# Integrating RL and Planning through Optimal Transport World Models

Willem Röpke*
Vrije Universiteit Brussel
willem.ropke@vub.be

Raphael Avalos*
Vrije Universiteit Brussel
raphael.avalos@vub.be

Roxana Rădulescu
Utrecht University
Vrije Universiteit Brussel

Ann Nowé
Vrije Universiteit Brussel

Diederik M. Roijers
City of Amsterdam
Vrije Universiteit Brussel

Florent Delgrange
Vrije Universiteit Brussel

## ABSTRACT

We introduce Optimal Transport MDPs (OT-MDPs), a framework for learning principled latent world models via optimal transport. Our approach formulates a generic optimal transport objective that trains a generative model of the environment by minimising a customisable cost function, which quantifies the discrepancy between latent and real trajectories. Through this perspective, we highlight the limitations of reconstruction-based methods and establish conditions on the cost function that enable theoretical guarantees. The quality of the learned model allows us to integrate reinforcement learning and planning methods. In particular, we leverage model-based value expansion to refine value estimates, providing rigorous theoretical justification. Additionally, we examine the use of Monte Carlo tree search and provide a theoretical analysis of the assumptions under which its application remains sound. Empirical evaluation across four MinAtar environments demonstrates that OT-MDPs yield high-fidelity models, leading to strong performance. Moreover, our results reveal challenges associated with planning in the latent model, suggesting critical directions for future research.

## KEYWORDS

Reinforcement learning, Optimal transport, Representation learning

## 1 INTRODUCTION

*Reinforcement learning* (RL) provides a framework for solving complex sequential decision-making problems by optimising policies through trial and error [36]. Scaling RL to high-dimensional environments requires effective *representation learning* to extract meaningful state encodings while discarding irrelevant details [8, 43]. As an alternative to direct representation learning, *model-based RL* employs an auxiliary model of the environment to generate imagined rollouts [26], facilitate planning [35], or refine learning targets [41]. In this work, we unify representation and model learning by training a principled latent model whose latent state representations are directly used to optimise the policy.

A key challenge in model learning is ensuring that the latent model accurately reflects the true environment. Prior work has addressed this by leveraging *bisimulation metrics* [17, 37], which enforce similarity between functionally equivalent states in the latent space. Such models provide strong theoretical guarantees [18, 43] and enable efficient policy learning [3, 28]. Among these, the *Wasserstein auto-encoded Markov decision processes* (WAE-MDP) framework has demonstrated both theoretical soundness and empirical success by minimising the optimal transport distance between the real environment and its latent reconstruction [11, 12].

We introduce *Optimal Transport MDPs* (OT-MDPs), a generalisation of WAE-MDPs that allows custom cost functions, providing greater flexibility in shaping latent representations for RL. This approach facilitates a deeper analysis of the limitations inherent to reconstruction-based methods and offers a direct means to structure representations that enhance policy learning. As a key instantiation, we propose *value equivalence*, which aligns the values of real and latent critics. Since the encoder is trained to map bisimilar states to similar representations, the policy can directly leverage it to enhance learning. To further exploit the latent model, we incorporate model-based value expansion (MVE) [16] into policy learning, yielding more accurate target values. We establish an upper bound on the discrepancy between true and MVE-derived values, which diminishes as model and critic quality improves. Furthermore, we present a corollary that underscores a limitation of our framework and suggests directions for future work. Additionally, we establish a theoretical foundation for applying Monte Carlo tree search (MCTS) in an OT-MDP. Extensive experiments show that OT-MDPs learn accurate world models and effectively combine planning with learning, resulting in strong performance in challenging environments.

## 2 RELATED WORK

**Representation Learning in RL.** *Bisimulation metrics* have been widely considered for learning state representations. Zhang et al. [43] propose learning a state representation for continuous control in pixel-based environments without relying on reconstruction terms, training the policy directly on a latent space; Rezaei-Shoshtari et al. [34] introduce a policy gradient theorem applicable to abstract MDPs. A common theme in these works is the assumption of Gaussian transition kernels and the use of Wasserstein-2 distances to maintain computational tractability, which is not required with our approach. Another common assumption is deterministic MDPs, where computing exact Wasserstein distances is straightforward and leads to the development of several algorithms for this setting [7, 39]. Gelada et al. [20] propose learning a latent model as an auxiliary task for representation learning, offering strong theoretical justification for their approach. However, in practice, the

*Equal contribution.

*Proc. of the Adaptive and Learning Agents Workshop (ALA 2025), Avalos, Aydeniz, Müller, Mohammedalamen (eds.), May 19 – 20, 2025, Detroit, Michigan, USA, ala-workshop.github.io.* 2025.

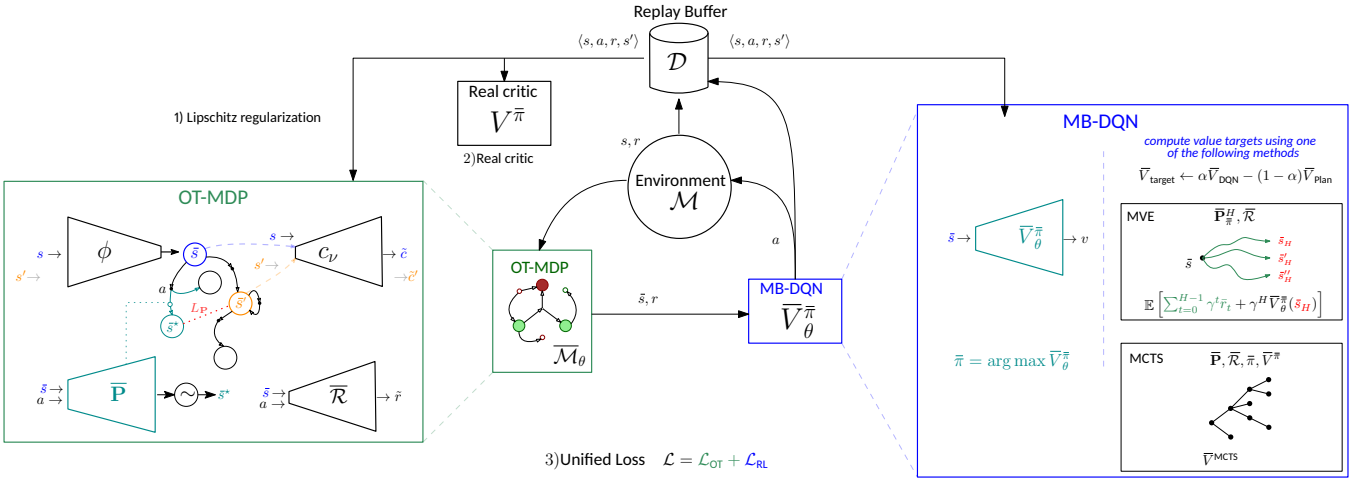



**Figure 1: A high-level overview of the OT-MDP framework. The OT-MDP component is presented in Section 5 and is responsible for learning the latent model and policy. The planning approaches are explained in Section 6 and provide improved learning targets for the policy.**

algorithm uses the deterministic assumption to avoid computing the Wasserstein distance. To overcome challenges in learning exact bisimulation metrics, Castro et al. [8] introduce MICo, a diffuse metric integrated with RL agents for effective representation learning. Finally, Agarwal et al. [1] present a contrastive representation learning technique that encodes states with similar optimal policies closer together in the latent space.

