Transitions Decompose as Passive and Agent Dynamics

Massimiliano Tamborski[1], **Carl Henrik Ek**[1], **David Abel**[2]
massimiliano.tamborski@cst.cam.ac.uk

[1]**University of Cambridge**
[2]**University of Edinburgh**

Abstract

In reinforcement learning (RL), the Markov decision process (MDP) is taken as the canonical definition of an environment. However, the MDP's transition function blends the inherent dynamics of the world with the influence of the agent's actions. To isolate this influence, we propose linearly decomposing the transition function into two components: (1) a passive environment probability and (2) an active deviation induced by the agent's actions. This decomposition reframes the MDP as the addition of a Markov chain and an agent-induced deviation. Specifically, we prove the universality of this additive decomposition, demonstrate the limitations of prior multiplicative approaches, and introduce an action influence lemma that bounds an agent's impact on expected returns based on its deviation from the passive dynamics.

1 Introduction

In reinforcement learning (RL), the Markov decision process (MDP; Bellman, 1957; Puterman, 2014) serves as the canonical formalism for modelling a sequential decision-making problem under full observability (Sutton & Barto, 2018). Central to the MDP is the transition function, $p(s' \mid s, a)$, which defines the probability of transitioning to state s' given the agent takes action a in state s. Conventionally, this function is treated as a monolithic entity intrinsic to the environment. That is, the choice of MDP (and thus, transition function) can be paired with any choice of agent, as they roughly correspond to the physical laws that govern the world of interest.

Transitions Involve Agents and Environments. However, this treatment of state transitions obscures a fundamental aspect of the interaction of agent and environment: the evolution of the state is a consequence of both the passive dynamics of the world and the active interventions of the agent. An ant, for instance, exerts minimal control over the weather of its biome. Here, a singular ant in the colony can take an action that influences its local niche—it might eat a delicious bit of fruit, or signal to its companion as to the fruit's whereabouts. The weather, by contrast, is best thought of as part of an unfolding background process that we might refer to as the *passive dynamics* of the environment, as suggested by Todorov (2006). In this way, the overall dynamics governing the ant's experience are due to both the passive dynamics and the actions taken by the ant (and its companions). Crucially, the passive dynamics of the world are still important to the ant: the weather can influence what the ant should do, despite being outside the ant's control.

In this way, the transition function of any MDP can be well thought of as combining (and obfuscating) the influence of both the environment *and* the agent. This perspective was first suggested by Todorov (2006), who proposed decomposing the transition function of an MDP into passive dynamics that evolve independently of an agent, alongside an *active component* where an agent's action modifies these passive probabilities. A related perspective developed by Dietterich et al. (2018) and expanded upon by Chitnis & Lozano-Pérez (2020) and Trimponias & Dietterich (2023) involves factoring

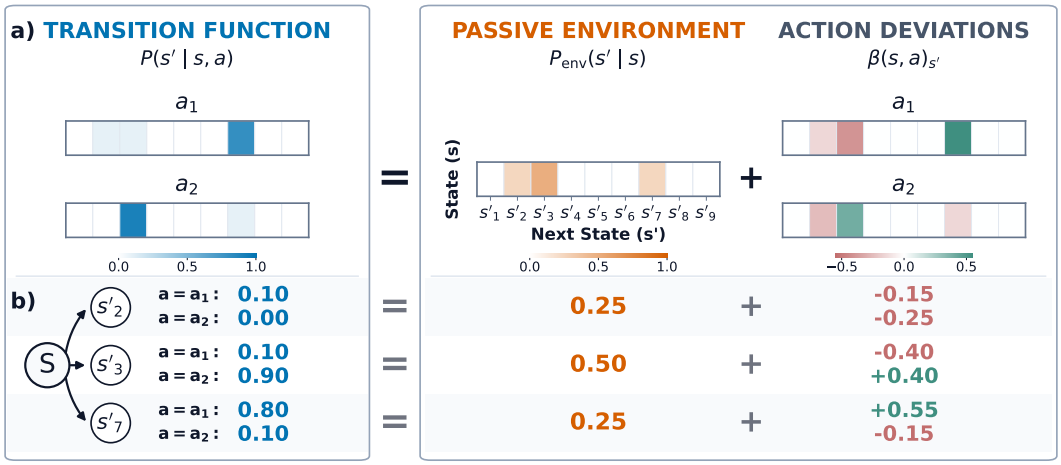

Figure 1: The transition function $p(s' \mid s, a)$ is decomposed into a passive environment probability $p_{\text{env}}(s' \mid s)$ and an action-induced deviation $\beta(s, a)_{s'}$. **(a) Top Row:** Visualisation of this decomposition for two actions a_1 and a_2. The horizontal bars represent transitions from an arbitrary state s to a set of possible next states $\{s'_1, \ldots, s'_9\}$. Cell shading indicates the probability or deviation magnitude. Action-induced deviations (red/green) are added to the passive dynamics (orange) to produce the transition probabilities (blue) **(b) Bottom Row:** A tabular representation of the same transitions, focusing only on the next states with non-zero values: $\{s'_2, s'_3, s'_7\}$. Values were generated using the FrozenLake environment (Towers et al., 2026) but are presented abstractly.

the state space into controllable (*endogenous*) and uncontrollable (*exogenous*) components. Recent advancements in this perspective have identified methods for agents to learn these partitions directly without *a priori* knowledge, and even accommodate partitions that change depending on the state (Rodriguez-Sanchez et al., 2026). While these works support the claim that transition functions combine influence from both agent and environment, they either require the state space to have a specific underlying structure or can only represent a subset of possible worlds.

The Additive View of Transition Decomposition. We here explore a general framing question central to RL: what is the right way to model the transition function of a stochastic process, given that the process is influenced by both the agent and environment? We here offer one answer to this question that explicitly separates the passive dynamics of the environment from the active influence of the agent's actions, drawing inspiration from Todorov (2006). Our proposal, as pictured in Figure 1, is simple: the next state of the environment can be described by (1) the passive effect of the environment, $p_{\text{env}}(s' \mid s)$, added together with (2) the deviation effect of the agent's action targeting that same state, $\beta(s, a)_{s'}$. The result is an additive view of the transition function of an MDP, where, for any state pair (s, s') and action a, the probability of transitioning to the next state s' is given by:

$$p(s' \mid s, a) = p_{\text{env}}(s' \mid s) + \beta(s, a)_{s'}. \tag{1}$$

This perspective aligns with classic optimal control theory. Kalman (1960) formalised linear dynamical systems as $dx/dt = F(t)x + G(t)u(t)$, explicitly separating the system's passive dynamics, $F(t)x$, from an additive modification scaled by the control input, $u(t)$. The proposed decomposition also shares the same spirit as residual learning (Silver et al., 2018). While residual learning is often applied at the policy level, it has also been applied to the transition function (Asri et al., 2024), where it has shown empirical advantages.

