SCALABLE MULTI-TASK INVERSE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

By learning transferable rewards, inverse reinforcement learning (IRL) enables counterfactual evaluation of agents under modified environments. Such transfer places strict requirements on coverage since target environments affect agents' state occupancy. We propose a multi-task IRL method that pools data across multiple agents with different rewards in the same environment under a low-rank assumption. In addition to alleviating coverage requirements, so each task need not visit every state as long as others do, the method offers scalable evaluation of multiple tasks under new environments as computationally intensive planning scales with rank rather than the number of tasks. We provide finite sample guarantees on reward recovery and on policy learning in new environments. Experiments show our method is robust to limited coverage, recovers rewards on and off of each task's support, transfers to target environments at lower regret than baselines, with its computational advantage over per-task methods widening as tasks grow.

1 INTRODUCTION

Inverse reinforcement learning aims to recover rewards that rationalize behavior in expert demonstrations (Russell, 1998). In that sense rewards are an agent's revealed preferences (Ziebart et al., 2008), enabling imitation learning (Abbeel & Ng, 2004; Ho & Ermon, 2016), and support alignment from human feedback (Christiano et al., 2017). Rewards are intrinsic to the agent's motives, as opposed to policies which are the result of planning given both the agent's rewards and the dynamics of the environment (Lucas, 1976). Hence we can use rewards to re-plan in perturbed environments and predict counterfactual behavior when agents face a new environment.

Many settings, such as online platforms and transportation networks, feature heterogeneous tasks. Following Ng & Russell (2000), we understand a task as mapping one-to-one to a (here agent-specific) reward function. Shutting down a train line, for instance, leads commuters with distinct preferences over speed, cost, and crowding to re-route accordingly. Therefore, a planner must anticipate this re-optimization despite limited state coverage by each commuter. We consider N independent tasks, that is N agent-specific reward functions, for each of which we observe the induced soft-optimal behavior in a source environment. We then seek to learn these rewards so as to anticipate how the behavior of agents would transfer to a target environment.

Single-task demonstrations typically concentrate on a narrow region of the state space, so partial coverage weakens reward identification (Lazzati et al., 2024). In particular, in the maximum-entropy RL framework, this materializes as temperature varies: high temperature improves coverage but flattens choice probabilities, while low temperature sharpens the reward signal but narrows coverage (Haarnoja et al., 2017). Fitting N tasks independently inherits the single-task support problem. Here, we break this temperature-indexed tradeoff by encoding an often-natural structural relatedness of tasks by a low-rank assumption. As we prove formally, this lessens per-agent coverage requirements.

Learning under a low rank assumption pools data from multiple source tasks to learn a common-across-tasks low-dimensional representation of the tasks' input, in our context, agents' preferences. The statistical gain can be understood as stemming from needing to only ensure good coverage in that compressed representation as opposed to in a fully unrestricted per-agent, per-state, and per-action sense. In the context of multi-task reward learning for planning in a target environment, we

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.

find that a natural-to-analyze setting is not low-rankness of rewards themselves but low-rankness of policy logits. This comes about from a representation of solutions to the maximum-entropy IRL problem in which rewards are affine in log policies (Cao et al., 2021). Shared dynamics and subgoals make such low-dimensional structure plausible (Barreto et al., 2017), and placing the assumption on choice probabilities rather than latent rewards makes it empirically testable (Section B.11).

The combination of low rank log policies and a reward operator that is affine in log policies implies that the typically expensive planning step required for IRL reduces from being per-task to per-basis function. k basis fixed points can be linearly composed to get N reward functions instead of N fixed points for N reward functions. When $k \ll N$, as can be expected in online settings with many users, this represents a substantial computational speed-up.

Our contributions are threefold. First, we introduce an offline, feature-free multi-task IRL method that pools coverage while retaining distinct per-task rewards and requiring only k basis reward-inversion solves. Second, we prove finite-sample guarantees for logit recovery, target-policy evaluation, and policy recovery under a coverage condition that permits missing task–state blocks. Third, a causal coverage intervention validates the predicted mechanism, and experiments across three domains show stronger off-support reward recovery, robust counterfactual transfer in RecSim, and wall-clock time growing as $N^{0.21}$, up to $13\times$ faster than per-task GenPQR at $N = 128$.

2 Related Literature

Single Task IRL Max-entropy IRL fits rewards by likelihood (Ziebart et al., 2008), while scalable variants use adversarial occupancy matching or soft Q-functions (Fu et al., 2018; Garg et al., 2021). Partial coverage limits offline reward identification (Lazzati et al., 2024). We build on the classify-then-regress view of van der Laan et al. (2025); the equivalent dynamic-discrete-choice literature estimates conditional choice probabilities and heterogeneous agent types (Rust, 1987; Arcidiacono & Miller, 2011).

Multi-task IRL Continuous latent-context methods such as PEMIRL infer task embeddings without known reward features, while hierarchical adversarial approaches learn shared multi-task structure (Yu et al., 2019; Chen et al., 2023). These methods require simulator interaction and adversarial optimization and provide neither our finite-sample guarantees under missing coverage nor rank-scaled reward inversion. Multi-intention methods recover K_{MI} discrete reward clusters and assign demonstrators to a mixture over them, often clustering directly on known reward features (Babeş-Vroman et al., 2011; Choi & Kim, 2012; Ramponi et al., 2020); we instead recover N distinct rewards through continuous loadings on k shared bases.

Low rank rewards and matrix completion The closest work imposes low rank on rewards. Shehab et al. (2025) recover one agent's time-varying reward by nuclear-norm minimization, with rank across time and a known tabular model; Guo & Moothedath (2026), extending Nagaraj et al. (2023), learn a low-rank reward matrix across tasks in forward RL, from sparsely observed rewards with online exploration and per-task planning. We instead impose low rank on log-policies and recover rewards from offline behavior alone, without reward features and with per-basis planning. Our analysis builds on nuclear-norm matrix completion (Negahban & Wainwright, 2012), low-rank multinomial-logit choice models (Kallus & Udell, 2020), and completion under arbitrary sampling (Foygel et al., 2011). We differ in allowing unobserved task–state blocks through a restricted coverage constant, in needing no sample-size condition, and in carrying the completion error through the Bellman reward operator to target-policy KL.

3 Method

3.1 Setup

We consider N tasks, indexed by $i \in [N] := \{1, \dots, N\}$. We observe demonstrations in a source environment and seek to predict how each task re-optimizes under target dynamics. Let $\mathcal{S}$ and $\mathcal{A}$ be finite state and action spaces where $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$. For each task i, define the source ($\ell = 1$) and target ($\ell = 2$) Markov decision processes (MDPs) by $M_{\ell,i} = (\mathcal{S}, \mathcal{A}, \mathsf{P}_\ell, r_i, d_0, \gamma_\ell)$. Here P_ℓ is the transition kernel, $r_i(s, a)$ is task i's finite mean reward at state s and action a, d_0 is

the initial state distribution, and $0 < \gamma_\ell < 1$ is the discount factor. For each task, the source and target MDPs share $\mathcal{S}, \mathcal{A}, r_i$, and d_0, while their transition kernels and discount factors may differ.

We assume that for each task i, the source policy $\pi^\star_{1,i}$ is soft optimal in $M_{1,i}$, that is, it maximizes entropy-regularized discounted return from every state with entropy coefficient 1 (Ziebart et al., 2008). For theoretical analysis, we assume that P_ℓ, d_0 and γ_ℓ are known. In practice, reward recovery uses observed source transitions through fitted Q iteration (FQI) and hence does not require a model of P_1. We assume target environment is specified by the practitioner, and re-planning requires known target dynamics. Our method first fits a shared policy model without known reward features, then recovers mean rewards by solving for a small set of shared basis functions.

3.2 Multi-task policy learning

We observe n tuples from the source environment (I_t, S_t, A_t, S'_t), generated by drawing a task-state pair $(I_t, S_t) \sim d$, an expert action $A_t \mid (I_t, S_t) \sim \pi^\star_{1,I_t}(\cdot \mid S_t)$, and a next state $S'_t \sim \mathsf{P}_1(\cdot \mid S_t, A_t)$. We let d denote the joint distribution over tasks and states in the source environments, with joint probability $d_{i,s} = \mathbb{P}(I_t = i, S_t = s)$ and $p_i := \sum_s d_{i,s} > 0$. Since rational agents optimize task-specific rewards, we expect that $d_{i,\cdot}$ will concentrate in some parts of the state space and hence may leave task-state pairs unobserved. We seek to learn rewards for all tasks on this offline dataset alone, without access to the true source transition kernel or reward features.

Tasks share source dynamics and may share subgoals with varying importance across tasks. Let $\Theta \in \mathbb{R}^{N\times SA}$ be a matrix of source policy logits, with source policies $\pi_i(a \mid s) = \mathrm{softmax}(\Theta_{i,(s,\cdot)})_a$. We parametrize these logits with k basis functions $\Phi(s,a) := (\phi_1(s,a), \ldots, \phi_k(s,a))^\top$ and task-specific coefficients $W_i \in \mathbb{R}^k$, so that $\Theta_{i,(s,a)} = W_i^\top \Phi(s,a)$. Stacking task coefficients and basis evaluations gives $\Theta = W\Phi \in \mathbb{R}^{N\times SA}$ of rank at most k.[1] Because softmax is invariant to a common shift of action logits, we fix a reference action $a_0 \in \mathcal{A}$ and enforce $\Theta_{i,(s,a_0)} = 0$. We implement this constraint by setting $\Phi(s,a_0) = 0$ for all $s \in \mathcal{S}$.

Let $D_i := \{(S_t, A_t, S'_t) : I_t = i\}$ denote task i's source transitions. Let $\|\cdot\|_F$ be the Frobenius norm and $\lambda > 0$ be the regularization strength. We estimate the factors jointly by a penalized maximum likelihood objective

$$\widehat{W}, \widehat{\Phi} \in \operatorname*{argmin}_{W,\Phi} -\frac{1}{n}\sum_{t=1}^{n} \log \pi_{I_t}(A_t|S_t) + \frac{\lambda}{2}\|W\|_F^2 + \frac{\lambda}{2}\|\Phi\|_F^2 \tag{1}$$

The fitted logit matrix is given by $\widehat{\Theta} := \widehat{W}\widehat{\Phi}$, resulting in a fitted log-policy matrix, $\widehat{u} \in \mathbb{R}^{N\times SA}$ where $\widehat{u}_{i,(s,a)} := \widehat{\Theta}_{i,(s,a)} - \mathrm{logsumexp}(\widehat{\Theta}_{i,(s,\cdot)})$. Sharing the basis allows tasks to pool information across their observed states. Individual tasks therefore need not be observed in every state, provided the pooled observations distinguish errors compatible with the low-rank structure. We make this coverage condition precise in Assumption 3, similar in spirit to task-diversity conditions in multi-task representation learning (Tripuraneni et al., 2020). Lemma 2 gives a sufficient spectral condition.

3.3 Efficient reward recovery through shared bases

In the max-entropy setting, different reward functions can induce the same source policy (Ng et al., 1999; Cao et al., 2021). To identify the reward of each task, we assume that its true reward, shared across the source and target environments, satisfies the following normalization.

Assumption 1. *For the fixed known $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ and $g : \mathcal{S} \to \mathbb{R}$ used in the procedure, every task's true reward satisfies*

$$\sum_{a\in\mathcal{A}} \mu(a \mid s) r_i(s,a) = g(s) \qquad \forall i \in [N], s \in \mathcal{S}. \tag{2}$$

Following the setup in van der Laan et al. (2025), write $(\mathsf{P}_1 V)(s,a) := \sum_{s'} \mathsf{P}_1(s' \mid s,a) V(s')$ and $(\mu f)(s) := \sum_b \mu(b \mid s) f(s,b)$, and define the operators $(\mathsf{P}_1 \mu\, f)(s,a) := \sum_{s',b} \mathsf{P}_1(s' \mid s,a)\, \mu(b \mid$

[1] This is a structural approximation, not a consequence of linear rewards: the maximum-entropy optimality map is nonlinear and can increase effective rank.

$s')\, f(s', b)$ and $(C_\mu f)(s, a) := f(s, a) - \sum_b \mu(b \mid s)\, f(s, b)$.

$$L_\mu := C_\mu (I - \gamma_1 \mathsf{P}_1 \mu)^{-1} \tag{3}$$

Note that the inverse exists because $\mathsf{P}_1\mu$ is row-stochastic and $\gamma_1 < 1$.

Lemma 1 (Normalized reward is affine in the log-policy)**.** *Let* $u_i^\star(s, a) := \log \pi_{1,i}^\star(a \mid s)$. *The unique reward satisfying the source soft Bellman equations and normalization* (2) *is*

$$r_i = L_\mu(u_i^\star - g) + g. \tag{4}$$

Since L_μ is bounded and linear, r_i is affine in $u_i^\star$. Appendix Lemma 10 proves uniqueness. The statistical theory below analyzes the exact normalized plug-in reward. Given fitted anchored logits

$$\widehat{\Theta}_{i,(\cdot,\cdot)} = \sum_{j=1}^{k} \widehat{W}_{ij} \widehat{\phi}_j, \qquad \widehat{c}_i(s) := \log \sum_a \exp\{\widehat{\Theta}_{i,(s,a)}\},$$

the fitted log-policy is $\widehat{u}_i = \widehat{\Theta}_i - \widehat{c}_i$, where $\widehat{c}_i$ is broadcast across actions. Its exact plug-in reward is therefore $\widehat{r}_i^{\,\mathrm{plug}} = L_\mu(\widehat{\Theta}_i - \widehat{c}_i - g) + g$. The normalizer $\widehat{c}_i$ depends nonlinearly on the task loadings, so the matrix of normalized log-policies need not remain rank k. Applying the reward operator exactly can consequently require one inversion per task, forfeiting the principal computational benefit of the low-rank representation.

Our implementation makes a deliberate basis-planning approximation: it applies L_μ directly to the low-rank anchored logits $\widehat{r}_i^{\,\mathrm{approx}} := L_\mu(\widehat{\Theta}_i - g) + g = \sum_{j=1}^{k} \widehat{W}_{ij} L_\mu \widehat{\phi}_j + (I - L_\mu)g$. This requires only k basis reward-inversion solves, rather than N task-specific solves. The approximation gap is explicit $\widehat{r}_i^{\,\mathrm{approx}} - \widehat{r}_i^{\,\mathrm{plug}} = L_\mu \widehat{c}_i$, which need not vanish under action-dependent transitions. Thus the theory isolates the statistical error of the normalized plug-in estimator, while the practical method adds this separately characterized computational approximation. All reported experiments use $\widehat{r}_i^{\,\mathrm{approx}}$ and demonstrate that the shortcut retains strong off-support recovery and transfer performance while delivering the intended rank-scaled computation. Appendix B.10 directly compares reward recovery between the exact method and its approximation and finds mild to null effects, with wall-clock savings that grow with N.

For a basis function ϕ and a state-action function v, define the affine Bellman operator

$$(T_\phi v)(s, a) := \big(\mathsf{P}_1\mu\big)(\gamma_1 v - \phi)(s, a) = \mathbb{E}_{s' \sim \mathsf{P}_1(\cdot \mid s, a)}\Big[\textstyle\sum_b \mu(b \mid s') \big(\gamma_1 v(s', b) - \phi(s', b)\big) \Big]. \tag{5}$$

Since $\mathsf{P}_1\mu$ is stochastic and $\gamma_1 < 1$, T_ϕ is a γ_1-contraction with a unique fixed point v_ϕ, and the basis reward of Lemma 1 is recovered as $L_\mu\phi = C_\mu(\phi - \gamma_1 v_\phi)$. We compute $v_\phi = T_\phi v_\phi$ by fitted iteration with a value network, regressing at each step onto the targets $T_\phi v$ evaluated at the observed next-states; we use the same neural evaluation across all environments.

3.4 Algorithm

Algorithm 1 has two stages: jointly fit rank-k anchored logits, and then run k FQI regressions on the learned basis functions and linearly recombine their rewards. A new task adds only k task-specific loadings and additional classification data; it does not add a reward-inversion solve, so the expensive stage scales with k not N. Section 5.4 verifies this computational advantage empirically.

For continuous-state environments, each basis function is represented by a neural network mapping a state to $\mathbb{R}^{|\mathcal{A}|}$ logits, with task logits formed as $\widehat{W}_i^\top \widehat{\Phi}(s, \cdot)$ before the softmax. Architecture, optimization, and regularization details are provided in Appendix B.4.

Note that if $g = 0$ then we can skip line 7 of the algorithm and simply set $\hat{b}_g = 0$.

4 Theoretical Guarantees

We first bound the estimation error of the shared source logit matrix under a pooled source coverage condition, and then show simultaneous bounds on target-policy value error, regret, and occupancy-weighted KL divergence from each task's soft-optimal target policy. Appendix A gives the full details.

Algorithm 1 Low-Rank Multi-Task IRL

Require: Demonstrations $D_{\text{total}} = \{D_i\}_{i=1}^N$, rank k, discount γ_1, reward normalization (μ, g).
1: **Initialize:** k basis networks for v_{ϕ_j} and v_g
2: **Classify:** Solve the joint softmax MLE (1) on $D_{\text{total}} \to \hat{W}, \hat{\Phi}$
3: **for** $j = 1, \dots, k$ **do**
4: **Regress:** Solve $\hat{v}_{\hat{\phi}_j} = T_{\hat{\phi}_j} \hat{v}_{\hat{\phi}_j}$ via fitted iteration on D_{total}
5: **Basis reward:** $\hat{r}_{\hat{\phi}_j} \leftarrow \hat{\phi}_j - \gamma_1 \hat{v}_{\hat{\phi}_j} + \mu(\gamma_1 \hat{v}_{\hat{\phi}_j} - \hat{\phi}_j)$
6: **end for**
7: **Offset:** Solve $\hat{v}_g = T_g \hat{v}_g$ via fitted iteration on D_{total}; set $\hat{b}_g \leftarrow g + \gamma_1(\hat{v}_g - \mu \hat{v}_g)$
8: **for** $i = 1, \dots, N$ **do**
9: $\hat{r}_i \leftarrow \sum_{j=1}^k \hat{W}_{ij} \hat{r}_{\hat{\phi}_j} + \hat{b}_g$ ▷ closed-form linear combination
10: **end for**
11: **return** $\{\hat{r}_i\}_{i=1}^N$

4.1 Statistical assumptions and estimator

Following the setting of Section 3, let $\Theta^\star$ be the anchored logit matrix of the true source policies, given by $\Theta^\star_{i,(s,a)} := \log \pi^\star_{1,i}(a \mid s) - \log(\pi^\star_{1,i}(a_0 \mid s))$. For analysis, the observations are i.i.d., and the source design d is known.

Assumption 2. *Assume that the true source logit matrix $\Theta^\star$ satisfies $\Theta^\star \neq 0$[2], $\operatorname{rank}(\Theta^\star) \leq k$ and $\|\Theta^\star\|_\infty \leq B$[3] where $k \leq \min\{N, SA\}$ and $B > 0$ is fixed independent of N, S and A.*

We analyze the bounded convex nuclear-norm formulation of the likelihood objective (1). For observation t, the negative log likelihood is $l_t(\Theta) = \log\left(\sum_{a \in \mathcal{A}} e^{\Theta_{I_t,(S_t,a)}}\right) - \Theta_{I_t,(S_t,A_t)}$, and the empirical loss is $L_n(\Theta) := \frac{1}{n}\sum_{t=1}^n l_t(\Theta)$. Fix a failure probability $0 < \delta < 1$, define $\nu := \max\left\{\left(1 - \frac{1}{A}\right)\max_i p_i,\ \frac{1}{2}\max_s \sum_i d_{i,s}\right\}$ and set the regularization parameter to

$$\lambda_n := 2\left\{\sqrt{\frac{2\nu}{n}\log\frac{2(N+SA)}{\delta}} + \frac{2\sqrt{2}}{3n}\log\frac{2(N+SA)}{\delta}\right\}. \tag{6}$$

We estimate the source logit matrix by the nuclear norm regularized maximum likelihood

$$\widehat{\Theta} \in \operatorname*{argmin}_{\Theta \in \mathcal{M}_B} \{L_n(\Theta) + \lambda_n \|\Theta\|_*\}^4, \tag{7}$$

where the feasible set is $\mathcal{M}_B := \left\{\Theta \in \mathbb{R}^{N \times SA} : \|\Theta\|_\infty \leq B, \Theta_{i,(s,a_0)} = 0 \text{ for every } i, s\right\}$.

We set $\widehat{u}_{i,(s,a)} := \widehat{\Theta}_{i,(s,a)} - \log\sum_{b \in \mathcal{A}} \exp\{\widehat{\Theta}_{i,(s,b)}\}$ and recover $\widehat{r}_i := L_\mu(\widehat{u}_i - g) + g$ via (4).

4.2 Pooled source coverage

Let $q := \operatorname{rank}(\Theta^\star) \leq k$ and let $\Theta^\star = U\Sigma V^\top$ be a singular value decomposition. Its tangent space is $T := \{UX^\top + YV^\top : X \in \mathbb{R}^{SA \times q},\ Y \in \mathbb{R}^{N \times q}\}$. Write P_T and $P_{T^\perp}$ for the Frobenius-orthogonal projections onto T and its complement. The regularized estimator's error lies, with high probability, in the anchored cone

$$\mathcal{C}(\Theta^\star) := \left\{\Delta : \Delta_{i,(s,a_0)} = 0\ \forall i, s, \quad \|P_{T^\perp}\Delta\|_* \leq 3\|P_T\Delta\|_*\right\}. \tag{8}$$

We measure source coverage on these possible errors by

$$c_0(d, \Theta^\star) := \inf_{\Delta \in \mathcal{C}(\Theta^\star)\setminus\{0\}} \frac{\sum_{i,s} d_{i,s}\left\|\Delta_{i,(s,\cdot)}\right\|_2^2}{\|\Delta\|_F^2}. \tag{9}$$

[2]The uniform-policy case $\Theta^\star = 0$ is treated separately in Proposition 1 in Appendix A.8.

[3]$\|\Theta\|_\infty := \max_{i,s,a} |\Theta_{i,(s,a)}|$ denotes the entrywise maximum norm.