**Principled representation**. In contrast, our work focuses on learning a theoretically grounded representation and an accurate abstraction of the environment, both with bisimulation guarantees. Learning latent space models with guarantees (and more particularly, using optimal transport formulation) can be approached via the Wasserstein-autoencoder [38] and -GAN [2, 23] frameworks in the (un)supervised setting. Variational and Wasserstein auto-encoded Markov decision processes [11, 13] have been considered for *distilling* RL policies with bisimulation guarantees into tractable controllers, amenable for formal verification. Incorporating distillation into the RL process has been considered in [3, 10, 13, 14] by alternating between optimizing DQN or A2C agents and WAE-MDPs in a round-robin fashion, however all those methods may suffer from stability issues and require reconstructing the input observations, which may paradoxically hinder bisimulation learning in general.

**Unifying representation learning and model-based approaches**. Other RL methods suggest that learning a latent model is not only beneficial for sample efficiency, but also to learn a representation that supports policy learning (even when guarantees lack) [9, 19, 25]. In particular, [9, 25] show that using a discrete representation may be practically beneficial. Our results demonstrate that this phenomenon is theoretically grounded when considering learning bisimulation metrics.

# 3 BACKGROUND

## 3.1 Markov Decision Processes

Markov decision processes (MDPs) are used to model sequential decision-making settings. An MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathcal{R}, s_I, \gamma \rangle$ is defined as a tuple consisting of a set of states $\mathcal{S}$, actions $\mathcal{A}$, a transition function $\mathbf{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, a bounded reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, an initial state $s_I \in \mathcal{S}$, and a discount factor $\gamma \in [0, 1)$. When $\mathcal{A}$ is a singleton, $\mathcal{M}$ is a purely stochastic process named a *Markov chain* (MC). A *trajectory* $(s_t, a_t)_{t \geq 0}$ is an infinite sequence of states and actions produced in $\mathcal{M}$, so that $s_0 = s_I$ and $s_{t+1} \sim \mathbf{P}(\cdot \mid s_t, a_t)$.

To act in an MDP, we consider stationary policies $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ mapping states to a distribution over actions. An MDP running under $\pi$ induces a MC $\mathcal{M}_\pi$ with reward and transition functions $\mathcal{R}_\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \mathcal{R}(s, a)$, $\mathbf{P}_\pi(\cdot \mid s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbf{P}(\cdot \mid s, a)$ along with a unique probability measure over the produced trajectories [33]. We write $\mathbb{E}_\pi$ for the associated expectation operator and $\mathbb{E}_\pi[\cdot \mid s_I = s]$ for the one where we fix the initial state of $\mathcal{M}$ to $s \in \mathcal{S}$. A *stationary measure* of a policy $\pi$ intuitively yields a distribution over states likely to be visited under $\pi$ and is formally defined as a solution of the equation $\xi_\pi(\cdot) = \mathbb{E}_{s \sim \xi_\pi} \mathbf{P}_\pi(\cdot \mid s)$. Such a measure is guaranteed to exist under reasonable assumptions (e.g, in episodic RL [27]). In RL, the goal is to learn a policy $\pi$ that maximises its discounted expected return from any state $s \in \mathcal{S}$, which is captured by the *value function* $V^\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \mid s_I = s \right]$.

## 3.2 Optimal Transport

*Optimal transport* (OT) concerns finding the most efficient way to transform one probability distribution into another, given a specified cost function [40]. Formally, for two probability measures $\mu$ and $\nu$ defined over spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and a cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$, the optimal transport problem is defined as

$$\mathrm{OT}_c(\mu, \nu) \triangleq \inf_{\gamma \in \Gamma(\mu, \nu)} \mathop{\mathbb{E}}_{(x,y) \sim \gamma} [c(x, y)],$$

where $\Gamma(\mu, \nu)$ denotes the set of all joint couplings of $\mu$ and $\nu$.

When the probability measures are defined over the same space $\mathcal{X}$ and the cost function is given by a metric $d$ on $\mathcal{X}$, the problem specialises to the computation of the *Wasserstein distance*. The Wasserstein(-1) distance can be expressed in its dual form as

$$\mathcal{W}_d(\mu, \nu) = \sup_{\|f\|_L \leq 1} \mathop{\mathbb{E}}_{x \sim \mu}[f(x)] - \mathop{\mathbb{E}}_{y \sim \nu}[f(y)] \tag{1}$$

where $\|f\|_L$ denotes the *Lipschitz norm* of $f$. A generative model with distribution $\nu$ can be trained to minimise the Wasserstein distance to $\mu$ using this dual formulation [2, 24].

### 3.3 Bisimulation

*Bisimulation* [22, 30] defines equivalence relations between states from which the agent exhibits similar behaviours, i.e., yielding identical immediate rewards and transition dynamics. The *coarsest* bisimulation relation (i.e., grouping the most states together) is denoted by $\sim$. The equivalence class of a state $s \in \mathcal{S}$ is denoted $[s]_\sim = \{ s' \in \mathcal{S} : s \sim s' \}$. Two MDPs $\mathcal{M}$ and $\mathcal{M}'$ are bisimilar (denoted $\mathcal{M} \sim \mathcal{M}'$) if their initial states are bisimilar, *which ensures identical optimal values*. While grouping states into equivalence classes is appealing, bisimulation is rigid: states with minor differences in rewards or transitions are deemed completely distinct. To relax this, bisimulation *(pseudo-)metrics* $d : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ are introduced to provide a numerical similarity measure [18]. The bisimulation distance $\widetilde{d}$ is the unique fixpoint of the operator $\mathcal{F} : \mathcal{P} \to \mathcal{P}$, defined by

$$\begin{aligned} \mathcal{F}(d)(s_1, s_2) = \max_{a \in \mathcal{A}} [\, & |\mathcal{R}(s_1, a) - \mathcal{R}(s_2, a)| \\ & + \gamma \cdot \mathcal{W}_d(\mathbf{P}(\cdot \mid s_1, a), \mathbf{P}(\cdot \mid s_2, a)) \,], \end{aligned} \tag{2}$$

for all $s_1, s_2 \in \mathcal{S}$, where $\mathcal{P}$ is the space of bounded pseudometrics. The kernel of $\widetilde{d}$ recovers the coarsest bisimulation relation, as $\widetilde{d}(s_1, s_2) = 0$ if and only if $s_1 \sim s_2$.

Since bisimulation metrics are challenging to compute online and do not account for non-optimal policies [7], we adopt the *on-policy* bisimulation distance, written $\widetilde{d}_\pi$, where bisimulation is computed in the induced MC $\mathcal{M}_\pi$. A crucial aspect of this distance is its direct influence on value functions:[1]

$$\left|V^\pi(s_1) - V^\pi(s_2)\right| \leq \widetilde{d}_\pi(s_1, s_2). \tag{3}$$

Thus, bisimilarly close states yield similar values under $\pi$, making latent encoders that capture this distance particularly beneficial.

### 3.4 Wasserstein Auto-Encoded MDPs

WAE-MDPs [11] are generative models that jointly learn a discrete latent abstraction of the environment $\overline{\mathcal{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \bar{\mathbf{P}}, \overline{\mathcal{R}}, \bar{s}_I, \gamma \rangle$, an encoder $\phi : \mathcal{S} \to \overline{\mathcal{S}}$, and a decoder $\psi : \overline{\mathcal{S}} \to \mathcal{S}$. While these models may also learn an action encoder, we assume finite action spaces and hence $\mathcal{A} = \overline{\mathcal{A}}$. In the following, we will also consider *latent policies* (i.e., policies for $\mathcal{M}$) $\bar{\pi} : \overline{\mathcal{S}} \to \Delta(\mathcal{A})$. Such policies can be executed in $\mathcal{M}$: given a state $s \in \mathcal{S}$, one just needs to encode $s$ to $\phi(s)$ and then draw $a \sim \bar{\pi}(\cdot \mid \phi(s))$. In the following, by slightly abusing notation, we may write $\xi_{\bar{\pi}}$ for the distribution of first drawing states $s$ from $\xi_{\bar{\pi}}$ and then actions $a$ from $\bar{\pi}(\cdot \mid \phi(s))$.

