Action-Free Offline RL via Demonstrator Diversity

Felix Schur [1]

Abstract

A central bottleneck in transitioning from offline representation learning to online decision-making is the "action gap": passive datasets (e.g., online videos) often lack the action labels required to ground latent representations in environment dynamics. We propose to bridge this gap by exploiting demonstrator diversity. Even when actions are unobserved, systematic variation in transitions across demonstrators can help disentangle latent action choice from environment stochasticity. We formalize this as a statewise column-stochastic non-negative matrix factorization (NMF) problem, where demonstrator-specific policies act as mixtures over shared latent transition kernels. Under "sufficiently scattered" policy diversity and rank conditions, we prove that latent actions and dynamics are identifiable up to a permutation. We extend these results to continuous observation spaces via a Gram-determinant minimum-volume criterion and prove that spatial continuity ensures a globally consistent action labeling. Our framework shows how heterogeneity in passive data can substitute for missing action labels, leaving limited interaction to resolve only the final action-label alignment.

1. Introduction

Recent progress in machine learning has been driven by massive internet-scale datasets. Large language models can be trained on web text because both inputs and prediction targets are directly observed. Reinforcement learning has not scaled in the same way: although the internet contains abundant sequential data—gameplay videos, robot videos, screen recordings, and human demonstrations—these data usually lack the action labels needed to train policies or action-conditioned dynamics models in the standard way (Seo et al., 2022; Baker et al., 2022). This lack of action annotations is a central obstacle to scaling RL from passive data (Nikulin et al., 2025). The difficulty is not only practical, but also statistical. From observation-only trajectories generated by a single behavior policy, action choice and environment stochasticity are confounded. For example, consider a video of a television whose channel changes over time. From the video alone, we cannot tell whether the TV is broken and changes channels by itself, or whether an unseen viewer is using a remote control. Both mechanisms can induce the same observed conditional distribution $p(o_{t+1} \mid o_t)$. More generally, many different latent action spaces, policies, and transition kernels can induce the same marginal transition law. Thus, latent actions and action-conditioned dynamics are not identifiable from a single stream of action-free data without additional structure. We propose to resolve this ambiguity by exploiting *demonstrator diversity*. The intuition is again simple. If a channel switch happens with the same probability for every viewer, then it is natural to attribute it to the TV itself rather than to viewer choice. But if transition probabilities differ systematically across viewers, then this variation is evidence of different action choices, provided that viewer identity has no direct effect on the TV beyond the chosen action. In our setting, demonstrator identity changes the latent policy while the action-conditioned transition dynamics remain shared. This turns the observable law $p(o_{t+1} \mid o_t, e)$ into a mixture of shared latent transition kernels with demonstrator-specific weights. If these policies are sufficiently diverse, the variation across demonstrators is rich enough to reveal the hidden action-conditioned transition structure, up to a relabeling of the latent actions.

Related work. Prior work on imitation and reinforcement learning from observation learns policies or action-like representations from state-only demonstrations, often using interaction, rewards, or limited action-labeled data to align latent actions with environment actions (Torabi et al., 2018; 2019; Edwards et al., 2019; Seo et al., 2022; Baker et al., 2022; Schmidt and Jiang, 2024; Zheng et al., 2023). We instead study a population-level identifiability question: when do action-free trajectories determine the latent action-conditioned transition kernels themselves? Closest to our setting, Levine et al. (2025) exploit heterogeneous action-free datasets to recover controllable structure in exogenous block MDPs. We use the same source of

[1]Department of Mathematics, ETH Zurich. Correspondence to: Felix Schur <felix.schur@stat.math.ethz.ch>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

signal—policy variation across demonstrators—but target a different object: identifiable latent actions and shared action-conditioned transition kernels. Technically, demonstrator identity makes the observable law $p(o_{t+1} \mid o_t, e)$ a stochastic mixture over shared action-conditioned kernels, yielding a statewise stochastic nonnegative matrix factorization problem. This also connects to instrumental-variable and invariant-learning views of offline decision-making, where exogenous variation changes action choices while leaving the environment mechanism fixed (Hartford et al., 2017; Arjovsky et al., 2019; Chen et al., 2022; Liao et al., 2024). Our results adapt identifiable NMF and minimum-volume arguments (Fu et al., 2018; Huang et al., 2018) to action-free offline RL.

2. Problem Setting

We consider a Markovian environment with discrete latent actions $A_t \in [k]$, where k is known, and observations $O_t \in \mathcal{O}$. Each trajectory is generated by one of m demonstrators, indexed by $E \in [m]$. Given $O_t = o$ and $E = e$, the demonstrator chooses an unobserved action $A_t \sim \pi_e^*(\cdot \mid o)$, and the environment evolves according to the shared action-conditioned kernel $O_{t+1} \sim p^*(\cdot \mid o, A_t)$. The data consist only of action-free tuples (o_t, o_{t+1}, e). We assume that the current observation and latent action are sufficient for the next observation, and that demonstrator identity affects the next observation only through the chosen action: $O_{t+1} \perp\!\!\!\perp \text{past} \mid (O_t, A_t)$ and $E \perp\!\!\!\perp O_{t+1} \mid (O_t, A_t)$. Thus demonstrators may differ in their policies, but the transition dynamics are shared across demonstrators. At the population level, we observe the conditional laws $p^*(\cdot \mid o, e)$ and seek to recover the latent transition kernels $p^*(\cdot \mid o, a)$ and policies $\pi_e^*(a \mid o)$. Marginalizing over the unobserved action gives, for every $o \in \mathcal{O}$ and $e \in [m]$,

$$p^*(\cdot \mid o, e) = \sum_{a=1}^{k} p^*(\cdot \mid o, a)\, \pi_e^*(a \mid o). \tag{1}$$

Hence, for each fixed observation o, the observable demonstrator-conditioned transition law is a mixture of latent action-conditioned transition kernels. Since latent action labels are arbitrary, recovery is possible only up to a permutation of the action indices.