[4]Here $\|\cdot\|_*$ is the nuclear norm. For $\Theta \in \mathbb{R}^{N \times SA}$ and $k \geq \operatorname{rank}(\Theta)$, $\|\Theta\|_* = \min_{W\Phi=\Theta} \frac{1}{2}(\|W\|_F^2 + \|\Phi\|_F^2)$, where the minimum is over $W \in \mathbb{R}^{N \times k}$ and $\Phi \in \mathbb{R}^{k \times SA}$. Thus the Frobenius penalties in (1) give a variational form of nuclear-norm regularization.

Assumption 3. $c_0(d, \Theta^\star) > 0$.

Intuitively, this condition holds exactly when every nonzero error in the cone changes a task–state block with $d_{i,s} > 0$ (Lemma 3 in Appendix A.4.1). This is a population coverage condition determined by the source design and the shared logit geometry. In task–state–action order, define $\mathbf{D}_d := \mathrm{diag}((d_{i,s})_{i,s,a})$ and define $\mathbf{P}_T \in \mathbb{R}^{NSA \times NSA}$ as the unique matrix satisfying $\mathbf{P}_T \,\mathrm{vec}(Z) = \mathrm{vec}(P_T Z)$ for every $Z \in \mathbb{R}^{N \times SA}$.

Lemma 2. *Under Assumption 2, let $\lambda_{\min}$ denote the smallest eigenvalue, then*

$$c_0(d, \Theta^\star) \geq \sup_{\tau \geq 0} \left\{ \lambda_{\min}(\mathbf{D}_d - \tau \mathbf{P}_T) + \frac{\tau}{1 + 18k} \right\} \geq \min_{i,s} d_{i,s}.$$

Appendix Lemma 5 proves this bound. At minimum, the tasks visiting each state must collectively span the latent directions; otherwise $c_0 = 0$. Example 7 constructs an instance with unobserved task–state block yet positive c_0, permitting full reward recovery and vanishing target regret as $n \to \infty$, even when target policies visit the unobserved source block (Theorem 1). Section 5.1 tests this directional coverage mechanism empirically.

4.3 Target values and finite-sample guarantees

Let Π be the set of all stationary target policies. For $\pi \in \Pi$, denote $d_2^\pi(s) := (1 - \gamma_2)\mathbb{E}_{\mathsf{P}_2, \pi, d_0}\left[\sum_{t \geq 0} \gamma_2^t \mathbf{1}\{S_t = s\}\right]$ the normalized discounted target state occupancy, and $\rho_2^\pi(s, a) := d_2^\pi(s)\pi(a \mid s)$ the normalized discounted target state-action occupancy.[5]

For each candidate $\pi \in \Pi$, we estimate its target soft value using the recovered reward $\widehat{r}_i$:

$$\widehat{J}_{2,i}(\pi) := J_{2,i}(\pi; \widehat{r}_i) = \frac{1}{1 - \gamma_2} \sum_{s,a} \rho_2^\pi(s, a)\{\widehat{r}_i(s, a) - \log \pi(a \mid s)\}, \tag{10}$$

The true value $J_{2,i}(\pi; r_i)$ uses r_i in place of $\widehat{r}_i$. For each task i, maximize the estimated target soft value:

$$\widehat{\pi}_i \in \operatorname*{argmax}_{\pi \in \Pi} \widehat{J}_{2,i}(\pi). \tag{11}$$

Let $\pi_{2,i}^*$ be the soft-optimal policy for $M_{2,i}$, characterized by its soft Bellman equations. In particular, $\pi_{2,i}^* \in \arg\max_{\pi'} J_{2,i}(\pi'; r_i)$. The logit-error radius ε_n and target-value radius $R_n(\pi)$ satisfy

$$\varepsilon_n \lesssim \frac{\sqrt{k\nu \log(2(N + SA)/\delta)/n} + \sqrt{k}\, \log(2(N + SA)/\delta)/n}{c_0(d, \Theta^\star)},$$
$$R_n(\pi) \lesssim \frac{\varepsilon_n}{1 - \gamma_2} \left\| L_\mu^\top \rho_2^\pi \right\|_2.$$

Here $\lesssim$ hides constants depending only on A and B. We give their exact definitions in (41) and (43).

Theorem 1. *Under the setting of Section 4.1 and Assumptions 1, 2, and 3, use the estimator* (7) *with tuning* (6)*, exact plug-in reward recovery from Section 3.3, and the selector* (11)*. With probability at least $1 - \delta$, the following hold simultaneously for every logit minimizer, every task i, every $\pi \in \Pi$, and every maximizing choice $\widehat{\pi}_i$.*

(i) Estimation and value: $\left\| \widehat{\Theta} - \Theta^\star \right\|_F \leq \varepsilon_n$, *and* $\left| \widehat{J}_{2,i}(\pi) - J_{2,i}(\pi; r_i) \right| \leq R_n(\pi)$.

(ii) Target regret: $J_{2,i}(\pi_{2,i}^\star; r_i) - J_{2,i}(\widehat{\pi}_i; r_i) \leq R_n(\pi_{2,i}^\star) + R_n(\widehat{\pi}_i)$.

(iii) Policy error: $\sum_s d_2^{\widehat{\pi}_i}(s)\ \mathrm{KL}\big(\widehat{\pi}_i(\cdot \mid s) \,\big\|\, \pi_{2,i}^\star(\cdot \mid s)\big) \leq (1 - \gamma_2)\{R_n(\pi_{2,i}^\star) + R_n(\widehat{\pi}_i)\}$.

The KL bound controls action discrepancies at states visited by the learned policy, weighted by its discounted occupancy. When $n\nu \geq \log(2(N + SA)/\delta)$, it simplifies up to constants to $\frac{1}{c_0(d, \Theta^\star)} \left(\|L_\mu^\top \rho_2^{\pi_{2,i}^\star}\|_2 + \|L_\mu^\top \rho_2^{\widehat{\pi}_i}\|_2 \right) \sqrt{\frac{k\nu}{n} \log \frac{2(N+SA)}{\delta}}$.

[5] Both d_2^π and ρ_2^π are known because the target dynamics and initial distribution are known.

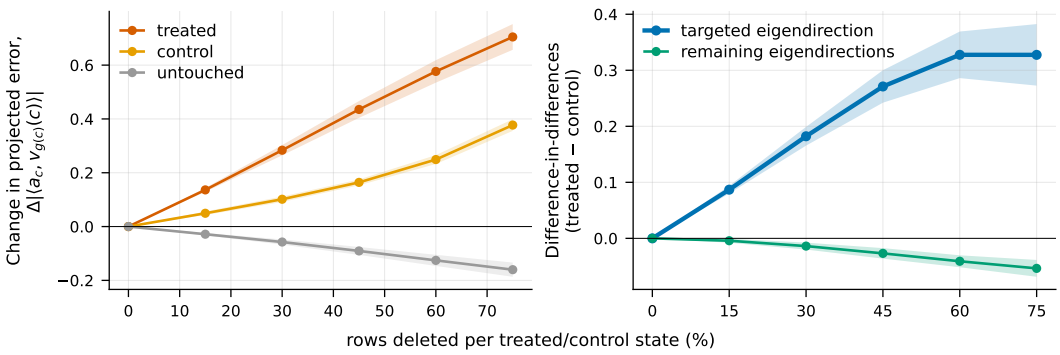

Figure 1: Causal test of directional coverage. Left: change in error projected onto the targeted eigendirection for treated, volume-matched control, and placebo cells. Right: treated-minus-control difference-in-differences for the targeted and remaining directions.

The bound separates three effects: c_0^{-1} penalizes latent directions weakly identified by the source demonstrations, the norm terms capture how reward recovery and target visitation amplify source logit estimation error, and the leading statistical term is of order $\sqrt{\frac{k\nu}{n}\log\frac{2(N+SA)}{\delta}}$.

5 Experiments

We evaluate whether LowRank (i) exploits pooled coverage, (ii) recovers rewards sample-efficiently on and off task support, (iii) transfers under target dynamics shift, and (iv) exhibits sublinear wall-clock growth as the task population increases fixing basis rank. The three domains are FourRooms, a partially observed sparse-reward navigation task; Highway, a continuous-state driving task; and RecSim, a tabular recommender with nonlinear topic fatigue. Rewards use four hidden task-preference dimensions; full environment and expert-training details appear in Appendix B.

We report Pearson correlation with μ-normalized ground-truth rewards on and off-support evaluation states. For continuous domains, these sets are defined over the pooled empirical evaluation states: a state is on-support if observed for that task and off-support if observed only for other tasks. This does not claim recovery at arbitrary unseen points in the continuous state space.

Baselines are per-task and pooled GenPQR and IQ-Learn, plus a feature-free four-cluster Multi-intent adaptation. Continuous latent meta-IRL and hierarchical adversarial methods require interactive simulator access and are therefore outside our offline setting. All applicable methods share architectures and held-out tuning; Appendix B.4 gives details.

5.1 Coverage drives off-support recovery

Our coverage condition is directional rather than a visit count: at state s, error along latent direction $v_k(s)$ enters the training loss only through the k-th eigenvalue $\lambda_k(s)$ of the visiting tasks' Gram matrix $\sum_i d_{i,s}(U^\star_{i,:})^\top U^\star_{i,:}$, so errors along directions the visiting tasks do not span, never appear in the training loss yet persist off-support; a singular Gram at any s forces $c_0 = 0$ (Appendix B.7).

We test this mechanism by progressively deleting observations from the tasks contributing most to selected states' leading eigendirection. Matched controls lose the same number of observations without targeting task alignment, while placebo states lose none. After refitting, the left panel reports the change from zero deletion in absolute error projected onto the targeted direction; the right reports treated-minus-control difference-in-differences for the targeted versus remaining $k-1$ directions.

Targeted projected error grows monotonically with deletion dose, and the difference-in-differences is concentrated in that direction. This isolates loss of directional coverage, rather than sample volume, as the source of the additional off-support error.

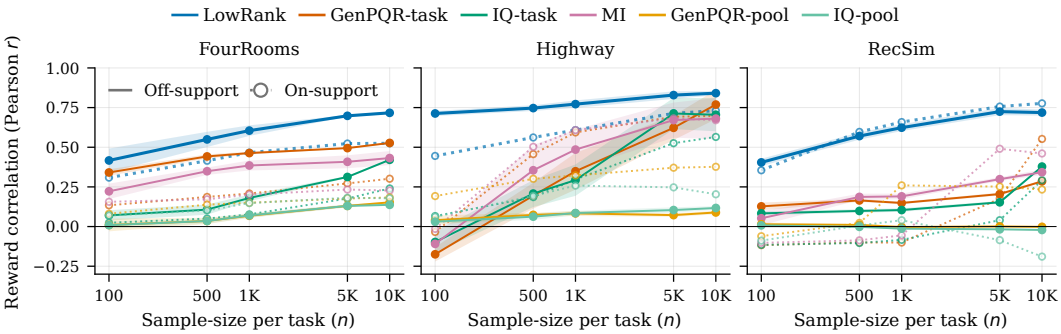

Figure 2: Reward recovery versus samples per task (mean and standard deviation over 10 seeds). LowRank leads off-support recovery at every budget; per-task methods narrow the gap only on Highway at high sample sizes.

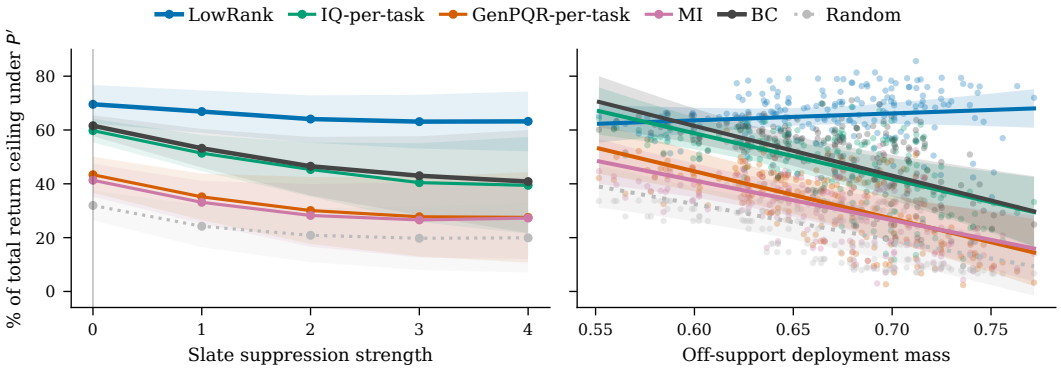

Figure 3: Policy recovery in target perturbed environments. Error bands are ± 1 s.d over ten seeds.

5.2 Sample efficient reward recovery

We vary the number of decisions per task from 100 to 10,000 while holding the task population and model architecture fixed. Figure 2 reports reward correlation separately on states observed for each task (on-support, dotted) and states supplied only by other tasks (off-support, solid). This isolates whether additional within-task data can substitute for coverage pooled across tasks.

LowRank achieves the highest off-support correlation at every data budget in all three domains, with the largest gains in the low-data regime. Per-task methods improve substantially with n and approach LowRank on Highway at 10,000 decisions, where sufficient within-task data eventually supplies broad coverage. The same gap persists in FourRooms and RecSim, whose partial observability and fatigue dynamics leave structured support gaps even at larger budgets. Naively pooled methods remain poor because they gain coverage only by discarding task heterogeneity, while discrete Multi-intent sharing remains less expressive than LowRank's continuous task loadings.

5.3 Policy recovery in perturbed environments

We evaluate counterfactual policy transfer in RecSim, where fatigue concentrates source behavior and target perturbations naturally push occupancy into rarely visited states. In our FourRooms and Highway configurations, ordinary trajectories already expose the reward-relevant features, so plausible dynamics changes did not produce a comparably clean off-support shift. Evaluating target return there would also conflate reward-transfer error with approximate neural planning, whereas RecSim's tabular dynamics permit exact target planning and oracle comparison. We therefore use RecSim to isolate the question addressed by our theory: whether recovered rewards remain valid when target occupancy moves beyond a task's source support. We increasingly suppress each task's preferred genre in target slates and compare the resulting policies with the oracle optimum.

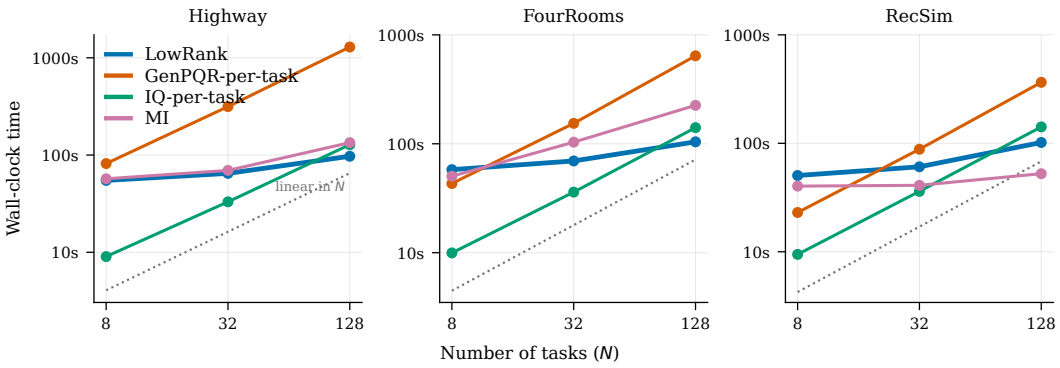

Figure 4: Wall-clock scaling with task population N at fixed basis rank k. LowRank exhibits substantially sublinear growth, while per-task methods scale approximately linearly

Figure 3 shows that LowRank achieves the highest normalized target return across perturbations. Behavior cloning and IQ-Learn are competitive at zero shift, but behavior cloning deteriorates as the target diverges from the source. In the right panel, LowRank's return is flat in off-support occupancy (slope $\hat{\beta} = +2.6$ percentage points per 0.1 off-support mass, 95% CI $[-3.2, 8.3]$), whereas IQ-Learn and GenPQR fall by 17.1 and 17.6 points per 0.1 of off-support mass. By Lemma 13, this return regret is equivalent to occupancy-weighted policy KL.

5.4 Computational Efficiency

At fixed basis rank k, LowRank should depend only weakly on the task population because its expensive reward-inversion step uses k rather than N fitted evaluations. We test this by increasing N from 8 to 128 with randomly sampled task coefficients. Figure 4 reports wall-clock time on log–log axes: the per-task baselines scale approximately linearly, whereas LowRank grows as $N^{0.21 \pm 0.025}$ across environments; Multi-intent is also sublinear but noisier at $N^{0.34 \pm 0.20}$. Relative to per-task GenPQR, which uses the same reward operator, LowRank is roughly $13\times$, $6\times$, and $4\times$ faster at $N = 128$ on Highway, FourRooms, and RecSim. IQ-Learn's cheaper per-task update makes it faster at small N, but it is overtaken by $N = 128$; at that scale no baseline is both faster and more accurate than LowRank (Table 2). Accuracy does not degrade with scale: at $N = 128$ ($n = 500$), LowRank's off-support reward correlation is 0.88, 0.53, and 0.78 on Highway, FourRooms, and RecSim, versus 0.74, 0.39, and 0.18 for per-task GenPQR. Multi-intent matches LowRank only on Highway (0.88) and trails on FourRooms and RecSim (0.24, 0.35; Table 2).

6 Conclusion

We develop a multi-task IRL method that assumes a low-rank structure on anchored source logits. This transforms multi-task IRL from N separate estimation problems into one shared problem, where identifiability does not require full per-task coverage of states. Instead, we require a mild restricted-eigenvalue condition on the source-data design rather than conditions on density ratios: demonstrations must span the low-rank cone. At a minimum, the tasks visiting each state must span the latent task directions and each task must visit states spanning the latent basis directions. Experiments demonstrate that pooling can recover the correct direction of reward variation where single-task estimates reverse it outside of their support. Pooling also reduces planning from per-task to per-basis, yielding sub-linear wall clock scaling in the number of tasks.

Our work suggests some avenues of further investigation. Our method and experiments used generic function approximation, but our finite-sample guarantees were limited to tabular MDPs with a known source kernel. Extending the theory to cover this approximation error, and the log-partition residual of the rank-k approximation, is a natural next step, as is relaxing the assumption of a shared reward normalization. Additionally, our environments are synthetic so that ground-truth rewards are available for clear comparisons between methods. Validation on real behavioral data remains open.

AI USE STATEMENT

In this work, we used generative AI tools for parts of coding up experiments (scaffolding for reproducibility from random seeds and tuning hyperparameters etc), figure generation and to flag mismatches in notation and typos in proofs and theorems remaining from multiple authors and the drafting process. We have reviewed all AI-assisted work: every proof, claim, and passage was checked by the authors, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Assumptions and full proofs of Lemmas and Theorem 1 are in Appendix A, and all experimental details (environments, reward construction, expert training, baselines, architectures, and the held-out tuning protocol) are in Appendix B. All environments are synthetic with known ground-truth rewards; every number is averaged over 10 seeds (mean and standard deviation), and anonymized code is included in the supplementary material.

REFERENCES

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

Peter Arcidiacono and Robert A Miller. Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica*, 79(6):1823–1867, 2011.

Monica Babeş-Vroman, Vukosi Marivate, Kaushik Subramanian, and Michael Littman. Apprenticeship learning about multiple intentions. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 897–904, 2011.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Jiayu Chen, Dipesh Tamboli, Tian Lan, and Vaneet Aggarwal. Multi-task hierarchical adversarial inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 4895–4920. PMLR, 2023.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & Miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 73383–73394, 2023.

Jaedeug Choi and Kee-Eung Kim. Nonparametric Bayesian inverse reinforcement learning for multiple reward functions. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Paul F Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Rina Foygel, Ohad Shamir, Nathan Srebro, and Ruslan Salakhutdinov. Learning with the weighted trace-norm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

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

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. IQ-Learn: Inverse soft-Q learning for imitation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4028–4039, 2021.

Yaoze Guo and Shana Moothedath. Provable multi-task reinforcement learning: A representation learning framework with low rank rewards. *arXiv preprint arXiv:2604.03891*, 2026.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. RecSim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019.

Nathan Kallus and Madeleine Udell. Dynamic assortment personalization in high dimensions. *Operations Research*, 68(4):1020–1037, 2020.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26085–26151. PMLR, 2024.

Robert E. Lucas, Jr. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46, 1976.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

Dheeraj Mysore Nagaraj, Suhas S Kowshik, Naman Agarwal, Praneeth Netrapalli, and Prateek Jain. Multi-user reinforcement learning with low rank rewards. In *International Conference on Machine Learning*, pp. 25627–25659. PMLR, 2023.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012.

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

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*, pp. 278–287, 1999.

Eduardo Pignatelli, Jarek Liesen, Robert Lange, Chris Lu, Pablo Samuel Castro, and Laura Toni. NAVIX: Scaling MiniGrid environments with JAX. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ram Rachum, Yotam Amitai, Yonatan Nakar, Reuth Mirsky, and Cameron Allen. BXRL: Behavior-explainable reinforcement learning. In *Workshop on Explainable Artificial Intelligence (XAI) at IJCAI-ECAI 2026*, 2026.

Giorgia Ramponi, Amarildo Likmeta, Alberto Maria Metelli, Andrea Tirinzoni, and Marcello Restelli. Truly batch model-free inverse reinforcement learning about multiple intentions. In *International Conference on Artificial Intelligence and Statistics*, pp. 2359–2369. PMLR, 2020.

Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pp. 101–103, 1998.

John Rust. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55(5):999–1033, 1987.

Mohamad Louai Shehab, Alperen Tercan, and Necmiye Ozay. Efficient reward identification in max entropy reinforcement learning with sparsity and rank priors. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 5104–5111. IEEE, 2025.

Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7852–7862, 2020.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

Lars van der Laan, Nathan Kallus, and Aurelien Bibaut. Inverse reinforcement learning with just classification and a few regressions. *arXiv preprint arXiv:2509.21172*, 2025.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2019.

Lantao Yu, Tianhe Yu, Chelsea Finn, and Stefano Ermon. Meta-inverse reinforcement learning with probabilistic context variables. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

A Theory appendix

Notations. For matrices, $\|\cdot\|_F$, $\|\cdot\|_*$, and $\|\cdot\|_{\mathrm{op}}$ denote the Frobenius, nuclear, and Euclidean operator norms, respectively. For matrices of the same size, $\langle\langle X, Y\rangle\rangle := \mathrm{tr}(X^\top Y)$ denotes the Frobenius inner product. Vector norms use double bars; $\|x\|_2$ is the Euclidean norm. We also write $\|\Theta\|_\infty := \max_{i,s,a} |\Theta_{i,(s,a)}|$ for the entrywise maximum norm. For $Z \in \mathbb{R}^{N\times SA}$, write $Z_{i,(s,\cdot)} := (Z_{i,(s,a)})_{a\in\mathcal{A}} \in \mathbb{R}^A$ for its action block at task i and state s.