[1] Namely, the value function is Lipschitz-continuous w.r.t. the pseudometric.

To minimise the bisimulation distance between the latent model and the environment, WAE-MDPs employ a *local reward loss* $L_{\mathcal{R}}$ and a *local transition loss* $L_{\mathbf{P}}$, defined as

$$L_{\mathcal{R}} = \mathop{\mathbb{E}}_{s,a \sim \xi_{\bar{\pi}}} \left| \mathcal{R}(s, a) - \overline{\mathcal{R}}(\phi(s), a) \right|, \tag{4}$$

$$L_{\mathbf{P}} = \mathop{\mathbb{E}}_{s,a \sim \xi_{\bar{\pi}}} \mathcal{W}_{\bar{d}} \left( \phi_\sharp \mathbf{P}(\cdot \mid s, a), \bar{\mathbf{P}}(\cdot \mid \phi(s), a) \right), \tag{5}$$

where $\bar{d}$ is the *discrete metric* and $\phi_\sharp \mathbf{P}$ denotes the *pushforward measure* of samples $s' \sim \mathbf{P}(\cdot \mid s, a)$ that are subsequently encoded to $\bar{s}' = \phi(s')$. The overall loss function, $\mathcal{L}_{\text{WAE}}$, is minimised:

$$\mathop{\mathbb{E}}_{s,a,s' \sim \xi_{\bar{\pi}}} \left[ d_{\mathcal{S}}(s, \psi \circ \phi(s)) + d_{\mathcal{S}}\left(s', \psi \circ \phi(s')\right) \right] + L_{\mathcal{R}} + \beta \cdot \left( L_{\xi_{\bar{\pi}}} + L_{\mathbf{P}} \right), \tag{6}$$

where $d_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ is a metric over $\mathcal{S}$, and $L_{\xi_{\bar{\pi}}}$ serves as a "steady-state regulariser," minimising the distance between the prior over latent transitions and the observed transitions, thereby ensuring sufficient spread over the latent space. Thus, $L_{\xi_{\bar{\pi}}}$ ensures sufficient coverage of the encoder $\phi$ over the latent state space $\overline{\mathcal{S}}$ further preventing collapse issues [12, 20]. Intuitively, the goal of $d_{\mathcal{S}}$ is to measure the distance between original states and those reconstructed via $\psi$. Importantly, WAE-MDPs are accompanied by bisimulation guarantees:

Theorem 3.1 ([13]). *Minimizing the reward and transition losses in WAE-MDPs almost surely (1) yields bisimilar models and (2) ensures that states with the same representation are bisimilarly close in $\mathcal{M}$:*

*(1)* (Abstraction quality) *For all states $s \in \mathcal{S}$,*

$$\mathop{\mathbb{E}}_{s \sim \xi_{\bar{\pi}}} \widetilde{d}_{\bar{\pi}}(s, \phi(s)) \leq \frac{L_{\mathcal{R}} + \gamma K \cdot L_{\mathbf{P}}}{1 - \gamma};$$

*(2)* (Representation quality) *For all states $s_1, s_2$ grouped to the same representation $\phi(s_1) = \phi(s_2)$,*

$$\widetilde{d}_{\bar{\pi}}(s_1, s_2) \leq \frac{L_{\mathcal{R}} + \gamma K \cdot L_{\mathbf{P}}}{1 - \gamma} \cdot \left( \frac{1}{\xi_{\bar{\pi}}(s_1)} + \frac{1}{\xi_{\bar{\pi}}(s_2)} \right),$$

*where $K = 2\|\mathcal{R}\|_\infty / (1 - \gamma)$.*

## 4 LEARNING A MODEL THROUGH OPTIMAL TRANSPORT

In this section, we describe how a model can be learned using an optimal transport formulation. We show that this formulation generalizes the one introduced in WAE-MDPs. Additionally, we discuss the issues introduced by the reconstruction terms prevalent in WAE-MDPs, which also appear in other model-based methods.

### 4.1 Problem Formulation

The fundamental objective of model-based reinforcement learning (MBRL) is to learn a model that facilitates policy improvement without relying on real-world transitions. This can be framed as learning a model whose generated transitions are, in a well-defined sense, "close" to those of the real environment. This viewpoint naturally aligns with an optimal transport formulation.

Let $\mathcal{M}$ and $\overline{\mathcal{M}}$ denote an MDP and a latent MDP, respectively and $\mathcal{T} = \{ \langle s, a, r, s' \rangle : s' \in \text{supp}(\mathbf{P}(\cdot \mid s, a)) \text{ and } r = \mathcal{R}(s, a) \}$ denote the set of transitions in $\mathcal{M}$. We define $\overline{\mathcal{T}}$ similarly in $\overline{\mathcal{M}}$. To measure how $\overline{\mathcal{M}}$ differs from $\mathcal{M}$ via the OT, let us consider

*the disjoint union* of the transition space of both models, denoted $\mathcal{T}^* = \mathcal{T} \uplus \overline{\mathcal{T}}$. To quantify the cost of transforming a real transition $\tau \in \mathcal{T}$ into a latent transition $\bar{\tau} \in \overline{\mathcal{T}}$, we introduce a cost function $c^* : \mathcal{T}^* \times \mathcal{T}^* \to \mathbb{R}_{\geq 0}$, as required in OT problems.

By a slight abuse of notation, we let $\xi_\pi$ and $\bar{\xi}_{\bar{\pi}}$ denote the stationary distributions over transitions induced by policies $\pi$ and $\bar{\pi}$ in $\mathcal{M}$ and $\overline{\mathcal{M}}$, respectively, rather than merely the distributions over states. Our objective is to learn a latent reward function $\overline{\mathcal{R}}$ and a latent transition function $\bar{\mathbf{P}}$ such that the induced stationary distributions minimise the OT:

$$\mathrm{OT}_{c^*}\left(\xi_\pi, \bar{\xi}_{\bar{\pi}}\right) = \inf_{\vartheta \in \Gamma\left(\xi_\pi, \bar{\xi}_{\bar{\pi}}\right)} \mathop{\mathbb{E}}_{(\tau,\bar{\tau})\sim\vartheta} c^*(\tau, \bar{\tau}) \tag{7}$$

This is in contrast to WAE-MDPs, for which the cost function coincides with a distance metric solely defined over the original spaces. WAE-MDPs formulation thus boils down to recovering the Wasserstein distance instead of the general OT formulation, which requires reconstructing latent transitions to the original space.

While Eq. (7) is theoretically elegant, finding the optimal coupling between both distributions is hardly tractable in practice. Furthermore, it requires distinct real and latent policies, which complicates the learning process. A common technique in MBRL is instead to learn an *encoder* $\phi : \mathcal{S} \to \overline{\mathcal{S}}$ that encodes real states into a latent representation. Importantly, as already mentioned, such an encoder allows executing a latent policy in the real environment by first mapping states to latent states. We will thus simply learn a single latent policy $\bar{\pi}$ and consider both $\xi_{\bar{\pi}}$ and $\bar{\xi}_{\bar{\pi}}$.

