consider $n = 72$ for comparison, while keeping the expert dataset size and per-gradient estimation procedure fixed. Each configuration is repeated over 10 independent random seeds, and Table 1 reports the mean errors $\pm$ one standard deviation, from which three observations are noteworthy.

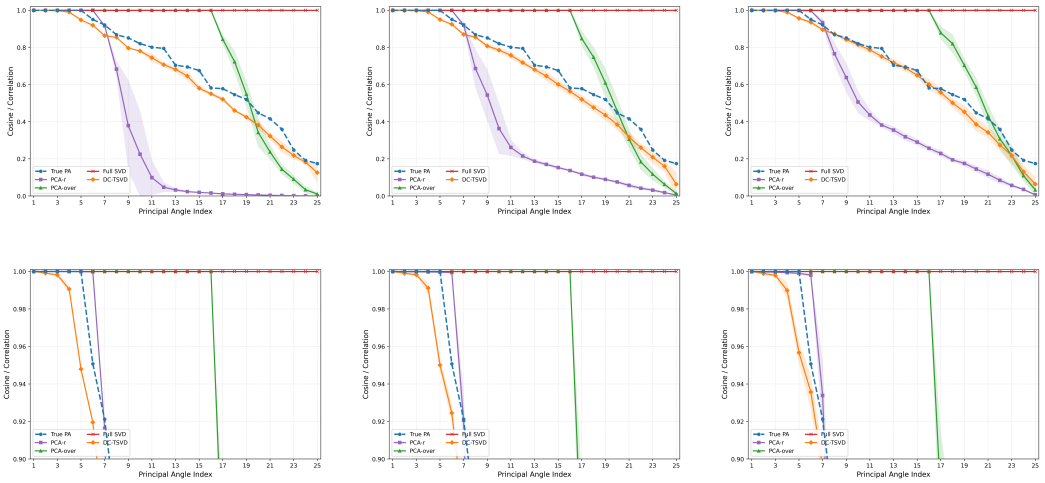
1. First, the Jacobian-induced leakage error remains small across most directions, consistent with the local stability condition in Assumption 2. The relatively larger leakage occurs only along the weakest signal direction, which is also the most susceptible to losing reliable identifiability.

2. Second, increasing the number of gradient observations n from 72 to 720 substantially reduces the average observation-induced error from 0.0426 to 0.0008. The observation-induced error is larger than the Jacobian-induced leakage at $n = 72$, but becomes nearly negligible at $n = 720$.



(a) $\eta = 0$ (b) $\eta = 0.25$ (c) $\eta = 0.5$

Figure 10: 5×5 windless vs. windy WindyGridWorld with $\beta = 0.5$



(a) $\eta = 0$ (b) $\eta = 0.25$ (c) $\eta = 0.5$

Figure 11: 5×5 windless vs. windy WindyGridWorld with $\beta = 1.0$

3. Finally, when the observation-induced error becomes sufficiently small, the total estimation error is numerically close to the Jacobian-induced leakage, indicating that the remaining error mainly arises from Jacobian-induced geometric distortion.

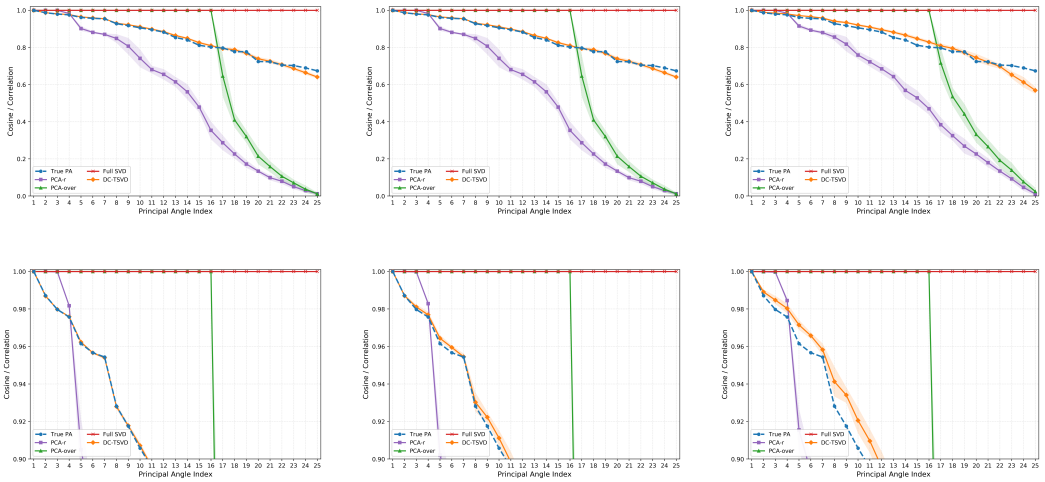
Nonlinear Local Geometry. To further evaluate DC-TSVD under a nonlinear reward representation, we consider

$$r_\theta = \theta + \alpha \tanh(\theta)$$

at a fixed parameter point θ_0, where $\tanh(\cdot)$ is applied element-wise. The corresponding Jacobian is