A.1 Setting

Let $\mathcal{S}$ and $\mathcal{A}$ be finite state and action spaces, with $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$. Assume there are N tasks. For each task $i \in [N]$, consider the MDPs:

$$M_{\ell,i} := (\mathcal{S}, \mathcal{A}, \mathsf{P}_\ell, r_i, d_0, \gamma_\ell), \qquad \ell \in \{1, 2\}.$$

We call $M_{1,i}$ the source environment for task i and call $M_{2,i}$ the target environment for task i.

We assume the MDPs share the initial state distribution d_0. For each i, the mean reward function r_i is shared across $M_{1,i}$ and $M_{2,i}$. We assume the transition kernels $\mathsf{P}_1, \mathsf{P}_2$, the initial state distribution d_0, and the discount factors $0 \le \gamma_1, \gamma_2 < 1$ are known.

We write $r_i(s, a)$ for the mean reward for task i at state s and action a. We assume each mean reward is finite.

Throughout the paper, we say a policy is soft optimal if it maximizes the Shannon entropy regularized value from every state with entropy coefficient 1, as characterized by the soft Bellman equations in (89). Let $\pi^\star_{1,i} : \mathcal{S} \to \Delta(\mathcal{A})$ be the soft-optimal policy in the source environment $M_{1,i}$.

Observations from source environments $M_{1,i}$. We observe independent tuples (I_t, S_t, A_t), $t = 1, \ldots, n$, generated by the following i.i.d. sampling procedure

$$\begin{aligned} \mathbb{P}(I_t = i, S_t = s) &= d_{i,s}, \\ A_t \mid (I_t = i, S_t = s) &\sim \pi^\star_{1,i}(\cdot \mid s). \end{aligned} \tag{12}$$

We assume the sampling probabilities $d_{i,s}$ for all $i \in [N]$ and $s \in \mathcal{S}$ are known and satisfy $d_{i,s} \ge 0$, $\sum_{i,s} d_{i,s} = 1$. We write $d := (d_{i,s})_{i\in[N], s\in\mathcal{S}}$.

We assume that each task $i \in [N]$ has positive marginal probability, where

$$p_i := \sum_s d_{i,s} > 0. \tag{13}$$

Candidate policies. Let Π be the set of all stationary target policies. For each $\pi \in \Pi$ where $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, define

$$d_2^\pi(s) := (1-\gamma_2)\mathbb{E}_{\mathsf{P}_2,\pi,d_0}\left[\sum_{t\geq 0}\gamma_2^t \mathbf{1}\{S_t = s\}\right], \tag{14}$$

$$\rho_2^\pi(s,a) := d_2^\pi(s)\pi(a \mid s). \tag{15}$$

We call d_2^π the normalized discounted target state occupancy, and ρ_2^π the normalized discounted target state-action occupancy. Note that both d_2^π and ρ_2^π are known because the target dynamics and initial distribution are known.

For a candidate policy $\pi \in \Pi$, we write $\mathbb{E}_{\mathsf{P}_2,\pi,d_0}$ for the expectation over a trajectory $(S_t, A_t)_{t\geq 0}$ generated by $S_0 \sim d_0$, $A_t \sim \pi(\cdot \mid S_t)$, and $S_{t+1} \sim \mathsf{P}_2(\cdot \mid S_t, A_t)$. Then, for any bounded function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathsf{P}_2,\pi,d_0}[\sum_{t\geq 0}\gamma_2^t f(S_t,A_t)] &= \sum_s \left(\sum_{t\geq 0}\gamma_2^t \mathbb{E}_{\mathsf{P}_2,\pi,d_0}[\mathbf{1}\{S_t = s\}]\right)\sum_a \pi(a \mid s) f(s,a) \\
&= \frac{1}{1-\gamma_2}\sum_s d_2^\pi(s)\sum_a \pi(a \mid s) f(s,a) \\
&= \frac{1}{1-\gamma_2}\sum_{s,a}\rho_2^\pi(s,a) f(s,a).
\end{aligned} \tag{16}$$

Source logits. We fix a reference action $a_0 \in \mathcal{A}$. Define $\Theta^\star \in \mathbb{R}^{N\times SA}$ by

$$\Theta^\star_{i,(s,a)} := \log\frac{\pi^\star_{1,i}(a \mid s)}{\pi^\star_{1,i}(a_0 \mid s)}, \qquad \Theta^\star_{i,(s,a_0)} = 0. \tag{17}$$

We call $\Theta^\star$ the true anchored source logit matrix. Note that the source soft-optimal policies have full support, so the log ratios $\Theta^\star_{i,(s,a)}$ are finite.

Fix $k \leq \min\{N, SA\}$ and $B > 0$.

Assumption 4. *Assume that the true source logit matrix $\Theta^\star$ satisfies $\Theta^\star \neq 0$ and*

$$\operatorname{rank}(\Theta^\star) \leq k, \qquad \|\Theta^\star\|_\infty \leq B. \tag{18}$$

Later in Section A.8 we explain how the case of $\Theta^\star = 0$, that is, if all source policies are uniform over actions independent of state, is resolved easily.

Reward identification. We use the reward normalization of van der Laan et al. (2025) and the recovery identity in their Theorem 1. We impose the following reward normalization and fix a known policy $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ and a known state function $g : \mathcal{S} \to \mathbb{R}$.

Assumption 5. *For the fixed known μ and g used in the procedure, every task's actual shared mean reward satisfies*

$$\sum_{a\in\mathcal{A}}\mu(a \mid s) r_i(s,a) = g(s) \qquad \forall i \in [N], s \in \mathcal{S}. \tag{19}$$

For a state function $V : \mathcal{S} \to \mathbb{R}$ and a state-action function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define linear operators

$$(\mathsf{P}_1 V)(s,a) := \sum_{s'}\mathsf{P}_1(s' \mid s,a)V(s'), \qquad (\mu f)(s) := \sum_b \mu(b \mid s) f(s,b).$$

For a state function $g : \mathcal{S} \to \mathbb{R}$, use the same symbol for its state-action extension $g(s,a) := g(s)$.

The composition $\mathsf{P}_1\mu$ and the centering operator C_μ act on state-action functions as

$$(\mathsf{P}_1\mu f)(s,a) := \sum_{s',b}\mathsf{P}_1(s' \mid s,a)\mu(b \mid s') f(s',b), \tag{20}$$

$$(C_\mu f)(s,a) := f(s,a) - \sum_b \mu(b \mid s) f(s,b). \tag{21}$$

Write $u^\star \in \mathbb{R}^{N\times SA}$ where $u_i^\star(s,a) := \log \pi_{1,i}^\star(a \mid s)$. In matrix formulas, state-action functions such as $u_i^\star$, r_i, and ρ_2^π are represented as column vectors in state then action order. Under Assumption 5, we can recover the reward by

$$r_i = C_\mu(I - \gamma_1 \mathsf{P}_1\mu)^{-1}(u_i^\star - g) + g. \tag{22}$$

Here I is the $SA \times SA$ identity. The inverse exists because $\mathsf{P}_1\mu$ is row-stochastic and $\gamma_1 < 1$.

Target objective. Candidate policies may assign zero probability to some actions and we use convention $0 \log 0 = 0$ throughout. For any target policy π and task i, define

$$J_{2,i}(\pi; r_i) := \mathbb{E}_{\mathsf{P}_2,\pi,d_0}\left[\sum_{t\geq 0} \gamma_2^t \{r_i(S_t, A_t) - \log \pi(A_t \mid S_t)\}\right]. \tag{23}$$

We call $J_{2,i}$ the target soft value of policy π, which is the expectation of the total discounted reward plus Shannon entropy by running policy π in $M_{2,i}$. By (16), we have that for any fixed $\pi \in \Pi$, and reward functions $\tilde{r}_i, r_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$J_{2,i}(\pi; \tilde{r}_i) - J_{2,i}(\pi; r_i) = \frac{1}{1-\gamma_2} \sum_{s,a} \rho_2^\pi(s,a)[\tilde{r}_i(s,a) - r_i(s,a)]. \tag{24}$$

Let $\pi_{2,i}^\star$ be the soft-optimal policy for $M_{2,i}$ from every state, characterized by its soft Bellman equations. In particular, $\pi_{2,i}^\star \in \arg\max_{\pi'} J_{2,i}(\pi'; r_i)$ and $\pi_{2,i}^\star \in \Pi$.

A.2 Procedure: single logit fit

Our procedure first fits the source logits, reconstructs the reward of each task, and selects the candidate policy with the largest estimated target soft value.

Logit estimation. Define the feasible set

$$\mathcal{M}_B := \left\{\Theta \in \mathbb{R}^{N\times SA} : \|\Theta\|_\infty \leq B, \quad \Theta_{i,(s,a_0)} = 0 \text{ for every } i, s\right\}. \tag{25}$$

For observation t, the negative log likelihood is given by $l_t(\Theta) = \log\left(\sum_{a\in\mathcal{A}} e^{\Theta_{I_t,(S_t,a)}}\right) - \Theta_{I_t,(S_t,A_t)}$, and the empirical loss is given by $L_n(\Theta) := \frac{1}{n}\sum_{t=1}^n l_t(\Theta)$.

Fix a failure probability $0 < \delta < 1$ and define

$$\nu := \max\left\{\left(1 - \frac{1}{A}\right)\max_i p_i,\ \frac{1}{2}\max_s \sum_i d_{i,s}\right\}, \tag{26}$$

$$\lambda_n := 2\left\{\sqrt{\frac{2\nu}{n}\log\frac{2(N+SA)}{\delta}} + \frac{2\sqrt{2}}{3n}\log\frac{2(N+SA)}{\delta}\right\}. \tag{27}$$

Note that this tuning uses total failure probability δ for a single logit matrix fit, so its error bound holds simultaneously across tasks i and candidate $\pi \in \Pi$.

We estimate the source logit matrix by

$$\widehat{\Theta} \in \operatorname*{argmin}_{\Theta\in\mathcal{M}_B} \left\{L_n(\Theta) + \lambda_n \|\Theta\|_*\right\}, \tag{28}$$

Note that a minimizer exists because the objective is continuous on the nonempty compact set $\mathcal{M}_B$.

Estimate the target value of each candidate policy Given $\widehat{\Theta}$, we define

$$\widehat{u}_{i,(s,a)} := \widehat{\Theta}_{i,(s,a)} - \log\sum_b e^{\widehat{\Theta}_{i,(s,b)}}, \tag{29}$$

$$\widehat{r}_i := C_\mu(I - \gamma_1 \mathsf{P}_1\mu)^{-1}(\widehat{u}_i - g) + g. \tag{30}$$

where $\widehat{u}_{i,(s,a)}$ is an estimate of the log probabilities $\log \pi^{\star}_{1,i}(a \mid s)$ and $\widehat{r}_i$ is an estimate of the reward of task i.

Next, for each $\pi \in \Pi$, define the estimated target soft value by evaluating at reward $\widehat{r}_i$, that is,

$$\widehat{J}_{2,i}(\pi) := J_{2,i}(\pi; \widehat{r}_i) = \frac{1}{1-\gamma_2} \sum_{s,a} \rho_2^{\pi}(s,a)[\widehat{r}_i(s,a) - \log \pi(a \mid s)]. \tag{31}$$

Plug-in policy selection. For each task i, choose

$$\widehat{\pi}_i \in \operatorname*{argmax}_{\pi \in \Pi} \widehat{J}_{2,i}(\pi). \tag{32}$$

A.3 Main result: single logit fit

Theorem 2. *Under the setting of Section A.1 and Assumptions 4, 5 and 6, use the procedure of Section A.2. With probability at least $1-\delta$, the following statements hold simultaneously for every minimizer in* (28)*, every task i, and every candidate $\pi \in \Pi$ and every maximizing choice in* (32)*.*

(i) Estimation and value. *The common logit estimate satisfies*

$$\left\| \widehat{\Theta} - \Theta^{\star} \right\|_F \le \varepsilon_n, \tag{33}$$

and each estimated target value satisfies

$$\left| \widehat{J}_{2,i}(\pi) - J_{2,i}(\pi; r_i) \right| \le R_n(\pi). \tag{34}$$

(ii) Regret. *The regret of $\widehat{\pi}_i$ satisfies*

$$J_{2,i}(\pi^{\star}_{2,i}; r_i) - J_{2,i}(\widehat{\pi}_i; r_i) \le R_n(\pi^{\star}_{2,i}) + R_n(\widehat{\pi}_i). \tag{35}$$

(iii) Policy error.

$$\begin{aligned} &\sum_s d_2^{\widehat{\pi}_i}(s)\, \mathrm{KL}\big(\widehat{\pi}_i(\cdot \mid s) \,\big\|\, \pi^{\star}_{2,i}(\cdot \mid s)\big) \\ &\le (1-\gamma_2)\{R_n(\pi^{\star}_{2,i}) + R_n(\widehat{\pi}_i)\} \\ &= \varepsilon_n L_B \left\{ \begin{array}{l} \left\| \left[C_\mu (I - \gamma_1 \mathsf{P}_1 \mu)^{-1}\right]^{\top} \rho_2^{\pi^{\star}_{2,i}} \right\|_2 \\ + \left\| \left[C_\mu (I - \gamma_1 \mathsf{P}_1 \mu)^{-1}\right]^{\top} \rho_2^{\widehat{\pi}_i} \right\|_2 \end{array} \right\}. \end{aligned} \tag{36}$$

Proof sketch. Part (i) follows from Lemmas 8 and 11; part (ii) from Lemma 12; and part (iii) from part (ii) and Lemma 13. All conclusions hold on the same estimation event. □

Note that the policy-error bounds (36) follow from (35) and the identity below

$$J_{2,i}(\pi^{\star}_{2,i}; r_i) - J_{2,i}(\pi; r_i) = \frac{1}{1-\gamma_2} \sum_s d_2^{\pi}(s)\, \mathrm{KL}\big(\pi(\cdot \mid s) \,\big\|\, \pi^{\star}_{2,i}(\cdot \mid s)\big)\,. \tag{37}$$

A.4 Coverage condition

Coverage condition. By Assumption 4, let $q := \mathrm{rank}(\Theta^{\star}) \le k$ and let $\Theta^{\star} = U \Sigma V^{\top}$ be a singular value decomposition of the true source logit matrix, with $U \in \mathbb{R}^{N \times q}$ and $V \in \mathbb{R}^{SA \times q}$ with orthonormal columns. We define the tangent space

$$T := \{UX^{\top} + YV^{\top} : X \in \mathbb{R}^{SA \times q}, Y \in \mathbb{R}^{N \times q}\}.$$

Let $P_T, P_{T^{\perp}} : \mathbb{R}^{N \times SA} \to \mathbb{R}^{N \times SA}$ denote orthogonal projection operators onto T and its orthogonal complement $T^{\perp}$, that is, for every matrix $Z \in \mathbb{R}^{N \times SA}$,

$$P_T Z := UU^{\top} Z + ZVV^{\top} - UU^{\top} ZVV^{\top}, \qquad P_{T^{\perp}} Z = (I - UU^{\top}) Z (I - VV^{\top}).$$

Note that $P_T, P_{T^\perp}$ are orthogonal projections with respect to the Frobenius inner product.

Define

$$\mathcal{C}(\Theta^\star) := \{\Delta : \Delta_{i,(s,a_0)} = 0 \,\forall i, s, \|P_{T^\perp}\Delta\|_* \le 3\,\|P_T\Delta\|_*\}, \tag{38}$$

and we call $\mathcal{C}(\Theta^\star)$ the anchored error cone.

Define

$$c_0(d, \Theta^\star) := \inf_{\Delta \in \mathcal{C}(\Theta^\star)\setminus\{0\}} \frac{\sum_{i,s} d_{i,s} \left\|\Delta_{i,(s,\cdot)}\right\|_2^2}{\|\Delta\|_F^2}. \tag{39}$$

We call c_0 the coverage coefficient. Note that the magnitude of c_0 is determined by the source design d and the low-rank geometry of $\Theta^\star$.

Assumption 6. $c_0(d, \Theta^\star) > 0$.

Under Assumption 6, for every $\Delta \in \mathcal{C}(\Theta^\star)$, $\sum_{i,s} d_{i,s} \left\|\Delta_{i,(s,\cdot)}\right\|_2^2 \ge c_0(d, \Theta^\star)\,\|\Delta\|_F^2$.

Confidence radii for logit estimation. Let

$$h_B := \int_0^1 (1-u)e^{-4Bu}du = \frac{4B - 1 + e^{-4B}}{16B^2} > 0. \tag{40}$$

Define

$$\varepsilon_n := \frac{A^2 e^{2B}(260B + 6/h_B)\sqrt{k}\lambda_n}{c_0(d, \Theta^\star)}, \tag{41}$$

The quantity ε_n bounds the source-logit error in Frobenius norm in Theorem 2. Substituting (27) gives

$$\varepsilon_n = \frac{A^2 e^{2B}(260B + 6/h_B)}{c_0(d, \Theta^\star)} \left\{ 2\sqrt{\frac{2k\nu \log(2(N+SA)/\delta)}{n}} + \frac{4\sqrt{2k}\log(2(N+SA)/\delta)}{3n} \right\}.$$

Target value radius. Let

$$L_B := \sqrt{\frac{Ae^{2B}}{e^{2B} + A - 1}}, \tag{42}$$

and define

$$R_n(\pi) := \frac{\varepsilon_n}{1-\gamma_2} L_B \left\| \left[C_\mu (I - \gamma_1 \mathsf{P}_1 \mu)^{-1}\right]^\top \rho_2^\pi \right\|_2. \tag{43}$$

We call $R_n(\pi)$ a confidence radius for the target soft value of candidate policy π.

Intuitively, ε_n measures uncertainty in the fitted logits, and the remaining factors in $R_n(\pi)$ describe how that error affects the target value of candidate policy π.

A.4.1 Interpreting the coverage condition

For any $Z \in \mathbb{R}^{N \times SA}$, define

$$(P_d Z)_{i,(s,\cdot)} := \mathbf{1}\{d_{i,s} > 0\} Z_{i,(s,\cdot)} \tag{44}$$

which denotes the projection onto observed blocks, and let $z = \mathrm{vec}(Z) \in \mathbb{R}^{NSA}$ be a vector that stacks the rows of Z in task, state, then action order. Define the symmetric matrices $\mathbf{D}_d, \mathbf{P}_T \in \mathbb{R}^{NSA \times NSA}$ by

$$\mathbf{D}_d := \mathrm{diag}\big((d_{i,s})_{i,(s,a)}\big), \qquad \mathbf{P}_T\,\mathrm{vec}(Z) := \mathrm{vec}(P_T Z) \quad \forall Z \in \mathbb{R}^{N \times SA}.$$

where each $d_{i,s}$ is repeated over all A actions and $\mathbf{P}_T$ is the matrix representation of the projection operator P_T.

Proofs of the coverage bounds

Restricted coverage and the observed blocks.

Lemma 3. *Under Assumption 4,*

$$c_0 > 0 \quad \Longleftrightarrow \quad \ker(P_d) \cap \mathcal{C}(\Theta^\star) = \{0\}. \tag{45}$$

where P_d is the projection onto observed blocks defined in (44)*.*

Proof.

$$\begin{aligned} c_0 &= \inf_{\Delta \in \mathcal{C}(\Theta^\star)\setminus\{0\}} \frac{\sum_{i,s} d_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2}{\|\Delta\|_F^2} \\ &= \inf_{\substack{\Delta \in \mathcal{C}(\Theta^\star) \\ \|\Delta\|_F = 1}} \sum_{i,s} d_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2 \\ &= \min_{\substack{\Delta \in \mathcal{C}(\Theta^\star) \\ \|\Delta\|_F = 1}} \sum_{i,s} d_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2. \end{aligned}$$

The first equality is (39). For the second, divide each nonzero Δ by $\|\Delta\|_F$ which leaves the ratio unchanged and preserves the cone constraints (38). For the last equality, the unit-norm constraint and (38) define a closed set. It is nonempty because $\Theta^\star \neq 0$ belongs to the cone. Every matrix in this set satisfies

$$\|\Delta\|_F^2 = \sum_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2 = 1 < \infty.$$

Thus the feasible set is closed and bounded in finite dimension, hence compact; the continuous objective therefore attains its minimum.

If $0 \neq \Delta \in \ker(P_d) \cap \mathcal{C}(\Theta^\star)$, then

$$0 \leq c_0 \leq \frac{\sum_{i,s} d_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2}{\|\Delta\|_F^2} = 0.$$

The second inequality follows from (39); the equality follows from $P_d\Delta = 0$, which makes every block with $d_{i,s} > 0$ vanish by (44). Hence $c_0 > 0$ implies $\ker(P_d) \cap \mathcal{C}(\Theta^\star) = \{0\}$.