Bousquet et al. [5] demonstrated that the optimal transport problem posed in Eq. (7) can be reformulated to avoid directly optimising the infimum coupling. Instead, given a space over latent variables $\mathcal{Z}$, one can optimise over encoders $q \colon \mathcal{T} \to \Delta(\mathcal{Z})$ whose marginal distribution $Q(\cdot) = \mathbb{E}_{\tau\sim\xi_{\bar{\pi}}}\, q(\cdot \mid \tau)$ matches a given prior $P \in \Delta(\mathcal{Z})$. As a specific instantiation of this theorem, we consider a latent space $\mathcal{Z} = \overline{\mathcal{S}} \times \mathcal{A} \times \overline{\mathcal{S}}$ with *fixed* prior $P$. We assume this prior is ruled by underlying latent dynamics $\bar{\mathbf{P}}$, from which one can "renconstruct" the rewards via the latent reward function $\overline{\mathcal{R}}$, obtaining

$$\mathrm{OT}_c\left(\xi_{\bar{\pi}}, \bar{\xi}_{\bar{\pi}}\right) = \inf_{q\colon Q=P} \mathop{\mathbb{E}}_{\tau\sim\xi_{\bar{\pi}}} \mathop{\mathbb{E}}_{\bar{s},\bar{a},\bar{s}'\sim q(\cdot|\tau)} c^*\left(\tau, \bar{\tau} = \left\langle \bar{s}, \bar{a}, \overline{\mathcal{R}}(\bar{s}, \bar{a}), \bar{s}' \right\rangle\right). \tag{8}$$

Note that $\overline{\mathcal{R}}$ and $\bar{\mathbf{P}}$ are learned. In practice, we consider deterministic encoders $q$ such that $q(s, a, s', r) = \langle \phi(s), a, \phi(s') \rangle$. Recall that the stationary distribution of the latent model $\bar{\xi}_{\bar{\pi}}$ depends on a learned transition function $\bar{\mathbf{P}}$ and reward function $\overline{\mathcal{R}}$. Intuitively, as $P$ is a prior over (rewardless) transitions drawn in $\overline{\mathcal{M}}$, requiring the marginal distribution of the encoded real transitions to match $P$ enforces the stationary distribution $\bar{\xi}_{\bar{\pi}}$ to mimic the dynamics of $\mathcal{M}$. This allows obtaining a robust abstraction of the real environment. Then, a well-defined cost $c^*$, that allows measuring the distance between original and latent rewards, permits to learn an appropriate generative model of the reward signal.

So, to obtain a strong model of the environment, all we require is to learn $\bar{\mathbf{P}}$, $\overline{\mathcal{R}}$, and an encoder $\phi$ that minimises the optimal transport as defined in Eq. (8). However, solving this OT problem remains challenging due to the strict constraint on the marginal distribution

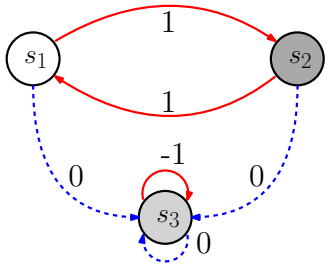



**Figure 2: An MDP illustrating the issues with the reconstruction terms in $c_{\mathrm{WAE}}$.**

of the encoder. Instead, we introduce a *regularised OT* formulation, replacing the infimum constraint with a regularisation term:

$$\mathrm{ROT}_c\left(\xi_{\bar{\pi}}, \bar{\xi}_{\bar{\pi}}\right) = \inf_{\phi,\bar{\mathbf{P}},\overline{\mathcal{R}}} \mathop{\mathbb{E}}_{\tau\sim\xi_{\bar{\pi}}} \left[c^*\left(\tau, \overline{\mathcal{R}} \oplus q(\tau)\right)\right] + \lambda \cdot D(Q, P) \tag{9}$$

where $\lambda > 0$ is a scale factor, $D$ is a discrepancy between the distributions, and $\overline{\mathcal{R}} \oplus q$ is the function mapping any transition $\tau = \langle s, a, r, s' \rangle$ to $q(\tau) = \left\langle \bar{s}, a, \bar{s}' \right\rangle$ and augmenting the resulting tuple with the latent reward $\bar{r} = \overline{\mathcal{R}}(\bar{s}, a)$. This yields the latent transition $\bar{\tau} = \left\langle \bar{s}, a, \bar{r}, \bar{s}' \right\rangle$. We define the discrepancy $D(Q, P)$ as the Wasserstein distance between $Q$ and $P$. As mentioned, $c^*$ should embed the distance between original and generated rewards. Accordingly, let $c^* = c + c_{\mathcal{R}}$ where $c$ is, again, some cost function over $\mathcal{T}^*$ and $c_{\mathcal{R}}(\tau, \tau') = \lambda \cdot \left|\mathrm{proj}_{\mathcal{R}}(\tau) - \mathrm{proj}_{\mathcal{R}}(\tau')\right|$, where $\mathrm{proj}_{\mathcal{R}}(\tau) = r$ is the projection of $\tau = \langle s, a, r, s' \rangle$ on the reward space. By a previous result from Delgrange et al. [11], we obtain the following final optimisation problem:

$$\inf_{\phi,\bar{\mathbf{P}},\overline{\mathcal{R}}} \mathop{\mathbb{E}}_{\tau\sim\xi_{\bar{\pi}}} \left[c\left(\tau, \overline{\mathcal{R}} \oplus q(\tau)\right)\right] + \lambda \left(L_{\xi_{\bar{\pi}}} + L_{\mathbf{P}} + L_{\mathcal{R}}\right). \tag{10}$$

As this loss includes both the local reward loss $L_{\mathcal{R}}$ and the local transition loss $L_{\mathbf{P}}$, minimising it results in learning a bisimilar model. Furthermore, since all theoretical guarantees for WAE-MDPs are based on these local losses (Thm. 3.1), they naturally extend to this novel formulation.

## 4.2 The Problem With Reconstruction

As shown, optimising Eq. (10) is sufficient to learn a bisimilar latent model of the real environment. Notably, there are no inherent constraints on the cost function, allowing it to be chosen freely. However, in practice, we ideally enforce that if a transition from the real MDP is encoded into a latent bisimilar MDP, the cost function assigns zero cost, ensuring high model fidelity. We define a cost function satisfying this property as *recognising* and formalise it below.

Property 4.1 (Recognising). *Given an optimal encoder $\phi^*$ such that $\phi^*([s]_\sim) = \{\bar{s}\}$,*

$$c\left(\langle s, a, s', r\rangle, \left\langle \phi^*(s), a, \phi^*(s'), r \right\rangle\right) = 0 \quad \forall \langle s, a, s', r\rangle \in \mathcal{T}. \tag{11}$$

Looking back to WAE-MDP, we can recover the implicit cost function as follows:

$$c_{\mathrm{WAE}}(\tau, \bar{\tau}) = d_{\mathcal{S}}\left(s, \psi(\bar{s})\right) + d_{\mathcal{S}}\left(s', \psi(\bar{s}')\right). \tag{12}$$

where $d_{\mathcal{S}}$ is usually assumed to be the Euclidean distance. This cost function measures the distance between a real state and a decoded latent state. While this cost function is conceptually straightforward, we demonstrate in the following theorem that it is not recognising.