3. Theoretical Guarantees

Our analysis proceeds in four steps. First, we show that without demonstrator diversity the problem is non-identifiable in general. Second, we establish *statewise* identifiability in finite observation spaces using identifiable stochastic (nonnegative matrix factorization) NMF. Third, we give an extension to continuous observation spaces. Fourth, we show that continuity and a no-collision condition upgrade statewise permutations to a *single global permutation*, which can then be fixed by a small amount of labeled action data.

Statewise identifiability in finite observation spaces. We first consider the case where $\mathcal{O}$ is finite. Fix $o \in \mathcal{O}$. For each demonstrator $e \in [m]$, define $p_e^o := p^*(\cdot \mid o, e) \in \mathbb{R}^{|\mathcal{O}|}$, and $t_a^o := p^*(\cdot \mid o, a) \in \mathbb{R}^{|\mathcal{O}|}$, and collect these into the matrices $P_o^* := \begin{bmatrix} p_1^o & \cdots & p_m^o \end{bmatrix} \in \mathbb{R}^{|\mathcal{O}| \times m}$,, $T_o^* := \begin{bmatrix} t_1^o & \cdots & t_k^o \end{bmatrix} \in \mathbb{R}^{|\mathcal{O}| \times k}$, and $\Pi_o^* := \begin{bmatrix} \pi_1^*(\cdot \mid o) & \cdots & \pi_m^*(\cdot \mid o) \end{bmatrix} \in \mathbb{R}^{k \times m}$. By (1),

$$P_o^* = T_o^* \Pi_o^*. \tag{2}$$

Each column of P_o^*, T_o^*, and Π_o^* is a probability vector. Define the set of column-stochastic matrices

$$\mathcal{M}_{r \times c} := \left\{ M \in [0,1]^{r \times c} \;\middle|\; \mathbb{1}^\top M = \mathbb{1}^\top \right\}.$$

Then $P_o^* \in \mathcal{M}_{|\mathcal{O}| \times m}$, $T_o^* \in \mathcal{M}_{|\mathcal{O}| \times k}$, and $\Pi_o^* \in \mathcal{M}_{k \times m}$. The finite-state identification problem is therefore: from the observable matrix P_o^*, recover the latent factors (T_o^*, Π_o^*) up to permutation. The factorization (2) is not unique in general: if $S \in \mathrm{GL}(k)$ is invertible, then

$$P_o^* = T_o^* \Pi_o^* = (T_o^* S)(S^{-1} \Pi_o^*).$$

Thus additional structure is needed to identify the latent factors. We use the sufficiently scattered condition of Fu et al. (2018), which ensures that among all feasible stochastic factorizations, the true one is the unique minimum-volume solution up to permutation.

Definition 3.1 (Sufficiently scattered)**.** Let $M \in \mathbb{R}^{u \times v}$ be nonnegative. Define $\mathrm{cone}(M) := \{M\theta : \theta \geq 0\}$ and $C^* := \{y : y^\top x \geq 0 \text{ for all } x \in C\}$ for any cone C. For $v \geq 2$, define the second-order cones $\mathcal{C}_v := \left\{ x \in \mathbb{R}^v : \mathbb{1}^\top x \geq \sqrt{v-1}\, \|x\|_2 \right\}$ and $\mathcal{C}_v^* := \left\{ x \in \mathbb{R}^v : \mathbb{1}^\top x \geq \|x\|_2 \right\}$. A matrix $H \in \mathcal{M}_{u \times v}$ is called *sufficiently scattered* if $\mathcal{C}_v \subseteq \mathrm{cone}(H^\top)$ and $\mathrm{cone}(H^\top)^* \cap \mathrm{bd}(\mathcal{C}_v^*) = \{\lambda e_j : j \in [v],\ \lambda \geq 0\}$.

Assumption 3.2 (Finite-state identifiability conditions)**.** For all $o \in \mathcal{O}$ we have $|\mathcal{O}| \geq k$ and $m \geq k$ and

(i) $\mathrm{rank}(T_o^*) = k$;

(ii) Π_o^* is sufficiently scattered.

The first condition says that the latent actions have linearly distinguishable transition effects at observation o. The second condition says that the demonstrator policy vectors are not all small perturbations of one another or concentrated near a low-dimensional subset of the simplex; rather, they point in sufficiently many different action-mixture directions that their conic hull spreads broadly through the positive orthant, making the latent actions geometrically separable. In this sense it formalizes *policy diversity*: different demonstrators must place meaningfully different relative masses on

the latent actions, so that each latent action leaves a distinct signature in the observed mixtures instead of being confounded with the others. We can now state the finite-state identifiability result.