Conversely, if $c_0 = 0$, let Δ_0 attain the minimum above, so $\Delta_0 \in \mathcal{C}(\Theta^\star)$ and $\|\Delta_0\|_F = 1$. Then

$$\sum_{i,s} d_{i,s} \|(\Delta_0)_{i,(s,\cdot)}\|_2^2 = 0.$$

Every summand is nonnegative, so each block with $d_{i,s} > 0$ must vanish:

$$(\Delta_0)_{i,(s,\cdot)} = 0 \qquad \text{whenever } d_{i,s} > 0.$$

By the definition (44) of P_d, this gives

$$P_d\Delta_0 = 0.$$

Since Δ_0 has unit norm, it is nonzero. Thus $c_0 = 0$ implies that the kernel and cone have a nonzero intersection, proving the reverse implication in (45). □

A spectral lower bound. Use the singular value decomposition from Section A.4 and the matrices $\mathbf{D}_d, \mathbf{P}_T$ from Section A.4.1. The entries of $\mathbf{P}_T$ are

$$\begin{aligned} &(\mathbf{P}_T)_{(i,(s,a)),(i',(s',a'))} \\ &= (UU^\top)_{i,i'} \mathbf{1}\{s = s',\ a = a'\} + \mathbf{1}\{i = i'\}(VV^\top)_{(s,a),(s',a')} \\ &\quad - (UU^\top)_{i,i'}(VV^\top)_{(s,a),(s',a')}. \end{aligned}$$

Lemma 4. *Under Assumption 4,*

$$c_0 \geq \sup_{\tau \geq 0} \left\{ \lambda_{\min}(\mathbf{D}_d - \tau \mathbf{P}_T) + \frac{\tau}{1 + 18k} \right\} \geq \min_{i,s} d_{i,s}. \tag{46}$$

Proof. Expanding the entries of $\mathbf{P}_T$ gives, for every $Z \in \mathbb{R}^{N\times SA}$,

$$\mathbf{P}_T \operatorname{vec}(Z) = \operatorname{vec}\big(UU^\top Z + ZVV^\top - UU^\top ZVV^\top\big) = \operatorname{vec}(P_T Z).$$

Since P_T is an orthogonal projection, we have

$$\mathbf{P}_T^\top = \mathbf{P}_T, \qquad \mathbf{P}_T^2 = \mathbf{P}_T, \qquad \operatorname{vec}(Z)^\top \mathbf{P}_T \operatorname{vec}(Z) = \|P_T Z\|_F^2 .$$

Take any $\Delta \in \mathcal{C}(\Theta^\star) \setminus \{0\}$ and set $Z = \Delta$ and $z = \operatorname{vec}(Z)$. Then

$$\|z\|_2^2 = \|Z\|_F^2 = \|\Delta\|_F^2 > 0,$$
$$z^\top \mathbf{D}_d z = \sum_{i,s} d_{i,s} \sum_a Z_{i,(s,a)}^2 = \sum_{i,s} d_{i,s} \sum_a \Delta_{i,(s,a)}^2 = \sum_{i,s} d_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2.$$

The projected matrix $P_T Z$ has rank at most $2k$ since

$$P_T Z = UU^\top Z + (I - UU^\top) ZVV^\top,$$
$$\operatorname{rank}(P_T Z) \le \operatorname{rank}(UU^\top Z) + \operatorname{rank}((I - UU^\top) ZVV^\top) \le 2\operatorname{rank}(\Theta^\star) \le 2k.$$

Therefore we have

$$\|P_{T^\perp} Z\|_F^2 \le \|P_{T^\perp} Z\|_*^2 \le 9 \|P_T Z\|_*^2 \le 9\operatorname{rank}(P_T Z) \|P_T Z\|_F^2 \le 18k \|P_T Z\|_F^2 .$$

By orthogonality,

$$\begin{aligned} \|z\|_2^2 = \|Z\|_F^2 &= \|P_T Z\|_F^2 + \|P_{T^\perp} Z\|_F^2 \\ &\le (1 + 18k) \|P_T Z\|_F^2 = (1 + 18k) z^\top \mathbf{P}_T z. \end{aligned} \tag{47}$$

For every $\tau \ge 0$,

$$\begin{aligned} \frac{\sum_{i,s} d_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2}{\|\Delta\|_F^2} &= \frac{z^\top \mathbf{D}_d z}{\|z\|_2^2} \\ &= \frac{z^\top (\mathbf{D}_d - \tau \mathbf{P}_T) z}{\|z\|_2^2} + \tau \frac{z^\top \mathbf{P}_T z}{\|z\|_2^2} \\ &\ge \lambda_{\min}(\mathbf{D}_d - \tau \mathbf{P}_T) + \frac{\tau}{1 + 18k}. \end{aligned}$$

The last inequality uses $z^\top M z / \|z\|_2^2 \ge \lambda_{\min}(M)$ with the symmetric matrix $M = \mathbf{D}_d - \tau \mathbf{P}_T$. The second term is at least $\tau/(1 + 18k)$ by (47) and $\tau \ge 0$. The right-hand side is independent of Δ, so taking the infimum in (39) gives

$$c_0 \ge \lambda_{\min}(\mathbf{D}_d - \tau \mathbf{P}_T) + \frac{\tau}{1 + 18k} \qquad \text{for every } \tau \ge 0.$$

Taking the supremum of the right-hand side over $\tau \ge 0$ proves the first inequality in (46). Choosing $\tau = 0$ gives

$$\sup_{\tau \ge 0} \left\{ \lambda_{\min}(\mathbf{D}_d - \tau \mathbf{P}_T) + \frac{\tau}{1 + 18k} \right\} \ge \lambda_{\min}(\mathbf{D}_d) = \min_{i,s} d_{i,s}.$$

□

Coverage characterization. The next lemma characterizes Assumption 6 and gives a lower bound on c_0.

Lemma 5. *Under Assumption 4,*

$$c_0(d, \Theta^\star) \ge \sup_{\tau \ge 0} \left\{ \lambda_{\min}\big(\mathbf{D}_d - \tau \mathbf{P}_T\big) + \frac{\tau}{1 + 18k} \right\} \ge \min_{i,s} d_{i,s}. \tag{48}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of a matrix, so $\lambda_{\min}(\mathbf{D}_d - \tau \mathbf{P}_T) = \min_{\substack{z \in \mathbb{R}^{NSA} \\ \|z\|_2 = 1}} z^\top (\mathbf{D}_d - \tau \mathbf{P}_T) z$. Moreover, $c_0(d, \Theta^\star) > 0$ if and only if $\ker(P_d) \cap \mathcal{C}(\Theta^\star) = \{0\}$.

Proof of Lemma 5. The coverage characterization (45) and spectral lower bound (46) yield Lemma 5. □

Intuitively, the coverage coefficient $c_0(d, \Theta^\star)$ is strictly positive when every nonzero error in the cone changes at least one observed block.

Remark 6. A positive lower bound in (48) gives a sufficient condition for identifying $\Theta^\star$ from the source population distribution against perturbations in $\mathcal{C}(\Theta^\star)$. Specifically, let Θ be an anchored source logit matrix that induces the same source population distribution as $\Theta^\star$, and suppose $\Delta := \Theta - \Theta^\star \in \mathcal{C}(\Theta^\star)$.

First, on each block with $d_{i,s} > 0$, equality of the population distributions implies equality of the conditional action probabilities: $\mathrm{softmax}(\Theta_{i,(s,\cdot)}) = \pi^\star_{1,i}(\cdot \mid s)$. Since $\Theta_{i,(s,a_0)} = 0$, taking action-probability ratios gives

$$\Theta_{i,(s,a)} = \log \frac{\pi^\star_{1,i}(a \mid s)}{\pi^\star_{1,i}(a_0 \mid s)} = \Theta^\star_{i,(s,a)} \qquad \forall a, \quad d_{i,s} > 0.$$

Thus $P_d \Delta = 0$: the logits agree on every observed block.

Second, because $\Delta \in \mathcal{C}(\Theta^\star)$ and $c_0 > 0$, definition (39) gives

$$0 = \sum_{i,s} d_{i,s} \left\| \Delta_{i,(s,\cdot)} \right\|_2^2 \geq c_0 \left\| \Delta \right\|_F^2 .$$

Hence $\Delta = 0$ and $\Theta = \Theta^\star$, which establishes identification within the error cone. The single-fit estimation error belongs to this cone on the "good event" which happens with high probability, as we later show in Step 2 of Section A.5.

Example with positive restricted coverage with an unobserved block. The following example shows how the lower bound certifies positive coverage coefficient for the single logit fit procedure even when some source task-state pairs are never observed.

Example 7. Take $N = S = A = 2$, $k = 1$, $\mathcal{A} = \{a_0, a_1\}$. Order matrix columns as $(1, a_0), (1, a_1), (2, a_0), (2, a_1)$ and set

$$(d_{i,s})_{i,s} = \begin{pmatrix} 1/4 & 1/4 \\ 1/2 & 0 \end{pmatrix}, \qquad \Theta^\star = \theta \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad 0 < \theta < B.$$

Then $p_1 = p_2 = 1/2$. The task-2, state-2 block is unobserved.

For an anchored error matrix Δ, let $Z = \Delta$. Its four potentially nonzero coordinates are u, v, w, x, one for action a_1 in each task-state block; x corresponds to the unobserved block $(2, 2)$. Write

$$Z = \begin{pmatrix} 0 & u & 0 & v \\ 0 & w & 0 & x \end{pmatrix}, \qquad z = \mathrm{vec}(Z) = (0, u, 0, v, 0, w, 0, x)^\top.$$

The tangent projection keeps the first row and the $(1, a_1)$ column:

$$P_T Z = \begin{pmatrix} 0 & u & 0 & v \\ 0 & w & 0 & 0 \end{pmatrix}.$$

Consequently, both matrices in Lemma 5 are diagonal:

$$\mathbf{D}_d = \mathrm{diag}(1/4, 1/4, 1/4, 1/4, 1/2, 1/2, 0, 0), \qquad \mathbf{P}_T = \mathrm{diag}(1, 1, 1, 1, 0, 1, 0, 0).$$

The missing coordinate x is perpendicular to the tangent space. For $\tau \geq 0$,

$$\mathbf{D}_d - \tau \mathbf{P}_T = \mathrm{diag}(1/4 - \tau, 1/4 - \tau, 1/4 - \tau, 1/4 - \tau, 1/2, 1/2 - \tau, 0, 0),$$
$$\lambda_{\min}(\mathbf{D}_d - \tau \mathbf{P}_T) = \min\{1/4 - \tau, 0\}.$$

Since $k = 1$, the expression in (48) is

$$\lambda_{\min}(\mathbf{D}_d - \tau \mathbf{P}_T) + \frac{\tau}{19} = \begin{cases} \tau/19, & 0 \leq \tau \leq 1/4, \\ 1/4 - 18\tau/19, & \tau \geq 1/4. \end{cases}$$

It increases up to $\tau = 1/4$ and decreases thereafter. Lemma 5 therefore gives

$$c_0 \geq \frac{1}{76} > 0$$

A.5 Estimation for the single logit fit

We prove the statistical bound for the single-fit estimator in Section A.2.

Write $D := N + SA$.

Logit estimation. Use ν, λ_n, and the estimator $\widehat{\Theta}$ from (26)–(28) in Section A.2, with failure probability $0 < \delta < 1$.

Coverage condition. Use q, T, P_T, and $P_{T^\perp}$ from Section A.4, the anchored error cone $\mathcal{C}(\Theta^\star)$ from (38), and the coverage coefficient $c_0(d, \Theta^\star)$ from (39). We will use this anchored cone throughout Section A.5 and assume Assumption 6. We abbreviate this coefficient as c_0 for fixed d and $\Theta^\star$.

Curvature condition. To relate coverage coefficient c_0 to likelihood curvature, for each task state pair i, s, define matrix $H_{i,s} \in \mathbb{R}^{A\times A}$

$$H_{i,s} := \operatorname{diag}\big(\pi^\star_{1,i}(\cdot \mid s)\big) - \pi^\star_{1,i}(\cdot \mid s)\pi^\star_{1,i}(\cdot \mid s)^\top. \tag{49}$$

Define

$$Q(\Delta) := \sum_{i,s} d_{i,s}\Delta^\top_{i,(s,\cdot)} H_{i,s}\Delta_{i,(s,\cdot)}, \tag{50}$$

$$Q_n(\Delta) := \frac{1}{n}\sum_{t=1}^{n} \Delta^\top_{I_t,(S_t,\cdot)} H_{I_t,S_t}\Delta_{I_t,(S_t,\cdot)}. \tag{51}$$

We call Q the Fisher population quadratic and Q_n the Fisher empirical quadratic.

In Step 3 below we prove that bounded logits together with Assumption 6 imply

$$Q(\Delta) \geq \frac{e^{-2B}c_0}{A^2}\,\|\Delta\|_F^2 \qquad \forall \Delta \in \mathcal{C}(\Theta^\star). \tag{52}$$

Confidence radius for fitted logits. Use h_B and ε_n from (40)–(41).

Logit estimation guarantee.

Lemma 8. *Under the source sampling model in Section A.1 and Assumption 4, suppose Assumption 6 holds. With probability at least $1-\delta$, every minimizer in* (28) *satisfies*

$$\begin{aligned}&\left\|\widehat{\Theta} - \Theta^\star\right\|_F \\ &\leq \frac{A^2 e^{2B}(260B + 6/h_B)}{c_0}\left\{2\sqrt{\frac{2k\nu\log(2(N+SA)/\delta)}{n}} + \frac{4\sqrt{k}\,\sqrt{2}\log(2(N+SA)/\delta)}{3n}\right\}.\end{aligned} \tag{53}$$

Proof of Lemma 8. Use the tangent space T defined in Section A.4. For any error matrix Δ, write

$$\Delta_T := P_T\Delta, \qquad \Delta_{T^\perp} := P_{T^\perp}\Delta.$$

Step 1: the good event and its probability. Let $e_i \in \mathbb{R}^N$ and $f_{(s,a)} \in \mathbb{R}^{SA}$ denote the canonical basis vectors and define

$$X(i,s,a) := e_i f^\top_{(s,a)}. \tag{54}$$

The gradient of the empirical loss at the truth is

$$\nabla L_n(\Theta^\star) = \frac{1}{n}\sum_{t=1}^{n}\sum_{a\in\mathcal{A}}\left\{\pi^\star_{1,I_t}(a \mid S_t) - \mathbf{1}\{A_t = a\}\right\} X(I_t, S_t, a). \tag{55}$$

Conditional on $(I_t, S_t) = (i, s)$,

$$\mathbb{E}[e_{A_t} \mid i, s] = \pi^\star_{1,i}(\cdot \mid s).$$

Define the score summand

$$Y_t := \sum_{a \in \mathcal{A}} \left\{ \pi^\star_{1,I_t}(a \mid S_t) - \mathbf{1}\{A_t = a\} \right\} X(I_t, S_t, a). \tag{56}$$

Thus $\mathbb{E}[Y_t] = 0$, and

$$\|\nabla L_n(\Theta^\star)\|_{\mathrm{op}} = \left\| \frac{1}{n} \sum_{t=1}^{n} Y_t \right\|_{\mathrm{op}}. \tag{57}$$

For any probability vector $q \in \mathbb{R}^A$ and a categorical random action $A \sim q$,

$$\mathbb{E}[(q - e_A)(q - e_A)^\top] = \mathrm{diag}(q) - qq^\top, \qquad \mathbb{E}\|q - e_A\|_2^2 = 1 - \|q\|_2^2. \tag{58}$$

Because Y_t has one nonzero row and one nonzero state block, direct multiplication gives

$$\left\|\mathbb{E}[Y_t Y_t^\top]\right\|_{\mathrm{op}} = \max_i \sum_s d_{i,s} \left(1 - \|\pi^\star_{1,i}(\cdot \mid s)\|_2^2\right),$$

$$\left\|\mathbb{E}[Y_t^\top Y_t]\right\|_{\mathrm{op}} = \max_s \left\| \sum_i d_{i,s} H_{i,s} \right\|_{\mathrm{op}}.$$

For the first variance term, Cauchy–Schwarz gives, for every probability vector q,

$$1 = \left(\sum_a q_a \right)^2 \le A \sum_a q_a^2, \qquad 1 - \|q\|_2^2 \le 1 - \frac{1}{A}.$$

Applying this to $q = \pi^\star_{1,i}(\cdot \mid s)$ in each summand yields

$$\left\|\mathbb{E}[Y_t Y_t^\top]\right\|_{\mathrm{op}} \le \left(1 - \frac{1}{A}\right) \max_i \sum_s d_{i,s}.$$

For the second variance term, the symmetric matrix $\mathrm{diag}(q) - qq^\top$ has operator norm at most its largest absolute row sum. Its diagonal entries are $q_a(1 - q_a)$ and its off-diagonal entries are $-q_a q_b$, so

$$\left\|\mathrm{diag}(q) - qq^\top\right\|_{\mathrm{op}} \le \max_a \left\{ q_a(1 - q_a) + \sum_{b \ne a} q_a q_b \right\} = \max_a 2q_a(1 - q_a) \le \frac{1}{2}. \tag{59}$$

Thus $\|H_{i,s}\|_{\mathrm{op}} \le 1/2$. The triangle inequality and the nonnegative weights then give

$$\left\|\mathbb{E}[Y_t^\top Y_t]\right\|_{\mathrm{op}} \le \max_s \sum_i d_{i,s} \|H_{i,s}\|_{\mathrm{op}} \le \frac{1}{2} \max_s \sum_i d_{i,s}.$$

Consequently, both matrix variance terms are bounded by

$$\nu = \max \left\{ \left(1 - \frac{1}{A}\right) \max_i \sum_s d_{i,s}, \quad \frac{1}{2} \max_s \sum_i d_{i,s} \right\}.$$

Furthermore, for every probability vector q and action a, $\|q - e_a\|_2^2 = 1 - 2q_a + \|q\|_2^2 \le 2$. Hence

$$\|Y_t\|_{\mathrm{op}} = \|\pi^\star_{1,I_t}(\cdot \mid S_t) - e_{A_t}\|_2 \le \sqrt{2}.$$

Rectangular matrix Bernstein (Tropp, 2015, Theorem 6.1.1) therefore gives, for every $x > 0$,

$$\mathbb{P}\left[\left\| \frac{1}{n} \sum_{t=1}^{n} Y_t \right\|_{\mathrm{op}} > \sqrt{\frac{2\nu x}{n}} + \frac{2\sqrt{2}x}{3n} \right] \le De^{-x}. \tag{60}$$

Following the terminology of Wainwright (Wainwright, 2019, Section 10.2.2, proof of Proposition 10.6), the *good event* is the event that the likelihood gradient at the truth satisfies

$$\|\nabla L_n(\Theta^\star)\|_{\mathrm{op}} \le \frac{\lambda_n}{2} = \sqrt{\frac{2\nu}{n} \log \frac{2(N + SA)}{\delta}} + \frac{2\sqrt{2}}{3n} \log \frac{2(N + SA)}{\delta}. \tag{61}$$

Blocks of zero source probability contribute zero to the variance sums; no division by $d_{i,s}$ was used. Taking $x = \log \frac{2(N+SA)}{\delta}$ in (60) shows that the good event (61) has probability at least $1 - \delta/2$.

Step 2: cone condition and upper bound on D_{L_n}. Work on the good event (61). Set

$$\Delta := \widehat{\Theta} - \Theta^\star.$$

Define the Bregman divergence

$$D_{L_n}(\Theta^\star + \Delta, \Theta^\star) := L_n(\Theta^\star + \Delta) - L_n(\Theta^\star) - \langle\!\langle \nabla L_n(\Theta^\star),\, \Delta \rangle\!\rangle\,. \tag{62}$$

Optimality of $\widehat{\Theta}$ gives

$$D_{L_n}(\Theta^\star + \Delta, \Theta^\star) \le -\langle\!\langle \nabla L_n(\Theta^\star),\, \Delta \rangle\!\rangle + \lambda_n \left\{ \|\Theta^\star\|_* - \|\Theta^\star + \Delta\|_* \right\}. \tag{63}$$

The good event (61) and nuclear/operator duality imply

$$|\langle\!\langle \nabla L_n(\Theta^\star),\, \Delta \rangle\!\rangle| \le \frac{\lambda_n}{2} \|\Delta\|_*\,. \tag{64}$$

By the definition of $T^\perp$, the row and column spaces of $\Delta_{T^\perp}$ are orthogonal to those of $\Theta^\star$. In orthonormal bases adapted to these spaces, their sum is block diagonal, so the nuclear norms add:

$$\|\Theta^\star + \Delta_{T^\perp}\|_* = \|\Theta^\star\|_* + \|\Delta_{T^\perp}\|_*\,.$$

Using $\Delta = \Delta_T + \Delta_{T^\perp}$ and the reverse triangle inequality,

$$\begin{aligned} \|\Theta^\star + \Delta\|_* &= \|\Theta^\star + \Delta_{T^\perp} + \Delta_T\|_* \\ &\ge \|\Theta^\star + \Delta_{T^\perp}\|_* - \|\Delta_T\|_* \\ &= \|\Theta^\star\|_* + \|\Delta_{T^\perp}\|_* - \|\Delta_T\|_*\,. \end{aligned}$$

Rearranging gives

$$\|\Theta^\star\|_* - \|\Theta^\star + \Delta\|_* \le \|\Delta_T\|_* - \|\Delta_{T^\perp}\|_*\,. \tag{65}$$

Combining (63)–(65),

$$D_{L_n}(\Theta^\star + \Delta, \Theta^\star) \le \frac{3\lambda_n}{2} \|\Delta_T\|_* - \frac{\lambda_n}{2} \|\Delta_{T^\perp}\|_*\,. \tag{66}$$

The upper bound (66) uses the good event (61) through the gradient estimate (64). Since the Bregman divergence is nonnegative by convexity of L_n and $\lambda_n > 0$, we obtain, on this good event,

$$\|\Delta_{T^\perp}\|_* \le 3\,\|\Delta_T\|_*\,. \tag{67}$$

A tangent-space matrix has rank at most $2k$, and therefore

$$\|\Delta\|_* \le 4\,\|\Delta_T\|_* \le 4\sqrt{2k}\,\|\Delta\|_F\,, \tag{68}$$

$$D_{L_n}(\Theta^\star + \Delta, \Theta^\star) \le \frac{3\lambda_n}{2}\sqrt{2k}\,\|\Delta\|_F\,. \tag{69}$$

On the good event (61), $\Delta \in \mathcal{C}(\Theta^\star)$, so

$$\mathcal{C}(\Theta^\star) = \{0\} \quad \Longrightarrow \quad \Delta = 0.$$

The remaining steps concern $\mathcal{C}(\Theta^\star) \neq \{0\}$.

Step 3: likelihood and population restricted strong convexity. Recall Q, Q_n from (50)–(51).

Claim 9. Every anchored matrix satisfies

$$Q(\Delta) \ge \frac{e^{-2B}}{A^2} \sum_{i,s} d_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2. \tag{70}$$

Proof of the claim. We first bound each block below, and then sum the inequalities under source weights.

