# Provably Efficient Third-Person Imitation from Offline Observation

**Aaron Zweig**
Courant Institute of Mathematical Sciences
New York University
az831@nyu.edu

**Joan Bruna**
Courant Institute of Mathematical Sciences
Center for Data Science
New York University
bruna@cims.nyu.edu

## Abstract

Domain adaptation in imitation learning represents an essential step towards improving generalizability. However, even in the restricted setting of third-person imitation where transfer is between isomorphic Markov Decision Processes, there are no strong guarantees on the performance of transferred policies. We present problem-dependent, statistical learning guarantees for third-person imitation from observation in an offline setting, and a lower bound on performance in an online setting.

## 1 INTRODUCTION

Imitation learning typically performs training and testing in the same environment. This is by necessity as the Markov Decision Process (MDP) formalism defines a policy on a particular state space. However, real world environments are rarely so cleanly defined and benign changes to the environment can induce a completely new state space. Although deep imitation learning (Ho and Ermon, 2016) still defines a policy on unseen states, it remains extremely difficult to effectively generalize (Duan et al., 2017).

Domain adaptation addresses how to generalize a policy defined in a source domain to perform the same task in a target domain (Higgins et al., 2017). Unfortunately, this objective is inherently ill-defined. One wouldn't expect to successfully transfer from a 2D gridworld to a self-driving car, but there is ambiguity in how to define a similarity measure on MDPs.

Third-person imitation (Stadie et al., 2017) resolves this ambiguity by considering transfer between isomorphic MDPs (formally defined in Section 2), where the objective is to observe a policy in the source domain, and imitate that policy in the target domain. In contrast to domain adaptation between unaligned distributions, the dynamics structure constrains the space of possible isomorphisms, and in some cases the source and target may be related by a unique isomorphism.

We consider an idealized setting for third-person imitation with complete information about the source domain, where we perfectly understand the dynamics and the policy to be imitated. This work offers a theoretical analysis, in particular demonstrating that restricting to isomorphic MDPs with complete knowledge does not trivialize the problem. Specifically, regarding how the agent may observe the target domain, we consider two regimes, summarized in Figure 1:

- In the **offline** regime (Section 4), an oracle perfectly transfers the source policy into the target domain, and the agent observes trajectories from the oracle policy (without seeing the oracle's actions). In this regime, we provide positive results establishing that with limited, state-only observations in the target domain, we can still efficiently imitate a policy defined in the source domain (Theorem 4.12).

- In the **online** regime (Section 5), the agent chooses policies in the target domain and draws trajectories. Our negative results in this setting (Theorem 5.1) prove that with full interaction in the target domain, imitation is extremely difficult in the presence of structural symmetry.

**A Motivating Example:** To clarify the setup and distinguish the two observation regimes, we elaborate upon an example. Suppose our source domain is a video game, where the state space corresponds to the monitor screen and the action space corresponds to key presses. And we wish to imitate an expert player of the game. The target domain is the same game played on a new monitor with higher screen brightness. Clearly the underlying game hasn't changed, and there is a natural bijection

*Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, PMLR volume 124, 2020.

from screen states of the target monitor to those of the source monitor, namely "dimming the screen".

On the one hand, in the offline setting, we're forbidden from playing on the new monitor ourselves. Instead we observe recordings of the expert, played on the brighter monitor. Again, as these are recordings, we see the states the expert visits but not their actions. On the other hand, in the online setting, we simply run transitions on the brighter monitor. Note that if the screen includes benign features which minimally impact the game (say the player's chosen name appears onscreen), it may be very difficult to learn the bijection between target and source monitor. Either way, through observations we guess a new policy to be played on the bright monitor, which hopefully mimics the expert's behavior.

**Summary of Contributions:** Our primary contribution in this work is a provably efficient algorithm for offline third-person imitation, with an polynomial upper bound for the sample complexity necessary to control the imitation loss. Our main technical novelty is a means of clipping the states of a Markov chain according to their stationary distribution, while preserving properties of a bijection between isomorphic chains. We also prove an algorithm-agnostic lower bound for online third-person imitation, through reduction to bandit lower bounds.

## 2 SETUP

### 2.1 Preliminaries

We consider a source MDP without reward $\mathcal{M} = \{S, A, P, p_0, \gamma\}$, and target MDP $\hat{\mathcal{M}} = \{\hat{S}, A, \hat{P}, \hat{p}_0, \gamma\}$. To characterize an isomorphism between $\mathcal{M}$ and $\hat{\mathcal{M}}$, we assume the existence of a bijective mapping $\pi_* : \hat{S} \to S$, such that $\hat{P}(s'|s,a) = P(\pi_*(s')|\pi_*(s), a)$ and $\hat{p}_0(s) = p_0(\pi_*(s))$. Note that in this notation, $\pi_*$ is not a policy.

We also fix an ordering of the states $\hat{S}$ so that $\pi_*$ may be written in matrix form $\Pi_*$ as a permutation matrix. In particular, we will overload notation to use $\pi_*$ as a permutation on $[|S|]$, such that $\pi_*(i) = j$ denotes that $\pi_*(\hat{s}_i) = s_j$. Let $\mathcal{P}$ denote the space of $\hat{S} \to S$ permutation matrices.

A policy $\phi$ maps states to distributions on actions, but for our purposes it will be convenient to consider the policy as a matrix $\Phi : S \to S \times A$. To relate the two notions, $\Phi$ is a block of diagonal matrices $\Phi_a : S \to S$ for each action, where $(\Phi_a)_{ii} = \phi(a|s_i)$, and $\Phi = [\Phi_{a_1}|\dots|\Phi_{a_{|A|}}]^T$.

The dynamics matrix is denoted $P : S \times A \to S$. It can also be decomposed into blocks $P_a : S \to S$ where $(P_a)_{ij} = p(s_j|s_i, a)$, and $P = [P_{a_1}|\dots|P_{a_{|A|}}]$.

Using this notation, $\Phi^T P^T$ forms the Markov chain on $S$ induced by following policy $\phi$. Explicitly,

$$P_\phi(s'|s) = \sum_a \phi(a|s) P(s'|s,a) = \left(\Phi^T P^T\right)_{s,s'} \quad (1)$$

Note that under this notation, the dynamics and initial distribution in $\hat{\mathcal{M}}$ can be written as $\hat{P} = \Pi_*^T P(I \otimes \Pi_*)$ and $\hat{p}_0 = \Pi_*^T p_0$ respectively. The occupancy measure $\rho_\phi$ is defined with regard to a policy, as well as the underlying dynamics and initial distribution. Specifically, $\rho_\phi(s,a) = (1-\gamma) E_{s_0 \sim p_0, \tau \sim \Phi}\left[\sum_{i=0}^\infty \gamma^i \phi(a|s) P(s_i = s)\right]$, where the dependence on the dynamics $P$ is through the sampling of a trajectory $\tau$.

Similarly, we introduce the state-only occupancy measure $\mu_\phi(s) := \sum_a \rho_\phi(s,a)$. We will make use of the identity $\rho_\phi(s,a) = \phi(a|s)\mu_\phi(s)$, as well as the fact that $\mu_\phi$ is the stationary distribution of the Markov chain $\Phi^T((1-\gamma)p_0 \mathbf{1}^T + \gamma P)^T$, which both follow from the constraint-based characterization of occupancy (Puterman, 1994).

The value function for a given policy $\phi$ and reward function $R$ is defined as

$$V_{\phi,R}(s) = E_{s_0 = s, \tau \sim \Phi}\left[\sum_{i=0}^\infty \gamma^i R(s_i, a_i)\right]. \quad (2)$$

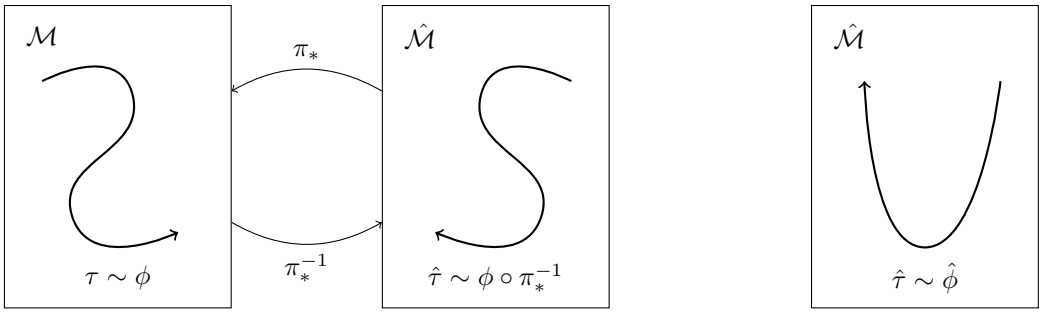
We note the very useful identity $(1-\gamma) E_{s_0 \sim p_0}[V_{\phi,R}(s_0)] = \langle \rho_\phi, R \rangle$.