Theorem 3.3 (Statewise identifiability in finite observation spaces)**.** *Assume Assumption 3.2. Then for every $o \in \mathcal{O}$, the optimization problem*

$$\begin{aligned} \min_{T_o \in \mathcal{M}_{|\mathcal{O}|\times k},\, \Pi_o \in \mathcal{M}_{k\times m}} \quad & \det(T_o^\top T_o) \\ \text{s.t.} \quad & P_o^* = T_o \Pi_o \end{aligned} \tag{3}$$

recovers the latent factors up to a permutation of the action labels. More precisely, there exists a permutation matrix $\Sigma_o \in \{0,1\}^{k\times k}$ such that $T_o = T_o^* \Sigma_o$, and $\Pi_o = \Sigma_o^\top \Pi_o^*$.

The minimum-determinant objective admits a simple geometric interpretation. Since $\det(T_o^\top T_o)$ is the squared k-dimensional volume spanned by the columns of T_o, minimizing it selects, among all stochastic factorizations of P_o^*, the latent transition kernels that are least spread out while still explaining the demonstrator-conditioned transition laws. Intuitively, this is a conservative principle: we attribute to the latent actions only the variation that is forced by the observed diversity across demonstrators, and do not introduce more extreme latent action effects than necessary. If the demonstrators are sufficiently diverse, this still forces recovery of all action effects that are genuinely distinguishable from the data. On the other hand, if two latent actions induce the same transition law at state o, then they are observationally equivalent there and cannot be identified separately (i.e. Assumption 3.2 (i) fails); in that case the true factorization becomes rank-deficient ($\operatorname{rank}(P_o^*) < k$) and $\det((T_o^*)^\top T_o^*) = 0$. Nevertheless, if one reduces the problem to the effective dimension $r_o := \operatorname{rank}(P_o^*)$, which under sufficient policy diversity coincides with the number of distinct transition effects at o, then the same minimum-volume argument identifies the remaining distinct latent transitions (and the corresponding aggregated policy masses) up to permutation. Theorem 3.3 is a *statewise* result: the permutation Σ_o may still depend on o. The next subsections address both continuous observation spaces and alignment of these statewise permutations across o.

Continuous-space identifiability. We now give a continuous-state analogue of the finite-state minimum-volume result that applies to *general transition measures*, including deterministic transitions represented by Dirac masses. Let $(\mathcal{O}, \mathcal{B})$ be a measurable space, let $\mathcal{P}(\mathcal{O})$ denote the set of probability measures on $(\mathcal{O}, \mathcal{B})$, and let $\mathcal{M}(\mathcal{O})$ denote the vector space of finite signed measures on $(\mathcal{O}, \mathcal{B})$. Fix a real Hilbert space $\mathcal{H}$ and a linear injective map $\Phi : \mathcal{M}(\mathcal{O}) \to \mathcal{H}$. A canonical example is a kernel mean embedding into an RKHS associated with a bounded characteristic kernel. Injectivity ensures that equality in embedding space implies equality of the underlying measures. Fix $o \in \mathcal{O}$. Define the latent transition measures $t_a^o := p^*(\cdot \mid o, a) \in \mathcal{P}(\mathcal{O}), a \in [k]$, and the observable demonstrator-conditioned transition measures $p_e^o := p^*(\cdot \mid o, e) \in \mathcal{P}(\mathcal{O}), e \in [m]$. By the mixture identity, $p_e^o = \sum_{a=1}^k \pi_e^*(a|o)\, t_a^o, e \in [m]$. Applying the linear map Φ yields $\Phi(p_e^o) = \sum_{a=1}^k \pi_e^*(a|o)\, \Phi(t_a^o)$. For any k-tuple of measures $(t_1, \dots, t_k)$, define the embedded Gram matrix $G_\Phi(t_1, \dots, t_k) := \left[\langle \Phi(t_i), \Phi(t_j)\rangle_{\mathcal{H}}\right]_{i,j=1}^k$. Its determinant measures the squared volume spanned by the embedded latent transition measures in $\mathcal{H}$. We also define the rank of a family of embedded observables by $\operatorname{rank}(P^*) := \dim \operatorname{span}\{\Phi(p_1^o), \dots, \Phi(p_m^o)\}$. Note that $\Pi_o^* = \left[\pi_1^*(\cdot \mid o) \quad \cdots \quad \pi_m^*(\cdot \mid o)\right] \in \mathbb{R}^{k\times m}$ remains a real-valued matrix and therefore the sufficiently scattered condition does not change.

Assumption 3.4 (Continuous-state identifiability conditions)**.** For all $o \in \mathcal{O}$ we have $|\mathcal{O}| \geq k$ and $m \geq k$ and

(i) $\operatorname{rank}\left(G_\Phi(t_1^o, \dots, t_k^o)\right) = k$;

(ii) Π_o^* is sufficiently scattered.

We now state the continuous-state identifiability result.