In this work, we prove that this decomposition is universal: every MDP can be equivalently defined according to this additive perspective (Theorem 3.3). We take this as a supporting criterion for any theory of decomposable transition functions, but speculate that the additive perspective we offer is but one point in a space of choices that accommodate this property. We further establish that

the decomposition from Todorov, while elegant, is not universal in the above sense (Theorem 3.4). We then draw out a number of salient consequences of the framing: there is a one-to-many map from MDPs to this decomposition (Proposition 3.6), foundational results such as the simulation lemma (Kearns & Singh, 2002) can be framed in terms of the agent's deviation (Lemma 3.8), and the framework can be generalised to multi-agent environments (Proposition 3.10). We suggest that viewing the transition function as the combined influence of agent and environment represents a compelling perspective for RL, with implications ranging from empowerment (Klyubin et al., 2005), to learning world models (Kearns & Singh, 2002), and agent-centric RL (Abel et al., 2024).

2 Preliminaries

Notation. We let calligraphic capital letters $\mathcal{X}, \mathcal{Y}$ denote sets, and let lowercase letters f, g, π denote functions on these sets. We let $\Delta(\mathcal{X})$ express the probability simplex on the set $\mathcal{X}$.

To focus on the transition function, we first abstract away rewards and discount factors. Thus, for simplicity and without loss of generality, we primarily study controlled Markov processes (CMPs; Puterman, 2014) rather than MDPs.

Definition 2.1 (Controlled Markov Process)**.** *A controlled Markov process (CMP) is a Markov decision process (MDP) without a specified reward function or discount factor. It is defined by the tuple $(\mathcal{S}, \mathcal{A}, p)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, and $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function.*

An agent interacts with the CMP by following a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which maps each state to a probability distribution over actions. To formalise our decomposition, we introduce two key concepts.

Definition 2.2. *A **passive environment** is a function $p_{\text{env}} : \mathcal{S} \to \Delta(\mathcal{S})$ that governs the transition dynamics of the world.*

Definition 2.3. *A **deviation** is a function $\beta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{S}|}$ that maps a state and action to a valid deviation vector over the state space.*

The passive environment outputs a distribution over subsequent states without knowing the agent's action, which then modifies these probabilities via the deviation function. While the passive environment p_{env} *could* equal the marginalisation over actions $p(s' \mid s) = \sum_a \pi(a \mid s) p(s' \mid s, a)$, it can be an arbitrary probability function. The deviation function β maps a state-action pair (s, a) to a vector $\beta(s, a)$ containing the probability deviations across all next states.

3 Additive Decomposition of Transition Functions

We propose linearly decomposing the transition function to isolate agent-driven changes from the passive dynamics. In this section, we formalise this decomposition, prove its universality compared to prior approaches, and explore its theoretical consequences.

3.1 Mathematical Formulation

We propose that the transition function of any CMP can be conceptually treated as the combination of a passive Markov chain and a deviation induced by the agent's actions.

Definition 3.1 (Transition Function Decomposition)**.** *The transition function $p(s' \mid s, a)$ of all CMPs with a finite state space $\mathcal{S}$ and action space $\mathcal{A}$ can be decomposed into a passive environment function p_{env} (Definition 2.2) and a deviation function β (Definition 2.3). Formally, there exists a pair (p_{env}, β) such that for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$:*

$$p(s' \mid s, a) = p_{\text{env}}(s' \mid s) + \beta(s, a)_{s'} \tag{2}$$

where $\beta(s, a)_{s'}$ is the entry of the deviation vector corresponding to the transition to state s'.

Remark 3.2. *While it is well known that fixing a policy reduces a CMP to a Markov chain, Definition 3.1 establishes that a CMP can be decomposed into a Markov chain and an action-induced deviation* even *in the absence of a policy.*

Because both p and p_{env} are valid probability distributions over $\mathcal{S}$, the deviations redistribute probability mass. Consequently, the entries of the deviation vector from any state s under action a must sum to zero:

$$\sum_{s' \in \mathcal{S}} \beta(s,a)_{s'} = 0, \quad \forall s \in \mathcal{S},\, a \in \mathcal{A}. \tag{3}$$

Conceptually, the deviation function cannot make a specific future state more probable without making other states less probable. To maintain a valid probability distribution, the deviation values are bounded by the passive dynamics:

$$-p_{\text{env}}(s' \mid s) \leq \beta(s,a)_{s'} \leq 1 - p_{\text{env}}(s' \mid s), \quad \forall s, s' \in \mathcal{S},\, a \in \mathcal{A}. \tag{4}$$

3.2 Universality and the Limits of Todorov's Framework

With these properties established, we can prove that this additive decomposition is universal.

Theorem 3.3 (Universality of Definition 3.1)**.** *The additive decomposition is universal: the transition function of any valid CMP can be expressed as in Equation 2.*

We defer all proofs to Section 5. We contrast our additive formulation with the framework proposed by Todorov (2006). While our approach treats actions abstractly as $a \in \mathcal{A}$, Todorov defines actions as control vectors $u \in \mathbb{R}^{|\mathcal{S}|}$ that directly scale the underlying Markov chain. To highlight this distinction, we adopt the notation $u(s)$, where $u(s)_{s'}$ denotes the entry specifying the control for transitioning to state s'. Under this formulation, the transition dynamics are defined as:

$$p(s' \mid s, u) = p_{\text{env}}(s' \mid s) \exp(u(s)_{s'}). \tag{5}$$

We demonstrate that this multiplicative assumption does not hold universally.

Theorem 3.4 (Non-universality of Todorov's Framework)**.** *The decomposition of passive and active dynamics proposed by Todorov (2006) is not universal. There exists at least one CMP that cannot be reconstructed for any choice of p_{env} and control vector u using this framework.*

While Todorov's framework lacks universality, it remains related to our decomposition.

Corollary 3.5. *The multiplicative decomposition proposed by Todorov (2006) is a special case of the additive decomposition in Equation 2, limited to environments where $p_{\text{env}}(s' \mid s) = 0$ implies $p(s' \mid s, u) = 0$, and where the action space is defined directly over log-probability deviations.*

3.3 Uniqueness and the Deviation Space

While any CMP can be decomposed into a (p_{env}, β) pair, the specific choice of values is not unique.

Proposition 3.6 (Non-uniqueness of a CMP)**.** *For any CMP with $|\mathcal{S}| \geq 2$, there are infinitely many combinations of p_{env} and β that give rise to the same CMP.*

Although infinitely many valid pairs exist for a given CMP, the two components are dependent on one another. Choosing a passive environment p_{env} always uniquely defines a valid deviation function via $\beta(s,a)_{s'} = p(s' \mid s, a) - p_{\text{env}}(s' \mid s)$. Conversely, an arbitrarily chosen β only defines a valid p_{env} if the difference $p(s' \mid s, a) - \beta(s,a)_{s'}$ evaluates to the exact same value for all actions $a \in \mathcal{A}$.