Lastly, we use the notation $\sigma_i(A)$ to denote the $i$th largest singular value of $A$.

### 2.2 Observation Settings

To begin, we're given full knowledge of the source domain $\mathcal{M}$, as well as $\Phi : S \times A \to S$ and $\rho_\Phi \in \mathbb{R}^{S \times A}$, the policy and corresponding occupancy measure we want to imitate. We consider two settings through which we can interact with the target domain, in order to learn how to adapt $\Phi$ into this new domain.

**Offline:** In the offline setting, we only observe the policy $\Phi_* := (I \otimes \Pi_*^T)\Phi\Pi_*$ being played in $\hat{\mathcal{M}}$. We can consider $\Phi_*$ as an oracle for third-person imitation, as this policy exactly maps from $\hat{\mathcal{M}}$ to $\mathcal{M}$, calls $\Phi$, and maps back. To guarantee the trajectories don't get trapped in a terminal state, we assume this agent has a $1-\gamma$ reset probability. Through these observations, we must output a policy $\hat{\Phi}$ to be played in $\hat{\mathcal{M}}$. We provide upper bounds for this setting in Section 4.




(a) (b)

Figure 1: The observation regimes. In the offline setting (a), the agent observes trajectories $\tau$ sampled from the policy $\phi$ that have been perfectly transferred into the isomorphic target domain. In the online setting (b), the MDPs are still isomorphic but the agent only observes trajectories after playing their own policy $\hat{\phi}$.

Crucially, in this setting we assume access to the states *but not actions* from observed trajectories, in the imitation from observation setting (Sun et al., 2019). This assumption is well-motivated. In practice, observed trajectories from an expert often come from video, where actions are difficult to infer (Liu et al., 2018). Additionally, the problem becomes trivial with observed actions, as one may mimic the oracle's actions at each state in $\hat{S}$ without trying to understand $\Pi_*$ at all.

**Online:** In the online setting, we define our own policy $\hat{\Phi}_t$ to play in $\hat{\mathcal{M}}$ at each timestep $t$, with full observation of the trajectories. After $T$ total transitions we output our final policy $\hat{\Phi}$. Intuitively, this setting allows for more varied observations in the target domain. But without an expert oracle to demonstrate the correct state distribution, an agent in this setting may be deceived by near-symmetry in the dynamics and predict the wrong alignment. We further highlight this difficulty in Section 5.

### 2.3 Imitation Objective

In either setting, through observations from the target domain we output a policy $\hat{\Phi}$. The corresponding occupancy measure we denote as $\rho_{\hat{\Phi},\Pi_*} \in \mathbb{R}^{\hat{S}\times A}$, where the subscript $\Pi_*$ reflects the dependence on the dynamics and initial distribution in $\hat{\mathcal{M}}$, namely $\Pi_*^T P(I \otimes \Pi_*)$ and $\Pi_*^T p_0$.

We measure imitation by comparing the correctly transferred policy $\Phi_*$ against the guessed policy $\hat{\Phi}$. Explicitly, our objective is

$$\inf_{\hat{\Phi}} G(\Phi, \Pi_*, \hat{\Phi}) := \inf_{\hat{\Phi}} TV\left((I \otimes \Pi_*^T)\rho_\Phi, \rho_{\hat{\Phi},\Pi_*}\right) \quad (3)$$

As a sanity check, we confirm that if we play $\hat{\Phi} = \Phi_* = (I \otimes \Pi_*^T)\Phi\Pi_*$, then indeed $\hat{\rho}_{\hat{\Phi},\Pi_*} = (I \otimes \Pi_*^T)\rho_\Phi$ and the occupancies are equal.

The total variation belongs to a larger family of distributional distances called integral probability metrics (IPM). A form of this objective with a more general IPM was introduced in (Ho and Ermon, 2016). To justify using this loss, note the objective can be equivalently written $\sup_{\|c\|_\infty \le 1} E_{s\sim\Pi_*^T p_0}[V_{\phi_*,c}(s) - V_{\hat{\phi},c}(s)]$. In other words, minimizing imitation objective guarantees $\Phi_*$ and $\hat{\Phi}$ perform nearly as well for any reward function with a bound on maximum magnitude.

## 3 RELATED WORK

The theory of imitation learning depends crucially on what interaction is available to the agent. Behavior cloning (Bain, 1995) learns a policy offline from supervised expert data. With online data, imitation learning can be cast as a measure matching problem on occupancy measures (Ho and Ermon, 2016). With an expert oracle, imitation learning has no-regret guarantees (Ross et al., 2011). Numerous of these algorithms for imitation learning can be adapted to the observation setting (Torabi et al., 2018a,b; Yang et al., 2019).

General domain adaptation for imitation learning has a rich applied literature (Ammar et al., 2015; Pastor et al., 2009; Tobin et al., 2017). Third-person imitation specifically was formalized in Stadie et al. (2017), extending the method of Ho and Ermon (2016) by learning domain-agnostic features. Other deep algorithms explicitly learn an alignment between the state spaces, based on multiple tasks in the same environments (Kim et al., 2019) or unsupervised image alignment (Gamrian and Goldberg, 2019).

The closest work to ours is Sun et al. (2019), which

shares the focus on imitation learning without access to actions, but differs in studying the first-person setting primarily with online feedback. This work also takes inspiration from literature on friendly graphs (Aflalo et al., 2015), which characterize robustly asymmetric structure.

# 4 OFFLINE IMITATION

## 4.1 Markov Chain Alignment

Because the offline setting only runs policy $\Phi_*$, and reveals no actions, it is equivalent to observing a trajectory of the state-only Markov chain induced by $\Phi_*$ in $\hat{\mathcal{M}}$. Let us elaborate on this fact.

Define the Markov chain $M := \Phi^T((1-\gamma)p_0\mathbf{1}^T+\gamma P)^T$, which is ergodic when restricted to the strongly connected components that intersect the initial distribution. In $\hat{\mathcal{M}}$, the dynamics are $\Pi_*^T P(I \otimes \Pi_*)$, the oracle policy is $(I\otimes\Pi_*)^T\Phi\Pi_*$, and the initial distribution is $\Pi_*^T p_0$. We also assume the oracle agent following $\Phi_*$ has a $1-\gamma$ reset probability.

All together, this implies our observations in the offline setting are drawn from a trajectory of $\Pi_*^T M\Pi_*$. In summary, given full knowledge of $M$ and a trajectory sampled from $\Pi_*^T M\Pi_*$, our algorithm will seek to learn the alignment $\Pi_*$ in order to approximate $\Phi_*$, hopefully leading to low imitation loss.

## 4.2 Symmetry without approximation

As a warmup, we consider the setting with no approximation where we observe $\Pi_*^T M\Pi_*$ exactly. To relate this chain to $M$, we can try to find symmetries, i.e. the minimizers of

$$\arg\min_{\Pi\in\mathcal{P}} \|\Pi^T M\Pi - \Pi_*^T M\Pi_*\|_F \ . \tag{4}$$

We can equivalently consider finding automorphisms of $M$, which may be posed as a minimization over permutation matrices $\Pi : S \to S$:

$$\arg\min_{\Pi} \|\Pi^T M\Pi - M\|_F \ . \tag{5}$$

Clearly both these objectives are minimized at 0. Intuitively, to recover $\Pi_*$ we'd like $\Pi_*$ to be the unique minimizer of (4), or equivalently $I$ to be the unique minimizer of (5). Hence, in order to make third-person imitation tractable, we will seek to bound (5) away from 0 when $\Pi \neq I$, or in other words focus on Markov chains which are robustly asymmetric.

We introduce notation:

**Definition 4.1** (Rescaled transition matrix)**.** For an ergodic Markov chain $M$ with stationary distribution $\mu$, let $D = diag(\mu)$ and define $L = D^{1/2}MD^{-1/2}$ as the *rescaled transition matrix* of $M$.

**Definition 4.2** (Friendly matrix)**.** A matrix $A$ is *friendly* if, given the singular value decomposition $A = U\Sigma V^T$, $\Sigma$ has distinct diagonal elements and $V^T\mathbf{1}$ has all nonzero elements. Similarly, a matrix $A$ is $(\alpha,\beta)$-friendly if $\sigma_\star := \min_i \sigma_i(A) - \sigma_{i+1}(A) > \alpha$ and $V^T\mathbf{1} > \beta\mathbf{1}$ elementwise. An ergodic Markov chain $M$ is *friendly* if its rescaled transition matrix $L$ is friendly.