Theorem 3.5 (State-wise identifiability in continuous observation spaces)**.** *Fix $o \in \mathcal{O}$ and suppress the dependence on o in the notation. Assume Assumption 3.4 and that the optimization problem*

$$\begin{aligned} \min_{\bar{t}_1, \dots, \bar{t}_k, \Pi} \quad & \det\left(G_\Phi(\bar{t}_1, \dots, \bar{t}_k)\right) \\ \text{s.t.} \quad & p_e^o = \sum_{a=1}^k \pi_e(a|o)\bar{t}_a, \qquad e \in [m], \\ & \Pi_o \in \mathcal{M}_{k\times m}, \\ & \bar{t}_a \in \mathcal{P}(\mathcal{O}), \qquad a \in [k], \end{aligned} \tag{4}$$

admits an optimizer. Then every optimizer of (4) is equal to the ground-truth factorization up to permutation: there exists a permutation matrix $\Sigma_o \in \{0,1\}^{k\times k}$ such that $\bar{t}_a = t_{\Sigma(a)}^o, a \in [k]$, and $\Pi_o = \Sigma_o^\top \Pi_o^*$.

In finite-dimensional stochastic NMF, optimizer existence is often automatic by compactness. In contrast, (4) is an optimization problem over an infinite-dimensional space of measures, so minimizing sequences need not converge without additional compactness assumptions. A standard sufficient route is to work on a compact Polish observation space and choose Φ to be weakly continuous (for example, a kernel mean embedding associated with a bounded continuous kernel). Then $\mathcal{P}(\mathcal{O})$ is weakly compact, $\mathcal{M}_{k\times m}$ is compact, the feasibility constraints are closed, and the objective $(t_1, \dots, t_k) \mapsto \det\left(G_\Phi(t_1, \dots, t_k)\right)$ is continuous, so an optimizer exists by the direct method.

From state-wise to global identifiability. The state-wise identifiability results identify the latent action model separately at each observation o, but only up to an observation-dependent permutation. We now show that, under continuity and connectedness, these local permutations must agree globally.

Throughout this subsection, let $\Phi : \mathcal{M}(\mathcal{O}) \to \mathcal{H}$ be the fixed injective linear embedding used in the continuous-space theorem, where $\mathcal{H}$ is a real Hilbert space. For each $o \in \mathcal{O}$ and $a \in [k]$, define the embedded true latent transitions $\phi_a^*(o) := \Phi\big(p^*(\cdot \mid o, a)\big) \in \mathcal{H}$, and, for any candidate latent model, define the embedded recovered latent transitions $\phi_a(o) := \Phi\big(p(\cdot \mid o, a)\big) \in \mathcal{H}$. We collect these into the operator-valued maps $T^*(o) := \begin{bmatrix}\phi_1^*(o) & \cdots & \phi_k^*(o)\end{bmatrix}$, and $T(o) := \begin{bmatrix}\phi_1(o) & \cdots & \phi_k(o)\end{bmatrix}$.

Assumption 3.6. Assume that:

(i) for each $a \in [k]$, the maps $o \mapsto \phi_a(o)$ and $o \mapsto \phi_a^*(o)$ are continuous from $\mathcal{O}$ into $\mathcal{H}$;

(ii) for every $o \in \mathcal{O}$ and every $i \neq j$, $\phi_i^*(o) \neq \phi_j^*(o)$.

Theorem 3.7 (From local to global permutations)**.** *Assume that $\mathcal{O}$ is a connected metric space and that Assumption 3.6 holds. Assume that for each $o \in \mathcal{O}$ there exists a permutation matrix $\Sigma(o)$ such that $T(o) = T^*(o)\Sigma(o)$, and $\Pi(o) = \Sigma(o)^\top \Pi^*(o)$. Then $\Sigma(o)$ is constant on $\mathcal{O}$. That is, there exists a single permutation matrix Σ such that for all $o \in \mathcal{O}$.*

$$T(o) = T^*(o)\Sigma, \qquad \Pi(o) = \Sigma^\top \Pi^*(o).$$

Theorem 3.7 shows that once the latent model is identified at each observation up to permutation, continuity and connectedness reduce the ambiguity to a single *global* permutation of the latent action labels. We now combine the state-wise identifiability theorem with Theorem 3.7 to obtain global identifiability.

Assumption 3.8 (Global identifiability conditions)**.** Assume the following.

(i) $\mathcal{O}$ is a connected metric space.

(ii) For every $o \in \mathcal{O}$, Assumption 3.4 is satisfied.

(iii) For each $a \in [k]$, the map $o \mapsto \Phi\big(p^*(\cdot \mid o, a)\big)$ is continuous from $\mathcal{O}$ into $\mathcal{H}$.

(iv) There exists a nonempty subset $S \subseteq \mathcal{O}$ on which the global permutation is known.

We give some intuition about the Assumption 3.8: Connectedness rules out choosing different action labelings on disconnected regions of the observation space and continuity prevents the local permutation ambiguity from changing abruptly across nearby observations (see Theorem 3.7). The sufficient-scattering condition guarantees state-wise identifiability by making the demonstrator policies diverse enough to separate the latent action-conditioned transition kernels and the positive-definiteness of the embedded Gram matrix ensures that the latent action effects are linearly independent, and hence distinguishable, at each observation (see Theorem 3.5).