Theorem 4.2. *The cost function $c_{WAE}$ as shown in Eq.* (12) *is not recognising.*

Proof. We illustrate that $c_{\text{WAE}}$ is not recognising using the MDP shown in Fig. 2. Suppose the initial state is $s_1$ and $\pi(\text{red}, s_1) = \pi(\text{red}, s_2) = 1$. Clearly, $s_1$ and $s_2$ are in $\pi$-bisimulation and can be represented by a single latent state $\bar{s}_1$. However, $\psi(\bar{s}_1)$ can at best reconstruct the true state correctly 50% of the time. Since $c_{\text{WAE}}$ measures the distance between $s$ and $\psi(\bar{s})$, $c_{\text{WAE}} > 0$ for half of the generated transitions and is therefore not recognising. □

In practice, this issue is further exacerbated due to WAE-MDPs minimising the $L_2$ distance between $s$ and $\psi \circ \phi(s)$, which can introduce unintended side effects. For instance, a decoder that minimises the reconstruction distance between the white $s_1$ and dark $s_2$ may decode $\bar{s}_1$ and $\bar{s}_2$ to grey $s_3$, as it minimises the $L_2$ distance to both white and dark. This hinders the cost function's ability to accurately evaluate whether the encoding is effective. Notably, while $V^*(s_1) = V^*(s_2) = \frac{1}{1-\gamma}$ and $V^*(s_3) = 0$, it follows that $V^*(\psi \circ \phi(s_1)) = V^*(s_3)$, further demonstrating the issue.

Consequently, cost functions that measure the distance between a real and reconstructed state can only be recognising for latent models whose state space permits a one-to-one mapping to the real state space. However, since the goal is typically to learn a more compact representation, reconstruction-based approaches are inherently unsuitable for this purpose.

### 4.3 Cost Function Analysis

We present a preliminary analysis of cost functions that may be more suitable for learning a bisimilar model than prior reconstruction-based approaches. Specifically, we consider cost functions that depend only on the state and latent state of transitions, defined as

$$c(\tau, \bar{\tau}) = h(f(s), g(\bar{s})). \tag{13}$$

where $h$ computes the cost after applying a transformation $f$ to the real state and another transformation $g$ to the latent state. For such a cost function, we establish the following result:

Theorem 4.3. *Let $c$ be a cost function as defined in Eq.* (13) *and $\phi^*$ be an optimal encoder such that $\phi^*([s]_\sim) = \bar{s}$. Then the following properties hold,*

1. *When $h$ is a metric, $c$ is recognising if and only if $f([s]_\sim) = \{z\}$ and $g \circ \phi^*(s) = z$ ;*
2. *For a recognising cost function $c$, it suffices to learn a deterministic encoder $\phi^*$ .*

Proof. By definition of a metric, $h(x, y) = 0$ if and only if $x = y$. From the definition of the recognising property, this implies that $c$ is recognising if and only if for all $s \in \mathcal{S}$, $f(s) = g(\phi^*(s))$. From this, (1) follows immediately, as it ensures that $c$ is recognising if and only if $f([s]_\sim) = \{z\}$ and $g \circ \phi^*(s) = z$. (2) is a straightforward consequence of the equivalence between lax bisimulation and MDP homomorphism [37] and the fact that the cost function is recognising and therefore attains a minimum with the optimal encoder. □

Although WAE-MDP also considered deterministic encoders, this choice was justified on the grounds of simplicity rather than principle. Moreover, when relying on reconstruction, as in WAE-MDPs, a deterministic encoder can perform significantly worse than a stochastic encoder as demonstrated in Section 4.2.

## 5 OPTIMAL TRANSPORT MARKOV DECISION PROCESSES

We introduce OT-MDPs, a generalisation and simplification of the WAE-MDP framework, that learns a principled model through our optimal transport formulation. In contrast to WAE-MDPs, which focus on policy distillation, our approach unifies model and policy learning and incorporates architectural improvements that stabilise training and reduce the number of hyperparameters.

### 5.1 OT-MDP Cost Functions

One advantage of allowing general cost functions, is that we can introduce additional feedback to guide the encoder during optimisation. Specifically, when training an RL agent within the model, the learned encodings should facilitate effective learning. To this end, we introduce the concept of *value equivalence*, which utilises the value function as a cost function.

$$c_{\text{value}}(s, \bar{s}) = \left| V^{\pi}(s) - \overline{V}^{\pi}(\bar{s}) \right| \tag{14}$$

Proposition 5.1. *The cost function $c_{value}$ as shown in Eq.* (14) *is recognising.*

This function assumes access to the value function for real states, $V^{\overline{\pi}}$, and measures its distance to a latent value function, $\overline{V}^{\overline{\pi}}$. Notably, when $\mathcal{M} \sim \overline{\mathcal{M}}$, it follows that $V^{\overline{\pi}}(s) = \overline{V}^{\overline{\pi}}(\phi(s))$ [22, 30], ensuring that $c_{\text{value}}$ is recognising. Crucially, this cost function eliminates the need to train additional reconstruction networks. In contrast, one could be tempted to only learn a single value function and incorporate a decoder network, i.e. $c(s, \bar{s}) = \left| V^{\overline{\pi}}(s) - V^{\overline{\pi}}(\psi(\bar{s})) \right|$. However, this cost function is also not recognising, as demonstrated in Section 4.2.

### 5.2 Learning a Model and a Policy

In this work, we propose using the representation generated by the OT-MDP for policy learning. Since the OT-MDP framework learns a bisimilar discrete abstraction of the real environment, we hypothesise that directly learning a policy on latent states leads to more efficient learning. Below, we outline the architecture of our algorithm and its training methodology. Complete pseudocode is provided in Algorithm 1. Additionally, we provide a schematic illustrating the different losses used to train OT-MDPs and the RL agent, highlighting their respective influences on various components of the framework.

**Architecture**. To map real states to latent states, we train an encoder $\phi$. The latent MDP consists of a latent reward function $\overline{\mathcal{R}} : \overline{\mathcal{S}} \times \mathcal{A} \to \mathbb{R}$ and a latent transition function $\overline{\mathbf{P}} : \overline{\mathcal{S}} \times \mathcal{A} \to \Delta(\overline{\mathcal{S}})$. In the OT-MDP framework, $\overline{\mathbf{P}}$ is only required for sampling, rather

**Algorithm 1** Optimal Transport Auto-Encoded MDP with DQN

**Input:** no. of maximiser iterations $N_{\max}$, regulariser scale factor $\lambda$ and OT scale factor $\beta$
**Output:** A latent MDP $\overline{\mathcal{M}}$ and policy $\bar{\pi}$

Initialise $\alpha = 1$
**while** not done **do**
  Collect transitions $\tau$ and store in replay buffer $\mathcal{D}$
  **for** $j = 1$ to $N_{\max}$ **do**
    Sample batch $\mathcal{B}_{\max}$
    Maximise $L_{\xi_{\bar{\pi}}} + L_{\mathbf{p}}$ ▷ To optimize Eq. 1, details in [11]
  Sample $\mathcal{B}_c$ and update parameters $\nu$ of the cost function
  Sample batch $\mathcal{B}_{\min}$
  Compute DQN targets $\overline{V}_{\text{DQN}}$ on $\mathcal{B}_{\min}$
  Compute planning targets $\overline{V}_{\text{plan}}$ on $\mathcal{B}_{\min}$ ▷ See Section 6
  $\overline{V}_{\text{target}} \leftarrow \alpha \overline{V}_{\text{DQN}} + (1-\alpha)\overline{V}_{\text{plan}}$
  Compute $L_{\text{RL}}$ using with $\overline{V}_{\text{target}}$
  $L_{\text{OT}} \leftarrow \mathbb{E}_{\tau \sim \xi_{\bar{\pi}}}\, c\Big(\tau, \overline{\mathcal{R}} \oplus q(\tau)\Big) + \lambda \left(L_{\xi_{\bar{\pi}}} + L_{\mathbf{p}} + L_{\mathcal{R}}\right)$
  Minimise combined loss: $(1-\beta)L_{\text{RL}} + \beta L_{\text{OT}}$
  $\alpha \leftarrow \text{clip}\left(L_{\mathbf{p}}, 0, 1\right)$