The significance of friendliness in graphs was studied in Aflalo et al. (2015), to characterize relaxations of the graph isomorphism problem. We first confirm several friendliness properties for Markov chains still hold.

**Proposition 4.3.** *For a permutation matrix* $\Pi$*,* $M = \Pi^T M\Pi$ *if and only if* $D = \Pi^T D\Pi$ *and* $L = \Pi^T L\Pi$*.*

*Proof.* Suppose $M = \Pi^T M\Pi$. If $\mu$ is the stationary distribution of $M$, then $(\mu^T\Pi)(\Pi^T M\Pi) = \mu^T\Pi$. So by uniqueness of the stationary distribution in an ergodic chain, $\mu = \Pi^T\mu$ and therefore $D = \Pi^T D\Pi$. Then clearly $D^{1/2} = \Pi^T D^{1/2}\Pi$ and therefore $L = \Pi^T L\Pi$.

For the reverse implication, $\Pi^T M\Pi = \Pi^T D^{-1/2}LD^{1/2}\Pi = D^{-1/2}LD^{1/2} = M$. □

**Proposition 4.4.** *If* $M$ *is friendly, then it has a trivial automorphism group.*

*Proof.* Suppose $M = \Pi^T M\Pi$. Then by Proposition 4.3, $\Pi^T L^T L\Pi = L^T L = V\Sigma^2 V^T$. In particular, choosing $v$ as a column of $V$, $L^T Lv = \sigma^2 v$ implies $L^T L\Pi v = \sigma^2\Pi v$. By friendliness, every eigenspace of $L^T L$ is one-dimensional, so $\Pi v = \pm v$. And $\mathbf{1}^T\Pi v = \mathbf{1}^T v > 0$, so $\Pi v = v$ and therefore $\Pi = I$. □

In what follows, for any SVD, we will always choose to orient $V$ such that $V^T\mathbf{1} \geq 0$ elementwise.

## 4.3 Exact Symmetry Algorithm

By Proposition 4.3, the automorphism group of $M$ is contained in the automorphism group of the rescaled transition matrix $L$. Interpreting $L$ as a weighted graph, determining its automorphisms is at least as computationally hard as the graph isomorphism problem (Aflalo et al., 2015).

In general, algorithms for graph isomorphisms optimize time complexity, whereas we are more interested in controlling sample complexity. Nevertheless, we have the following result:

**Theorem 4.5.** *Given $M$ and $\Pi_*^T M \Pi_*$, if $M$ is a friendly Markov chain, there is an algorithm to exactly recover $\Pi_*$ in $O(|S|^3)$ time.*

This result is a simple extension of the main result in Umeyama (1988), applying the friendliness property to Markov chains rather than adjacency matrices. But the characterization of automorphisms will be used again later to control sample complexity, when we only observe $\Pi_*^T M \Pi_*$ through sampled trajectories.

We begin with the following:

**Proposition 4.6.** *Given two friendly matrices decomposed as $L_1 = U_1 \Sigma V_1^T$ and $L_2 = U_2 \Sigma V_2^T$, suppose $L_2 = \Pi_*^T L_1 \Pi_*$. Then $\Pi_*$ is the unique permutation which satisfies $V_2 = \Pi^T V_1$.*

*Proof.* Clearly $L_2^T L_2 = \Pi_*^T L_1^T L_1 \Pi_*$. Rewriting with the SVD gives $V_2 \Sigma^2 V_2^T = \Pi_*^T V_1 \Sigma^2 V_1^T \Pi_*$.

Rearranging, this implies $V_2^T \Pi_*^T V_1$ commutes with $\Sigma^2$. Commuting with a diagonal matrix with distinct elements implies $V_2^T \Pi_*^T V_1$ is diagonal. As this product is also unitary and real, it must be that $V_2^T \Pi_*^T V_1 = S$ where $S$ is diagonal and $S^2 = I$.

Again rearranging, this implies $\mathbf{1}^T V_1 = \mathbf{1}^T \Pi_*^T V_1 = \mathbf{1}^T V_2 S$. By the assumption on the SVD orientation, $S$ must preserve signs, therefore $S = I$, and $V_2 = \Pi_*^T V_1$.

Now, suppose $V_2 = \Pi^T V_1$. Then $L_2^T L_2 = \Pi^T L_1^T L_1 \Pi$, so $\Pi_*^T \Pi$ is an automorphism of $L_2^T L_2$ and therefore $\Pi = \Pi_*$. □

*Proof of Theorem 4.5.* Let $L_1$ and $L_2$ be the rescaled transition matrices of $M$ and $\Pi_*^T M \Pi_*$ respectively. Reusing the same SVD notation, by Proposition 4.3 and 4.6, $V_2 = \Pi_*^T V_1$. Consider the linear assignment problem $\min_{\Pi \in \mathcal{P}} \|V_2 - \Pi^T V_1\|_F$, which may be solved in $O(|S|^3)$ time using the Hungarian algorithm (Kuhn, 1955). Again by Proposition 4.6, this linear program is minimized at 0 and recovers $\Pi_*$ as the unique minimizer. □

### 4.4 Symmetry with approximation

With finite sample complexity, we still know the base chain $M$ exactly, but we get empirical estimates of the permuted chain $\Pi_*^T M \Pi_*$ by running trajectories. Specifically, $m$ samples $(X_1, \ldots, X_m)$ are drawn from $\Pi_*^T M \Pi_*$, with $X_1 \sim \Pi_*^T p_0$.

Call the empirical estimate $\hat{M}$, i.e. $\hat{M}_{ij} = \frac{N_{ij}}{N_i}$ where $N_{ij}$ counts the number of observed $i \to j$ transitions and $N_i = \sum_j N_{ij}$. And the empirical stationary distribution is $\hat{\mu}$ where $\hat{\mu}_i = \frac{N_i}{\sum_j N_j}$ and $\hat{D} = diag(\hat{\mu})$.

We can characterize the approximation error of the chain and stationary distribution as $E := \Pi_* \hat{M} \Pi_*^T - M$ and $\Delta = \Pi_* \hat{D} \Pi_*^T - D$ respectively. Note these error terms are defined in the original state space $S$.

Our goal is to use $\hat{M}$ to produce a good policy in the target space. Say we predict the bijection is $\Pi$, and play the policy $\hat{\Phi} = (I \otimes \Pi^T)\Phi\Pi$, whereas the correct policy in the target space is $\Phi_* = (I \otimes \Pi_*^T)\Phi\Pi_*$. We'd like to be able to control the imitation distance between these two policies when $\Pi \approx \Pi_*$.

For that purpose, define $I_t(M) = \{i \in S : \mu_i \geq t, \mu^T = \mu^T M\}$, where $\mu$ is the stationary distribution of $M$, so these states will be visited "sufficiently" often. We first show correctness of the bijection on these states suffices for good imitation.

**Lemma 4.7** (Policy Difference Lemma (Kakade and Langford, 2002))**.** *For two policies $\phi_1, \phi_2$ in the MDP defined by $\{S, A, P, R, p_0\}$,*

$$E_{s \sim p_0}[V_{\phi_1,R}(s) - V_{\phi_2,R}(s)]. = E_{\tau \sim \phi_1, p_0}\left[\sum_{t=0} \gamma^t A^R_{\phi_1,\phi_2}(s_t)\right] = \frac{1}{1-\gamma}\langle \mu_{\phi_1}, A^R_{\phi_1,\phi_2}\rangle ,$$

*where $A^R_{\phi_1,\phi_2}(s) = E_{a \sim \phi_1(\cdot|s)}[E_{s' \sim P}[R(s,a) + \gamma V_{\phi_2,R}(s') - V_{\phi_2,R}(s)]]$ is the average advantage function.*

**Theorem 4.8.** *Suppose $\pi^{-1}(s_i) = \pi_*^{-1}(s_i)$ for $i \in I_t(M)$. Then $G(\Phi, \Pi_*, \hat{\Phi}) \leq \frac{2t|S|}{(1-\gamma)^2}$.*

*Proof.* First we decompose the objective

$$\begin{aligned} G(\Phi, \Pi_*, \hat{\Phi}) &= TV((I \otimes \Pi_*^T)\rho_\Phi, \rho_{\hat{\Phi},\Pi_*}) \\ &= \sup_{\|c\|_\infty \leq 1} \langle \rho_{\Phi_*,\Pi_*} - \rho_{\hat{\Phi},\Pi_*}, c\rangle \\ &= \sup_{\|c\|_\infty \leq 1} E_{\hat{s} \sim \Pi_*^T p_0}[V_{\phi_*,c}(\hat{s}) - V_{\hat{\phi},c}(\hat{s})] . \end{aligned}$$

From the assumption and the definition of $\Phi_*$ and $\hat{\Phi}$, we have $\phi_*(\cdot|\hat{s}_i) = \hat{\phi}(\cdot|\hat{s}_i)$ whenever $i \in \pi_*^{-1}(I_t(M))$. Equivalently, since $\mu_\phi$ is the stationary distribution of $M$ in the original space, and $\mu_{\phi_*} = \Pi_*^T \mu_\phi$, we have $\phi_*(\cdot|\hat{s}) = \hat{\phi}(\cdot|\hat{s})$ whenever $\mu_{\phi_*}(\hat{s}) \geq t$.