Corollary 3.9 (Global identifiability)**.** *Assume Assumption 3.8(i)-(iii) hold. Let $o \mapsto \big(p(\cdot \mid o, 1), \ldots, p(\cdot \mid o, k), \Pi(o)\big)$ be any continuous family such that for every $o \in \mathcal{O}$ the tuple $\big(p(\cdot \mid o, 1), \ldots, p(\cdot \mid o, k), \Pi(o)\big)$ is feasible for (4), and it attains the minimum in (4). Then there exists a single permutation matrix Σ such that for all $o \in \mathcal{O}$.*

$$p(\cdot \mid o, a) = p^*(\cdot \mid o, \Sigma(a)), \qquad \Pi(o) = \Sigma^\top \Pi^*(o)$$

If, in addition, Assumption 3.8(iv) holds, then $\Sigma = I_k$, and therefore for all $o \in \mathcal{O},\ a \in [k]$

$$p(\cdot \mid o, a) = p^*(\cdot \mid o, a), \qquad \Pi(o) = \Pi^*(o).$$

In particular, the latent action-conditioned transition kernels and the demonstrator-specific policies are globally identifiable.

Assumption 3.8 do not need to hold uniformly on all of $\mathcal{O}$. If they hold on a connected subset of observations, the same argument gives global identifiability on that subset; if $\mathcal{O}$ has multiple connected components, the result holds componentwise unless anchor labels are available to align them. Once latent action-conditioned transition kernels are identifiable up to a global permutation, a small amount of online interaction or action-labeled data can anchor the permutation and map latent actions to environment actions. This suggests a route from large passive datasets to online adaptation with minimal real-world interaction.

4. Conclusion and Future Work

We have established that demonstrator diversity provides a principled mechanism to resolve the "action gap" in action-free offline RL. By framing heterogeneous transitions as a state-wise mixture problem, we proved that latent transition kernels are identifiable up to a global permutation under sufficiently scattered policy diversity. Theoretically, this shows that systematic variation in behavior can substitute for missing action labels to recover environment dynamics.

Future research will focus on deriving finite-sample error bounds for minimum-volume NMF, investigating robustness under model misspecification or near-identical demonstrator policies, and scaling these geometric principles to high-dimensional visual inputs.

References

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

B. Baker, I. Akkaya, P. Zhokhov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro, and J. Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. In *Advances in Neural Information Processing Systems*, 2022.

Y. Chen, L. Xu, C. Gulcehre, T. Le Paine, A. Gretton, N. De Freitas, and A. Doucet. On instrumental variable regression for deep offline policy evaluation. *Journal of Machine Learning Research*, 23(302):1–40, 2022.

A. Edwards, H. Sahni, Y. Schroecker, and C. Isbell. Imitating latent policies from observation. In *International conference on machine learning*, pages 1755–1763. PMLR, 2019.

X. Fu, K. Huang, and N. D. Sidiropoulos. On identifiability of nonnegative matrix factorization. *IEEE Signal Processing Letters*, 25(3):328–332, 2018.

J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1414–1423. PMLR, 2017.

K. Huang, X. Fu, and N. D. Sidiropoulos. Learning hidden markov models from pairwise co-occurrences with application to topic modeling. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2068–2077. PMLR, 2018.

A. Levine, P. Stone, and A. Zhang. Offline action-free learning of ex-bmdps by comparing diverse datasets. *arXiv preprint arXiv:2503.21018*, 2025.

L. Liao, Z. Fu, Z. Yang, Y. Wang, D. Ma, M. Kolar, and Z. Wang. Instrumental variable value iteration for causal offline reinforcement learning. *Journal of Machine Learning Research*, 25(303):1–56, 2024.

A. Nikulin, I. Zisman, D. Tarasov, N. Lyubaykin, A. Polubarov, I. Kiselev, and V. Kurenkov. Latent action learning requires supervision in the presence of distractors. *arXiv preprint arXiv:2502.00379*, 2025.

D. Schmidt and M. Jiang. Learning to act without actions. In *International Conference on Learning Representations*, 2024.

Y. Seo, K. Lee, S. L. James, and P. Abbeel. Reinforcement learning with action-free pre-training from videos. In *International Conference on Machine Learning*, pages 19561–19579. PMLR, 2022.

F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4950–4957. IJCAI, 2018. doi: 10.24963/ijcai.2018/687.

F. Torabi, G. Warnell, and P. Stone. Generative adversarial imitation from observation. In *ICML Workshop on Imitation, Intent, and Interaction (I3)*, 2019. PMLR Workshop Proceedings.

Q. Zheng, M. Henaff, B. Amos, and A. Grover. Semi-supervised offline reinforcement learning with action-free trajectories. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42339–42362. PMLR, 2023.

A. Proofs

A.1. Proof of Theorem 3.3

Theorem A.1 (Theorem 1 of Fu et al. (2018)). *Let $u, v, z \in \mathbb{N}$ with $u, z \geq v$. Suppose*

$$X^* \in \mathbb{R}^{u \times z}, \qquad W^* \in \mathbb{R}^{u \times v}, \qquad H^* \in \mathbb{R}^{v \times z}$$

satisfy

$$X^* = W^* H^*, \qquad \operatorname{rank}(X^*) = \operatorname{rank}(H^*) = v,$$

and assume that H^* is sufficiently scattered. Then the optimization problem

$$\begin{aligned} \min_{W \in \mathbb{R}^{u \times v}, H \in \mathbb{R}^{v \times z}} \quad & \det(W^\top W) \\ \text{s.t.} \quad & X^* = WH \\ & H\mathbb{1}^\top = \mathbb{1}, H \geq 0 \end{aligned} \tag{5}$$

has a unique solution up to permutation: there exists a permutation matrix $\Sigma \in \{0,1\}^{v \times v}$ and a full-rank permutation matrix $D \in \mathbb{R}^{v \times v}$ such that

$$W = W^* \Sigma D, \qquad H = D^{-1} \Sigma^\top H^*.$$

Lemma A.2 (Column-stochasticity removes diagonal scaling). *Let $T, T' \in \mathcal{M}_{u \times k}$ and $\Pi, \Pi' \in \mathcal{M}_{k \times m}$ satisfy*

$$T' = TD, \qquad \Pi' = D^{-1}\Pi$$

for some invertible diagonal matrix $D \in \mathbb{R}^{k \times k}$. Then $D = I_k$.