Fix (i, s). The bounded logits give, for every action a,

$$\pi^\star_{1,i}(a \mid s) = \frac{e^{\Theta^\star_{i,(s,a)}}}{\sum_b e^{\Theta^\star_{i,(s,b)}}} \ge \frac{e^{-B}}{Ae^{B}} = \frac{e^{-2B}}{A}.$$

The quadratic form $\Delta_{i,(s,\cdot)}^\top H_{i,s}\Delta_{i,(s,\cdot)}$ equals the variance of $\Delta_{i,(s,a)}$ under $a \sim \pi_{1,i}^\star(\cdot \mid s)$. Writing this variance as the minimum squared deviation from a constant gives

$$\begin{aligned}\Delta_{i,(s,\cdot)}^\top H_{i,s}\Delta_{i,(s,\cdot)} &= \min_{c\in\mathbb{R}} \sum_a \pi_{1,i}^\star(a \mid s)(\Delta_{i,(s,a)} - c)^2 \\ &\geq \frac{e^{-2B}}{A} \min_{c\in\mathbb{R}} \sum_a (\Delta_{i,(s,a)} - c)^2 \\ &= \frac{e^{-2B}}{A}\left[\|\Delta_{i,(s,\cdot)}\|_2^2 - \frac{1}{A}\left(\sum_a \Delta_{i,(s,a)}\right)^2\right].\end{aligned}$$

Anchoring controls the subtracted term: since $\Delta_{i,(s,a_0)} = 0$, Cauchy–Schwarz gives

$$\left(\sum_a \Delta_{i,(s,a)}\right)^2 = \left(\sum_{a\neq a_0} \Delta_{i,(s,a)}\right)^2 \leq (A-1)\sum_{a\neq a_0} \Delta_{i,(s,a)}^2 = (A-1)\|\Delta_{i,(s,\cdot)}\|_2^2.$$

Substitution therefore yields the required bound for each block:

$$\Delta_{i,(s,\cdot)}^\top H_{i,s}\Delta_{i,(s,\cdot)} \geq \frac{e^{-2B}}{A}\left(1 - \frac{A-1}{A}\right)\|\Delta_{i,(s,\cdot)}\|_2^2 = \frac{e^{-2B}}{A^2}\|\Delta_{i,(s,\cdot)}\|_2^2. \tag{71}$$

Finally, multiply by $d_{i,s} \geq 0$ and sum to obtain (70). For a cone direction, definition (39) supplies the stated comparison with the Frobenius norm. □

When $\mathcal{C}(\Theta^\star) \neq \{0\}$, every $\Delta \in \mathcal{C}(\Theta^\star)$ therefore satisfies

$$Q(\Delta) \geq \frac{e^{-2B}c_0}{A^2}\,|\!|\!|\Delta|\!|\!|_F^2\,,$$

by the definition (39). This proves (52).

We next prove the Hessian comparison (73) for the log-sum-exp function along any line segment in $[-B,B]^A$. Let $z, z' \in [-B,B]^A$, $q_u = \mathrm{softmax}(z + u(z'-z))$, and $q_0 = \mathrm{softmax}(z)$. Since every coordinate of $z'-z$ lies in $[-2B,2B]$,

$$\frac{q_u(a)}{q_0(a)} = \frac{e^{u(z'_a - z_a)}}{\sum_b q_0(b)e^{u(z'_b - z_b)}} \geq e^{-4Bu}. \tag{72}$$

Using

$$v^\top[\mathrm{diag}(q) - qq^\top]v = \min_c \sum_a q_a(v_a - c)^2,$$

coordinatewise domination in (72) implies

$$\mathrm{diag}(q_u) - q_u q_u^\top \succeq e^{-4Bu}\{\mathrm{diag}(q_0) - q_0 q_0^\top\}. \tag{73}$$

To apply this comparison to the empirical loss, fix an observation t and take $z = \Theta^\star_{I_t,(S_t,\cdot)}$ and $z' = \Theta^\star_{I_t,(S_t,\cdot)} + \Delta_{I_t,(S_t,\cdot)}$. Both blocks lie in $[-B,B]^A$ because the truth and the estimate are feasible. The linear term in l_t has zero second derivative, while the Hessian of log-sum-exp is the categorical covariance. Thus, with q_u corresponding to this block,

$$\frac{d^2}{du^2} l_t(\Theta^\star + u\Delta) = \Delta_{I_t,(S_t,\cdot)}^\top[\mathrm{diag}(q_u) - q_u q_u^\top]\Delta_{I_t,(S_t,\cdot)} \geq e^{-4Bu}\Delta_{I_t,(S_t,\cdot)}^\top H_{I_t,S_t}\Delta_{I_t,(S_t,\cdot)}.$$

Averaging over t gives

$$\frac{d^2}{du^2} L_n(\Theta^\star + u\Delta) \geq \frac{e^{-4Bu}}{n}\sum_{t=1}^n \Delta_{I_t,(S_t,\cdot)}^\top H_{I_t,S_t}\Delta_{I_t,(S_t,\cdot)} = e^{-4Bu}Q_n(\Delta).$$

The Bregman divergence is exactly the integral remainder after subtracting the first-order term. Taylor's theorem therefore gives

$$\begin{aligned} D_{L_n}(\Theta^\star + \Delta, \Theta^\star) &= L_n(\Theta^\star + \Delta) - L_n(\Theta^\star) - \langle\!\langle \nabla L_n(\Theta^\star),\, \Delta \rangle\!\rangle \\ &= \int_0^1 (1-u) \frac{d^2}{du^2} L_n(\Theta^\star + u\Delta)\, du \\ &\geq \left[\int_0^1 (1-u) e^{-4Bu}\, du \right] Q_n(\Delta) \\ &= \frac{4B - 1 + e^{-4B}}{16B^2} Q_n(\Delta). \end{aligned} \tag{74}$$

The inequality holds by the preceding second-derivative bound and $1 - u \geq 0$; the last equality is h_B from (40). For each fixed feasible Δ, taking expectations gives $D_L(\Theta^\star + \Delta, \Theta^\star) \geq h_B Q(\Delta)$ for the population loss $L := \mathbb{E}L_n$. Together with (52), this yields

$$D_L(\Theta^\star + \Delta, \Theta^\star) \geq \frac{h_B e^{-2B} c_0}{A^2} \|\Delta\|_F^2, \qquad \Delta \in \mathcal{C}(\Theta^\star), \quad \Theta^\star + \Delta \in \mathcal{M}_B,$$

which establishes population restricted strong convexity (RSC) on feasible cone directions; see Wainwright (Wainwright, 2019, Chapter 9, Section 9.3.1).

Step 4: empirical-process control at one deterministic radius. We bound the empirical-process error uniformly over eligible errors of Frobenius norm at most a fixed deterministic t, with failure probability at most $\delta/2$. Fix a deterministic $t > 0$. Define

$$\mathcal{B}(t) := \{\Delta \in \mathcal{C}(\Theta^\star) : \|\Delta\|_F \leq t, \quad \|\Delta\|_\infty \leq 2B\}. \tag{75}$$

Write $f_\Delta(i,s) = \Delta_{i,(s,\cdot)}^\top H_{i,s} \Delta_{i,(s,\cdot)}$. Let $q_a := \pi_{1,i}^\star(a \mid s)$. By the definition $H_{i,s} = \mathrm{diag}(q) - qq^\top$,

$$\begin{aligned} f_\Delta(i,s) &= \Delta_{i,(s,\cdot)}^\top [\mathrm{diag}(q) - qq^\top] \Delta_{i,(s,\cdot)} \\ &= \sum_a q_a \Delta_{i,(s,a)}^2 - \left(\sum_a q_a \Delta_{i,(s,a)} \right)^2 \\ &= \mathbb{E}_{a\sim q}[\Delta_{i,(s,a)}^2] - \left(\mathbb{E}_{a\sim q}[\Delta_{i,(s,a)}]\right)^2 = \mathrm{Var}_{a\sim q}(\Delta_{i,(s,a)}). \end{aligned}$$

Here (i,s) and Δ are fixed. Thus $f_\Delta(i,s) \geq 0$, and dropping the nonnegative squared mean gives $f_\Delta(i,s) \leq \sum_a q_a \Delta_{i,(s,a)}^2$. For $\Delta \in \mathcal{B}(t)$, the constraint $\|\Delta\|_\infty \leq 2B$ implies $\Delta_{i,(s,a)}^2 \leq 4B^2$ for every action. Since $\sum_a q_a = 1$,

$$\sum_a q_a \Delta_{i,(s,a)}^2 \leq 4B^2 \sum_a q_a = 4B^2.$$

Combining these three observations yields

$$0 \leq f_\Delta(i,s) \leq \sum_a \pi_{1,i}^\star(a \mid s) \Delta_{i,(s,a)}^2 \leq 4B^2. \tag{76}$$

Bounding the Fisher operator norm. We next justify $\|H_{i,s}\|_{\mathrm{op}} \leq 1/2$ directly. For a fixed block, retain $q_a = \pi_{1,i}^\star(a \mid s)$. The entries of $H_{i,s} = \mathrm{diag}(q) - qq^\top$ are

$$(H_{i,s})_{aa} = q_a(1 - q_a), \qquad (H_{i,s})_{ab} = -q_a q_b \quad (a \neq b).$$

Thus the sum of absolute values in row a satisfies

$$\begin{aligned} \sum_b |(H_{i,s})_{ab}| &= q_a(1-q_a) + \sum_{b \neq a} q_a q_b \\ &= q_a(1-q_a) + q_a(1-q_a) \\ &= 2q_a(1-q_a) \leq \frac{1}{2}. \end{aligned}$$

The second equality uses $\sum_{b\neq a} q_b = 1-q_a$; the last inequality uses $q_a(1-q_a) = 1/4-(q_a-1/2)^2 \leq 1/4$.

For any real matrix M and any vector x of the appropriate dimension, the triangle inequality and Cauchy–Schwarz give, for each row a,

$$\begin{aligned}
|(Mx)_a|^2 &= \left|\sum_b M_{ab}x_b\right|^2 \\
&\leq \left(\sum_b |M_{ab}|\,|x_b|\right)^2 \\
&= \left(\sum_b \sqrt{|M_{ab}|}\,\sqrt{|M_{ab}|}\,|x_b|\right)^2 \\
&\leq \left(\sum_b |M_{ab}|\right)\left(\sum_b |M_{ab}|\,|x_b|^2\right).
\end{aligned}$$

Summing over rows,

$$\begin{aligned}
\|Mx\|_2^2 &= \sum_a |(Mx)_a|^2 \\
&\leq \sum_a \left(\sum_b |M_{ab}|\right)\left(\sum_b |M_{ab}|\,|x_b|^2\right) \\
&\leq \left(\max_a \sum_b |M_{ab}|\right)\sum_a\sum_b |M_{ab}|\,|x_b|^2 \\
&= \left(\max_a \sum_b |M_{ab}|\right)\sum_b |x_b|^2\left(\sum_a |M_{ab}|\right) \\
&\leq \left(\max_a \sum_b |M_{ab}|\right)\left(\max_b \sum_a |M_{ab}|\right)\|x\|_2^2.
\end{aligned}$$

For the Euclidean operator norm, taking the supremum over $x \neq 0$ yields

$$\|M\|_{\mathrm{op}}^2 = \sup_{x\neq 0}\frac{\|Mx\|_2^2}{\|x\|_2^2} \leq \left(\max_a \sum_b |M_{ab}|\right)\left(\max_b \sum_a |M_{ab}|\right).$$

Since $H_{i,s}$ is symmetric,

$$\|H_{i,s}\|_{\mathrm{op}} \leq \max_a \sum_b |(H_{i,s})_{ab}| \leq \frac{1}{2}.$$

Converting source weights to the Frobenius norm. For any matrix Δ, Cauchy–Schwarz and the definition of the operator norm now give, block by block,

$$\begin{aligned}
\Delta_{i,(s,\cdot)}^\top H_{i,s}\Delta_{i,(s,\cdot)} &\leq \|\Delta_{i,(s,\cdot)}\|_2\,\|H_{i,s}\Delta_{i,(s,\cdot)}\|_2 \\
&\leq \|H_{i,s}\|_{\mathrm{op}}\,\|\Delta_{i,(s,\cdot)}\|_2^2 \leq \frac{1}{2}\|\Delta_{i,(s,\cdot)}\|_2^2.
\end{aligned}$$

Multiply these inequalities by $d_{i,s}$ and take the sum, we get

$$\begin{aligned} Q(\Delta) &= \sum_{i,s} d_{i,s} \Delta_{i,(s,\cdot)}^{\top} H_{i,s} \Delta_{i,(s,\cdot)} \\ &\le \frac{1}{2} \sum_{i,s} d_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2 \\ &\le \frac{1}{2} \left(\max_{i,s} d_{i,s} \right) \sum_{i,s} \|\Delta_{i,(s,\cdot)}\|_2^2 \\ &= \frac{1}{2} \left(\max_{i,s} d_{i,s} \right) \|\Delta\|_F^2 . \end{aligned}$$

Finally, for each fixed state s, nonnegativity of the design probabilities gives

$$d_{i,s} \le \sum_j d_{j,s} \qquad \text{for every task } i.$$

Taking maxima over i, s, and then using the second term in the maximum defining ν in (26), yields

$$\frac{1}{2} \max_{i,s} d_{i,s} \le \frac{1}{2} \max_s \sum_i d_{i,s} \le \nu.$$

Combining the preceding bounds proves, for every matrix Δ,

$$Q(\Delta) \le \frac{1}{2} \left(\max_{i,s} d_{i,s} \right) \|\Delta\|_F^2 \le \nu \|\Delta\|_F^2 . \tag{77}$$

Consequently, for every $\Delta \in \mathcal{B}(t)$,

$$Q(\Delta) \le \nu t^2, \qquad \mathrm{Var}(f_\Delta(I,S)) \le \mathbb{E} f_\Delta(I,S)^2 \le 4B^2 Q(\Delta) \le 4B^2 \nu t^2.$$

In addition, $0 \le Q(\Delta) \le 4B^2$, so $|f_\Delta(i,s) - Q(\Delta)| \le 4B^2$.

For $\eta \in \{-1, +1\}$ set

$$U_\eta := \sup_{\Delta \in \mathcal{B}(t)} \eta \{ Q_n(\Delta) - Q(\Delta) \} \ge 0.$$

For each fixed sign η and $\Delta \in \mathcal{B}(t)$, under $(I,S) \sim d$ the random variable $\eta\{f_\Delta(I,S) - Q(\Delta)\}$ has mean zero, absolute value at most $4B^2$, and variance at most $4B^2 \nu t^2$. Moreover,

$$U_\eta = \sup_{\Delta \in \mathcal{B}(t)} \frac{1}{n} \sum_{\ell=1}^{n} \eta \{ f_\Delta(I_\ell, S_\ell) - Q(\Delta) \}.$$

Apply Talagrand's concentration inequality in Wainwright (Wainwright, 2019, Eq. (3.86), p. 90) with $Z = U_\eta$, $b = 4B^2$, $\sigma^2 \le 4B^2 \nu t^2$, $\epsilon = 1$, and tail parameter x/n. Using the constants stated there, for each fixed $\eta \in \{-1, +1\}$ and $x > 0$, with probability at least $1 - e^{-x}$,

$$U_\eta \le 2\mathbb{E} U_\eta + 2Bt \sqrt{\frac{2\nu x}{n}} + \frac{28 B^2 x}{3n}. \tag{78}$$

For each $x > 0$, a union bound over $\eta \in \{-1, +1\}$ yields, with probability at least $1 - 2e^{-x}$,

$$\sup_{\Delta \in \mathcal{B}(t)} |Q_n(\Delta) - Q(\Delta)| = \max_{\eta \in \{-1,+1\}} U_\eta \le 2 \max_{\eta \in \{-1,+1\}} \mathbb{E} U_\eta + 2Bt \sqrt{\frac{2\nu x}{n}} + \frac{28 B^2 x}{3n}.$$

We next control the expectations in (78). Let σ_t be independent scalar Rademacher variables and $\epsilon_t \in \{-1, +1\}^A$ independent vectors with independent Rademacher coordinates, all independent of the sample. Let $(I_t', S_t')_{t=1}^n$ be an independent copy of the sample, and let $\mathbb{E}'$ denote expectation over this copy. For each fixed $\eta \in \{-1, +1\}$, since $Q(\Delta) = \mathbb{E}' f_\Delta(I_t', S_t')$,

$$\begin{aligned} \mathbb{E} U_\eta &= \mathbb{E} \sup_{\Delta \in \mathcal{B}(t)} \mathbb{E}' \left[\frac{1}{n} \sum_t \eta \{ f_\Delta(I_t, S_t) - f_\Delta(I_t', S_t') \} \right] \\ &\le \mathbb{E}\mathbb{E}' \sup_{\Delta \in \mathcal{B}(t)} \frac{1}{n} \sum_t \eta \{ f_\Delta(I_t, S_t) - f_\Delta(I_t', S_t') \}. \end{aligned} \tag{79}$$

The inequality uses $\sup_\Delta \mathbb{E}'[\,\cdot\,] \le \mathbb{E}'[\sup_\Delta(\,\cdot\,)]$ with the original sample held fixed.

For each t, interchanging (I_t, S_t) and (I'_t, S'_t) preserves the sampling law and reverses the difference for every Δ simultaneously. Hence the collection of differences has the same joint distribution after multiplication by the independent signs σ_t, giving

$$\begin{aligned}
&\mathbb{E}\sup_{\Delta\in\mathcal{B}(t)}\frac{1}{n}\sum_t \eta\{f_\Delta(I_t,S_t)-f_\Delta(I'_t,S'_t)\}\\
=&\mathbb{E}\sup_{\Delta\in\mathcal{B}(t)}\frac{1}{n}\sum_t \eta\sigma_t\{f_\Delta(I_t,S_t)-f_\Delta(I'_t,S'_t)\}\\
\le&2\mathbb{E}\sup_{\Delta\in\mathcal{B}(t)}\frac{1}{n}\sum_t \sigma_t f_\Delta(I_t,S_t). \qquad (80)
\end{aligned}$$

For the final inequality, bound the supremum of the two sums by the sum of their suprema. Their expectations coincide because the samples have the same law and both $\eta\sigma_t$ and $-\eta\sigma_t$ are Rademacher signs.

For every $\Delta \in \mathcal{B}(t)$ and (i,s), (76) gives

$$\|H_{i,s}^{1/2}\Delta_{i,(s,\cdot)}\|_2^2 = f_\Delta(i,s) \le 4B^2.$$

Consequently, for any $\Delta, \Gamma \in \mathcal{B}(t)$, Cauchy–Schwarz yields

$$\begin{aligned}
|f_\Delta(i,s)-f_\Gamma(i,s)| &\le \left(\|H_{i,s}^{1/2}\Delta_{i,(s,\cdot)}\|_2+\|H_{i,s}^{1/2}\Gamma_{i,(s,\cdot)}\|_2\right)\|H_{i,s}^{1/2}(\Delta-\Gamma)_{i,(s,\cdot)}\|_2\\
&\le 4B\|H_{i,s}^{1/2}(\Delta-\Gamma)_{i,(s,\cdot)}\|_2.
\end{aligned}$$

Conditional on the sample, apply the vector contraction inequality (Maurer, 2016, Theorem 3) to the scalar-valued functions $\Delta \mapsto f_\Delta(I_\ell, S_\ell)$ and vector-valued functions $\Delta \mapsto 4BH_{I_\ell,S_\ell}^{1/2}\Delta_{I_\ell,(S_\ell,\cdot)}$ on $\mathcal{B}(t)$. This gives

$$\mathbb{E}_\sigma\sup_{\Delta\in\mathcal{B}(t)}\frac{1}{n}\sum_{\ell=1}^n \sigma_\ell f_\Delta(I_\ell,S_\ell) \le 4\sqrt{2}B\,\mathbb{E}_\epsilon\sup_{\Delta\in\mathcal{B}(t)}\frac{1}{n}\sum_{\ell=1}^n\langle\epsilon_\ell, H_{I_\ell,S_\ell}^{1/2}\Delta_{I_\ell,(S_\ell,\cdot)}\rangle. \qquad (81)$$

Here $\mathbb{E}_\sigma$ and $\mathbb{E}_\epsilon$ denote expectations over the Rademacher variables $(\sigma_\ell)_{\ell=1}^n$ and vectors $(\epsilon_\ell)_{\ell=1}^n$, respectively, with the sample $(I_\ell, S_\ell)_{\ell=1}^n$ held fixed. We denote

$$\Xi_t := \sum_a [H_{I_t,S_t}^{1/2}\epsilon_t]_a X(I_t,S_t,a). \qquad (82)$$

Combining (79)–(81) yields

$$\begin{aligned}
\mathbb{E}U_\eta &\le 8\sqrt{2}B\,\mathbb{E}\sup_{\Delta\in\mathcal{B}(t)}\frac{1}{n}\sum_t\langle\epsilon_t, H_{I_t,S_t}^{1/2}\Delta_{I_t,(S_t,\cdot)}\rangle\\
&= 8\sqrt{2}B\,\mathbb{E}\sup_{\Delta\in\mathcal{B}(t)}\left|\left\langle\!\left\langle\frac{1}{n}\sum_t\Xi_t,\,\Delta\right\rangle\!\right\rangle\right|\\
&\le 64B\sqrt{k}\,t\,\mathbb{E}\left\|\frac{1}{n}\sum_t\Xi_t\right\|_{\mathrm{op}}. \qquad (83)
\end{aligned}$$

The last inequality uses (68). Now fix the observations $(I_\ell, S_\ell)_{\ell=1}^n$ and the Rademacher vectors $(\epsilon_\ell)_{\ell=1}^n$, and define

$$\mathcal{L}(\Delta) := \left\langle\!\left\langle\frac{1}{n}\sum_{\ell=1}^n\Xi_\ell,\,\Delta\right\rangle\!\right\rangle.$$

The constraints in (75) are unchanged under $\Delta \mapsto -\Delta$. Thus

$$\Delta\in\mathcal{B}(t) \iff -\Delta\in\mathcal{B}(t), \qquad \mathcal{L}(-\Delta) = -\mathcal{L}(\Delta).$$

Consequently,

$$\begin{aligned}\sup_{\Delta\in\mathcal{B}(t)}|\mathcal{L}(\Delta)| &= \max\left\{\sup_{\Delta\in\mathcal{B}(t)}\mathcal{L}(\Delta),\ \sup_{\Delta\in\mathcal{B}(t)}[-\mathcal{L}(\Delta)]\right\}\\ &= \max\left\{\sup_{\Delta\in\mathcal{B}(t)}\mathcal{L}(\Delta),\ \sup_{\Delta\in\mathcal{B}(t)}\mathcal{L}(-\Delta)\right\}\\ &= \sup_{\Delta\in\mathcal{B}(t)}\mathcal{L}(\Delta).\end{aligned}$$