Note that $\phi_*(\cdot|\hat{s}) = \hat{\phi}(\cdot|\hat{s})$ implies $A^R_{\phi_*,\hat{\phi}}(\hat{s}) = 0$ for any $R$. Hence,

$$
\begin{aligned}
&(1-\gamma)E_{\hat{s}\sim\Pi_*^T p_0}[V_{\phi_*,c}(\hat{s}) - V_{\hat{\phi},c}(\hat{s})] \\
&\quad= \sum_{i\in\pi_*^{-1}(I_t(M))} \mu_{\phi_*}(\hat{s}_i) A^c_{\phi_*,\hat{\phi}}(\hat{s}_i) \\
&\qquad+ \sum_{i\notin\pi_*^{-1}(I_t(M))} \mu_{\phi_*}(\hat{s}_i) A^c_{\phi_*,\hat{\phi}}(\hat{s}_i) \\
&\quad\leq \sum_{i\notin\pi_*^{-1}(I_t(M))} t|A^c_{\phi_*,\hat{\phi}}(\hat{s}_i)| \\
&\quad\leq \frac{2t|S|}{1-\gamma},
\end{aligned}
$$

following from the simple bound $\max_s |A^c_{\phi_1,\phi_2}(s)| \leq \frac{2}{1-\gamma}$.

□

The bound in Theorem 4.8 depends on $\Pi$ in a very discrete sense, controlled by the states where $\Pi$ and $\Pi_*$ agree. Say $\Pi$ contains a single error, $\hat{s} = \pi^{-1}(s) = \pi_*^{-1}(s')$ for $s \neq s'$. Then at $\hat{s}$ we mistakenly play the action distribution $\hat{\phi}(\cdot|\hat{s}) = \phi(\cdot|s)$, rather than the correct distribution $\phi_*(\cdot|\hat{s}) = \phi(\cdot|s')$. Because we never observe actions from the oracle, $\phi$ could be arbitrarily different at $s$ and $s'$, yielding a very suboptimal occupancy measure.

Finally, we also remark on an approximate version of our setting, where the source and target MDPs are nearly isomorphic. We measure nearness via TV distance of the dynamics distributions, informed by the Simulation Lemma:

**Lemma 4.9** (Lemma 4 in Kearns and Singh (2002))**.** *Consider MDPs* $\mathcal{M}_1 := \{S, A, P_1, p_0, R, \gamma\}$ *and* $\mathcal{M}_2 := \{S, A, P_2, p_0, R, \gamma\}$. *If* $\sup_{s,a} TV(P_1(\cdot|s,a), P_2(\cdot|s,a)) \leq \epsilon$*, then for any policy* $\phi$*:*

$$
\|V_{\phi,\mathcal{M}_1} - V_{\phi,\mathcal{M}_2}\|_\infty \leq \frac{\gamma\epsilon R_{max}}{2(1-\gamma)^2}
$$

Let $\pi_*^\sharp$ be the pushforward of $\pi_*$. Then the following is immediate, granting a bound on the imitation objective that additively depends on the closeness of the source and target domains:

**Corollary 4.10.** *Suppose the target MDP* $\hat{\mathcal{M}}$ *only obeys the bound* $\sup_{s,a} TV(P(\cdot|\pi_*(s),a), \pi_*^\sharp\hat{P}(\cdot|s,a)) \leq \epsilon$*. Then under the same assumptions as Theorem 4.8,* $G(\Phi, \Pi_*, \hat{\Phi}) \leq \frac{2t|S|}{(1-\gamma)^2} + \frac{\gamma\epsilon}{2(1-\gamma)^2}$

### 4.5 Approximate Symmetry Algorithm

In light of Theorem 4.8, an algorithm could either seek to recover $\Pi_*$ exactly, or find a $\Pi$ which agrees with $\Pi_*$ on high occupancy states. We consider a learning algorithm for both objectives, and bound its sample complexity. The trick will be carefully setting the threshold $t$ that defines what constitutes high occupancy.

**Algorithm 1:** Permuted Policy Learning

**Input:** $P$, $\Phi$, $\gamma$, $p_0$, $t$, $(X_1, \ldots, X_m)$
**Output:** A policy $\hat{\Phi} : \hat{S} \to \hat{S} \times A$
$M \leftarrow \Phi^T((1-\gamma)p_0\mathbf{1}^T + \gamma P)^T$
$\mu \leftarrow$ STATIONARY$(M)$
**for** $(i,j) \in [|S|] \times [|S|]$ **do**
  $N_{ij} \leftarrow 0$
**end**
**for** $t \in [m-1]$ **do**
  $N_{X_t,X_{t+1}} \leftarrow N_{X_t,X_{t+1}} + 1$
**end**
$\hat{\mu} \leftarrow 0$
$\hat{M} \leftarrow 0$
**for** $i \in [|S|]$ **do**
  $\hat{\mu}_i \leftarrow \sum_j N_{ij}/(m-1)$
  **for** $j \in [|S|]$ **do**
    $\hat{M}_{ij} \leftarrow N_{ij}/\sum_k N_{ik}$
  **end**
**end**
$D \leftarrow$ DIAG$(\mu)$
$\hat{D} \leftarrow$ DIAG$(\hat{\mu})$
$I_t \leftarrow \{i \in [|S|] : \mu_i \geq t\}$
$\hat{I}_t \leftarrow \{i \in [|S|] : \hat{\mu}_i \geq t\}$
$M \leftarrow$ SUBMATRIX$(M, I_t, I_t)$
$\hat{M} \leftarrow$ SUBMATRIX$(M, \hat{I}_t, \hat{I}_t)$
$D \leftarrow$ SUBMATRIX$(D, I_t, I_t)$
$\hat{D} \leftarrow$ SUBMATRIX$(\hat{D}, \hat{I}_t, \hat{I}_t)$
$U, \Sigma, V \leftarrow \text{SVD}(D^{1/2}MD^{-1/2})$
$\hat{U}, \hat{\Sigma}, \hat{V} \leftarrow \text{SVD}(\hat{D}^{1/2}\hat{M}\hat{D}^{-1/2})$
$\Pi' \leftarrow$ HUNGARIAN$(V, \hat{V})$
Choose any $\Pi \in \mathcal{P}$ such that $\forall i \in \hat{I}_t$, $\pi(i) = \pi'(i)$
**return** $(I \otimes \Pi)^T \Phi \Pi$

To state the theorem, we introduce the subscript $t$ notation to denote the principle submatrix defined by the indices of $I_t$, and $g := \min_i |\mu_i - t|$ is the gap between the threshold and stationary values. Lastly, we define:

**Definition 4.11** (Pseudospectral gap)**.** The *pseudospectral gap* of an ergodic Markov chain $M$ is $\gamma_{ps}(M) = \max_{k\geq 1} \frac{1-\lambda_2((D^{-1}M^TD)^kM^k)}{k}$, where $\lambda_2$ denotes the second largest eigenvalue.

If $M$ is not ergodic, we will take $\gamma_{ps}(M)$ to mean the pseudospectral gap of $M$ restricted to the strongly connected components that intersect $p_0$.

**Theorem 4.12.** *The policy learning algorithm in Algorithm 1 satisfies the following: for* $1 \geq \delta \geq 0$, $t > 0$, *if* $D_t^{1/2} M_t D_t^{-1/2}$ *is* $(\alpha, \beta)$*-friendly and* $m = poly\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{t}, |I_t|, \frac{1}{g}, \frac{1}{\gamma_{ps}(M)}, \log \frac{1}{1-\gamma}, \log |S|, \log \frac{1}{\delta}\right)$ *then with probability at least* $1 - \delta$*, the output policy* $\hat{\Phi}$ *satisfies* $G(\Phi, \Pi_*, \hat{\Phi}) \leq \frac{2t|S|}{(1-\gamma)^2}$. *In particular, if* $\min_i \mu_i > t$, $\hat{\Phi} = \Phi_*$.