Proof. Write $D = \operatorname{diag}(d_1, \ldots, d_k)$. Since T and T' are column-stochastic,

$$\mathbb{1}^\top T' = \mathbb{1}^\top TD = \mathbb{1}^\top D = \mathbb{1}^\top.$$

Hence $d_a = 1$ for every $a \in [k]$, so $D = I_k$. □

Under Assumption 3.2, Theorem A.1 implies that any feasible factorization of P_o^* is unique up to permutation and diagonal scaling. Concretely, there exist a permutation matrix Σ_o and an invertible diagonal matrix D_o such that

$$T_o = T_o^* \Sigma_o D_o, \qquad \Pi_o = D_o^{-1} \Sigma_o^\top \Pi_o^*.$$

By Lemma A.2, column-stochasticity forces $D_o = I_k$. Therefore

$$T_o = T_o^* \Sigma_o, \qquad \Pi_o = \Sigma_o^\top \Pi_o^*,$$

as claimed.

A.2. Proof of Theorem 3.5

Lemma A.3 (Correct determinant bound). *Let $\Pi^* \in \mathbb{R}_+^{k \times m}$ have rank k, and define*

$$K := \operatorname{cone}(\Pi^*) \subset \mathbb{R}^k.$$

Assume

$$\mathcal{C}_k \subseteq K, \qquad K^* \cap \operatorname{bd}(\mathcal{C}_k^*) = \{\lambda e_j : j \in [k],\ \lambda \geq 0\}.$$

Let $A \in \mathrm{GL}(k)$ satisfy

$$A\Pi^* \geq 0, \qquad A^\top \mathbb{1} = \mathbb{1}.$$

Then

$$|\det(A)| \leq 1.$$

Moreover, equality holds if and only if A is a permutation matrix.

Proof. Let $r_1^\top, \ldots, r_k^\top$ denote the rows of A.

Since $A\Pi^* \geq 0$, for each i we have

$$r_i^\top \Pi^* \geq 0.$$

Equivalently, for every column π_j^* of Π^*,

$$r_i^\top \pi_j^* \geq 0.$$

Hence each row vector r_i lies in the dual cone K^*.

Because $\mathcal{C}_k \subseteq K$, duality of cones gives

$$K^* \subseteq \mathcal{C}_k^* = \{x \in \mathbb{R}^k : \mathbb{1}^\top x \geq \|x\|_2\}.$$

Therefore each row satisfies

$$\|r_i\|_2 \leq \mathbb{1}^\top r_i.$$

By Hadamard's inequality applied to the rows of A,

$$|\det(A)| \leq \prod_{i=1}^k \|r_i\|_2 \leq \prod_{i=1}^k \mathbb{1}^\top r_i.$$

Now

$$\mathbb{1}^\top r_i = (A\mathbb{1})_i,$$

so

$$\prod_{i=1}^k \mathbb{1}^\top r_i = \prod_{i=1}^k (A\mathbb{1})_i.$$

Also,

$$\sum_{i=1}^k (A\mathbb{1})_i = \mathbb{1}^\top A \mathbb{1} = (A^\top \mathbb{1})^\top \mathbb{1} = \mathbb{1}^\top \mathbb{1} = k.$$

Since each $(A\mathbb{1})_i \geq \|r_i\|_2 \geq 0$, AM–GM yields

$$\prod_{i=1}^{k}(A\mathbb{1})_i \leq 1.$$

Hence

$$|\det(A)| \leq 1.$$

Now suppose equality holds. Then equality must hold in:

(i) Hadamard's inequality, so the rows $r_1, \ldots, r_k$ are pairwise orthogonal;

(ii) the bound $\|r_i\|_2 \leq \mathbb{1}^\top r_i$, so each $r_i \in \mathrm{bd}(\mathcal{C}_k^*)$;

(iii) AM–GM, so each $(A\mathbb{1})_i = 1$, i.e. each row sum equals 1.

Thus each row r_i lies in

$$K^* \cap \mathrm{bd}(\mathcal{C}_k^*),$$

so by assumption,

$$r_i = \lambda_i e_{\ell_i}$$

for some $\lambda_i \geq 0$ and $\ell_i \in [k]$. Since the row sum is 1, we get $\lambda_i = 1$. Thus every row is a standard basis vector. Because A is invertible, these basis vectors must be distinct, so A is a permutation matrix.

Conversely, any permutation matrix satisfies the assumptions and has determinant of absolute value 1. □

Proof of Theorem 3.5. Fix $o \in \mathcal{O}$ and suppress the dependence on o in the notation. Write

$$t_a \coloneqq t_a^o, \qquad p_e \coloneqq p_e^o, \qquad \Pi^* \coloneqq \Pi_o^*.$$

Let $(\bar{t}_1, \ldots, \bar{t}_k, \Pi)$ be any feasible solution of (4).