We next bound $\mathbb{E}\left\|\frac{1}{n}\sum_t \Xi_t\right\|_{\mathrm{op}}$ in (83) using matrix Bernstein. Conditional on $(I_t, S_t) = (i, s)$,

$$\mathbb{E}_\epsilon \Xi_t = 0, \qquad \mathbb{E}_\epsilon[(H_{i,s}^{1/2}\epsilon_t)(H_{i,s}^{1/2}\epsilon_t)^\top] = H_{i,s}.$$

By (82), the entries of Ξ_t are

$$[\Xi_t]_{j,(\tilde{s},a)} = \begin{cases}[H_{i,s}^{1/2}\epsilon_t]_a, & j = i \text{ and } \tilde{s} = s,\\ 0, & \text{otherwise.}\end{cases}$$

Thus only row i and columns $(s,1),\dots,(s,A)$ can contain nonzero entries. The two matrix products are

$$[\Xi_t\Xi_t^\top]_{j,k} = \sum_{\tilde{s},a}[\Xi_t]_{j,(\tilde{s},a)}[\Xi_t]_{k,(\tilde{s},a)} = \mathbf{1}\{j=i,\ k=i\}\epsilon_t^\top H_{i,s}\epsilon_t,$$

$$[\Xi_t^\top\Xi_t]_{(\tilde{s},a),(\tilde{s}',b)} = \sum_j [\Xi_t]_{j,(\tilde{s},a)}[\Xi_t]_{j,(\tilde{s}',b)} = \mathbf{1}\{\tilde{s}=s,\ \tilde{s}'=s\}[H_{i,s}^{1/2}\epsilon_t]_a[H_{i,s}^{1/2}\epsilon_t]_b.$$

Taking expectations over the signs uses

$$\mathbb{E}_\epsilon[\epsilon_t^\top H_{i,s}\epsilon_t] = \mathrm{tr}(H_{i,s}), \qquad \mathbb{E}_\epsilon\left[[H_{i,s}^{1/2}\epsilon_t]_a[H_{i,s}^{1/2}\epsilon_t]_b\right] = (H_{i,s})_{a,b}.$$

Using the basis vectors $e_i, f_{(s,a)}$ from (54) and $\mathbb{P}\{(I_t,S_t)=(i,s)\} = d_{i,s}$, we obtain

$$\mathbb{E}[\Xi_t\Xi_t^\top] = \sum_{i,s} d_{i,s}\mathrm{tr}(H_{i,s})e_ie_i^\top = \mathrm{diag}_{i=1}^N\left(\sum_s d_{i,s}\mathrm{tr}(H_{i,s})\right),$$

$$\mathbb{E}[\Xi_t^\top\Xi_t] = \sum_{i,s} d_{i,s}\sum_{a,b}(H_{i,s})_{a,b}f_{(s,a)}f_{(s,b)}^\top = \mathrm{blockdiag}_{s=1}^S\left(\sum_i d_{i,s}H_{i,s}\right).$$

Using $\mathrm{tr}(H_{i,s}) = 1 - \sum_a \pi^\star_{1,i}(a\mid s)^2 \le 1 - 1/A$ gives

$$\left\|\mathbb{E}[\Xi_t\Xi_t^\top]\right\|_{\mathrm{op}} = \max_i\sum_s d_{i,s}\mathrm{tr}(H_{i,s}) \le \left(1-\frac{1}{A}\right)\max_i\sum_s d_{i,s} \le \nu.$$

On the other hand,

$$\left\|\mathbb{E}[\Xi_t^\top\Xi_t]\right\|_{\mathrm{op}} = \max_s\left\|\sum_i d_{i,s}H_{i,s}\right\|_{\mathrm{op}} \le \max_s\sum_i d_{i,s}\left\|H_{i,s}\right\|_{\mathrm{op}} \le \frac{1}{2}\max_s\sum_i d_{i,s} \le \nu.$$

Both final inequalities use the definition of ν in (26). Also,

$$\left\|\Xi_t\right\|_{\mathrm{op}}^2 = \epsilon_t^\top H_{I_t,S_t}\epsilon_t = 1 - \left(\sum_a \pi^\star_{1,I_t}(a\mid S_t)\epsilon_{t,a}\right)^2 \le 1 \le 2.$$

Thus each summand Ξ_t is independent, mean zero, bounded by $\sqrt{2}$, with sum variance statistic at most $n\nu$. The expectation part of rectangular matrix Bernstein (Tropp, 2015, Theorem 6.1.1) gives

$$\mathbb{E}\left\|\frac{1}{n}\sum_t\Xi_t\right\|_{\mathrm{op}} \le \sqrt{\frac{2\nu\log D}{n}} + \frac{\sqrt{2}\log D}{3n}. \tag{84}$$

Combining (78)–(84), with probability at least $1 - 2e^{-x}$,

$$\sup_{\Delta \in \mathcal{B}(t)} |Q_n(\Delta) - Q(\Delta)| \le 128B\sqrt{k}\, t \left[\sqrt{\frac{2\nu \log D}{n}} + \frac{\sqrt{2}\log D}{3n} \right] + 2Bt\sqrt{\frac{2\nu x}{n}} + \frac{28B^2 x}{3n}. \quad (85)$$

Set $x = \log(4/\delta)$, so $1 - 2e^{-x} = 1 - \delta/2$. Since $D = N + SA \ge 2$, we have $x = \log \frac{4}{\delta} \le \log \frac{2D}{\delta}$ and $\log D \le \log \frac{2D}{\delta}$. The definition of λ_n in (27) therefore gives

$$\sqrt{\frac{2\nu \log D}{n}} + \frac{\sqrt{2}\log D}{3n} \le \frac{\lambda_n}{2},$$
$$\sqrt{\frac{2\nu x}{n}} \le \frac{\lambda_n}{2}.$$

Using $k \ge 1$, the first two terms in (85) are thus bounded by

$$64B\sqrt{k}\,\lambda_n t + B\lambda_n t \le 65B\sqrt{k}\,\lambda_n t.$$

Therefore, for the fixed deterministic radius $t > 0$, with probability at least $1 - \delta/2$,

$$\sup_{\Delta \in \mathcal{B}(t)} |Q_n(\Delta) - Q(\Delta)| \le 65B\sqrt{k}\lambda_n t + \frac{28B^2}{3n} \log \frac{4}{\delta}. \quad (86)$$

This is a statement at the one fixed deterministic t. It is not a simultaneous statement over radii.

Step 5: empirical curvature beyond a deterministic radius. We first derive a sufficient lower bound on a deterministic radius t, then choose t_0 accordingly. Fix $t > 0$ and consider any Γ satisfying

$$\Gamma \in \mathcal{B}(t), \qquad \|\Gamma\|_F = t.$$

By (52),

$$Q(\Gamma) \ge \frac{e^{-2B} c_0}{A^2} \|\Gamma\|_F^2 = \frac{e^{-2B} c_0}{A^2} t^2 > 0.$$

On the fixed-radius event (86), with $x = \log(4/\delta)$,

$$\begin{aligned} \frac{|Q_n(\Gamma) - Q(\Gamma)|}{Q(\Gamma)} &\le \frac{65B\sqrt{k}\,\lambda_n t + 28B^2 x/(3n)}{Q(\Gamma)} && \text{by (86)}, \\ &\le \frac{A^2 e^{2B}\{65B\sqrt{k}\,\lambda_n t + 28B^2 x/(3n)\}}{c_0 t^2} && \text{by (52) and } \|\Gamma\|_F = t, \\ &= \frac{65A^2 e^{2B} B\sqrt{k}\,\lambda_n}{c_0 t} + \frac{28A^2 e^{2B} B^2 x}{3nc_0 t^2}. \end{aligned}$$

Setting each term to be at most $1/4$ gives

$$\frac{65A^2 e^{2B} B\sqrt{k}\,\lambda_n}{c_0 t} \le \frac{1}{4} \iff t \ge \frac{260A^2 e^{2B} B\sqrt{k}\,\lambda_n}{c_0},$$
$$\frac{28A^2 e^{2B} B^2 x}{3nc_0 t^2} \le \frac{1}{4} \iff t \ge \sqrt{\frac{112A^2 e^{2B} B^2 x}{3nc_0}}.$$

For any $\Gamma \in \mathcal{C}(\Theta^\star) \setminus \{0\}$, equations (52) and (77) imply

$$\frac{e^{-2B} c_0}{A^2} \|\Gamma\|_F^2 \le Q(\Gamma) \le \nu \|\Gamma\|_F^2 \quad \Longrightarrow \quad \frac{e^{-2B} c_0}{A^2} \le \nu.$$

The definition of λ_n in (27), $k \ge 1$, and $x = \log(4/\delta) \le \log(2(N + SA)/\delta)$ give

$$k\lambda_n^2 \ge \frac{8k\nu}{n} \log \frac{2(N + SA)}{\delta} \ge \frac{8e^{-2B} c_0 x}{A^2 n}.$$

Consequently, the second lower bound on t satisfies

$$\sqrt{\frac{112A^2e^{2B}B^2x}{3nc_0}} \le \sqrt{\frac{14}{3}}\,\frac{A^2e^{2B}B\sqrt{k}\,\lambda_n}{c_0}.$$

Since $\sqrt{14/3} < 260$, the first condition implies the second. We therefore impose

$$t \ge \frac{260A^2e^{2B}B\sqrt{k}\,\lambda_n}{c_0}.$$

We also require that t bound the estimation error $\Delta = \widehat{\Theta} - \Theta^\star$. On the good event (61), once $Q_n(\Delta) \ge Q(\Delta)/2$ has been established for $\|\Delta\|_F \ge t$, the bounds (52), (74), and (69) give

$$\frac{h_Be^{-2B}c_0}{2A^2}\,\|\Delta\|_F^2 \le D_{L_n}(\Theta^\star+\Delta,\Theta^\star) \le \frac{3\sqrt{2}}{2}\sqrt{k}\,\lambda_n\,\|\Delta\|_F\,.$$

For $\|\Delta\|_F \ge t > 0$, division by $\|\Delta\|_F$ shows that the additional requirement

$$t \ge \frac{3\sqrt{2}A^2e^{2B}\sqrt{k}\,\lambda_n}{h_Bc_0}$$

ensures $\|\Delta\|_F \le t$. The case $\|\Delta\|_F < t$ already satisfies this bound. Since $3\sqrt{2} \le 6$, the following radius satisfies both requirements:

$$t_0 := \frac{A^2e^{2B}(260B + 6/h_B)\sqrt{k}\,\lambda_n}{c_0}. \tag{87}$$

Apply (86) once with $t = t_0$; its event holds uniformly over $\mathcal{B}(t_0)$.

For any $\Gamma \in \mathcal{C}(\Theta^\star)$ with $\|\Gamma\|_\infty \le 2B$ and $\|\Gamma\|_F \ge t_0$, scale to that radius:

$$\Gamma_{t_0} := \frac{t_0}{\|\Gamma\|_F}\Gamma \in \mathcal{C}(\Theta^\star), \qquad \|\Gamma_{t_0}\|_F = t_0, \qquad \|\Gamma_{t_0}\|_\infty \le \|\Gamma\|_\infty \le 2B.$$

Thus $\Gamma_{t_0} \in \mathcal{B}(t_0)$. We evaluate the Step 4 bound at this scaled matrix Γ_{t_0}:

$$\begin{aligned}|Q_n(\Gamma_{t_0}) - Q(\Gamma_{t_0})| &\le \sup_{\Delta\in\mathcal{B}(t_0)} |Q_n(\Delta) - Q(\Delta)| \\ &\le 65B\sqrt{k}\,\lambda_n t_0 + \frac{28B^2}{3n}\log\frac{4}{\delta} \qquad \text{by (86).}\end{aligned}$$

The empirical-process bound holds simultaneously for every $\Delta \in \mathcal{B}(t_0)$ on this event. Also, (52) gives $Q(\Gamma_{t_0}) \ge \frac{e^{-2B}c_0}{A^2}t_0^2$. To transfer the resulting relative-error bound to the original matrix Γ, use quadratic homogeneity:

$$Q_n(\Gamma_{t_0}) = \frac{t_0^2}{\|\Gamma\|_F^2}Q_n(\Gamma), \qquad Q(\Gamma_{t_0}) = \frac{t_0^2}{\|\Gamma\|_F^2}Q(\Gamma).$$

Consequently, on the same fixed-radius event,

$$\frac{|Q_n(\Gamma) - Q(\Gamma)|}{Q(\Gamma)} = \frac{|Q_n(\Gamma_{t_0}) - Q(\Gamma_{t_0})|}{Q(\Gamma_{t_0})} \le \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

We obtain the empirical curvature bound

$$Q_n(\Gamma) \ge \frac{1}{2}Q(\Gamma) \ge \frac{e^{-2B}c_0}{2A^2}\,\|\Gamma\|_F^2\,, \qquad \Gamma \in \mathcal{C}(\Theta^\star), \quad \|\Gamma\|_\infty \le 2B, \quad \|\Gamma\|_F \ge t_0. \tag{88}$$

On the event in (86) with $t = t_0$, which has probability at least $1 - \delta/2$, inequality (88) holds simultaneously for every $\Gamma \in \mathcal{C}(\Theta^\star)$ satisfying $\|\Gamma\|_\infty \le 2B$ and $\|\Gamma\|_F \ge t_0$.

Step 6: apply optimality and obtain the error bound. On the intersection of (61) and (86) at $t = t_0$, fix any minimizer $\widehat{\Theta}$ and set $\Delta = \widehat{\Theta} - \Theta^\star$. If $\|\Delta\|_F < t_0$, the desired bound already holds by the choice (87) of t_0. Otherwise, (67) and feasibility give $\Delta \in \mathcal{C}(\Theta^\star)$ and $\|\Delta\|_\infty \le 2B$. Combining (88) with the Bregman bounds (74) and (69) yields

$$\frac{h_B e^{-2B} c_0}{2A^2} \|\Delta\|_F^2 \le h_B Q_n(\Delta) \le D_{L_n}(\Theta^\star + \Delta, \Theta^\star) \le \frac{3\sqrt{2}}{2} \sqrt{k}\, \lambda_n \|\Delta\|_F .$$

Dividing by $h_B e^{-2B} c_0 \|\Delta\|_F / (2A^2) > 0$ and using (87), we get

$$\|\Delta\|_F \le \frac{3\sqrt{2} A^2 e^{2B} \sqrt{k}\, \lambda_n}{h_B c_0} \le t_0.$$

We next verify that the good event and the fixed-radius event hold simultaneously with probability at least $1 - \delta$.

Denote the good event (61) by

$$\mathcal{E}_{\text{good}} := \left\{ \|\nabla L_n(\Theta^\star)\|_{\text{op}} \le \frac{\lambda_n}{2} \right\}.$$

By (60) and (27),

$$\mathbb{P}(\mathcal{E}_{\text{good}}^c) \le D \exp\left[-\log \frac{2(N + SA)}{\delta} \right] = \frac{\delta}{2}.$$

Use the deterministic t_0 from (87) and define

$$x := \log \frac{4}{\delta}, \qquad \mathcal{E}_{\text{loc}} := \left\{ \sup_{\Gamma \in \mathcal{B}(t_0)} |Q_n(\Gamma) - Q(\Gamma)| \le 65 B \sqrt{k}\, \lambda_n t_0 + \frac{28 B^2 x}{3n} \right\}.$$

Equations (85)–(86), at $t = t_0$, give

$$\mathbb{P}(\mathcal{E}_{\text{loc}}^c) \le 2e^{-x} = 2\frac{\delta}{4} = \frac{\delta}{2}.$$

The union bound gives

$$\mathbb{P}(\mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{loc}}) \ge 1 - \mathbb{P}(\mathcal{E}_{\text{good}}^c) - \mathbb{P}(\mathcal{E}_{\text{loc}}^c) \ge 1 - \delta.$$

Write the full set of minimizers in (28) as

$$\widehat{\mathcal{M}} := \operatorname*{argmin}_{\Theta \in \mathcal{M}_B} \{ L_n(\Theta) + \lambda_n \|\Theta\|_* \}.$$

On $\mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{loc}}$, equations (67) and (69) give, for every $\widehat{\Theta} \in \widehat{\mathcal{M}}$,

$$\Delta := \widehat{\Theta} - \Theta^\star \in \mathcal{C}(\Theta^\star), \qquad D_{L_n}(\Theta^\star + \Delta, \Theta^\star) \le \frac{3\lambda_n}{2} \sqrt{2k}\, \|\Delta\|_F .$$

Because $\Theta^\star \neq 0$ and $\Theta^\star \in T$, the cone is nontrivial, so Step 6 applies to each such minimizer and gives $\|\Delta\|_F \le t_0$. Thus, using (87) and (41), on $\mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{loc}}$,

$$\sup_{\widehat{\Theta} \in \widehat{\mathcal{M}}} \left\| \widehat{\Theta} - \Theta^\star \right\|_F \le \varepsilon_n = \frac{A^2 e^{2B} (260B + 6/h_B) \sqrt{k}\, \lambda_n}{c_0}.$$

Consequently,

$$\mathbb{P}\left[\sup_{\widehat{\Theta} \in \widehat{\mathcal{M}}} \left\| \widehat{\Theta} - \Theta^\star \right\|_F \le \varepsilon_n \right] \ge \mathbb{P}(\mathcal{E}_{\text{good}} \cap \mathcal{E}_{\text{loc}}) \ge 1 - \delta,$$

which is (53). □

A.6 Reward identification and value transfer

Section A.5 bounds the error in the fitted source logits. To turn this into a target regret bound, we first identify the rewards and then bound the resulting errors in target-policy values. Under the normalization in Assumption 5, each task's source log-policy uniquely determines its reward (Lemma 10). The Lipschitz log-softmax map and the affine reward-recovery formula thus propagate logit error into reward error. For a fixed target policy, the value error is the expected discounted sum of reward errors along its trajectories. Lemma 11 combines these relations to obtain simultaneous value-error bounds for all tasks and candidate policies.

Lemma 12 in Section A.7 applies these simultaneous value-error bounds to plug-in selection. It bounds the value loss relative to a comparator by the sum of that comparator's radius and the selected policy's radius. Taking the comparator to be $\pi^\star_{2,i}$ gives the regret bound.

A.6.1 Reward recovery under the normalization

Lemma 10 (Normalized reward is affine in the log-policy)**.** *Under Assumption 5, let* $u_i^\star(s,a) := \log \pi^\star_{1,i}(a \mid s)$ *and*

$$L_\mu := C_\mu(I - \gamma_1 \mathsf{P}_1 \mu)^{-1}.$$

The unique reward inducing the source soft-optimal policy $\pi^\star_{1,i}$ *and satisfying* (19) *is*

$$r_i = L_\mu(u_i^\star - g) + g = L_\mu u_i^\star + (I - L_\mu)g.$$

Thus, for fixed source dynamics, discount, and normalization (μ, g)*, reward recovery is affine in the source log-policy.*

Proof. Fix task i. First determine all reward–value pairs $(r_i, V_{1,i})$ consistent with the source policy $\pi^\star_{1,i}$. The unit-temperature soft Bellman equations are

$$\begin{aligned} Q_{1,i} &= r_i + \gamma_1 \mathsf{P}_1 V_{1,i}, \\ V_{1,i}(s) &= \log \sum_a e^{Q_{1,i}(s,a)}, \qquad \pi^\star_{1,i}(a \mid s) = e^{Q_{1,i}(s,a) - V_{1,i}(s)}. \end{aligned} \tag{89}$$

Taking logarithms of the policy equation gives $Q_{1,i} = u_i^\star + V_{1,i}$. Substituting into the first Bellman equation shows that every solution belongs to the family

$$\left\{ \left(u_i^\star + V - \gamma_1 \mathsf{P}_1 V,\, V\right) : V \in \mathbb{R}^S \right\}.$$

Conversely, choose any state function V, set $r_i = u_i^\star + V - \gamma_1 \mathsf{P}_1 V$, and let $Q_{1,i} = u_i^\star + V$. Then $Q_{1,i} = r_i + \gamma_1 \mathsf{P}_1 V$, and $\sum_a e^{u_i^\star(s,a)} = 1$ gives

$$\begin{aligned} \log \sum_a e^{Q_{1,i}(s,a)} &= V(s) + \log \sum_a e^{u_i^\star(s,a)} = V(s), \\ \operatorname{softmax}(Q_{1,i}(s,\cdot))_a &= e^{u_i^\star(s,a)} = \pi^\star_{1,i}(a \mid s). \end{aligned}$$

Thus every pair in the displayed family solves the soft Bellman equations for the observed policy.

Now impose $\mu r_i = g$. Within this family, the constraint is

$$g = \mu r_i = \mu u_i^\star + V_{1,i} - \gamma_1 \mu \mathsf{P}_1 V_{1,i}$$

Rearranging the preceding equality gives

$$(I - \gamma_1 \mu \mathsf{P}_1) V_{1,i} = g - \mu u_i^\star.$$

Since $\mu \mathsf{P}_1$ is a stochastic state-to-state kernel and $\gamma_1 < 1$, this determines a unique state value

$$V_{1,i} = (I - \gamma_1 \mu \mathsf{P}_1)^{-1}(g - \mu u_i^\star),$$

and hence a unique reward $r_i = u_i^\star + V_{1,i} - \gamma_1 \mathsf{P}_1 V_{1,i}$.