The most important feature of this bound is the dependence on $|S|$. In the sample complexity it only appears through a log term, and all other terms can be independent of $|S|$ depending on the choice of $t$ and the structural properties of $M$. The error is still linear in $|S|$, but this term appears necessary. If some occupancy mass leaves the well-supported states $\pi_*^{-1}(I_t)$, it could cover all the negligible states, and either incur error linear in $|S|$, or require exploration of every state and therefore sample complexity linear in $|S|$.

*Proof sketch.* Here we give the main ideas of the proof, full details are provided in the Appendix.

Remind that $\hat{M} = \Pi_*^T (M + E)\Pi_*$ and $\hat{D} = \Pi_*^T (D + \Delta)\Pi_*$. We also define $\tilde{M} = M + E$ as the empirical chain permuted back into the original MDP. Likewise define $\tilde{D} = D + \Delta$, and $\tilde{\mu}$ to be the diagonal of $\tilde{D}$.

Given $\hat{M}$ and $\hat{D}$, the immediate choice for an estimator of the rescaled transition matrix would be $\hat{D}^{1/2} \hat{M} \hat{D}^{-1/2}$. However, this will not be well-defined if our samples don't visit every state of $\hat{S}$. Furthermore, if $M$ is only ergodic when restricted to a subset of $S$, then $\hat{D}^{-1}$ won't be defined even with infinite sample complexity. Similarly, if $\mu_* := \min_i \mu_i$ is vanishingly small, $m$ will become prohibitively large in order to guarantee that $\hat{D}^{-1}$ is well-defined.

Our primary technical novelty addresses both these issues by setting a threshold $t$ on stationary mass, and discarding states below the threshold. Define $I_t = \{i \in [|S|] : \mu_i \geq t\}$ and $\hat{I}_t = \{i \in [|S|] : \hat{\mu}_i \geq t\}$. We restate the notation that a subscript $t$ denotes taking the principle submatrix corresponding to $I_t$ or $\hat{I}_t$ depending on the matrix's domain. So for example, $M_t$ is $M$ restricted to rows and columns given by $I_t$, and likewise $\hat{M}_t$ is $\hat{M}$ restricted to $\hat{I}_t$.

Several concentration results for empirical Markov chain transitions and stationary distributions control the convergence of our estimators (Wolfer and Kontorovich, 2019a,b). Our main assumption is that the gap $g = \min_i |\mu_i - t|$ is non-negligible. Then with high probability and sample complexity depending on $g$ but not $\min_i \mu_i$, $\mu_i \geq t$ iff $\tilde{\mu}_i \geq t$. In other words, no empirical stationary estimates will "cross" the threshold, or put another way $\pi_*^{-1}(I_t) = \hat{I}_t$. We can then restrict our attention to the states above the threshold, such that the sample complexity necessary for concentration $\tilde{M} \approx M$ depends on $t$ but not $\min_i \mu_i$ (and only logarithmically on $|S|$).

For $t > 0$, the restricted rescaled transition matrix $L_t = D_t^{1/2} M_t D_t^{-1/2}$ is well-defined. And with high probability we can define our estimator $\hat{L}_t = \hat{D}_t^{1/2} \hat{M}_t \hat{D}_t^{-1/2}$. Appealing to a strong friendliness assumption on $L_t$, singular value perturbation inequalities imply that $\hat{L}_t$ is also friendly.

Finally, the asymmetric properties of friendly matrices given in Proposition 4.6 enable exact recovery of the submatrix of $\Pi_*$ restricted to the indices $I_t$ and $\hat{I}_t$. And by Theorem 4.8, determining the alignment on all high-occupancy states still yields a bound on the imitation loss.

□

| | $m = 10^3$ | $m = 10^4$ | $m = 10^5$ |
|---|---|---|---|
| Garnet MDP | $.88 \pm .03$ | $.78 \pm .11$ | $.23 \pm .18$ |
| Planted MDP | $.15 \pm .30$ | $.07 \pm .22$ | $3 * 10^{-4} \pm 10^{-4}$ |

Table 1: Mean and standard deviation of imitation loss of permuted policy learning on synthetic data

### 4.6 Experiments

To confirm the efficacy of Algorithm 1, we consider some simple experiments on a synthetic dataset. We consider Garnet MDPs (Bhatnagar et al., 2009), a model for generating small random environments, as well as a modified "planted" setting that specifically enforces concentration of the occupancy mass under the expert policy to a small set of states. We give complete details for these models in the Appendix.

The results are given in Table 1. Crucially, we observe this algorithm performs poorly if the underlying MDP is not friendly or has a large gap in the stationary values. Indeed, on the Garnet MDPs, the algorithm may fail if the sample complexity isn't sufficient to cover all state transitions. The planted MDPs ensure the desired properties and as expected yield much improved imitation loss.

## 5 ONLINE IMITATION

### 5.1 MDP Alignment

In the online setting, we're still seeking to imitate $\Phi$, or equivalently $\rho_\Phi$. However, we no longer observe trajectories of the correct policy $(I \otimes \Pi_*)^T \Phi \Pi_*$ played in $\hat{\mathcal{M}}$.

Instead, we are in a setting similar to a bandit, but without reward. At time $t$, we play a policy $\hat{\Phi}_t$ defined on $\hat{\mathcal{M}}$ and observe a transition. We allow resets to the initial distribution. After $T$ plays, where $T$ may be a random variable, we choose a final policy $\hat{\Phi}$ and receive instantaneous regret given by $G(\Phi, \Pi_*, \hat{\Phi})$.

One simple algorithm might treat each possible bijection as an arm, where pulling $\Pi$ is akin to running a trajectory using the policy $\Pi^T \Phi \Pi$, and then infer which alignment best matches the behavior policy. Or one could consider algorithms which don't play policies of the form $\Pi^T \Phi \Pi$ but simply explore the target space in a principled way.

Nevertheless, we derive a lower bound on the imitation loss of any algorithm in the online setting, demonstrating even complete knowledge of the source domain doesn't trivialize third-person imitation.

### 5.2 Lower Bound Counterexample

Consider a small bandit-like MDP (Figure 2a). Red corresponds to action $r$, blue corresponds to action $b$, and purple corresponds to both. The numbers on the edges give transition probabilities when taking the associated action. Let the initial distribution be $p_0(x_0) = p_0(y_0) = 1/2$. In other words, the initial state is either $x_0$ or $y_0$. Starting at $x_0$, the initial action is deterministic: playing $r$ leads to $x_1$, playing $b$ leads to $y_1$. Starting at $y$ the actions lead to the opposite states. Then the choice of action is irrelevant, and the transition to a terminal state is determined by $\alpha$ at $x_1$ and $\beta$ at $y_1$.

This characterizes $\mathcal{M}$. To introduce $\hat{\mathcal{M}}$, let's consider two possible bijections $\Pi_1$ and $\Pi_2$, which correspond to the possible target MDPs in Figure 2b and Figure 2c (note the values of $\alpha$ and $\beta$ are swapped given $\Pi_2$). These correspond to two possible dynamics on our target space. $\Pi_1$ is essentially the identity map, preserving states up to hats. Whereas $\Pi_2(\hat{x}_i) = y_i$ and $\Pi_2(\hat{y}_i) = x_i$.

Finally, suppose the behavior policy we want to imitate in $\mathcal{M}$ is defined by $\phi(r|x_0) = 1$ and $\phi(b|y_0) = 1$. In other words, the agent always travels in the first step to $x_1$. That means, under $\Pi_1$ we want to travel to $\hat{x}_1$, and under $\Pi_2$ we want to travel to $\hat{y}_1$. Intuitively, because $\rho_\Phi$ is highly asymmetric, but the MDP is nearly symmetric, one cannot choose a policy that performs well in multiple permutations of the MDP. We formalize this intuition below.

**Theorem 5.1.** *Choose any positive values $\epsilon < \epsilon_0$ and $\delta < \delta_0$, where $\epsilon_0$ and $\delta_0$ are universal constants, and let $\alpha = 1/2 + \epsilon$ and $\beta = 1/2 - \epsilon$. Consider any algorithm $\mathcal{A}$ that achieves $\gamma/4$-optimal imitation loss on the above MDP with probability at least $1 - \delta$. Then $E[T|\Pi_* = \Pi_i] = \Omega\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$ for some $i \in \{1, 2\}$.*

*Proof.* Fix a policy $\hat{\phi}$, and we will write $\rho_{\hat{\phi},\Pi}$ as simply $\rho_\Pi$.