This perspective provides a clean separation between the dynamics of a world and the specific capabilities of an agent operating within it. If, for example, we assume we have a specific passive environment p_{env} representing what would happen in the agent's absence, governed solely by the physical laws of the environment, different agents or action spaces simply represent different valid

deviations from that baseline. To formalise this influence, we distinguish between the deviations that the *environment* allows the agent to make and the deviations that a specific *agent* exhibits.

If we only fix a specific passive environment p_{env}, but not a CMP, there remains an infinite set of possible deviation vectors that can be added to it to result in valid transition probabilities. We define this landscape of valid options as the **allowable deviation space**, $\mathcal{D}_{\text{env}}(s) \subseteq \mathbb{R}^{|\mathcal{S}|}$. All vectors within this space must satisfy the bounds in Equation 4 (a formal characterisation is provided in Section 5.2).

Under this framing, the deviation function β serves as the mechanism that selects a specific valid vector from this space of deviations for each state-action pair; formally, $\beta(s,a) \in \mathcal{D}_{\text{env}}(s)$. By choosing different β functions, we can construct infinitely many distinct CMPs from the same p_{env}. Two different agents might operate within the exact same passive environment, but their differing capabilities translate into distinct β functions, and thus different deviation vectors within $\mathcal{D}_{\text{env}}(s)$. For example, if we place two agents on a frozen lake, both might select the action 'step forward'; however, one may slip, whilst the other (perhaps equipped with ice cleats) successfully advances. Unlike Todorov's formulation, where the agent *directly* scales the passive probabilities, we take the view that the environment establishes a space of valid interventions, and the agent's actions map to this space via β. The actions $a \in \mathcal{A}$ act as abstractions of the deviation induced by the β function.

3.4 The Action Influence Lemma

Up to this point, we have restricted our analysis to CMPs to focus on the transition function. However, to analyse how an agent's deviation function affects its performance, we reintroduce the reward signal. We therefore extend our focus to the MDP.

Definition 3.7 (Markov Decision Process)**.** *A Markov decision process (MDP) is defined by the tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$, which augments a CMP with a reward function r and a discount factor $\gamma \in [0,1)$.*

In this section, we will restrict our focus to a state-dependent reward function, $r : \mathcal{S} \to \mathbb{R}$. Under the transition function p, the expected discounted return of a policy π starting from an initial state distribution μ is $V^{\pi}(\mu) = \mathbb{E}_{\pi,p}[\sum_{t=0}^{\infty} \gamma^t r(s_t) \mid s_0 \sim \mu]$. Similarly, the expected discounted return under the passive environment dynamics p_{env} alone is $V^{\text{env}}(\mu) = \mathbb{E}_{p_{\text{env}}}[\sum_{t=0}^{\infty} \gamma^t r(s_t) \mid s_0 \sim \mu]$.

By extending our additive decomposition (Definition 3.1) to this setting, we can explicitly bound how an agent's deviation from the passive environment impacts its expected return.

Lemma 3.8 (Action Influence Lemma[1])**.** *Let $\epsilon(\pi, \beta)$ denote the maximum deviation when transitioning from s to s' via a stationary policy π,*

$$\epsilon(\pi,\beta) = \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \left| \sum_{a \in \mathcal{A}} \pi(a \mid s) \beta(s,a)_{s'} \right| \tag{6}$$

Given an initial state distribution μ and a state-dependent reward function r with bounded norm $\|r\|_{\infty} = R_{\max}$, the difference between the expected return under policy π and that under the passive environment p_{env}, denoted $V^{\pi}(\mu)$ and $V^{\text{env}}(\mu)$ respectively, is bounded by:

$$|V^{\pi}(\mu) - V^{\text{env}}(\mu)| \leq \frac{\gamma \epsilon(\pi,\beta) R_{\max}}{(1-\gamma)^2}. \tag{7}$$

This result resembles the simulation lemma (Kearns & Singh, 2002), which compares the difference in expected return of a single policy when evaluated in two different MDPs. Similarly, Lemma 3.8 bounds the difference in expected return resulting from the agent's actions on the passive environment. The main difference is that in the simulation lemma, the ϵ term is the difference in the transition function across the two MDPs, whereas in Lemma 3.8, it is a function of how the agent's policy and deviation function interact to influence the passive dynamics.

[1] We name this result as a lemma to underscore its conceptual link to the simulation lemma.

The simulation lemma establishes performance bounds for transferring a policy between environments; the action influence lemma bounds the maximum impact an agent can exert on the passive environment, constrained by the magnitude of its deviation function β.

3.5 The Advantage Function under Additive Decomposition

To isolate how an agent's action drives performance, we apply our decomposition to the advantage function, $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$. Standard notation blends passive dynamics and agent control within a single transition function, but substituting $p(s' \mid s,a) = p_{\text{env}}(s' \mid s) + \beta(s,a)_{s'}$ into Q^π and V^π reveals that the passive environment p_{env} cancels out in the *immediate* transition. While the passive dynamics still influence the advantage function indirectly via downstream state values, the immediate advantage depends on the agent's deviation function β.

Proposition 3.9 (Advantage as Expected Deviation)**.** *Under the decomposition in Equation 2 and a state-dependent reward function r, the advantage function can be computed as:*

$$A^\pi(s,a) = \gamma \sum_{s' \in \mathcal{S}} \Big(\beta(s,a)_{s'} - \bar{\beta}^\pi(s)_{s'}\Big) V^\pi(s')\,, \tag{8}$$

where $\bar{\beta}^\pi(s)_{s'} = \sum_a \pi(a \mid s)\beta(s,a)_{s'}$ is the expected deviation under policy π.

Equation 8 demonstrates that under this additive decomposition, the advantage function loses its *explicit* dependence on p_{env}, though it remains *implicitly* dependent through V^π. We extend this formulation to action-dependent rewards and provide its proof in Section 5.4.

3.6 Generalisation to Multiple Components

Definition 3.1 splits the transition function into one passive component and one active deviation component. Yet a more general statement also holds, which is potentially useful for scenarios involving multiple agents.

Proposition 3.10 (Generalised Transition Function Decomposition)**.** *The transition function $p(s' \mid s,a)$ of any CMP can be decomposed into a passive probability distribution p_{env} and an arbitrary (finite) number N of deviation functions, β_i. Formally, for any integer $N \geq 1$:*

$$p(s' \mid s,a) = p_{\text{env}}(s' \mid s) + \sum_{i=1}^{N} \beta_i(s,a)_{s'}\,, \tag{9}$$

where $\sum_{s' \in \mathcal{S}} \beta_i(s,a)_{s'} = 0$ for all $i \in \{1,\dots,N\}$.

While given a CMP and a passive environment for $N = 1$, this decomposition is uniquely determined, we lose this uniqueness property for systems where $N > 1$.