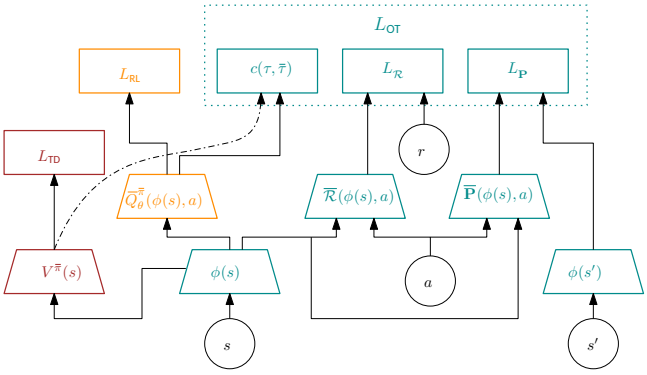



**Figure 3: A diagram of the different losses used to train the components of the latent MDP $\overline{\mathcal{M}}$ and policy $\bar{\pi}$.**

than computing transition probabilities explicitly. Therefore, we learn a generative model of $\overline{\mathbf{P}}$ using an Inverse Autoregressive Flow (IAF, 29). IAFs leverage masked autoregressive distribution estimators [21], allowing them to approximate arbitrary distributions instead of being restricted to normal or deterministic ones, as is common in related work. Moreover, IAFs enable efficient sampling with a single forward pass through the network, significantly reducing computational overhead.

For reinforcement learning, OT-MDPs can in principle be paired with any algorithm. In this work, we employ DQN [31], trained on latent states produced by the encoder. Specifically, we train a latent critic $\overline{V}^{\bar{\pi}}$ and use epsilon-greedy action selection during training. The cost function $c_{\text{value}}$ in the OT objective computes the difference between a learned value of the real state $V^{\bar{\pi}}(s)$ and the value of the latent state $\overline{V}^{\bar{\pi}}(\bar{s})$. Notably, the critic learned by the reinforcement learning algorithm can be reused, eliminating the need for training an additional latent critic.

**Training**. The OT-MDP training process alternates between collecting transitions from the policy and performing joint updates for the model and policy. Before updating these components, we train the cost function $c_{\text{value}}$ that depends on a real and latent critic. While OT-MDPs theoretically require the latent states to be discrete, in practice, we adopt the same relaxation as Delgrange et al. [11] for stability. These continuous approximations ensure that the entire architecture remains differentiable. Crucially, we gradually anneal this relaxation so that the latent states progressively approach discreteness. Subsequently, we optimise the encoder $\phi$, the latent transition and reward networks $\overline{\mathbf{P}}, \overline{\mathcal{R}}$, and the policy and critic networks $\bar{\pi}, \overline{V}^{\bar{\pi}}$. Following Castro et al. [8], we merge agent and model training into a single step by combining all relevant losses, thereby enabling the state encoder to be influenced by both the RL agent's performance and the model quality.

**Learning the Wasserstein estimator**. Recall that we learn the latent transition function by minimising its Wasserstein distance to the real transition function. Crucially, this requires learning a Lipschitz-1 function $f$ to approximate the supremum in the dual formulation. In previous work, particularly in WAE-MDPs [11], this was achieved by adding a gradient penalty that forces the network to belong to this class [24]. We propose an architectural alternative: discretising the relaxed latent state while applying a gradient using the straight-through estimator [4]. This approach ensures a discretised state space, where we obtain the convenient result that optimal transport collapses to the total variation distance, $d_{\text{TV}}$ [40]. Moreover, this distance has a convenient formulation, $d_{\text{TV}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \left|\mathbb{E}_{x \sim \mu}\left[f(x)\right] - \mathbb{E}_{y \sim \nu}\left[f(y)\right]\right|$ where $\mathcal{F}$ is the set of $1/2$-bounded functions [32]. Thus to learn $f$, rather than imposing a gradient penalty to enforce 1-Lipschitzness, we ensure that it is $1/2$-bounded through a penalty term that encourages the network to remain within the required bounds.

## 6 INTEGRATING PLANNING AND LEARNING

In this section, we propose two principled approaches to integrate learning and planning within the OT-MDP framework. Specifically, we demonstrate how the model can be used as a simulator and, under certain conditions, how it can facilitate direct policy search.

### 6.1 Model-Based Value Expansion

OT-MDPs learn a latent MDP that is $\pi$-bisimilarly close to the real MDP under policy $\pi$. Since the policy is trained on latent states, an intuitive approach is to perform additional rollouts within the learned model, effectively bypassing the need for real environment interactions and thereby improving sample efficiency. However, exact $\pi$-bisimulation is guaranteed only when local losses converge to zero, and conducting complete policy rollouts in an approximate model can lead to suboptimal results due to compounding model errors. Model-Based Value Expansion (MVE) offers a compromise by proposing finite-depth policy rollouts [16]. The resulting model-based value is then used as a value target (see Algorithm 1).

Let us first define the process through which we can sample from a model. Let $\mathbf{P}_\pi$ be the transition function of the Markov chain induced by $\pi$. Then we define sampling from this distribution

at depth $h$ as,

$$\mathbf{P}_\pi^0(\cdot \mid s) = \delta_s \quad \text{where } \delta_s \text{ is the Dirac measure with impulse } s;$$
$$\mathbf{P}_\pi^{h+1}(\cdot \mid s) = \mathop{\mathbb{E}}_{s' \sim \mathbf{P}_\pi(\cdot|s)} \left[ \mathbf{P}_\pi^h(\cdot \mid s') \right] \quad \forall h \geq 0.$$

The model-based value obtained by unrolling the policy with the latent transition function $\bar{\mathbf{P}}$ and reward function $\bar{\mathcal{R}}$ until a given depth $H$ is then defined as follows:

**Definition 6.1.** Let $\overline{V}_\theta^{\overline{\pi}}(\cdot)$ be a latent critic with parameters $\theta$. We define the model-based value function $\overline{V}_{\theta,H}^{\overline{\pi}}(\cdot)$ as the $H$-step bootstrapping of $\overline{V}_\theta^{\overline{\pi}}$ using the latent reward function $\bar{\mathcal{R}}$ and latent transition function $\bar{\mathbf{P}}$ as,

$$\overline{V}_{\theta,H}^{\overline{\pi}}(\bar{s}) = \sum_{t=0}^{H-1} \gamma^t \mathop{\mathbb{E}}_{\bar{s}_t, a_t \sim \bar{\mathbf{P}}_{\overline{\pi}}^t} \left[ \bar{\mathcal{R}}(\bar{s}_t, a_t) \right] + \gamma^H \mathop{\mathbb{E}}_{\bar{s}_H, a_H \sim \bar{\mathbf{P}}_{\overline{\pi}}^H} \left[ \overline{V}_\theta^{\overline{\pi}}(\bar{s}_H) \right] \tag{15}$$

The goal of incorporating MVE within our framework is to obtain improved value estimates for latent states $\bar{s} = \phi(s)$. Consequently, it is important to theoretically analyse the difference between the model-based value $\overline{V}_{\theta,H}^{\overline{\pi}}(\phi(s))$ and real value $V^{\overline{\pi}}(s)$. Let the *local value loss* $L_V$ be defined as the difference between the learned latent critic and the true value,

$$L_V = \mathop{\mathbb{E}}_{s \sim \xi_{\overline{\pi}}} \left[ \left| \overline{V}_\theta^{\overline{\pi}}(\phi(s)) - V^{\overline{\pi}}(s) \right| \right]. \tag{16}$$

In Theorem 6.1, we show that the difference between the true and the model-based value can be bounded by the model losses and the value loss. Importantly, all of these values can be well-approximated since they are *local*, as opposed to the global losses that are generally intractable [20]. We provide a formal proof in the appendix.