Again use the variational form of total variation to say $TV(\rho_1, \rho_2) = \sup_{\|c\|_\infty \leq 1} \langle \rho_1 - \rho_2, c \rangle$. Choose $c$ so that $c(\hat{x}_i, a) = 1$ for $i \in \{1, 2, 3\}$ and $a \in A$, and 0 elsewhere. Then a direct calculation gives $G(\Phi, \Pi_1, \hat{\Phi}) = TV((I \otimes \Pi_1^T)\rho_\Phi, \rho_{\Pi_1}) \geq \gamma - \gamma(\hat{\phi}(r|\hat{x}_0) + \hat{\phi}(b|\hat{y}_0))/2$.

Now we proceed by a reduction to multi-armed bandits with known biases. Consider a two-armed bandit with Bernoulli rewards, where the hypotheses for arm biases are $H_1 = \{\alpha, \beta\}$ and $H_2 = \{\beta, \alpha\}$. We define the following algorithm $\mathcal{B}$ for the two-armed bandit. First run algorithm $\mathcal{A}$ on our MDP, where we couple pulls from arm 1 with transitions from $\hat{x}_1$ and pulls from arm 2 with transitions from $\hat{y}_1$. Call $\hat{\phi}$ the policy output by $\mathcal{A}$. Then output arm 1 if $\hat{\phi}(r|\hat{x}_0) > 1/2$, otherwise arm 2.

Under hypothesis $H_1$, $\Pi_* = \Pi_1$, so by our assumptions on $\mathcal{A}$, with probability at least $1 - \delta$ we have $\gamma/4 > \gamma - \gamma(\hat{\phi}(r|\hat{x}_0) + \hat{\phi}(b|\hat{y}_0))/2$, which implies $\hat{\phi}(r|\hat{x}_0) > 1/2$. Similar reasoning implies $\hat{\phi}(r|\hat{x}_0) \leq 1/2$ under $H_2$, hence $\mathcal{B}$ outputs the optimal arm with probability at least $1 - \delta$. Because $(\alpha, \beta) = (1/2 + \epsilon, 1/2 - \epsilon)$, and the sample complexity of $\mathcal{A}$ is lower bounded by that of $\mathcal{B}$, the result then follows from Theorem 13 in Mannor and Tsitsiklis (2004).

□

This bound illustrates why imitation is substantially more challenging than seeking high reward. In a regular RL problem with reward at the terminal states, if $\alpha \approx \beta$ then the expected reward changes very slightly depending on the policy. But in the imitation setting, the value of $\alpha$ and $\beta$ are essentially features of the states, which the agent must (very inefficiently) distinguish in order to achieve error lower than $\gamma/4$. Likewise, this counterexample captures why the online setting is the more challenging one studied in this work. In the offline regime, an oracle would only visit states on one half of the MDP and easily break the symmetry.

One may attribute this pessimistic bound to the choice of total variation distance. Indeed, among IPMs, total variation has very poor generalization properties (Sun et al., 2019). However, an alternative choice of IPM corresponds to a non-uniform prior over reward functions that the behavior policy is truly optimizing. If the prior strongly favored reward functions that smoothly depend on the local dynamics, then $c(\hat{x}_i, a) \approx c(\hat{y}_i, a)$ and this counterexample would no longer hold. But this is a somewhat unnatural assumption, precluding for example a 2D gridworld with positive reward only at one state (since the gridworld would have many symmetries).

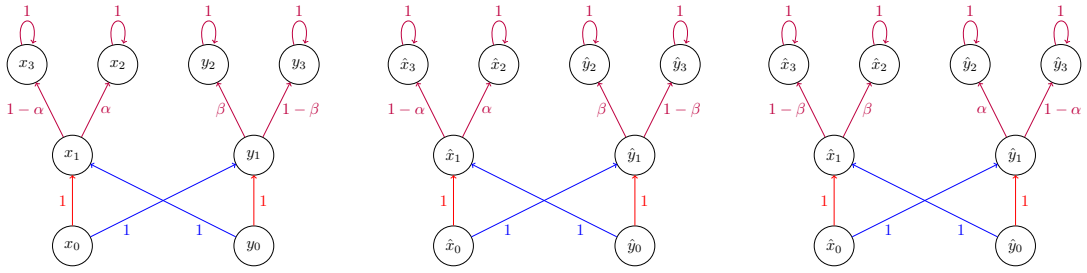



Figure 2: The bandit-like MDP, where (a) is the source domain, (b) is the target domain given $\Pi_1$ and (c) is the target domain given $\Pi_2$.

## 6 CONCLUSION AND FUTURE WORK

In this paper, we introduced a theoretical analysis of third-person imitation, as an initial step in more fully understanding generalization in RL. We demonstrated upper bounds for imitation learning across isomorphic domains under offline and state-only assumptions, and a lower bound for the online setting. These bounds depend heavily on the structural properties of the dynamics and behavior policy, as well as the setting of third-person imitation where the domain adaptation is across isomorphic environments.

The upper bound dependence on structural and spectral properties is likely not optimal, although the dependence on $|S|$ in the error likely cannot be improved. The lower bound is somewhat more robust, and any MDP with symmetry such that this bandit-like MDP can be embedded will suffer a similar lower bound on sample complexity.

The isomorphism assumption is certainly too strict in general. However, weakening the assumption requires a characterization of MDP similarity, in order to decide when one should expect policy transfer through imitation to be feasible. MDPs with features (Krishnamurthy et al., 2016) could better characterize similarity, where the spectral features studied in this work could be combined with observed state features for more effective alignment through linear assignment. Future work may include studying third-person imitation in the online setting for upper bounds, or exploiting MDP asymmetry in deep imitation.

### Acknowledgements

We are extremely grateful to David Brandfonbrener, Min Jae Song, and Raghav Singhal, who gave feedback and insightful suggestions throughout the work.

This work partially supported by the Alfred P. Sloan Foundation, NSF RI-1816753, NSF CAREER CIF 1845360, NSF CHS-1901091, Samsung Electronics, and the Institute for Advanced Study.

### References

Yonathan Aflalo, Alexander Bronstein, and Ron Kimmel. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences*, 112 (10):2942–2947, 2015.

Haitham Bou Ammar, Eric Eaton, Paul Ruvolo, and Matthew E Taylor. Unsupervised cross-domain transfer in policy gradient reinforcement learning via manifold alignment. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Michael Bain. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129, 1995.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Yan Duan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning. In *Advances in neural information processing systems*, pages 1087–1098, 2017.

Shani Gamrian and Yoav Goldberg. Transfer learning for related reinforcement learning tasks via image-to-image translation. In *International Conference on Machine Learning*, pages 2063–2072, 2019.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1480–1490, 2017.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.

Alan J Hoffman and Helmut W Wielandt. The variation of the spectrum of a normal matrix. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 118–120. World Scientific, 2003.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

Kun Ho Kim, Yihong Gu, Jiaming Song, Shengjia Zhao, and Stefano Ermon. Cross domain imitation learning. *arXiv preprint arXiv:1910.00105*, 2019.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955.

YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1118–1125. IEEE, 2018.

Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun): 623–648, 2004.

Peter Pastor, Heiko Hoffmann, Tamim Asfour, and Stefan Schaal. Learning and generalization of motor skills by learning from demonstration. In *IEEE International Conference on Robotics and Automation*, pages 763–768, 2009.

Daniel Paulin et al. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20, 2015.

Martin L Puterman. Markov decision processes: Discrete stochastic dynamic programming, 1994.

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

Bradly C Stadie, Pieter Abbeel, and Ilya Sutskever. Third-person imitation learning. *arXiv preprint arXiv:1703.01703*, 2017.

Wen Sun, Anirudh Vemula, Byron Boots, and Drew Bagnell. Provably efficient imitation learning from observation alone. In *International Conference on Machine Learning*, pages 6036–6045, 2019.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30, 2017.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4950–4957, 2018a.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018b.

Shinji Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE transactions on pattern analysis and machine intelligence*, 10 (5):695–703, 1988.

Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. In *Conference on Learning Theory*, pages 3120–3159, 2019a.

Geoffrey Wolfer and Aryeh Kontorovich. Minimax learning of ergodic markov chains. In *Algorithmic Learning Theory*, pages 904–930, 2019b.