Proposition 3.11. *Given a target CMP and a fixed passive environment p_{env}, the individual deviation components β_i are not uniquely identifiable for $N > 1$.*

Interestingly, given valid deviation functions β_i and a CMP, p_{env} is still uniquely determined.

4 Discussion

We have here presented a new perspective on transitions in RL: every Markov decision process can be understood as the confluence of a passive environment together with an agent that intervenes on that environment via action. This proposal is not altogether new, as its intellectual lineage traces throughout the history of RL, causality, and optimal control. We develop an additive decomposition described by Equation 2 that is universal (Theorem 3.3) in the sense that every MDP can be modelled in this way. However, there are often many such choices for this decomposition for any one MDP (Proposition 3.6). We further show that this additive decomposition can accommodate arbitrarily

many (finite) deviation functions (Proposition 3.10), suggesting the perspective can extend to multi-agent RL. Lastly, we introduce the action influence lemma (Lemma 3.8), a sibling of the simulation lemma (Kearns & Singh, 2002) adapted to the case when agents are understood as deviating the environment away from some passive dynamics.

A Note on Decomposing Transition Probabilities. While we have decomposed the transition function, an alternative approach is to decompose how the state itself evolves. For instance, one might assume the next state is governed by an additive function, such as $s' = f(s) + g(s, a) + \epsilon$ for two arbitrary functions f and g alongside random noise ϵ, and then map this result to a probability measure. However, modelling the direct evolution of the state requires explicit knowledge of *how* the agent and environment interact, which may be highly non-linear. Because decomposing the transition function is universal (Theorem 3.3), it inherently encompasses any underlying state dynamics.

Embodiment and the Agent-Environment Boundary. The precise location of the agent-environment boundary represents a long-standing conceptual puzzle, with no definitive consensus on where the agent ends and its world begins (Jiang, 2019; Harutyunyan, 2020).

By focusing on the observable *effects* of an action, our mathematical decomposition remains agnostic to whether an agent is conceptualised as an external entity or deeply embedded within its world (Demski & Garrabrant, 2019; Lewandowski et al., 2026). While shifting this boundary alters the practical definitions of p_{env} and β, the underlying framework naturally accommodates any perspective.

One possible interpretation is to view the deviation function β as the mathematical representation of the agent's embodiment, where varying morphology dictates how a chosen action is translated into physical reality. For instance, in the frozen lake scenario (Section 3.3), two distinct agents select the identical action, yet only one slips; here, their differing bodies perfectly map to the asymmetric shifts captured by β. The deviation function β need not always have this meaning. For a human playing a video game (Demski & Garrabrant, 2019), where the virtual game can be seen as external to the agent (the player), β still captures how the agent's actions influence the dynamics of the game.

Open Questions. Several compelling open questions remain. First, we note that the additive form of Equation 2 is but one choice of how to decompose an agent's and environment's influence on the underlying stochastic process. We suggest that this form can be generalised through an operator, $\oplus$, such that

$$p(s' \mid s, a) = p_{\text{env}}(s' \mid s) \oplus \beta(s, a)_{s'} \,, \tag{10}$$

where $\oplus$ represents the broad family of ways in which the two forces can be combined. We take it as a significant direction for future work to analyse the plausible choices of $\oplus$, and the properties we may want to hold of this operator. For instance, we suggest that universality (Theorem 3.3) is likely a desirable property for such an operator to have, but it is not obviously the only one (and perhaps may be dispensed with to make way for other critical properties). Second, we conjecture that the proposed decomposition links to empowerment (Klyubin et al., 2005), a measure of the agent's capacity to influence its future state. Since an agent whose β is zero across all states and actions cannot influence its surroundings, it likely has no empowerment. Third, we propose extending this decomposable view beyond Markov processes to partially observable settings. Here, the same basic premise applies: the stream of observations is determined jointly by the influence of the agent and the environment. Thus, we can decompose the generative process that produces the observation in the same way.

Acknowledgments

The authors are grateful to Ahmed Khan, Logan Mondal Bhamidipaty, Raphael Trumpp, and the anonymous reviewers for their thoughtful comments on a draft of the paper.

References

David Abel, Mark K Ho, and Anna Harutyunyan. Three dogmas of reinforcement learning. *Reinforcement Learning Journal*, 1, 2024.

Zakariae EL Asri, Olivier Sigaud, and Nicolas Thome. Physics-informed model and hybrid planning for efficient Dyna-style reinforcement learning. *arXiv preprint arXiv:2407.02217*, 2024.

Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, pp. 679–684, 1957.

Rohan Chitnis and Tomás Lozano-Pérez. Learning compact models for planning with exogenous processes. In *Conference on Robot Learning*, 2020.

Abram Demski and Scott Garrabrant. Embedded agency. *arXiv preprint arXiv:1902.09469*, 2019.

Thomas Dietterich, George Trimponias, and Zhitang Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning. In *International Conference on Machine Learning*, 2018.

Anna Harutyunyan. What is an agent? https://anna.harutyunyan.net/wp-content/uploads/2020/09/What_is_an_agent.pdf, 2020.

Nan Jiang. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

Rudolf Emil Kalman. Contributions to the theory of optimal control. *Bol. Soc. Mat. Mexicana*, 5(2), 1960.

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

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135. IEEE, 2005.

Alex Lewandowski, Aditya Ramesh, Edan Meyer, Dale Schuurmans, and Marlos C Machado. The world is bigger! A computationally-embedded perspective on the big world hypothesis. In *Advances in Neural Information Processing Systems*, 2026.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Rafael Rodriguez-Sanchez, Cameron Allen, and George Konidaris. From pixels to factors: Learning independently controllable state variables for reinforcement learning. *Reinforcement Learning Journal*, 3, 2026.

Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

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

Emanuel Todorov. Linearly-solvable markov decision problems. In *Advances in Neural Information Processing Systems*, 2006.

Mark Towers, Ariel Kwiatkowski, John Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Kallinteris Andreas, Markus Krimmel, Arjun Kg, Rodrigo Perez-Vicente, et al. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems*, 2026.

George Trimponias and Thomas G Dietterich. Reinforcement learning with exogenous states and rewards. *arXiv preprint arXiv:2303.12957*, 2023.

Supplementary Materials

The following content was not necessarily subject to peer review.

5 Proofs and Definitions

We provide the proofs for each result below, following the order of subsections in Section 3.

5.1 Proofs for Section 3.2 (Universality and Limits of Todorov's Framework)

Theorem 3.3 (Universality of Definition 3.1)**.** *The additive decomposition is universal: the transition function of any valid CMP can be expressed as in Equation 2.*

Proof of Theorem 3.3.