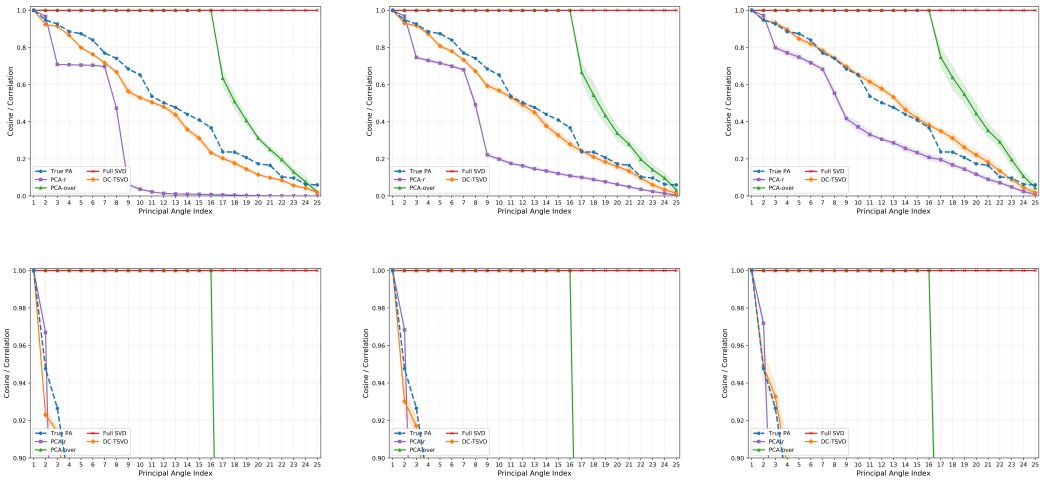
$$J_{\theta_0} = \operatorname{diag}\big(1 + \alpha \operatorname{sech}^2(\theta_{0,j})\big)_{j=1}^{d}.$$

We consider the 3×3 horizontal-to-vertical setting with $\beta = 0.5$ and $\alpha \in \{0.25, 0.5\}$. For each α, the same θ_0, and hence the same local Jacobian J_{θ_0}, is fixed across all gradient observations and experimental configurations.



(a) $\eta = 0$ (b) $\eta = 0.25$ (c) $\eta = 0.5$

Figure 12: 5×5 horizontal vs. vertical wind WindyGridWorld with $\beta = 0.5$



(a) $\eta = 0$ (b) $\eta = 0.25$ (c) $\eta = 0.5$

Figure 13: 5×5 horizontal vs. vertical wind WindyGridWorld with $\beta = 1.0$

For an expert dataset $\mathcal{D}$ consisting of state–action observations, consider the likelihood-based IRL objective

$$\ell_P(\theta; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(s,a)\in\mathcal{D}} \log \pi_\theta(a \mid s), \qquad \nabla_\theta \ell_P(\theta; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(s,a)\in\mathcal{D}} \nabla_\theta \log \pi_\theta(a \mid s).$$

Thus, each local policy score $\nabla_\theta \log \pi_{\theta_0}(a \mid s)$ is a per-observation contribution to the empirical IRL gradient. Moreover, these scores lie in the theoretical gradient space $\mathcal{G}_P^{\text{theo}}$. Since perturbations along any $u \in \mathcal{U}_P$ leave the induced policy unchanged, the corresponding directional derivative vanishes:

$$u^\top \nabla_r \log \pi_r(a \mid s) = 0, \qquad \forall\, u \in \mathcal{U}_P,$$

and hence

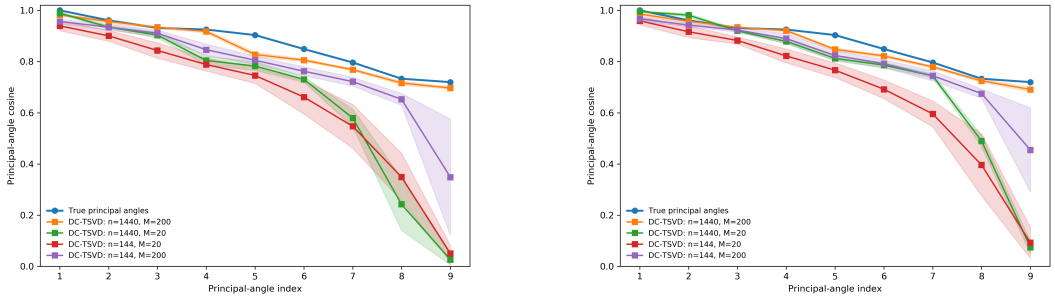
$$\nabla_r \log \pi_r(a \mid s) \in \mathcal{U}_P^\perp.$$

By the chain rule,

$$\nabla_\theta \log \pi_\theta(a \mid s) = J_\theta^\top \nabla_r \log \pi_r(a \mid s) \in J_\theta^\top \mathcal{U}_P^\perp = \mathcal{G}_P^{\text{theo}}.$$

Table 1: Empirical diagnostics of the two error components on the 3×3 WindyGridWorld (horizontal vs. vertical, $\beta = 0.5$, $\eta = 0.5$) with different numbers of gradient observations.