Chao Yang, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Huaping Liu, Junzhou Huang, and Chuang Gan. Imitation learning from observations by minimizing inverse dynamics disagreement. In *Advances in Neural Information Processing Systems*, pages 239–249, 2019.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

# A Proof of Theorem 4.12

We define $D_{p_0} = \sum_i (p_0)_i^2/\mu_i$, where we interpret $0/0 = 0$. Note that $p_0$ is absolutely continuous with respect to $\mu$, so this term is well-defined.

We now state the necessary concentration results. Note the theorems are slightly altered from their statements in the literature, but follow immediately from their original proofs. The first has a better dependence on $|S|$ by considering $L_2$ norm rather than $L_1$, and with slightly loose sample complexity. The second is exactly an intermediate statement made in the theorem's original proof.

**Theorem A.1** (Theorem 1 in Wolfer and Kontorovich (2019b))**.** *If* $m = O\left(\frac{1}{\gamma_{ps}\epsilon^2\mu_i}\log\left(\frac{|S|\sqrt{D_{p_0}}}{\delta}\right)\right)$ *then with probability at least* $1-\delta/2$, $\|M(i,\cdot) - \tilde{M}(i,\cdot)\|_2 \le \epsilon$.

**Theorem A.2** (Theorem 5.1 in Wolfer and Kontorovich (2019a))**.** *If* $m = O\left(\frac{1}{\gamma_{ps}\epsilon^2\mu_i}\log\left(\frac{|S|\sqrt{D_{p_0}}}{\delta}\right)\right)$ *then with probability at least* $1-\delta/2$, $|\mu_i - \tilde{\mu}_i| \le \epsilon\mu_i$.

We observe a simple consequence of the definition of the rescaled transition matrix:

**Proposition A.3.** *For an ergodic Markov chain* $M$ *with rescaled transition matrix* $L$, $\sigma_1(L) = 1$ *and* $\gamma_{ps}(M) \ge 1 - \sigma_2(L)^2$.

*Proof.* Choosing $k = 1$ in the definition of $\gamma_{ps}$ gives the product $D^{-1}M^TDM = D^{-1/2}L^TLD^{1/2}$. The first term itself is a Markov chain called the multiplicative reversiblization (Paulin et al., 2015). Because the chain has maximum eigenvalue 1 and the eigenvalues of $L^TL$ are the squares of the singular values, it follows $\sigma_1(L) = 1$ and $\gamma_{ps} \ge 1 - \sigma_2(L)^2$. □

We set the occupancy threshold via $t$, and consider properties of the empirical estimators:

**Lemma A.4.** *Let* $g := \min_i |\mu_i - t|$. *Assume* $g > 0$, $t > 0$, *and* $1/4 > \epsilon > 0$. *If* $m = O\left(\max\left(\frac{1}{\epsilon^2 t}, \frac{1}{g^2}\right)\frac{1}{\gamma_{ps}}\log\left(\frac{|S|\sqrt{D_{p_0}}}{\delta}\right)\right)$, *then with probability at least* $1-\delta$, *we have the following:*

*(a)* $\pi_*^{-1}(I_t) = \hat{I}_t$

*(b)* $\hat{D}_t^{-1}$ *is well-defined*

*(c)* $\|E_t\|_F \le \epsilon\sqrt{|I_t|}$

*(d)* $\|f_+(\Delta_t)^{1/2}\|_F \le \sqrt{\epsilon}$ *where* $f_+(\cdot)$ *is the elementwise absolute value.*

*(e)* $\|(D_t + \Delta_t)^{-1/2} - D_t^{-1/2}\|_F \le 2\sqrt{\frac{\epsilon|I_t|}{t}}$

*Proof.* Note that for all $i \in I_t$, $\mu_i \ge t$. So choosing precision $\epsilon$ and confidence $\frac{\delta}{2|I_t|}$ in Theorem A.1 and Theorem A.2, taking a union bound over all $i \in I_t$, and noting $|I_t| \le |S|$, we have that when $m = O\left(\frac{1}{\gamma_{ps}\epsilon^2 t}\log\left(\frac{|S|\sqrt{D_{p_0}}}{\delta}\right)\right)$, with probability at least $1-\delta/2$, $\|M(i,\cdot) - \tilde{M}(i,\cdot)\|_2 \le \epsilon$ and $|\mu_i - \tilde{\mu}_i| \le \epsilon\mu_i$.

Additionally, choosing precision $\frac{g}{2\mu_i}$ and confidence $\frac{\delta}{2|S|}$ in Theorem A.2, and taking a union bound over all $i \in [|S|]$, when $m = O\left(\frac{1}{\gamma_{ps}g^2}\log\left(\frac{|S|\sqrt{D_{p_0}}}{\delta}\right)\right)$, with probability at least $1-\delta/2$ we have $|\mu_i - \tilde{\mu}_i| \le g/2$.

By the second application of the concentration results, for all $i \in [|S|]$, $|\mu_i - \hat{\mu}_{\pi_*^{-1}(i)}| = |\mu_i - \tilde{\mu}_i| \le g/2 < g$. So from the definition of $g$ it's clear that $\mu_i \ge t$ iff $\hat{\mu}_{\pi_*^{-1}}(i) \ge t$. Hence, $i \in I_t$ iff $\pi_*^{-1}(i) \in \hat{I}_t$.

If $i \in I_t$, $\mu_i \ge t$. Hence $\hat{\mu}_{\pi_*^{-1}(i)} \ge \mu_i - g/2 > \mu_i - g \ge t > 0$. This means each diagonal element of $\hat{D}_t$ is positive, hence it's invertible.

By part (a), if we define $\Pi_{t^*}$ to be the restriction of $\Pi_*$ to the indices $I_t$ and $\hat{I}_t$, then $\Pi_{t^*}$ is still a permutation matrix. Furthermore, $E_t = (\Pi_* \hat{M} \Pi_*^T - M)_t = \Pi_{t^*} \hat{M}_t \Pi_{t^*}^T - M_t = \tilde{M}_t - M_t$.

Then $\|E_t\|_F^2 = \sum_{i \in I_t} \|\tilde{M}_t(i,\cdot) - M_t(i,\cdot)\|_2^2 \le \sum_{i \in I_t} \|\tilde{M}(i,\cdot) - M(i,\cdot)\|_2^2 \le \epsilon^2 |I_t|$.

Similarly, $\|f_+(\Delta_t)^{1/2}\|_F^2 = \sum_{i \in I_t} |\mu_i - \tilde{\mu}_i| \le \sum_{i \in I_t} \epsilon \mu_i \le \epsilon$.

To derive the last inequality, note that $|\mu_i - \tilde{\mu}_i| \le \epsilon \mu_i$ implies $(1-\epsilon)\mu_i \le \tilde{\mu}_i \le (1+\epsilon)\mu_i$. Therefore

$$\begin{aligned}
\|(D_t + \Delta_t)^{-1/2} - D_t^{-1/2}\|_F^2 &= \sum_{i \in I_t} \left( \frac{1}{\sqrt{\tilde{\mu}_i}} - \frac{1}{\sqrt{\mu_i}} \right)^2 \\
&= \sum_{i \in I_t} \frac{\tilde{\mu}_i + \mu_i - 2\sqrt{\tilde{\mu}_i \mu_i}}{\tilde{\mu}_i \mu_i} \\
&\le \sum_{i \in I_t} \frac{(1+\epsilon)\mu_i + \mu_i - 2\sqrt{(1-\epsilon)\mu_i \mu_i}}{(1-\epsilon)\mu_i \mu_i} \\
&\le \frac{|I_t|}{t} * \frac{2 + \epsilon - 2\sqrt{1-\epsilon}}{1-\epsilon} \\
&\le \frac{|I_t|}{t} * \frac{3\epsilon}{1-\epsilon} \\
&\le \frac{4\epsilon |I_t|}{t}
\end{aligned}$$

where the second last inequality uses $\sqrt{1-\epsilon} \ge 1 - \epsilon$ for $1 > \epsilon > 0$.

□

**Lemma A.5.** *If $L_t$ is $(\alpha, \beta)$-friendly for sufficiently large $\alpha$ and $\beta$, the matrix $\hat{L}_t$ is friendly if it is well-defined.*

*Proof.* Observe that $\tilde{L}_t := \Pi_{t^*} \hat{L}_t \Pi_{t^*}^T = (D_t + \Delta_t)^{1/2}(M_t + E_t)(D_t + \Delta_t)^{-1/2}$, so it suffices to show this matrix is friendly.