Substituting this value of $V_{1,i}$ directly gives

$$r_i = u_i^\star + (I - \gamma_1 \mu \mathsf{P}_1)^{-1}(g - \mu u_i^\star) - \gamma_1 \mathsf{P}_1 (I - \gamma_1 \mu \mathsf{P}_1)^{-1}(g - \mu u_i^\star).$$

Since $V_{1,i}$ and g depend only on the state and $\sum_a \mu(a \mid s) = 1$, averaging over actions gives $\mu V_{1,i} = V_{1,i}$ and $\mu g = g$. The definition (21) of C_μ and linearity of μ then give

$$C_\mu V_{1,i} = V_{1,i} - \mu V_{1,i} = 0, \qquad C_\mu g = g - \mu g = 0,$$
$$g - \mu u_i^\star = \mu g - \mu u_i^\star = -\mu(u_i^\star - g).$$

Using these identities and $\mu r_i = g$, we obtain

$$\begin{aligned} r_i &= g + C_\mu r_i = g + C_\mu(u_i^\star - g - \gamma_1 \mathsf{P}_1 V_{1,i}) \\ &= g + C_\mu \left[(u_i^\star - g) + \gamma_1 \mathsf{P}_1 (I - \gamma_1 \mu \mathsf{P}_1)^{-1} \mu (u_i^\star - g)\right] \\ &= g + C_\mu \left[I + \gamma_1 \mathsf{P}_1 (I - \gamma_1 \mu \mathsf{P}_1)^{-1} \mu\right] (u_i^\star - g). \end{aligned}$$

The operator in brackets equals $(I - \gamma_1 \mathsf{P}_1 \mu)^{-1}$ since direct multiplication gives

$$\begin{aligned} &(I - \gamma_1 \mathsf{P}_1 \mu) \left[I + \gamma_1 \mathsf{P}_1 (I - \gamma_1 \mu \mathsf{P}_1)^{-1} \mu\right] \\ &= I - \gamma_1 \mathsf{P}_1 \mu + \gamma_1 \mathsf{P}_1 (I - \gamma_1 \mu \mathsf{P}_1)(I - \gamma_1 \mu \mathsf{P}_1)^{-1} \mu \\ &= I. \end{aligned}$$

Therefore

$$r_i = C_\mu (I - \gamma_1 \mathsf{P}_1 \mu)^{-1} (u_i^\star - g) + g.$$

Since L_μ is linear and $(I - L_\mu)g$ does not depend on $u_i^\star$, the recovery map is affine in $u_i^\star$. □

A.6.2 From fitted logits to target-policy values

The following lemma converts the logit-error bound into simultaneous bounds on the value errors of all candidate policies.

Lemma 11. *Suppose Assumption 5 holds and $\Theta^\star \in \mathcal{M}_B$. For any feasible matrix $\widehat{\Theta} \in \mathcal{M}_B$, construct $\widehat{u}_i$, $\widehat{r}_i$, and $\widehat{J}_{2,i}$ by* (29)–(31)*. For any $\epsilon \geq 0$, the implication*

$$\begin{aligned} &\left\|\left|\widehat{\Theta} - \Theta^\star\right|\right\|_F \leq \epsilon \\ &\implies \quad \left|\widehat{J}_{2,i}(\pi) - J_{2,i}(\pi; r_i)\right| \leq \frac{L_B \epsilon}{1 - \gamma_2} \left\{ \sum_s \left\| \left[\left[C_\mu (I - \gamma_1 \mathsf{P}_1 \mu)^{-1} \right]^\top \rho_2^\pi \right]_{(s,\cdot)} \right\|_2^2 \right\}^{1/2} \end{aligned} \tag{90}$$

holds simultaneously for every task i and candidate $\pi \in \Pi$. Here L_B is the constant in (42)*.*

Proof. Fix i and π and write

$$\Delta := \widehat{\Theta} - \Theta^\star.$$

Step 1: the blockwise log-softmax difference. For each state s, the fundamental theorem of calculus applied to log-sum-exp gives

$$\begin{aligned} &\log \sum_b e^{\widehat{\Theta}_{i,(s,b)}} - \log \sum_b e^{\Theta^\star_{i,(s,b)}} = \int_0^1 \sum_b \mathrm{softmax}(\Theta^\star_{i,(s,\cdot)} + t\Delta_{i,(s,\cdot)})_b \Delta_{i,(s,b)}\, dt \\ &= \bar{\pi}_{i,(s,\cdot)}^\top \Delta_{i,(s,\cdot)}, \end{aligned}$$

where we denote

$$\bar{\pi}_{i,(s,\cdot)} := \int_0^1 \mathrm{softmax}(\Theta^\star_{i,(s,\cdot)} + t\Delta_{i,(s,\cdot)})\, dt.$$

The vector $\bar{\pi}_{i,(s,\cdot)}$ is a probability vector. Subtracting the log-normalizing constants from the logits therefore yields

$$\widehat{u}_{i,(s,\cdot)} - u_i^\star(s,\cdot) = (I_A - \mathbf{1}\bar{\pi}_{i,(s,\cdot)}^\top)\Delta_{i,(s,\cdot)}. \tag{91}$$

Step 2: the log-softmax operator bound. We now proceed to bound the operator norm of the matrix in (91). Fix a state s. For any $x \in \mathbb{R}^A$, expanding the squared norm and completing the square gives

$$\begin{aligned}\|(I_A - \mathbf{1}\bar{\pi}_{i,(s,\cdot)}^\top)x\|_2^2 &= \|x\|_2^2 - 2(\mathbf{1}^\top x)(\bar{\pi}_{i,(s,\cdot)}^\top x) + A(\bar{\pi}_{i,(s,\cdot)}^\top x)^2 \\ &= \|x\|_2^2 - \frac{(\mathbf{1}^\top x)^2}{A} + A\left[\left(\bar{\pi}_{i,(s,\cdot)} - \frac{1}{A}\mathbf{1}\right)^\top x\right]^2.\end{aligned}$$

Dropping the nonpositive term $-(\mathbf{1}^\top x)^2/A$ and applying Cauchy–Schwarz yields

$$\|(I_A - \mathbf{1}\bar{\pi}_{i,(s,\cdot)}^\top)x\|_2^2 \le \left(1 + A\left\|\bar{\pi}_{i,(s,\cdot)} - \frac{1}{A}\mathbf{1}\right\|_2^2\right)\|x\|_2^2 = A\|\bar{\pi}_{i,(s,\cdot)}\|_2^2\|x\|_2^2,$$

where the last equality uses $\mathbf{1}^\top\bar{\pi}_{i,(s,\cdot)} = 1$. Taking the supremum over $\|x\|_2 = 1$ gives

$$\left\|\left\|I_A - \mathbf{1}\bar{\pi}_{i,(s,\cdot)}^\top\right\|\right\|_{\mathrm{op}} \le \sqrt{A}\|\bar{\pi}_{i,(s,\cdot)}\|_2. \tag{92}$$

Since $\Theta^\star, \widehat{\Theta} \in \mathcal{M}_B$, every coordinate of $\Theta^\star + t\Delta = (1-t)\Theta^\star + t\widehat{\Theta}$ lies in $[-B, B]$ for all $t \in [0,1]$.

$$\mathrm{softmax}(z)_a = \frac{1}{1 + \sum_{b\neq a} e^{z_b - z_a}} \le \frac{1}{1 + (A-1)e^{-2B}} = \frac{e^{2B}}{e^{2B} + A - 1}.$$

The same bound holds for every coordinate of $\bar{\pi}_{i,(s,\cdot)}$. Since $\|\bar{\pi}_{i,(s,\cdot)}\|_2^2 \le \|\bar{\pi}_{i,(s,\cdot)}\|_\infty \sum_a \bar{\pi}_{i,(s,a)} = \|\bar{\pi}_{i,(s,\cdot)}\|_\infty$, applying (92) gives

$$\left\|\left\|I_A - \mathbf{1}\bar{\pi}_{i,(s,\cdot)}^\top\right\|\right\|_{\mathrm{op}} \le \sqrt{\frac{Ae^{2B}}{e^{2B} + A - 1}} = L_B. \tag{93}$$

Combining (91) and (93),

$$\|\widehat{u}_{i,(s,\cdot)} - u_i^\star(s,\cdot)\|_2 \le L_B\|\Delta_{i,(s,\cdot)}\|_2. \tag{94}$$

In particular, for each fixed task i,

$$\begin{aligned}\sum_s \|\widehat{u}_{i,(s,\cdot)} - u_i^\star(s,\cdot)\|_2^2 &\le L_B^2 \sum_s \|\Delta_{i,(s,\cdot)}\|_2^2 \\ &\le L_B^2 \sum_j \sum_s \|\Delta_{j,(s,\cdot)}\|_2^2 = L_B^2 \,|\!|\!|\Delta|\!|\!|_F^2 .\end{aligned} \tag{95}$$

Step 3: reward recovery and the occupancy identity. Subtracting (22) from (30) gives

$$\widehat{r}_i - r_i = C_\mu(I - \gamma_1 \mathsf{P}_1\mu)^{-1}(\widehat{u}_i - u_i^\star). \tag{96}$$

For a fixed target policy, its entropy term is the same for both rewards, so it cancels in their value difference. By the occupancy identity (16), we therefore have

$$\begin{aligned}\widehat{J}_{2,i}(\pi) - J_{2,i}(\pi; r_i) &= \frac{1}{1-\gamma_2}\sum_{s,a}\rho_2^\pi(s,a)\{\widehat{r}_i(s,a) - r_i(s,a)\} \\ &= \frac{1}{1-\gamma_2}\left\langle \left[C_\mu(I - \gamma_1\mathsf{P}_1\mu)^{-1}\right]^\top \rho_2^\pi, \widehat{u}_i - u_i^\star \right\rangle .\end{aligned} \tag{97}$$

We upper bound the inner product by

$$\begin{aligned}
&\left|\left\langle \left[C_\mu(I-\gamma_1\mathsf{P}_1\mu)^{-1}\right]^\top \rho_2^\pi, \widehat{u}_i - u_i^\star \right\rangle\right| \\
&= \left|\sum_s \left\langle \left[\left[C_\mu(I-\gamma_1\mathsf{P}_1\mu)^{-1}\right]^\top \rho_2^\pi\right]_{(s,\cdot)}, \widehat{u}_{i,(s,\cdot)} - u_i^\star(s,\cdot)\right\rangle\right| \\
&\le \sum_s \left\|\left[\left[C_\mu(I-\gamma_1\mathsf{P}_1\mu)^{-1}\right]^\top \rho_2^\pi\right]_{(s,\cdot)}\right\|_2 \|\widehat{u}_{i,(s,\cdot)} - u_i^\star(s,\cdot)\|_2 \\
&\le \left\{\sum_s \left\|\left[\left[C_\mu(I-\gamma_1\mathsf{P}_1\mu)^{-1}\right]^\top \rho_2^\pi\right]_{(s,\cdot)}\right\|_2^2\right\}^{1/2} \left\{\sum_s \|\widehat{u}_{i,(s,\cdot)} - u_i^\star(s,\cdot)\|_2^2\right\}^{1/2} \\
&\le L_B \,|\!|\!|\Delta|\!|\!|_F \left\{\sum_s \left\|\left[\left[C_\mu(I-\gamma_1\mathsf{P}_1\mu)^{-1}\right]^\top \rho_2^\pi\right]_{(s,\cdot)}\right\|_2^2\right\}^{1/2}.
\end{aligned}$$

The first equality decomposes the inner product by state. The first inequality uses the triangle inequality and Cauchy–Schwarz on the actions. The second inequality is Cauchy–Schwarz over s, and the last inequality follows by (95). Now use $|\!|\!|\Delta|\!|\!|_F \le \epsilon$ and divide by $1-\gamma_2$ to obtain (90). □

A.7 Plug-in selection and soft regret

We apply the simultaneous value-error bounds from Section A.6 to plug-in selection, then use the soft regret–KL identity to bound policy error.

A.7.1 Selection from simultaneous value certificates

The following lemma bounds the loss from maximizing estimated values by the sum of the comparator's and selected policy's error radii.

Lemma 12. *Fix a task i. Suppose $\widetilde{J}_{2,i} : \Pi \to \mathbb{R}$ and nonnegative radii $R_i(\pi)$ satisfy*

$$|\widetilde{J}_{2,i}(\pi) - J_{2,i}(\pi; r_i)| \le R_i(\pi) \qquad \forall \pi \in \Pi. \tag{98}$$

Then for any $\widetilde{\pi}_i$ such that

$$\widetilde{\pi}_i \in \operatorname*{argmax}_{\pi\in\Pi} \widetilde{J}_{2,i}(\pi),$$

every $\pi \in \Pi$ satisfies

$$J_{2,i}(\pi; r_i) - J_{2,i}(\widetilde{\pi}_i; r_i) \le R_i(\pi) + R_i(\widetilde{\pi}_i). \tag{99}$$

Proof. Fix a comparator $\pi \in \Pi$, then we have

$$\begin{aligned}
J_{2,i}(\pi; r_i) &\le \widetilde{J}_{2,i}(\pi) + R_i(\pi) \\
&\le \widetilde{J}_{2,i}(\widetilde{\pi}_i) + R_i(\pi) \\
&\le J_{2,i}(\widetilde{\pi}_i; r_i) + R_i(\widetilde{\pi}_i) + R_i(\pi).
\end{aligned}$$

The first and last inequalities use (98) for π and $\widetilde{\pi}_i$, respectively. The middle inequality holds because $\widetilde{\pi}_i$ maximizes $\widetilde{J}_{2,i}$. Subtracting $J_{2,i}(\widetilde{\pi}_i; r_i)$ proves (99). □

A.7.2 Soft regret as KL divergence

Take $\pi = \pi_{2,i}^\star$ in (99) and apply the following identity to $\widetilde{\pi}_i$ to bound its occupancy-weighted KL divergence from the optimal target policy.

Lemma 13. *For every task i and every stationary target policy π, let d_2^π be its normalized occupancy as in* (14)*, then*

$$(1-\gamma_2)\left[J_{2,i}(\pi_{2,i}^\star; r_i) - J_{2,i}(\pi; r_i)\right] = \sum_s d_2^\pi(s)\,\mathrm{KL}\big(\pi(\cdot \mid s) \,\big\|\, \pi_{2,i}^\star(\cdot \mid s)\big). \tag{100}$$

Proof. Fix i, and let $V^\star_{2,i}(s)$ be the optimal target soft value starting from state s under reward r_i. The soft Bellman equations with entropy coefficient 1 give

$$V^\star_{2,i}(s) = \log \sum_a \exp\left\{ r_i(s,a) + \gamma_2 \sum_{s'} \mathsf{P}_2(s' \mid s,a) V^\star_{2,i}(s') \right\},$$
$$\pi^\star_{2,i}(a \mid s) = \exp\left\{ r_i(s,a) + \gamma_2 \sum_{s'} \mathsf{P}_2(s' \mid s,a) V^\star_{2,i}(s') - V^\star_{2,i}(s) \right\}.$$

The rewards are finite, $\gamma_2 < 1$, and entropy is bounded by $\log A$, so $V^\star_{2,i}$ is finite and the policy formula gives $\pi^\star_{2,i}(a \mid s) > 0$. Taking logarithms yields

$$\log \pi^\star_{2,i}(a \mid s) = r_i(s,a) + \gamma_2 \sum_{s'} \mathsf{P}_2(s' \mid s,a) V^\star_{2,i}(s') - V^\star_{2,i}(s).$$

For every action with $\pi(a \mid s) > 0$, substituting this expression into the regularized reward under π gives

$$r_i(s,a) - \log \pi(a \mid s) = V^\star_{2,i}(s) - \gamma_2 \sum_{s'} \mathsf{P}_2(s' \mid s,a) V^\star_{2,i}(s') - \log \frac{\pi(a \mid s)}{\pi^\star_{2,i}(a \mid s)}.$$

Averaging the last term over $a \sim \pi(\cdot \mid s)$ gives

$$\sum_a \pi(a \mid s) \log \frac{\pi(a \mid s)}{\pi^\star_{2,i}(a \mid s)} = \mathrm{KL}\big(\pi(\cdot \mid s) \,\big\|\, \pi^\star_{2,i}(\cdot \mid s)\big),$$

where actions of zero probability contribute zero.

Now generate a trajectory using P_2, π, and initial distribution d_0. For each finite horizon T, conditional expectation gives

$$\begin{aligned}
&\mathbb{E}_{\mathsf{P}_2,\pi,d_0} \sum_{t=0}^{T-1} \gamma_2^t \left[V^\star_{2,i}(S_t) - \gamma_2 \sum_{s'} \mathsf{P}_2(s' \mid S_t, A_t) V^\star_{2,i}(s') \right] \\
&= \sum_{t=0}^{T-1} \gamma_2^t \mathbb{E}_{\mathsf{P}_2,\pi,d_0} V^\star_{2,i}(S_t) - \sum_{t=0}^{T-1} \gamma_2^{t+1} \mathbb{E}_{\mathsf{P}_2,\pi,d_0} V^\star_{2,i}(S_{t+1}) \\
&= \sum_s d_0(s) V^\star_{2,i}(s) - \gamma_2^T \mathbb{E}_{\mathsf{P}_2,\pi,d_0} V^\star_{2,i}(S_T).
\end{aligned}$$

Since the state and action spaces are finite, the rewards are finite, and $\pi^\star_{2,i}$ is positive, $V^\star_{2,i}$ and the statewise KL terms are bounded. Thus the second term vanishes as $T \to \infty$, and the discounted sums converge. Also,

$$\sum_s d_0(s) V^\star_{2,i}(s) = J_{2,i}(\pi^\star_{2,i}; r_i).$$

Substituting the regularized-reward identity above and rearranging therefore gives

$$\begin{aligned}
&J_{2,i}(\pi^\star_{2,i}; r_i) - J_{2,i}(\pi; r_i) \\
&= \mathbb{E}_{\mathsf{P}_2,\pi,d_0} \sum_{t \geq 0} \gamma_2^t \,\mathrm{KL}\big(\pi(\cdot \mid S_t) \,\big\|\, \pi^\star_{2,i}(\cdot \mid S_t)\big) \\
&= \frac{1}{1-\gamma_2} \sum_s d_2^\pi(s) \,\mathrm{KL}\big(\pi(\cdot \mid s) \,\big\|\, \pi^\star_{2,i}(\cdot \mid s)\big).
\end{aligned}$$

The last equality follows by the occupancy identity (16). Multiplying by $1 - \gamma_2$ proves (100). □

A.8 The zero-truth case

This section treats the case when $\Theta^\star = 0$ that was excluded in Assumption 4. Then $\pi^\star_{1,i}(a \mid s) = \frac{1}{A}$ for all i, s, a.

Proposition 1 (Exact recovery when $\Theta^\star = 0$)**.** *Suppose $\Theta^\star = 0$. Choose λ_n by* (27) *with $0 < \delta < 1$. With probability at least $1 - \delta/2$, every minimizer $\widehat{\Theta}$ of* (28) *equals zero. No restricted-coverage assumption is required. If Assumption 5 holds, the estimates in* (29)–(31) *also satisfy*

$$\widehat{r}_i = r_i, \qquad \widehat{J}_{2,i}(\pi) = J_{2,i}(\pi; r_i) \qquad \forall i \in [N],\ \pi \in \Pi$$

on the same event.

Proof. At $\Theta^\star = 0$, the concentration bound (60) with $x = \log(2(N + SA)/\delta)$ and the tuning (27) show that the event (61) holds with probability at least $1 - \delta/2$:

$$\|\nabla L_n(0)\|_{\mathrm{op}} \le \frac{\lambda_n}{2}.$$

On this event, fix any minimizer $\widehat{\Theta}$ of (28). Since $0 \in \mathcal{M}_B$,

$$\begin{aligned} 0 &\ge L_n(\widehat{\Theta}) - L_n(0) + \lambda_n \left\|\widehat{\Theta}\right\|_* \\ &\ge \left\langle\left\langle \nabla L_n(0),\, \widehat{\Theta} \right\rangle\right\rangle + \lambda_n \left\|\widehat{\Theta}\right\|_* \\ &\ge \frac{\lambda_n}{2} \left\|\widehat{\Theta}\right\|_* \ge 0. \end{aligned}$$

The first inequality is optimality of $\widehat{\Theta}$; the second is convexity of L_n; the third follows from (64) with $\Theta^\star = 0$ and $\Delta = \widehat{\Theta}$. Since $\lambda_n > 0$, these inequalities force $\widehat{\Theta} = 0$.

Equation (29) now gives $\widehat{u}_i(s, a) = -\log A = u_i^\star(s, a)$ for every i, s, a. Under Assumption 5, Lemma 10 and (30) therefore give $\widehat{r}_i = r_i$. Substituting into (31) proves the claimed value equality for every $i \in [N]$ and $\pi \in \Pi$. □

Proposition 1 applies to the single-fit estimator in (28).

B Experiments

B.1 Environments

All three environments share the linear-reward structure $r_i(s, a) = W_i^\top \psi(s, a)$ with $K = 4$ features ψ, and the feature map ψ is unavailable to every learning method.

FourRooms. A 19×19 grid of four rooms joined by single-cell doorways, with a fixed wall layout, implemented in NAVIX (Pignatelli et al., 2025), a JAX reimplementation of MiniGrid (Chevalier-Boisvert et al., 2023). Colored objects of $K = 4$ types (12 per color, 48 total) are scattered on reachable cells; the agent auto-collects an object by entering its cell, and collected objects respawn at random empty cells. The per-step feature $\psi \in \{0, 1\}^4$ is the color one-hot of any object collected that step (zero otherwise), so rewards are sparse. Observability is local: the agent sees an $11 \times 11 \times 4$ object window centered on its position plus its normalized coordinates, and we stack the last three (observation, action) frames for trajectory-aware classification. There are four cardinal navigation actions and episodes last 100 steps. We use $N = 8$ tasks (four pure color-seekers and four mixtures). Expert policies are trained by PPO with a CNN encoder.

Highway. A multi-lane driving environment (Rachum et al., 2026) with continuous kinematic state. The observation is the five nearest vehicles, each with five features (presence, longitudinal and lateral position, and velocities), for a 25-dimensional ego-centric state; the action set is five discrete meta-actions (change lane left/right, idle, accelerate, decelerate). The $K = 4$ features are speed, collision, headway (proximity to the lead vehicle), and lane change, and tasks differ in their signed weights over these (e.g., preferring speed, tailgating, or penalizing lane changes). NPC traffic is stochastic. We use $N = 5$ tasks, with expert policies trained by PPO with an MLP encoder.

RecSim. A tabular sequential recommender (Ie et al., 2019) over $T = 4$ topics ($K = 4$ features). The state is the vector of cumulative per-topic click counts, which is monotone within an episode and takes $9^4 = 6561$ discrete values. At each of the steps the user is shown 10 candidate items plus a skip action (11 actions in total) and chooses via a forward-looking soft-Q policy at temperature 0.1. Each topic has a burnout threshold (we use $[5, 6, 7, 8]$); once a topic's cumulative clicks reach its threshold, its appeal drops to zero for the remainder of the episode. The feature ψ is the clicked item's quality times its topic one-hot times the current topic appeal. We use $N = 8$ users (four pure and four 50/50 mixtures). Because the state space is small, expert policies are computed exactly by tabular soft value iteration, one per (user, discount) pair; this makes RecSim the setting where we can evaluate transfer against oracle policies without approximation error (Section 5.3).