	Leakage	Observation-induced error		Total error	
Angle	error	$n = 72$	$n = 720$	$n = 72$	$n = 720$
1	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
2	0.0070 ± 0.0049	0.0073 ± 0.0050	0.0004 ± 0.0003	0.0063 ± 0.0041	0.0074 ± 0.0050
3	0.0178 ± 0.0067	0.0081 ± 0.0069	0.0004 ± 0.0004	0.0145 ± 0.0066	0.0180 ± 0.0062
4	0.0074 ± 0.0044	0.0114 ± 0.0089	0.0006 ± 0.0004	0.0102 ± 0.0077	0.0076 ± 0.0042
5	0.0096 ± 0.0082	0.0230 ± 0.0182	0.0008 ± 0.0006	0.0190 ± 0.0200	0.0096 ± 0.0081
6	0.0139 ± 0.0091	0.0235 ± 0.0145	0.0006 ± 0.0006	0.0208 ± 0.0133	0.0139 ± 0.0095
7	0.0148 ± 0.0081	0.0427 ± 0.0289	0.0009 ± 0.0011	0.0369 ± 0.0361	0.0148 ± 0.0088
8	0.0216 ± 0.0178	0.0439 ± 0.0378	0.0019 ± 0.0013	0.0499 ± 0.0414	0.0215 ± 0.0178
9	0.0612 ± 0.0400	0.2235 ± 0.2507	0.0018 ± 0.0013	0.2825 ± 0.2487	0.0602 ± 0.0399
Average	0.0170	0.0426	0.0008	0.0489	0.0170



$\alpha = 0.25$ $\alpha = 0.5$

Figure 14: DC-TSVD estimates under the nonlinear reward parameterization for $\alpha = 0.25$ and 0.5. The four curves correspond to different (n, M) configurations, where n is the number of gradient observations and M is the number of Monte Carlo trajectories per gradient observation; the ground-truth PS-space geometry is shown for comparison.

Therefore, local policy scores provide valid observations of the theoretical gradient space.

We accordingly construct the gradient observations from local policy scores evaluated at θ_0. Each gradient observation is a Monte Carlo estimate of $\nabla_\theta \log \pi_{\theta_0}(a \mid s)$ for a selected state–action pair (s, a), obtained from M Monte Carlo trajectories using geometric termination. We use balanced state–action sampling as a controlled coverage design to isolate the effect of gradient-estimation noise from insufficient state–action coverage. We vary both the number n of gradient observations and the number M of trajectories per gradient estimate, considering the four combinations $n \in \{144, 1440\}$ and $M \in \{20, 200\}$. Each configuration is repeated over 10 independent random seeds.

For each configuration, we apply DC-TSVD to the resulting gradient matrices and compare the estimated transfer-relevant principal-angle cosines $\hat{\rho}_{q+j}$ with the ground-truth PS-space cosines ρ_j^{PS}, $j = 1, \ldots, r$. Figure 14 reports the mean estimates over the 10 seeds, with shaded regions indicating one standard deviation. Increasing either n or M generally improves the agreement with the PS-space ground truth, while increasing both yields the closest agreement. These results demonstrate that, under controlled gradient coverage, DC-TSVD can estimate the transfer-relevant principal-angle geometry from noisy gradient observations under a nonlinear reward representation and a fixed local Jacobian.

I Limitations and Future Work

Our analysis relies on the Local Reward Representation Assumption, requiring sufficient reward expressivity over the transfer-relevant identifiable subspaces. Consequently, the current geometric characterization does not directly cover reward representations that fail to represent these subspaces.

An important extension is to move beyond the local gradient-space formulation considered in this work. The theoretical gradient spaces are defined at a fixed parameter point θ_0, or more generally under varying parameters for which the Jacobian-induced image of the identifiable subspace remains invariant. When this image varies with the reward parameter, the theoretical gradient space itself becomes parameter-dependent, requiring explicit control of Jacobian variation across parameter points.

Another direction is to develop statistically calibrated diagnostics for transfer geometry. The transfer bound in (1) motivates the separation quantity

$$\Delta_2 := \sin^2\left(\frac{\theta_2^{\mathrm{PS}}}{2}\right) = \frac{1-\rho_2^{\mathrm{PS}}}{2},$$

whose reciprocal determines the geometric amplification of reward estimation error. Rather than testing whether the second principal angle is exactly zero, for a prescribed separation threshold $\Delta_0 > 0$, one could consider

$$H_0 : \Delta_2 \le \Delta_0 \qquad \text{versus} \qquad H_1 : \Delta_2 > \Delta_0.$$

While the present analysis provides deterministic perturbation bounds for DC-TSVD, constructing a level-controlled test would additionally require probabilistic calibration of the gradient-estimation error. Such an extension would require concentration results tailored to the underlying gradient-sampling mechanism and, for general reward parameterizations, explicit control of Jacobian-induced geometric distortion.