We need the following bounds, utilizing the inequality $\sqrt{a+b} - \sqrt{a} \le \sqrt{|b|}$:

$$\begin{aligned}
\|(D_t + \Delta_t)^{1/2} - D_t^{1/2}\|_F \le \|f_+(\Delta_t)^{1/2}\|_F &\le \sqrt{\epsilon} \\
\|D_t^{-1/2}\|_F &\le \sqrt{\frac{|I_t|}{t}} \\
\|(D_t + \Delta_t)^{1/2}\|_F \le \|D_t^{1/2}\|_F + \|f_+(\Delta_t)^{1/2}\|_F &\le 1 + \sqrt{\epsilon} \\
\|M_t\|_F &\le \sqrt{|I_t|} \\
\|M_t + E_t\|_F &\le (1+\epsilon)\sqrt{|I_t|}
\end{aligned}$$

Decompose the perturbation of $L_t$ as

$$\begin{aligned}
\tilde{L}_t - L_t = &(D_t + \Delta_t)^{1/2}(M_t + E_T)((D_t + \Delta_t)^{-1/2} - D_t^{-1/2}) \\
&+ (D_t + \Delta_t)^{1/2}(M_t + E_t - M_t) D_t^{-1/2} \\
&+ ((D_t + \Delta_t)^{1/2} - D_t^{1/2}) M_t D_t^{-1/2}
\end{aligned}$$

Then we apply the inequalities above, using the triangle inequality and submultiplicativity to obtain

$$\begin{aligned}\|\tilde{L}_t - L_t\|_F &\leq \frac{16\sqrt{\epsilon}|I_t|}{\sqrt{t}} + \frac{2\epsilon|I_t|}{\sqrt{t}} + \frac{\sqrt{\epsilon}|I_t|}{\sqrt{t}} \\ &\leq \frac{19\sqrt{\epsilon}|I_t|}{\sqrt{t}}\end{aligned}$$

Now decompose $L_t = U\Sigma V^T$ and $\tilde{L}_t = \tilde{U}\tilde{\Sigma}\tilde{V}^T$.

By the Wielandt-Hoffman inequality (Hoffman and Wielandt, 2003), $\sum_i(\sigma_i(\tilde{L}_t)-\sigma_i(L_t))^2 \leq \|\tilde{L}_t - L_t\|_F^2$. Therefore, if $\alpha = \min_i \sigma_i(L_t) - \sigma_{i+1}(L_t) > 2\|\tilde{L}_t - L_t\|_F$, then $\sigma_i(\tilde{L}_t) - \sigma_{i+1}(\tilde{L}_t) > 0$.

By the Cauchy interlacing theorem and Propostion A.3, $\sigma_1(L_t) \leq \sigma_1(L) = 1$. And $\min_i \sigma_i(L_t)^2 - \sigma_{i+1}(L_t)^2 \geq \alpha^2$.

Therefore, we can apply the Davis-Kahn theorem (Yu et al., 2015) to conclude $1 - |\tilde{v}_i^T v_i| \leq \zeta$ where

$$\zeta := \left( \frac{2\left(2 + \frac{19\sqrt{\epsilon}|I_t|}{\sqrt{t}}\right)\frac{19\sqrt{\epsilon}|I_t|}{\sqrt{t}}}{\alpha^2} \right)$$

Orienting $\tilde{V}$ so that $\tilde{V}^T\mathbf{1} \geq 0$, it follows $|\tilde{v}_i^T v_i| = \tilde{v}_i^T v_i$.

If $\beta > \sqrt{2|I_t|\zeta}$, then the friendliness assumption implies $v_i^T\mathbf{1} > \sqrt{2|I_t|\zeta}$ and therefore

$$\begin{aligned}\tilde{v}_i^T\mathbf{1} &\geq v_i^T\mathbf{1} - |v_i^T\mathbf{1} - \tilde{v}_i^T\mathbf{1}| \\ &> \sqrt{2|I_t|\zeta} - \|\mathbf{1}\|_2\|v_i - \tilde{v}_i\|_2 \\ &= \sqrt{2|I_t|\zeta} - \sqrt{|I_t|}\sqrt{1 + 1 - 2\tilde{v}_i^T v} \\ &> \sqrt{2|I_t|\zeta} - \sqrt{2|I_t|\zeta} \\ &> 0\end{aligned}$$

□

Now we can recover $\Pi_{t^*}$, using the decomposition $\hat{L}_t = \hat{U}\tilde{\Sigma}\hat{V}^T$.

**Lemma A.6.** *Under the same assumptions as Lemma A.5, if $|I_t|\zeta < \frac{1}{2}$, the unique permutation matrix $\Pi$ such that $\|\Pi^T V - \hat{V}\|_F \leq \sqrt{2|I_t|\zeta}$ is $\Pi_{t^*}$.*

*Proof.* By Proposition 4.6, and the friendliness of $\hat{L}_t$ and $\tilde{L}_t$, we have that $\Pi_{t^*}^T\tilde{V} = \hat{V}$. It follows that

$$\begin{aligned}\|\Pi_{t^*}^T V - \hat{V}\|_F &\leq \|\Pi_{t^*}^T V - \Pi_{t^*}^T\tilde{V}\|_F + \|\Pi_{t^*}^T\tilde{V} - \hat{V}\|_F \\ &= \|V - \tilde{V}\|_F\end{aligned}$$

And note that $\|V - \tilde{V}\|_F^2 = \sum_i \|v_i - \tilde{v}_i\|_2^2 = \sum_i 2 - 2v_i^T\tilde{v}_i \leq 2|I_t|\zeta$.

Conversely,

$$\begin{aligned}\|\Pi - \Pi_{t^*}\|_F &= \|\Pi^T\tilde{V} - \Pi_{t^*}^T\tilde{V}\|_F \\ &\leq \|\Pi^T\tilde{V} - \Pi^T V\|_F + \|\Pi^T V - \hat{V}\|_F + \|\hat{V} - \Pi_{t^*}^T\tilde{V}\|_F \\ &\leq \sqrt{2|I_t|\zeta} + \sqrt{2|I_t|\zeta} \\ &< \sqrt{2}\end{aligned}$$

Because distinct permutation matrices differ in Frobenius norm by at least $\sqrt{2}$, this guarantees $\Pi = \Pi_{t^*}$. □

*Proof of Theorem 4.12.* For a given $t$, suppose $D_t^{1/2} M_t D_t^{-1/2}$ is $(\alpha, \beta)$-friendly. Then we choose $\epsilon$ to satisfy the following:

1. $\alpha > 2\|\tilde{L}_t - L_t\|_F$
2. $\beta > \sqrt{2|I_t|\zeta}$
3. $|I_t|\zeta < 1/2$

We observe these are all satisfied at $\sqrt{\epsilon} = O\left(\frac{\alpha^2\beta^2\sqrt{t}}{|I_t|^2}\right)$

If $m = poly\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{t}, |I_t|, \frac{1}{g}, \frac{1}{\gamma_{ps}(M)}, \log D_{p_0}, \log|S|, \log\frac{1}{\delta}\right)$, then Lemma A.4 and A.5 imply the estimator $\hat{L}_t$ is friendly with probability at least $1-\delta$. So by Lemma A.6, we conclude the permutation $\Pi'$ recovered from the Hungarian algorithm in Algorithm 1 agrees with $\Pi_*$ on $I_t$ and $\hat{I}_t$. Finally, Theorem 4.8 bounds the imitation objective.

Lastly, we rewrite the sample complexity, using the fact that $D_{p_0} \leq \frac{1}{1-\gamma}$ from the definition of $\mu$. We also note that from Proposition A.3, in the exact recovery setting $\min_i \mu_i > t$, we may replace $\gamma_{ps}$ in the sample complexity with $1 - \sigma_2(L_t)^2$.

□

# B Experimental Details

For all trials we consider an MDP with $|S| = 100$, $|A| = 5$, $\gamma = 0.95$. To sample Garnet MDPs (Bhatnagar et al., 2009) we set the branching parameter, which dictates the support of the next-state distribution at each state-action pair, as $b = 5$. Given the support, the next-state distribution is uniform over the $b$ states.

The planted MDPs are determined by setting 5 "alive" states which are chosen to have almost all the occupancy mass. Explicitly, we first sample a deterministic policy $\phi$ to be the expert policy. The initial distribution is restricted to the alive states. For sampling the dynamics distributions $P(\cdot|s,a)$, if $\phi(a|s) = 0$, we sample according to the Garnet MDP branching. If $\phi(a|s) = 1$, we sample a distribution $b$ alive states with $1-\epsilon$ total mass, and a distribution on $b$ "dead" states with $\epsilon$ total mass, where $\epsilon = 0.0001$.