Define the embedded latent and observable elements

$$u_a \coloneqq \Phi(t_a) \in \mathcal{H},$$
$$\bar{u}_a \coloneqq \Phi(\bar{t}_a) \in \mathcal{H},$$
$$v_e \coloneqq \Phi(p_e) \in \mathcal{H}.$$

and collect them into linear maps

$$T^* = \begin{bmatrix} u_1 & \cdots & u_k \end{bmatrix},$$
$$\bar{T} = \begin{bmatrix} \bar{u}_1 & \cdots & \bar{u}_k \end{bmatrix},$$
$$P = \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix}.$$

By the ground-truth mixture identity, feasibility, and linearity of Φ,

$$P = \bar{T}\Pi = T^*\Pi^*.$$

By Assumption 3.4(i),

$$\mathrm{rank}(G_\Phi(t_1, \ldots, t_k)) = k,$$

so the Gram matrix $G_\Phi(t_1, \ldots, t_k)$ is positive definite. Hence $u_1, \ldots, u_k$ are linearly independent, and therefore

$$\dim \mathrm{span}\{u_1, \ldots, u_k\} = k.$$

Moreover, since Π^* is sufficiently scattered, the cone

$$K \coloneqq \mathrm{cone}(\Pi^*)$$

contains $\mathcal{C}_k$, which is full-dimensional in $\mathbb{R}^k$. Hence K is full-dimensional, so $\mathrm{rank}(\Pi^*) = k$.

Therefore the columns of Π^* span $\mathbb{R}^k$, and thus

$$\mathrm{span}\{v_1, \ldots, v_m\} = T^*(\mathbb{R}^k) = \mathrm{span}\{u_1, \ldots, u_k\}.$$

In particular,

$$\dim \mathrm{span}\{v_1, \ldots, v_m\} = k.$$

On the other hand, since $P = \bar{T}\Pi$, we have

$$\mathrm{span}\{v_1, \ldots, v_m\} \subseteq \mathrm{span}\{\bar{u}_1, \ldots, \bar{u}_k\}.$$

The left-hand side has dimension k, while the right-hand side is spanned by at most k vectors, so in fact

$$\mathrm{span}\{\bar{u}_1, \ldots, \bar{u}_k\} = \mathrm{span}\{u_1, \ldots, u_k\}.$$

Therefore there exists a unique invertible matrix $S \in \mathrm{GL}(k)$ such that

$$\bar{T} = T^*S.$$

Substituting into $P = \bar{T}\Pi = T^*\Pi^*$ gives

$$T^*S\Pi = T^*\Pi^*.$$

Since the columns of T^* are linearly independent, T^* is injective on $\mathbb{R}^k$, and hence

$$S\Pi = \Pi^*, \qquad \text{so} \qquad \Pi = S^{-1}\Pi^*.$$

Let

$$A \coloneqq S^{-1}.$$

Then

$$A\Pi^* = \Pi \geq 0.$$

Next, because Φ is linear and injective, the identity $\bar{T} = T^*S$ implies that for each $a \in [k]$,

$$\bar{t}_a = \sum_{b=1}^{k} S_{ba}\, t_b$$

as finite signed measures in $\mathcal{M}(\mathcal{O})$. Since both $\bar{t}_a$ and each t_b are probability measures, taking total mass yields

$$1 = \bar{t}_a(\mathcal{O}) = \sum_{b=1}^{k} S_{ba}\, t_b(\mathcal{O}) = \sum_{b=1}^{k} S_{ba}.$$

Thus every column of S sums to one, i.e.

$$\mathbb{1}^\top S = \mathbb{1}^\top.$$

Equivalently,

$$A^\top \mathbb{1} = \mathbb{1}.$$

We may therefore apply Lemma A.3 to conclude that

$$|\det(A)| \le 1, \qquad \text{hence} \qquad |\det(S)| \ge 1,$$

with equality if and only if S is a permutation matrix.

Finally, the embedded Gram matrix transforms as

$$G_\Phi(\bar{t}_1, \dots, \bar{t}_k) = S^\top G_\Phi(t_1, \dots, t_k) S,$$

and therefore

$$\det\left(G_\Phi(\bar{t}_1, \dots, \bar{t}_k)\right) = \det\left(G_\Phi(t_1, \dots, t_k)\right) \det(S)^2.$$

Since $|\det(S)| \ge 1$, we obtain

$$\det\left(G_\Phi(\bar{t}_1, \dots, \bar{t}_k)\right) \ge \det\left(G_\Phi(t_1, \dots, t_k)\right).$$

The ground-truth factorization $(t_1, \dots, t_k, \Pi^*)$ is feasible, so it achieves the value

$$\det\left(G_\Phi(t_1, \dots, t_k)\right).$$

Hence every optimizer must satisfy $|\det(S)| = 1$. By Lemma A.3, this happens if and only if S is a permutation matrix. Therefore there exists a permutation matrix Σ such that

$$\bar{T} = T^* \Sigma, \qquad \Pi = \Sigma^\top \Pi^*.$$

Because Φ is injective, $\bar{T} = T^* \Sigma$ implies

$$\bar{t}_a = t_{\Sigma(a)}, \qquad a \in [k].$$

Restoring the dependence on o gives exactly the claimed conclusion:

$$\bar{t}_a = t^o_{\Sigma(a)}, \qquad a \in [k], \qquad \text{and} \qquad \Pi_o = \Sigma_o^\top \Pi^*_o.$$

□

A.3. Proof of Theorem 3.7

Fix $o_0 \in \mathcal{O}$, and let $\Sigma_0 := \Sigma(o_0)$. Write σ_0 for the corresponding permutation of $[k]$, so that

$$\phi_a(o_0) = \phi^*_{\sigma_0(a)}(o_0) \qquad \text{for all } a \in [k].$$

Because the true embedded latent transitions at o_0 are pairwise distinct, the quantity

$$\delta_0 := \min_{i \ne j} \left\| \phi^*_i(o_0) - \phi^*_j(o_0) \right\|_{\mathcal{H}}$$

is strictly positive.