We show that given any valid Markov chain p_{env}, there exists a β capable of reconstructing any arbitrary CMP, p. Because both p and p_{env} are valid probability distributions, the difference $p(s' \mid s, a) - p_{\text{env}}(s' \mid s) \in [-1, 1]$. By defining the deviation as $\beta(s, a)_{s'} = p(s' \mid s, a) - p_{\text{env}}(s' \mid s)$, we satisfy the bounds established in Equation 4, confirming that a valid deviation always exists to bridge the passive environment and the full CMP. □

Proposition 3.4 (Non-universality of Todorov's Framework)**.** *The decomposition of passive and active dynamics proposed by Todorov (2006) is not universal. There exists at least one CMP that cannot be reconstructed for any choice of p_{env} and control vector u using this framework.*

Proof of Proposition 3.4.

Consider a CMP with at least two states, $s_1, s_2 \in \mathcal{S}$, where a transition from s_1 to s_2 is impossible under the passive dynamics, meaning $p_{\text{env}}(s_2 \mid s_1) = 0$. Suppose, however, that this transition can happen if the agent takes an action u_1, such that $p(s_2 \mid s_1, u_1) > 0$. Because Todorov assumes an agent cannot induce a transition if the passive probability is zero (if $p_{\text{env}}(s' \mid s) = 0$ then $p(s' \mid s, u) = 0$ for all control vectors u), the framework cannot recover this CMP. □

Corollary 3.5. *The multiplicative decomposition proposed by Todorov (2006) is a special case of the additive decomposition in Equation 2, limited to environments where $p_{\text{env}}(s' \mid s) = 0$ implies $p(s' \mid s, u) = 0$, and where the action space is defined directly over log-probability deviations.*

Proof of Corollary 3.5.

Under Todorov's framework, the transition function is defined as

$$p(s' \mid s, u) = p_{\text{env}}(s' \mid s) \exp(u(s)_{s'}). \tag{11}$$

By adding and subtracting the passive environment $p_{\text{env}}(s' \mid s)$, we can rewrite this equation as

$$p(s' \mid s, u) = p_{\text{env}}(s' \mid s) + p_{\text{env}}(s' \mid s) \exp(u(s)_{s'}) - p_{\text{env}}(s' \mid s). \tag{12}$$

Grouping the last two terms allows us to isolate the specific structural deviation $\beta(s, u)_{s'}$

$$p(s' \mid s, u) = p_{\text{env}}(s' \mid s) + \underbrace{p_{\text{env}}(s' \mid s) \left(\exp(u(s)_{s'}) - 1\right)}_{\beta(s,u)_{s'}}. \tag{13}$$

We have an explicit mapping for the deviation vector entries

$$\beta(s, u)_{s'} = p_{\text{env}}(s' \mid s) \left(\exp(u(s)_{s'}) - 1\right). \tag{14}$$

To satisfy Definition 3.1, we must verify that this deviation vector sums to zero across all next states $s' \in \mathcal{S}$

$$\begin{aligned}\sum_{s'\in\mathcal{S}} \beta(s,u)_{s'} &= \sum_{s'\in\mathcal{S}} \left(p_{\text{env}}(s' \mid s)\exp(u(s)_{s'}) - p_{\text{env}}(s' \mid s)\right) \\ &= \sum_{s'\in\mathcal{S}} \underbrace{p_{\text{env}}(s' \mid s)\exp(u(s)_{s'})}_{p(s'|s,u)} - \sum_{s'\in\mathcal{S}} p_{\text{env}}(s' \mid s). \end{aligned} \tag{15}$$

Because both the transition probability $p(s' \mid s, u)$ and the passive environment $p_{\text{env}}(\cdot \mid s)$ are valid probability distributions, their sums over $\mathcal{S}$ must equal 1. We then have

$$\sum_{s'\in\mathcal{S}} \beta(s,u)_{s'} = 1 - 1 = 0. \tag{16}$$

Thus, we can represent any transition functions we can express in Todorov's framework with our additive decomposition. This completes the proof. □

5.2 Definition for Section 3.3 (Uniqueness and the Deviation Space)

The allowable deviation space defines the limits of the possible deviations as follows:

Definition 5.1. *The **allowable deviation space** $\mathcal{D}_{\text{env}}(s)$ for a given state s is the set of all mathematically valid deviation vectors $x \in \mathbb{R}^{|\mathcal{S}|}$ whose components sum to zero and satisfy Equation 4:*

$$\begin{aligned}\mathcal{D}_{\text{env}}(s) = \Bigg\{ x \in \mathbb{R}^{|\mathcal{S}|} \;\Bigg|\; &\sum_{s'\in\mathcal{S}} x_{s'} = 0 \text{ and} \\ &\forall s' \in \mathcal{S}, -p_{\text{env}}(s' \mid s) \le x_{s'} \le 1 - p_{\text{env}}(s' \mid s) \Bigg\}. \end{aligned} \tag{17}$$

5.3 Proofs for Section 3.4 (The Action Influence Lemma)

Let $\mathbb{P}^{\pi}_{\mu}$ denote the probability distribution induced by a stationary policy π and an initial state distribution μ.

The primary goal of this appendix is to prove the Action Influence Lemma (Lemma 3.8). To establish this final value bound, we must first prove a series of supporting lemmas.

Lemma 5.2. *The unnormalised state occupancy measure $d^{\pi}_{\mu}(s')$ for the decomposition in Definition 3.1 can be written as follows*

$$d^{\pi}_{\mu}(s') = \mu(s') + \gamma \sum_{s\in\mathcal{S}} d^{\pi}_{\mu}(s) \sum_{a\in\mathcal{A}} \pi(a \mid s)\left[p_{\text{env}}(s' \mid s) + \beta(s,a)_{s'}\right], \tag{18}$$

where $\mu(s')$ is the initial state distribution, $\mu(s') = \mathbb{P}^{\pi}_{\mu}(S_0 = s')$.

Proof of Lemma 5.2.

The unnormalised state occupancy measure $d^{\pi}_{\mu}(s')$ is defined as

$$d^{\pi}_{\mu}(s') = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^{\pi}_{\mu}(S_t = s'). \tag{19}$$

If we isolate the initial step $t = 0$, the occupancy measure expands as

$$d_\mu^\pi(s') = \gamma^0 \mathbb{P}_\mu^\pi(S_0 = s') + \sum_{t=1}^{\infty} \gamma^t \mathbb{P}_\mu^\pi(S_t = s')$$
$$= \mu(s') + \sum_{t=1}^{\infty} \gamma^t \mathbb{P}_\mu^\pi(S_t = s'). \tag{20}$$