B.2 Reward Construction and Expert Training

For each environment we set task weights $W \in \mathbb{R}^{N\times 4}$ as a combination of pure tasks (weight on a single feature) and mixed tasks (weight spread across features), chosen so that the rows span $\mathbb{R}^4$. Choosing four reward features does not force the rank k of the log-policy matrix to be four; the forward soft-RL map is nonlinear, so we use $k = 4$ throughout and probe misspecification in Section B.11. Given each task's reward, we train an expert (PPO for FourRooms and Highway, exact soft value iteration for RecSim), roll out trajectories for offline training, and μ-normalize the ground-truth rewards to match our identification constraint (with μ uniform and $g = 0$).

B.3 Baselines

We compare against two single-agent methods and one multi-task method.

GenPQR (van der Laan et al., 2025). The single-agent classify-then-regress method our approach extends: a softmax classifier estimates the log-policy, which the reward operator maps to a reward. We run it in two multi-task variants: *per-task* (fit independently on each task's data) and *pooled* (fit a single reward on all tasks' data, ignoring task identity).

IQ-Learn (Garg et al., 2021). A non-adversarial occupancy-matching method that learns a soft Q-function directly, from which a reward is extracted by the inverse soft-Bellman operator. Run in the same *per-task* and *pooled* variants as GenPQR.

Multi-intent IRL. A feature-free, model-free adaptation of Babeş-Vroman et al. (2011). The original clusters demonstrations using known reward features and a model-based inner loop, neither of which is available in our setting. We keep the core structure: a soft-clustering EM over log-policies with $K_{MI} = 4$ clusters, where each cluster's reward is recovered from the data of the tasks assigned to it (the same model-free operator as our method) and planning is performed once per cluster. Each task's reward is its responsibility-weighted combination of the K_{MI} cluster rewards.

B.4 Architectures, Hyperparameters, and Tuning

Within an environment all methods share the same encoder and width, so differences reflect the estimation strategy rather than model capacity: a CNN over the local object view for FourRooms, an MLP for Highway, and an MLP over the tabular click state for RecSim. The classifier is trained with Adam.

Regularization differs by method, and each was given a matched opportunity to be tuned. Our classifier uses the variational nuclear-norm penalty with a scalar $\lambda_c/\sqrt{n}$, so the penalty weakens as the per-agent budget grows; the schedule form follows the penalised-MLE rate $\lambda \sim \sqrt{\log d/n}$ and the constant was selected on Highway over $c \in \{0, 0.1, 0.2, 0.4\}$, with $c = 0.2$ winning at every budget. We transfer $c = 0.2$ unchanged to FourRooms and RecSim rather than retuning per environment. The per-task and pooled GenPQR baselines were swept over the same schedule ($c \in \{0, 0.05, 0.2, 0.5\}$) and no value improved them, so they are reported at $c = 0$; their off-support failure is structural rather than a regularization artifact. The IQ-Learn baselines do not use the classifier, so the schedule does not apply; their free parameter is the soft-Q temperature τ, which we fix at 1.0, the temperature at which the expert log-policy is defined.

	FourRooms	Highway	RecSim
Encoder	CNN	MLP	MLP
Hidden width	512	128	256
Residual blocks	0	0	2
FQI iterations	16	16	4
FQI regression epochs	50	50	40
Basis rank k	4	4	4
Nuclear-norm scalar c	0.2	0.2	0.2
MI clusters K_{MI}	4	4	4
MI E-step holdout	0.2	0.2	0.2

Table 1: IRL pipeline settings per environment. All methods within an environment use the same encoder and width, so the comparison isolates the estimation strategy.

B.5 METRICS

We report Pearson correlation between recovered and μ-normalized ground-truth rewards, which is invariant to the global scale and shift that IRL cannot identify. For each task, on-support states are those the task visits and off-support states are the remainder, where the state space is the union of states visited across all tasks.

For transfer (Section 5.3) we report return in the perturbed target as a fraction of the oracle ceiling, where the ceiling is the return of the optimal tabular policy for the true reward under the target dynamics. Perturbations lower the probability that a task's preferred topic appears in the candidate slate, swept over the degree of suppression; a random-uniform policy and a behavior-cloning policy bound the baselines.

B.6 COMPUTATIONAL EFFICIENCY

Reward weights are sampled independently per agent. In Highway we draw $w \sim \mathcal{N}(0, I)$ over the four features (speed, collision, headway, lane change), constrain the collision weight to be non-positive since a positive one encodes seeking crashes, and normalize each row to unit norm. In FourRooms we draw $|\mathcal{N}(0, I)|$ over the four object colors and scale each row so its largest entry is one, confining the population to the positive orthant, RecSim samples mixing non-negative coefficients over a fixed rank-4 basis.

B.7 COVERAGE MECHANISM: DERIVATION OF THE WEIGHTED-NORM DECOMPOSITION

This appendix derives the eigenvalue decomposition of the weighted-norm penalty stated inline in Section 5.1.

Let Θ^* be the true logits governing choices and U^* its left singular vectors from the SVD $\Theta^* = U^*\Sigma V^{*\prime}$. For each state s, define the ground-truth design Gram matrix

$$\Sigma_{col}(s) = \sum_i d_{i,s}\,(U^\star_{i,:})^\top U^\star_{i,:},$$

with eigenpairs $\{\lambda_k(s), v_k(s)\}_{k=1}^k$ from its eigendecomposition, where $U^\star_{i,:}$ is the i-th row of U^* and $w_i(s)$ is task i's visit weight at state s.

Let Δ be the estimation error after the classification step, and let $U^*U^{*\prime}\Delta \equiv U^*A'$ be its projection onto the column space of Θ^*, with coordinates $A' = U^{*\prime}\Delta$. Writing $a_{s,a}$ for the (s, a) column of

Domain	Method	N=8	N=32	N=128
Highway, n=500	**LowRank**	$+0.912 \pm 0.010$	$+0.885 \pm 0.011$	$+0.884 \pm 0.010$
	GenPQR-per-task	$+0.787 \pm 0.039$	$+0.742 \pm 0.052$	$+0.739 \pm 0.026$
	IQ-per-task	$+0.677 \pm 0.040$	$+0.649 \pm 0.042$	$+0.645 \pm 0.024$
	MI	$+0.849 \pm 0.014$	$+0.846 \pm 0.082$	$+0.878 \pm 0.003$
Highway, n=100	**LowRank**	$+0.879 \pm 0.018$	$+0.847 \pm 0.017$	$+0.844 \pm 0.016$
	GenPQR-per-task	-0.109 ± 0.082	-0.113 ± 0.029	-0.097 ± 0.020
	IQ-per-task	-0.112 ± 0.058	-0.085 ± 0.028	-0.072 ± 0.020
	MI	$+0.304 \pm 0.204$	$+0.645 \pm 0.057$	$+0.815 \pm 0.028$
FourRooms, n=500	**LowRank**	$+0.539 \pm 0.033$	$+0.529 \pm 0.027$	$+0.528 \pm 0.026$
	GenPQR-per-task	$+0.403 \pm 0.009$	$+0.384 \pm 0.007$	$+0.387 \pm 0.003$
	IQ-per-task	$+0.127 \pm 0.030$	$+0.122 \pm 0.010$	$+0.117 \pm 0.007$
	MI	$+0.323 \pm 0.032$	$+0.259 \pm 0.030$	$+0.243 \pm 0.021$
FourRooms, n=100	**LowRank**	$+0.396 \pm 0.037$	$+0.394 \pm 0.037$	$+0.397 \pm 0.035$
	GenPQR-per-task	$+0.307 \pm 0.011$	$+0.275 \pm 0.010$	$+0.283 \pm 0.005$
	IQ-per-task	$+0.069 \pm 0.032$	$+0.064 \pm 0.010$	$+0.055 \pm 0.006$
	MI	$+0.241 \pm 0.076$	$+0.198 \pm 0.021$	$+0.198 \pm 0.010$
RecSim, n=500	**LowRank**	$+0.776 \pm 0.015$	$+0.775 \pm 0.018$	$+0.776 \pm 0.015$
	GenPQR-per-task	$+0.196 \pm 0.010$	$+0.190 \pm 0.004$	$+0.183 \pm 0.002$
	IQ-per-task	$+0.134 \pm 0.010$	$+0.124 \pm 0.006$	$+0.117 \pm 0.002$
	MI	$+0.351 \pm 0.025$	$+0.348 \pm 0.015$	$+0.348 \pm 0.016$
RecSim, n=100	**LowRank**	$+0.744 \pm 0.027$	$+0.746 \pm 0.022$	$+0.743 \pm 0.022$
	GenPQR-per-task	$+0.136 \pm 0.013$	$+0.138 \pm 0.006$	$+0.134 \pm 0.003$
	IQ-per-task	$+0.099 \pm 0.011$	$+0.100 \pm 0.003$	$+0.096 \pm 0.002$
	MI	$+0.336 \pm 0.028$	$+0.337 \pm 0.021$	$+0.343 \pm 0.015$

Table 2: Reward-recovery quality (off-support global Pearson r, mean $\pm$ s.d. over 10 seeds) against population size N. LowRank and both per-task baselines are flat in N (largest change is 0.048); the small step from N=8 to N=32 on Highway and FourRooms is taken by all methods alike and reflects those populations sampling reward weights per agent, so agents 8–31 differ from the first 8. MI is the exception: it gains $+0.51$ on Highway at n=100, where its clusters are starved at N=8 (s.d. 0.204, falling to 0.028 by N=128), but loses 0.08 on FourRooms at n=500 with tight error bars.

A', the per-agent error is $\Delta_{i,(s,a)} = U^\star_{i,:} a_{s,a}$. A state's contribution to the weighted norm is then

$$\begin{aligned}
\sum_{i,a} d_i(s)\, \Delta^2_{i,(s,a)} &= \sum_{i,a} d_i(s)\, \left(U^\star_{i,:} a_{s,a}\right)^2 \\
&= \sum_a a'_{s,a} \left(\sum_i d_i(s)\, (U^\star_{i,:})^\top U^\star_{i,:} \right) a_{s,a} \\
&= \sum_a a'_{s,a}\, \Sigma_{col}(s)\, a_{s,a} \\
&= \sum_j \lambda_j(s) \sum_a \langle a_{s,a}, v_j(s) \rangle^2,
\end{aligned}$$

where the last equality applies the spectral theorem to $\Sigma_{col}(s)$. Low coverage along a direction is a small eigenvalue $\lambda_j(s) = \sum_i w_i(s) \left\langle (U_{i,:})^\top, v_j(s) \right\rangle^2$: when $\lambda_j(s)$ is small, a large error component $\langle a_{s,a}, v_j(s) \rangle$ contributes little to the penalty and hides in the weighted norm, though it can be large in the unweighted Frobenius norm that governs off-support recovery.

B.8 Scaling Tasks

Another possible strategy to scale across many tasks is a transfer learning approach: learning the basis, Φ, for a relatively small pool of tasks (phase 1) and then for remainder or newly arriving tasks (phase 2), fitting the basis loadings W via cross entropy loss for a frozen basis. The key is to curate

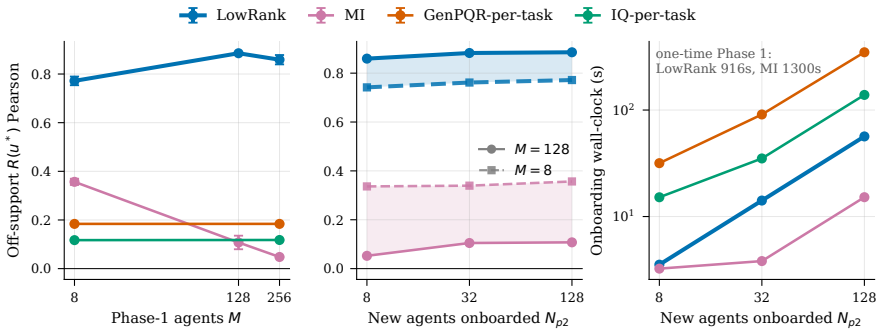

Figure 5: Transfer learning is effective: training basis on M agents in Phase 1 and transferring to N_{p2} Phase 2 agents

the tasks for phase 1 carefully: choosing tasks that span the basis space for identification and have maximal data for statistical efficiency.

Figure 5 shows results from an experiment where we vary both the number of tasks in Phase 1, M, and the number of tasks in Phase 2, N_{p2}. Increasing M by definition has no effect on per-task variants of GenPQR and IQ but has sizeable effects on LowRank and MI. For LowRank, increasing M from 8 to 128 leads to a ≈ 0.15 increase in off-support Pearson correlation for new tasks in Phase 2 whereas the opposite is true of MI leading to a -0.25 decrease.

B.9 Logit finite sample bound rate scaling

To isolate each predicted scaling dependence in Theorem 1, we vary one dimension at a time holding fixed the others around the reference configuration: $N = 8, S = 81, A = 5, k = 2$, and $n = 2^{22}$, at a regularization parameter chosen on a held-out dataset, $\lambda/\lambda_{\text{theorem}} = 1/32$. In the rank k sweep, the truth is $(\Theta_k^\star = k^{-1/2} \sum_{j=1}^{k} U_{:,j}(V_{:,j})^\top)$, with k mutually orthogonal rank-one components, giving exact rank k while preserving its Frobenius norm. The state-space sweep uses finer or coarser discretizations of the same fixed state domain, normalized to preserve signal strength. In the action-space sweep, we vary the number of slate positions and assign additional positions a balanced set of latent action profiles, rescaled to preserve total signal strength.

Figure 6 shows the empirical scaling rate against that predicted by the theory. We match on sample-size and state-space cardinality while scaling more conservatively than the theory suggests for action-space cardinality and rank.

B.10 Raw logit approximation

In Section 3.3, we make an approximation by applying the reward operator to the raw unnormalized logits, which represents large computational savings by allowing us to perform k per-basis function versus N per-task planning steps. To measure the effect of this approximation, we compare performance against the exact method, where planning is performed on the true log-policies per task.

Figure 7 shows mild effects of using the approximation instead of the per-task normalized logits exactly. Using paired-bootstrap 95% confidence intervals, for the 12 pairs of comparisons across data budget, environment and on vs off-support, 5 favor the exact method, 5 favor the approximation and 2 are indistinguishable from zero. The exact method is also slower: it takes approximately $1.1\times$ the wall-clock of the approximation for Highway ($k = 4, N = 5$) and $1.9\times$ and $1.8\times$ for FourRooms and RecSim ($k = 4, N = 8$). Wall-clock grows almost exactly with N, consistent with the exact path scaling per task (N inversions) versus per basis column k for the approximation.

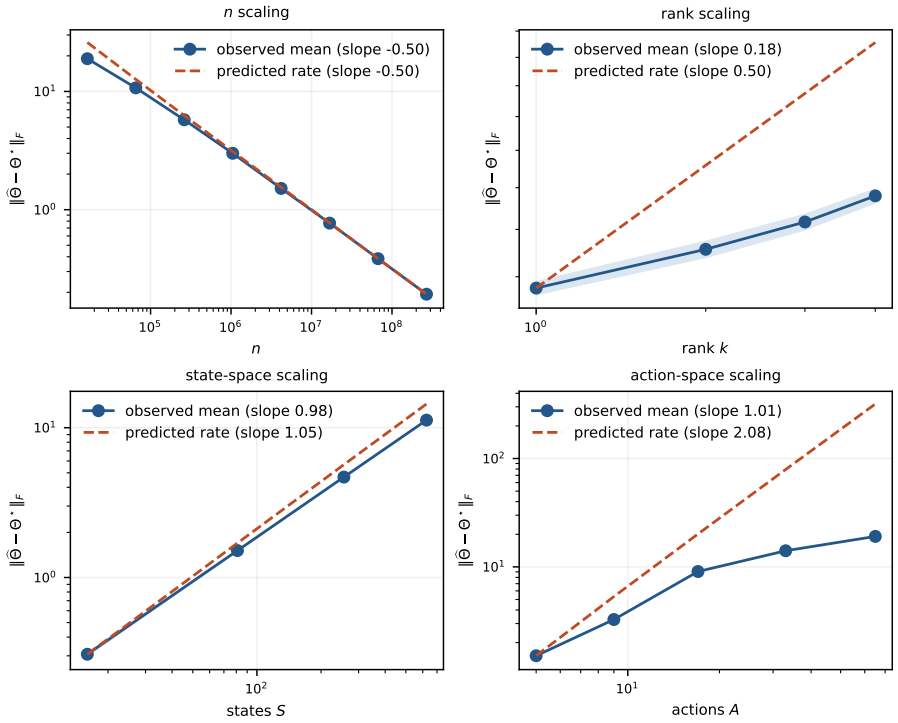

Figure 6: Empirical scaling in RecSim matches the theory in n and S and is more conservative in k and A.

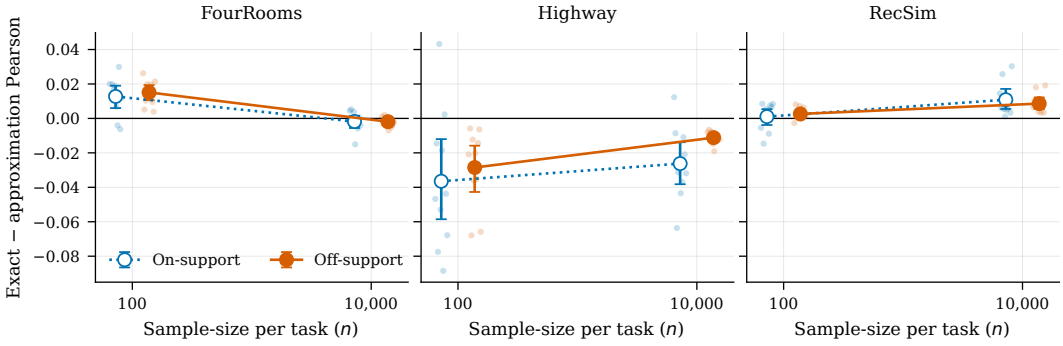

Figure 7: Difference between reward recovery between reward operator applied to exact log normalized policies versus approximation

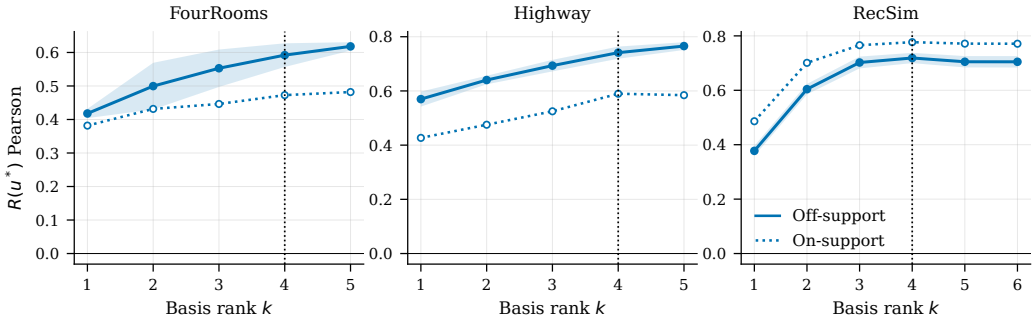

Figure 8: Misspecification of rank leads to graceful degradation

	$n = 100$			$n = 10^4$		
K_{MI}	Highway	FourRooms	RecSim	Highway	FourRooms	RecSim
1 †	0.04 ± 0.03	0.01 ± 0.04	0.01 ± 0.01	0.09 ± 0.01	0.15 ± 0.00	-0.00 ± 0.00
2	$\mathbf{0.23 \pm 0.13}$	0.13 ± 0.08	0.04 ± 0.06	0.54 ± 0.03	0.37 ± 0.06	0.18 ± 0.01
3	0.04 ± 0.07	0.18 ± 0.07	0.08 ± 0.06	0.60 ± 0.13	0.41 ± 0.03	0.30 ± 0.03
4	-0.11 ± 0.06	0.22 ± 0.05	0.05 ± 0.04	0.68 ± 0.13	0.43 ± 0.04	$\mathbf{0.34 \pm 0.02}$
6	—	0.27 ± 0.04	0.10 ± 0.03	—	0.47 ± 0.03	0.32 ± 0.01
N ‡	-0.17 ± 0.05	$\mathbf{0.34 \pm 0.02}$	$\mathbf{0.13 \pm 0.02}$	$\mathbf{0.77 \pm 0.09}$	$\mathbf{0.53 \pm 0.00}$	0.29 ± 0.01
LowRank (ours)	0.71 ± 0.02	0.42 ± 0.08	0.40 ± 0.02	0.84 ± 0.01	0.72 ± 0.01	0.72 ± 0.02

Table 3: MI-Cluster across cluster count K_{MI} (off-support Pearson r, mean $\pm$ s.d. over 10 seeds); N=5 for Highway, 8 otherwise. Bold marks each column's best K. Ours beats MI at its best K in every domain at both budgets. Shapes are budget-dependent: Highway rises to K=N at n=10^4 but falls to it at n=100, and RecSim's peak at the generative rank K=4 disappears.

B.11 Rank Misspecification

A key assumption is that the log-policy matrix is approximately rank k. Because we learn k basis functions and run FQI on each, we study robustness to rank misspecification: how recovery behaves when the true rank differs from k. This also probes the linear-reward assumption, since a reward that is nonlinear in the underlying features raises the true log-policy rank, one way the assumed k can be too low. We expect graceful degradation, under-ranking underfits smoothly while over-ranking is benign, rather than a sharp failure.

Figure 8 shows that as we both under and over-specify rank, recovery performance is smooth. Choosing k lower than true rank leads to graceful degradation while k higher leads to a plateau in performance. RecSim plateaus at $k = 4$ matching the rank of the reward features whereas both FourRooms and Highway performance increases with k, related to the use of continuous states in those environments and hence greater rank of true log policies. This also demonstrates the value of applying low rank to log policies, instead of rewards, as we can check robustness to rank.

B.12 Multi-intent scaling number of clusters

Table 3 tests how scaling the number of clusters affects the multi-intent baseline, at a sample-size of 100 and $10K$. The effect on scaling is mixed across environments and sample-sizes. Scaling clusters to the same number of tasks sometimes has the best performance (e.g FourRooms) but there are instances when an intermediate K performs best (e.g Highway at $n = 100$ and RecSim at $N = 10^4$). Performance is lower than LowRank even for the best choice of K_{MI} per cell.