Theorem 6.1. *For any policy $\overline{\pi}$ with local losses $L_{\mathcal{R}}$, $L_{\mathbf{P}}$ and $L_V$, we obtain for any rollout horizon $H$,*

$$\mathop{\mathbb{E}}_{s \sim \xi_{\overline{\pi}}} \left[ \left| \overline{V}_{\theta,H}^{\overline{\pi}}(\phi(s)) - V^{\overline{\pi}}(s) \right| \right] \leq \frac{1-\gamma^H}{1-\gamma} \left( L_{\mathcal{R}} + \gamma K L_{\mathbf{P}} \right) + \gamma^H L_V \tag{17}$$

*where $K$ is a constant related to the model-based value function.*

This result is significant as it provides a quantifiable measure of the model's performance. However, since the local losses and value loss are computed as expectations over the stationary distribution $\xi_{\overline{\pi}}$, they provide only an aggregate estimate of the model and critic quality. This necessitates that the value gap is also obtained in expectation rather than state-dependent. Consequently, determining the appropriate planning depth involves an aggregate approach, balancing the benefits of using the model against relying on the value estimate. A straightforward corollary is that the planning horizon $H$ that minimises the bound is always either 0 or $\infty$.

Corollary 6.2. *The rollout horizon $H$ minimising the bound from Theorem 6.1, denoted as $H_{\min}$ is given by,*

$$H_{\min} = \begin{cases} 0 & \text{if} \quad L_{\mathcal{R}} + \gamma K L_{\mathbf{P}} >= L_V \\ \infty & \text{otherwise} \end{cases} \tag{18}$$

As a result, Corollary 6.2 dictates that values should be computed either entirely within the model or solely from the value function. However, model-based values may still provide useful estimates up to a limited depth. To reconcile this, we define the final value target as $\overline{V}_{\text{target}}^{\overline{\pi}}(\bar{s}) = \alpha \overline{V}_{\text{DQN}}^{\overline{\pi}}(\bar{s}) + (1-\alpha)\overline{V}_{\text{MVE}}^{\overline{\pi}}(\bar{s})$, where $\overline{V}_{\text{DQN}}^{\overline{\pi}}$ is the standard DQN value target, $\overline{V}_{\text{MVE}}^{\overline{\pi}}$ is the value target from MVE, and $\alpha$ controls the relative influence of the model. We dynamically adjust $\alpha$ based on the transition loss $L_{\mathbf{P}}$, a strong indicator of model quality, defining $\alpha = \text{clip}\left(L_{\mathbf{P}}, 0, 1\right)$. Intuitively, lower transition losses suggest a more accurate model, allowing for greater reliance on $\overline{V}_{\theta,H}^{\overline{\pi}}$. This strategy also mitigates excessive dependence on the model during early learning. We emphasise that extending Theorem 6.1 to be state-specific could offer significant algorithmic advantages.

## 6.2 Monte-Carlo Tree Search

Given that we learn a model of the environment, it is natural to consider moving beyond imagined policy rollouts to direct policy search. However, OT-MDPs are trained to approximate a $\pi$-bisimilar model, which depends on the data-collection policy. Moreover, all theoretical guarantees relate the real and latent MDPs under the behaviour policy. Consequently, deviating from this policy may yield inaccurate value estimates for candidate policies. Additionally, if certain regions of the real MDP are absent from the stationary distribution of the behaviour policy, relying on the model in these regions is unlikely to be effective.

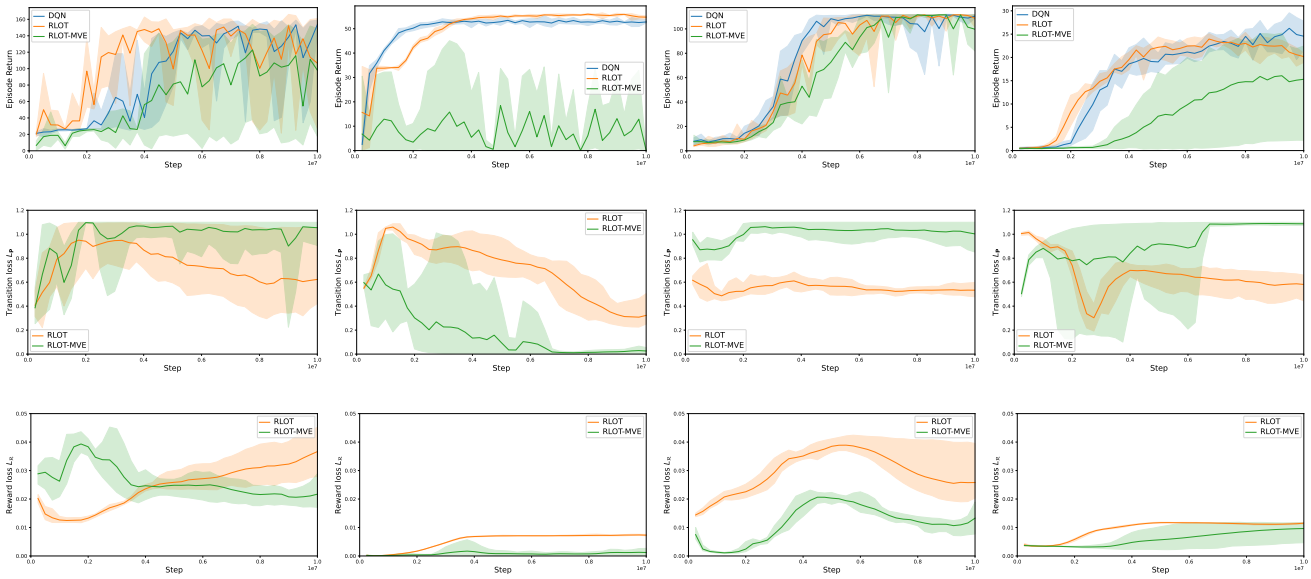
To address this, we show that ensuring full support in each state is sufficient to theoretically guarantee that all states of the real MDP are represented in the support of the stationary distribution. This, in turn, implies that when all losses converge to zero, the model achieves exact bisimulation with the real MDP. Unlike $\pi$-bisimulation, exact bisimulation preserves the values of the optimal policy, making direct policy search viable. Notably, the full support assumption is also common in maximum entropy RL and robust RL [15], highlighting intriguing connections for future work. We provide a complete proof for Theorem 6.3 in Appendix B

Theorem 6.3. *Let $\mathcal{M}$ be an ergodic MDP and assume that the latent policy $\overline{\pi}$ has full support in every state, i.e. $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$, $\overline{\pi}(a|\phi(s)) > 0$. For a latent MDP $\overline{\mathcal{M}}$, the local losses defined in Eq. (4) are zero if and only if $\mathcal{M} \sim \overline{\mathcal{M}}$.*

While this theorem guarantees exact bisimulation in the limit, it is unlikely to hold throughout training. Therefore, we adopt a similar principle as in Section 6.1, balancing the regular DQN target and the MCTS target using the transition loss. This approach intuitively encourages agents to act according to the optimal policy while mitigating overreliance on the model.