By continuity of $o \mapsto \phi_a(o)$ and $o \mapsto \phi^*_a(o)$, there exists a neighborhood U of o_0 such that for all $o \in U$ and all $a \in [k]$,

$$\|\phi_a(o) - \phi_a(o_0)\|_{\mathcal{H}} < \delta_0/4, \qquad \|\phi^*_a(o) - \phi^*_a(o_0)\|_{\mathcal{H}} < \delta_0/4.$$

Fix $o \in U$. For each $a \in [k]$, we then have

$$\begin{aligned} \|\phi_a(o) - \phi^*_{\sigma_0(a)}(o)\|_{\mathcal{H}} &\le \|\phi_a(o) - \phi_a(o_0)\|_{\mathcal{H}} \\ &\quad + \|\phi^*_{\sigma_0(a)}(o_0) - \phi^*_{\sigma_0(a)}(o)\|_{\mathcal{H}} \\ &< \delta_0/2. \end{aligned}$$

On the other hand, for any $b \ne \sigma_0(a)$,

$$\begin{aligned} \|\phi_a(o) - \phi^*_b(o)\|_{\mathcal{H}} &\ge \|\phi^*_{\sigma_0(a)}(o_0) - \phi^*_b(o_0)\|_{\mathcal{H}} \\ &\quad - \|\phi_a(o) - \phi_a(o_0)\|_{\mathcal{H}} \\ &\quad - \|\phi^*_b(o) - \phi^*_b(o_0)\|_{\mathcal{H}} \\ &> \delta_0 - \delta_0/4 - \delta_0/4 \\ &= \delta_0/2. \end{aligned}$$

Thus, for every $a \in [k]$, the vector $\phi_a(o)$ is strictly closer than $\delta_0/2$ to $\phi^*_{\sigma_0(a)}(o)$ and strictly farther than $\delta_0/2$ from every other true component $\phi^*_b(o)$. Therefore the only permutation relating $T(o)$ and $T^*(o)$ is Σ_0. Hence

$$\Sigma(o) = \Sigma_0 \qquad \text{for all } o \in U.$$

So $\Sigma(\cdot)$ is locally constant.

Since $\Sigma(o)$ takes values in the finite discrete set of $k \times k$ permutation matrices, any locally constant map is constant on each connected component of $\mathcal{O}$. Because $\mathcal{O}$ is connected, $\Sigma(o)$ must be constant on all of $\mathcal{O}$.

A.4. Proof of Corollary 3.9

Fix any $o \in \mathcal{O}$. By Assumption 3.8 (ii), Theorem 3.5 applies at that observation. Hence there exists a permutation matrix $\Sigma(o)$ such that

$$p(\cdot \mid o, a) = p^*(\cdot \mid o, \Sigma(o)(a)), \qquad \Pi(o) = \Sigma(o)^\top \Pi^*(o).$$

Because Φ is injective, this is equivalent to

$$T(o) = T^*(o)\Sigma(o), \qquad \Pi(o) = \Sigma(o)^\top \Pi^*(o),$$

where

$$\begin{aligned} T(o) &= \begin{bmatrix} \Phi(p(\cdot \mid o, 1)) & \cdots & \Phi(p(\cdot \mid o, k)) \end{bmatrix}, \\ T^*(o) &= \begin{bmatrix} \Phi(p^*(\cdot \mid o, 1)) & \cdots & \Phi(p^*(\cdot \mid o, k)) \end{bmatrix}. \end{aligned}$$

By Assumption 3.8 (iii) and the continuity assumed for the candidate family, the maps $o \mapsto T(o)$ and $o \mapsto T^*(o)$ are continuous. Moreover, Assumption 3.8 (ii) implies that the embedded true latent transitions are linearly independent at every o, and in particular pairwise distinct. Therefore the

hypotheses of Theorem 3.7 are satisfied, so there exists a single permutation matrix Σ such that

$$p(\cdot \mid o, a) = p^*(\cdot \mid o, \Sigma(a)),$$
$$\Pi(o) = \Sigma^\top \Pi^*(o), \qquad \text{for all } o \in \mathcal{O}.$$

Finally, by Assumption 3.8 (iv), the permutation is known on the anchor subset S. Since the same permutation Σ applies globally, it is uniquely determined there and hence uniquely determined everywhere. If the anchor identifies the true labeling, then $\Sigma = I_k$, yielding

$$p(\cdot \mid o, a) = p^*(\cdot \mid o, a),$$
$$\Pi(o) = \Pi^*(o), \qquad \text{for all } o \in \mathcal{O},\ a \in [k].$$

This proves the claim.