For $t \geq 1$, we marginalise over the previous state s and action a

$$\begin{aligned}
\mathbb{P}_\mu^\pi(S_t = s') &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \mathbb{P}_\mu^\pi(S_t = s', A_{t-1} = a, S_{t-1} = s) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \mathbb{P}_\mu^\pi(S_t = s' \mid A_{t-1} = a, S_{t-1} = s) \\
&\qquad \times \mathbb{P}_\mu^\pi(A_{t-1} = a \mid S_{t-1} = s)\, \mathbb{P}_\mu^\pi(S_{t-1} = s) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} p(s' \mid s, a)\pi(a \mid s) \mathbb{P}_\mu^\pi(S_{t-1} = s) \\
&= \sum_{s\in\mathcal{S}} \mathbb{P}_\mu^\pi(S_{t-1} = s) \sum_{a\in\mathcal{A}} \pi(a \mid s) p(s' \mid s, a).
\end{aligned} \tag{21}$$

Plugging Equation 21 into Equation 20 and rearranging the terms, we obtain

$$\begin{aligned}
d_\mu^\pi(s') &= \mu(s') + \sum_{t=1}^{\infty} \gamma^t \left(\sum_{s\in\mathcal{S}} \mathbb{P}_\mu^\pi(S_{t-1} = s) \sum_{a\in\mathcal{A}} \pi(a \mid s) p(s' \mid s, a) \right) \\
&= \mu(s') + \gamma \sum_{s\in\mathcal{S}} \left(\sum_{a\in\mathcal{A}} \pi(a \mid s) p(s' \mid s, a) \right) \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{P}_\mu^\pi(S_{t-1} = s).
\end{aligned} \tag{22}$$

Applying a change of variables ($k = t - 1$) to the right-most sum, we recover Equation 19

$$\sum_{k=0}^{\infty} \gamma^k \mathbb{P}_\mu^\pi(S_k = s) = d_\mu^\pi(s). \tag{23}$$

Finally, substituting $d_\mu^\pi(s)$ and our transition decomposition $p(s' \mid s, a) = p_{\text{env}}(s' \mid s) + \beta(s, a)_{s'}$, we obtain the stated result

$$d_\mu^\pi(s') = \mu(s') + \gamma \sum_{s\in\mathcal{S}} d_\mu^\pi(s) \sum_{a\in\mathcal{A}} \pi(a \mid s) \left[p_{\text{env}}(s' \mid s) + \beta(s, a)_{s'} \right]. \tag{24}$$

This concludes the proof. □

Lemma 5.3. *Let $\epsilon(\pi, \beta)$ denote the maximum deviation when transitioning from s to s'*

$$\epsilon(\pi, \beta) = \max_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} \left| \sum_{a\in\mathcal{A}} \pi(a \mid s) \beta(s, a)_{s'} \right|. \tag{25}$$

The L_1 distance between the occupancy measures due to the policy and to the environment is

$$\left\| d_\mu^\pi - d_\mu^{env} \right\|_1 \leq \frac{\gamma \epsilon(\pi, \beta)}{(1 - \gamma)^2}. \tag{26}$$

Proof of Lemma 5.3.

By Lemma 5.2, the occupancy measure due to the policy is given by

$$d_\mu^\pi(s') = \mu(s') + \gamma \sum_{s\in\mathcal{S}} d_\mu^\pi(s) \sum_{a\in\mathcal{A}} \pi(a \mid s) \left[p_{\text{env}}(s' \mid s) + \beta(s,a)_{s'}\right]. \tag{27}$$

Similarly, the environment's occupancy measure (where β is zero everywhere) is

$$d_\mu^{\text{env}}(s') = \mu(s') + \gamma \sum_{s\in\mathcal{S}} d_\mu^{\text{env}}(s) p_{\text{env}}(s' \mid s). \tag{28}$$

Subtracting Equation 28 from Equation 27, and noting that $\sum_{a\in\mathcal{A}} \pi(a \mid s) = 1$, we obtain

$$\begin{aligned} d_\mu^\pi(s') - d_\mu^{\text{env}}(s') &= \left(\gamma \sum_{s\in\mathcal{S}} d_\mu^\pi(s) p_{\text{env}}(s' \mid s) - \gamma \sum_{s\in\mathcal{S}} d_\mu^{\text{env}}(s) p_{\text{env}}(s' \mid s)\right) \\ &\quad + \gamma \sum_{s\in\mathcal{S}} d_\mu^\pi(s) \sum_{a\in\mathcal{A}} \pi(a \mid s)\beta(s,a)_{s'} \\ &= \gamma \sum_{s\in\mathcal{S}} \left(d_\mu^\pi(s) - d_\mu^{\text{env}}(s)\right) p_{\text{env}}(s' \mid s) + \gamma \sum_{s\in\mathcal{S}} d_\mu^\pi(s) \sum_{a\in\mathcal{A}} \pi(a \mid s)\beta(s,a)_{s'}. \end{aligned} \tag{29}$$

We want to find a bound for the L_1 norm

$$\left\|d_\mu^\pi - d_\mu^{\text{env}}\right\|_1 = \sum_{s'\in\mathcal{S}} \left|d_\mu^\pi(s') - d_\mu^{\text{env}}(s')\right| \tag{30}$$

$$\begin{aligned} &= \sum_{s'\in\mathcal{S}} \Bigg| \gamma \sum_{s\in\mathcal{S}} \left(d_\mu^\pi(s) - d_\mu^{\text{env}}(s)\right) p_{\text{env}}(s' \mid s) \\ &\quad + \gamma \sum_{s\in\mathcal{S}} d_\mu^\pi(s) \sum_{a\in\mathcal{A}} \pi(a \mid s)\beta(s,a)_{s'} \Bigg|. \end{aligned} \tag{31}$$

Applying the triangle inequality to separate the two main terms, we have

$$\begin{aligned} &\leq \sum_{s'\in\mathcal{S}} \left| \gamma \sum_{s\in\mathcal{S}} \left(d_\mu^\pi(s) - d_\mu^{\text{env}}(s)\right) p_{\text{env}}(s' \mid s) \right| \\ &\quad + \sum_{s'\in\mathcal{S}} \left| \gamma \sum_{s\in\mathcal{S}} d_\mu^\pi(s) \sum_{a\in\mathcal{A}} \pi(a \mid s)\beta(s,a)_{s'} \right|. \end{aligned} \tag{32}$$

We apply the triangle inequality again to move the absolute value inside the sums over s. Noting that both $p_{\text{env}}(s' \mid s)$ and $d_\mu^\pi(s)$ are non-negative, this gives

$$\begin{aligned} &\leq \gamma \sum_{s'\in\mathcal{S}} \sum_{s\in\mathcal{S}} \left|d_\mu^\pi(s) - d_\mu^{\text{env}}(s)\right| p_{\text{env}}(s' \mid s) \\ &\quad + \gamma \sum_{s'\in\mathcal{S}} \sum_{s\in\mathcal{S}} d_\mu^\pi(s) \left| \sum_{a\in\mathcal{A}} \pi(a \mid s)\beta(s,a)_{s'} \right|. \end{aligned} \tag{33}$$

Finally, swapping the order of summation for both terms

$$\begin{aligned} &= \gamma \sum_{s\in\mathcal{S}} \left|d_\mu^\pi(s) - d_\mu^{\text{env}}(s)\right| \underbrace{\sum_{s'\in\mathcal{S}} p_{\text{env}}(s' \mid s)}_{=1} \\ &\quad + \gamma \sum_{s\in\mathcal{S}} d_\mu^\pi(s) \sum_{s'\in\mathcal{S}} \left| \sum_{a\in\mathcal{A}} \pi(a \mid s)\beta(s,a)_{s'} \right|. \end{aligned} \tag{34}$$

The first half of the equation matches Equation 30. We then bound the rest of the equation by the maximum value of $\sum_{s' \in \mathcal{S}} \left| \sum_{a \in \mathcal{A}} \pi(a \mid s) \beta(s,a)_{s'} \right|$. Because this maximum is now a constant independent of the outer summation over $d_\mu^\pi(s)$, it can be factorised out

$$\begin{aligned} &\leq \gamma \left\| d_\mu^\pi - d_\mu^{\text{env}} \right\|_1 \\ &\quad + \gamma \left(\max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \left| \sum_{a \in \mathcal{A}} \pi(a \mid s) \beta(s,a)_{s'} \right| \right) \sum_{s \in \mathcal{S}} d_\mu^\pi(s). \end{aligned} \tag{35}$$

Letting $\epsilon(\pi, \beta) = \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \left| \sum_{a \in \mathcal{A}} \pi(a \mid s) \beta(s,a)_{s'} \right|$, we have

$$\left\| d_\mu^\pi - d_\mu^{\text{env}} \right\|_1 \leq \gamma \left\| d_\mu^\pi - d_\mu^{\text{env}} \right\|_1 + \gamma \epsilon(\pi, \beta) \sum_{s \in \mathcal{S}} d_\mu^\pi(s). \tag{36}$$

Since $\sum_{s \in \mathcal{S}} d_\mu^\pi(s) = \frac{1}{1-\gamma}$, we substitute and rearrange to isolate the norm

$$(1-\gamma) \left\| d_\mu^\pi - d_\mu^{\text{env}} \right\|_1 \leq \frac{\gamma \epsilon(\pi, \beta)}{1-\gamma}. \tag{37}$$

Dividing by $(1-\gamma)$

$$\left\| d_\mu^\pi - d_\mu^{\text{env}} \right\|_1 \leq \frac{\gamma \epsilon(\pi, \beta)}{(1-\gamma)^2}. \tag{38}$$

This concludes the proof. □

Lemma 5.4. *Under a finite Markov Decision Process (MDP) with a stationary policy π, a bounded state-dependent reward function $r : \mathcal{S} \to \mathbb{R}$, and a discount factor $\gamma \in [0, 1)$, the expected return $V^\pi(\mu)$ from an initial state distribution μ is equal to the inner product of the reward function and the unnormalised state occupancy measure $d_\mu^\pi(s)$, such that*

$$V^\pi(\mu) = \sum_{s \in \mathcal{S}} d_\mu^\pi(s) r(s)\,, \tag{39}$$

where the unnormalised state occupancy measure is defined as $d_\mu^\pi(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\mu^\pi\left(S_t = s\right)$.

Proof of Lemma 5.4.

By definition, the expected infinite-horizon discounted return $V^\pi(\mu)$ for an initial state distribution $S_0 \sim \mu$ under policy π is given by the expectation over trajectories[a]

$$V^\pi(\mu) = \mathbb{E}_\mu^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(S_t) \right]. \tag{40}$$

We then have

$$V^\pi(\mu) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\mu^\pi \left[r(S_t) \right]. \tag{41}$$

Since the state space $\mathcal{S}$ is finite, we can iterate through all possible states $s \in \mathcal{S}$

$$\begin{aligned} V^\pi(\mu) &= \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \mathbb{P}_\mu^\pi(S_t = s) r(s) \\ &= \sum_{s \in \mathcal{S}} \left(\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\mu^\pi(S_t = s) \right) r(s). \end{aligned} \tag{42}$$

By substituting the definition of the unnormalised state occupancy measure, $d_\mu^\pi(s)$

$$V^\pi(\mu) = \sum_{s \in \mathcal{S}} d_\mu^\pi(s) r(s). \tag{43}$$

This completes the proof. □

[a]The following proof is inspired by the one available at https://rltheory.github.io/lecture-notes/planning-in-mdps/lec2/

Lemma 3.8. *Let $\epsilon(\pi, \beta)$ denote the maximum deviation when transitioning from s to s' via a stationary policy π,*

$$\epsilon(\pi, \beta) = \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \left| \sum_{a \in \mathcal{A}} \pi(a \mid s) \beta(s, a)_{s'} \right| \tag{44}$$

Given an initial state distribution μ and a state-dependent reward function r with bounded norm $\|r\|_\infty = R_{\max}$, the difference between the expected returns $V^\pi(\mu)$ and $V^{\text{env}}(\mu)$ under the policy and the base environment, respectively, is bounded by

$$|V^\pi(\mu) - V^{\text{env}}(\mu)| \leq \frac{\gamma \epsilon(\pi, \beta) R_{\max}}{(1-\gamma)^2}. \tag{45}$$

Proof of Lemma 3.8.

By Lemma 5.4, we have

$$V^\pi(\mu) = \sum_{s \in \mathcal{S}} d_\mu^\pi(s) r(s). \tag{46}$$

Similarly, the expected return under the passive environment's dynamics is given by

$$V^{\text{env}}(\mu) = \sum_{s \in \mathcal{S}} d_\mu^{\text{env}}(s) r(s). \tag{47}$$

Subtracting Equation 47 from Equation 46, we factor out the common reward term

$$\begin{aligned} V^\pi(\mu) - V^{\text{env}}(\mu) &= \sum_{s \in \mathcal{S}} d_\mu^\pi(s) r(s) - \sum_{s \in \mathcal{S}} d_\mu^{\text{env}}(s) r(s) \\ &= \sum_{s \in \mathcal{S}} \left(d_\mu^\pi(s) - d_\mu^{\text{env}}(s) \right) r(s) \,. \end{aligned} \tag{48}$$

We want to find a bound for the absolute difference of these expected returns

$$|V^\pi(\mu) - V^{\text{env}}(\mu)| = \left| \sum_{s \in \mathcal{S}} \left(d_\mu^\pi(s) - d_\mu^{\text{env}}(s) \right) r(s) \right|. \tag{49}$$

By applying the triangle inequality, followed by Hölder's inequality with $p = 1$ and $q = \infty$, and noting that $\|r\|_\infty = \max_{s \in \mathcal{S}} |r(s)| = R_{\max}$, we obtain

$$\begin{aligned} |V^\pi(\mu) - V^{\text{env}}(\mu)| &\leq \left\| d_\mu^\pi - d_\mu^{\text{env}} \right\|_1 \|r\|_\infty \\ &= \left\| d_\mu^\pi - d_\mu^{\text{env}} \right\|_1 \cdot R_{\max}. \end{aligned} \tag{50}$$

By Lemma 5.3

$$\left\| d_\mu^\pi - d_\mu^{\text{env}} \right\|_1 \leq \frac{\gamma \epsilon(\pi, \beta)}{(1-\gamma)^2}. \tag{51}$$

Substituting the distance bound from Equation 51 into Equation 50

$$|V^\pi(\mu) - V^{\text{env}}(\mu)| \leq \frac{\gamma \epsilon(\pi, \beta) R_{\max}}{(1-\gamma)^2}. \tag{52}$$

This concludes the proof. □

5.4 Proofs for Section 3.5 (The Advantage Function under Additive Decomposition)

Proposition 3.9 (Advantage as Expected Deviation)**.** *Under the linear transition decomposition in Definition 3.1, the advantage function is given by*

$$A^\pi(s,a) = \big(r(s,a) - r_\pi(s)\big) + \gamma \sum_{s' \in \mathcal{S}} \Big(\beta(s,a)_{s'} - \bar{\beta}^\pi(s)_{s'}\Big) V^\pi(s'), \tag{53}$$

where $r_\pi(s) = \sum_a \pi(a|s) r(s,a)$ *is the expected immediate reward, and* $\bar{\beta}^\pi(s)_{s'} = \sum_a \pi(a|s)\beta(s,a)_{s'}$ *is the expected deviation under policy* π*.*

Furthermore, if the reward is state-dependent such that $r(s,a) = r(s)$*, the immediate reward terms cancel, and the advantage reduces to the expected value of the relative deviation*

$$A^\pi(s,a) = \gamma \sum_{s' \in \mathcal{S}} \Big(\beta(s,a)_{s'} - \bar{\beta}^\pi(s)_{s'}\Big) V^\pi(s'). \tag{54}$$

Proof of Proposition 3.9.

We begin by expanding the action-value function $Q^\pi(s,a)$ using our decomposition

$$\begin{aligned} Q^\pi(s,a) &= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) V^\pi(s') \\ &= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \Big(p_{\text{env}}(s' \mid s) + \beta(s,a)_{s'}\Big) V^\pi(s') \\ &= r(s,a) + \underbrace{\gamma \sum_{s' \in \mathcal{S}} p_{\text{env}}(s' \mid s) V^\pi(s')}_{\alpha(s)} + \gamma \sum_{s' \in \mathcal{S}} \beta(s,a)_{s'} V^\pi(s'). \end{aligned} \tag{55}$$

Here, $\alpha(s)$ captures the expected downstream value driven by the immediate passive environment. We then compute the state-value function $V^\pi(s)$ by taking the expectation over actions $a \sim \pi(\cdot|s)$

$$\begin{aligned} V^\pi(s) &= \sum_a \pi(a|s) Q^\pi(s,a) \\ &= \sum_a \pi(a|s) r(s,a) + \alpha(s) \sum_a \pi(a|s) + \gamma \sum_{s' \in \mathcal{S}} \left(\sum_a \pi(a|s) \beta(s,a)_{s'} \right) V^\pi(s') \\ &= r_\pi(s) + \alpha(s) + \gamma \sum_{s' \in \mathcal{S}} \bar{\beta}^\pi(s)_{s'} V^\pi(s'). \end{aligned} \tag{56}$$

Subtracting $V^\pi(s)$ from $Q^\pi(s,a)$ returns the advantage. The state-dependent immediate passive term $\alpha(s)$ cancels out

$$\begin{aligned} A^\pi(s,a) &= Q^\pi(s,a) - V^\pi(s) \\ &= \big(r(s,a) - r_\pi(s)\big) + \big(\alpha(s) - \alpha(s)\big) + \gamma \sum_{s'} \Big(\beta(s,a)_{s'} - \bar{\beta}^\pi(s)_{s'}\Big) V^\pi(s') \\ &= \big(r(s,a) - r_\pi(s)\big) + \gamma \sum_{s'} \Big(\beta(s,a)_{s'} - \bar{\beta}^\pi(s)_{s'}\Big) V^\pi(s'). \end{aligned} \tag{57}$$

Finally, if the reward is state-dependent such that $r(s,a) = r(s)$, the expected immediate reward simplifies to $r_\pi(s) = \sum_a \pi(a|s) r(s) = r(s)$. Consequently

$$A^\pi(s,a) = \gamma \sum_{s'} \Big(\beta(s,a)_{s'} - \bar{\beta}^\pi(s)_{s'}\Big) V^\pi(s'). \tag{58}$$

This completes the proof. □

5.5 Proofs for Section 3.6 (Generalisation to Multiple Components)

Proposition 3.10 (Generalised Transition Function Decomposition)**.** *The transition function $p(s' \mid s, a)$ of any CMP can be decomposed into a passive probability distribution $p_{\text{env}}(s' \mid s)$ and an arbitrary (finite) number N of deviation functions, β. Formally, for any integer $N \geq 1$:*

$$p(s' \mid s, a) = p_{\text{env}}(s' \mid s) + \sum_{i=1}^{N} \beta_i(s, a)_{s'} \,, \tag{59}$$

where $\sum_{s' \in \mathcal{S}} \beta_i(s, a)_{s'} = 0$ for all $i \in \{1, \ldots, N\}$.

Proof of Proposition 3.10.

We proceed to prove its universality by induction on the number of components N.
Base case ($N = 1$): This holds as a direct consequence of Theorem 3.3.
Inductive step: Assume the decomposition holds for k components, such that $p(s' \mid s, a) = p_{\text{env}}(s' \mid s) + \sum_{i=1}^{k} \beta_i(s, a)_{s'}$ is a valid transition function. To establish the decomposition for $k + 1$ components, we isolate the k-th component, $\beta_k(s, a)_{s'}$. We can decompose this single valid deviation into two new valid deviations, β_{k_a} and β_{k_b}, such that $\beta_{k_a} + \beta_{k_b} = \beta_k$ and both sum to zero over s'. Replacing β_k with these two new components in the summation gives a valid decomposition with exactly $k + 1$ deviations. Thus, the property holds for $N = k + 1$. □

Proposition 3.11. *Given a target CMP and a fixed passive environment p_{env}, the individual deviation components β_i are not uniquely identifiable for $N > 1$.*

Proof of Proposition 3.11.

Consider an environment with two agents ($N = 2$). Let $\beta(s, a)_{s'} = p(s' \mid s, a) - p_{\text{env}}(s' \mid s)$ be the total required probability shift across the two agents. Any valid shift can be arbitrarily distributed between β_1 and β_2 such that they sum to $\beta(s, a)_{s'}$ and give rise to a valid probability distribution. Therefore, the influence of each agent's action is non-unique. □