# 7 EARLY-STAGE EVALUATION

We compare the learning dynamics of a standard DQN agent to a DQN agent trained jointly with OT-MDPs, which operates directly on the learned latent representation and uses the value equivalence cost function (see Eq. (14)). We refer to this approach as Reinforcement Learning with Optimal Transport (RLOT). To evaluate its




**(a) Space Invaders** **(b) Freeway** **(c) Breakout** **(d) Asterix**

**Figure 4: The evaluation return of the greedy latent policy $\bar{\pi}$ (top); the local transition loss $L_{\mathbf{P}}$ (middle); and the local reward loss $L_{\mathcal{R}}$ (bottom).**

performance, we test both agents across four MinAtar environments [42]. Our implementation is completely in JAX [6], ensuring efficient, batched execution on the GPU. We evaluate all experiments across five seeds on all environments.

## 7.1 Learning a Model and Policy

We present our results on Space Invaders, Freeway, Breakout, and Asterix in Fig. 4. In all environments, RLOT matches the performance of DQN while simultaneously learning a strong model of the environment. However, Asterix proves significantly more challenging, likely due to suboptimal hyperparameter selection. We do observe difficulties in model learning for RLOT in Space Invaders, alongside instability in episode returns. Addressing this challenge remains an important direction for future work.

## 7.2 Evaluating Planning

We conduct an initial evaluation of model-based value expansion (MVE) within our framework, deferring the evaluation of MCTS to future work. To isolate the impact of MVE, we perform 32 model-based rollouts with $H = 1$, ensuring the smallest possible incremental difference between using MVE and not. Notably, we find that the agent leveraging MVE achieves performance comparable to RLOT, except in Freeway where it fails to find an effective policy. However, upon inspecting the model losses, it is clear it only reaches similar performance to RLOT and DQN when the transition loss, $L_{\mathbf{P}}$, is approximately equal to 1. This is due to our introduction of the parameter $\alpha = \text{clip}\left(L_{\mathbf{P}}, 0, 1\right)$, which controls the influence of MVE on the value target. When $L_{\mathbf{P}} = 1$, MVE is effectively excluded and has no impact on learning. This outcome suggests potential issues, such as premature reliance on the model or a faulty implementation, warranting further investigation. For future work, we aim to evaluate this further.

## 8 CONCLUSION

We introduce OT-MDPs, a principled extension of WAE-MDPs that moves beyond the limitations of reconstruction-based objectives. By formulating model-based RL through optimal transport, we provide a flexible framework that enables the design of custom cost functions to directly shape latent representations for downstream tasks. This approach not only enhances model quality but also facilitates direct policy learning within the latent space. A key contribution of OT-MDPs is the unification of model and policy learning, allowing for a consolidated training process. Additionally, we integrate two planning methods, model-based value expansion and Monte Carlo tree search, within OT-MDPs and provide rigorous guarantees that justify their use. Empirical results demonstrate that OT-MDPs, when combined with DQN, yield strong policies while leveraging the learned model effectively. Notably, our experimental results with model-based value expansion highlight the importance of balancing model quality and exploitation during training. In future work, we aim to expand our experimental evaluation and explore architectural refinements that further stabilise learning.

## ACKNOWLEDGMENTS

WR and RA are supported by the Research Foundation – Flanders (FWO), grant numbers 1197622N and 11F5721N. This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

# REFERENCES


[1] Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G. Bellemare. 2021. Contrastive Behavioral Similarity Embeddings for Generalization in Reinforcement Learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

[2] Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 214–223. http://proceedings.mlr.press/v70/arjovsky17a.html

[3] Raphaël Avalos, Florent Delgrange, Ann Nowe, Guillermo Perez, and Diederik M Roijers. 2024. The Wasserstein Believer: Learning Belief Updates for Partially Observable Environments through Reliable Latent Space Models. In *The Twelfth International Conference on Learning Representations*.

[4] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. 2013. Estimating or Propagating Gradients through Stochastic Neurons for Conditional Computation. *CoRR* abs/1308.3432 (2013). arXiv:1308.3432

[5] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. 2017. From Optimal Transport to Generative Modeling: The VEGAN Cookbook. arXiv:1705.07642 [stat.ML]

[6] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. *JAX: composable transformations of Python+NumPy programs*. http://github.com/jax-ml/jax

[7] Pablo Samuel Castro. 2020. Scalable Methods for Computing State Similarity in Deterministic Markov Decision Processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10069–10076.

[8] Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. 2021. MICo: Improved Representations via Sampling-Based State Similarity for Markov Decision Processes. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 30113–30126.

[9] Dane S. Corneil, Wulfram Gerstner, and Johanni Brea. 2018. Efficient Model-Based Deep Reinforcement Learning with Variational State Tabulation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1057–1066. http://proceedings.mlr.press/v80/corneil18a.html

[10] Florent Delgrange, Guy Avni, Anna Lukina, Christian Schilling, Ann Nowe, and Guillermo Perez. 2025. Composing Reinforcement Learning Policies, with Formal Guarantees. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*.

[11] Florent Delgrange, Ann Nowe, and Guillermo Perez. 2023. Wasserstein Auto-Encoded MDPs: Formal Verification of Efficiently Distilled RL Policies with Many-Sided Guarantees. In *The Eleventh International Conference on Learning Representations*.

[12] Florent Delgrange, Ann Nowé, and Guillermo A. Pérez. 2022. Distillation of RL Policies with Formal Guarantees via Variational Abstraction of Markov Decision Processes. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 6 (June 2022), 6497–6505. https://doi.org/10.1609/aaai.v36i6.20602

[13] Florent Delgrange, Ann Nowé, and Guillermo A. Pérez. 2022. Distillation of RL Policies with Formal Guarantees via Variational Abstraction of Markov Decision Processes. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 6497–6505. https://doi.org/10.1609/AAAI.V36I6.20602

[14] Florent Delgrange, Mathieu Reymond, Ann Nowe, and Guillermo A. Pérez. 2023. WAE-PCN: Wasserstein-autoencoded Pareto Conditioned Networks. In *Proc. of the Adaptive and Learning Agents Workshop (ALA 2023)* (15 ed.), Francisco Cruz, Conor F. Hayes , Caroline Wang, and Connor Yates (Eds.), Vol. https://alaworkshop2023.github.io/. 1–7. https://alaworkshop2023.github.io 2023 Adaptive and Learning Agents Workshop at AAMAS, ALA 2023 ; Conference date: 29-05-2023 Through 30-05-2023.

[15] Benjamin Eysenbach and Sergey Levine. 2022. Maximum Entropy RL (Provably) Solves Some Robust RL Problems. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

[16] Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I. Jordan, Joseph E. Gonzalez, and Sergey Levine. 2018. Model-Based Value Estimation for Efficient Model-Free Reinforcement Learning. arXiv:1803.00101 [cs.LG]

[17] Norm Ferns, Prakash Panangaden, and Doina Precup. 2004. Metrics for Finite Markov Decision Processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI '04)*. AUAI Press, Arlington, Virginia, USA, 162–169.

[18] Norm Ferns, Prakash Panangaden, and Doina Precup. 2011. Bisimulation Metrics for Continuous Markov Decision Processes. *SIAM J. Comput.* 40, 6 (2011), 1662–1714. https://doi.org/10.1137/10080484X arXiv:https://doi.org/10.1137/10080484X

[19] Vincent François-Lavet, Yoshua Bengio, Doina Precup, and Joelle Pineau. 2019. Combined Reinforcement Learning via Abstract Representations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 3582–3589. https://doi.org/10.1609/AAAI.V33I01.33013582

[20] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. 2019. DeepMDP: Learning Continuous Latent Space Models for Representation Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2170–2179.

[21] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. 2015. MADE: Masked Autoencoder for Distribution Estimation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings, Vol. 37)*, Francis R. Bach and David M. Blei (Eds.). JMLR.org, 881–889.

[22] Robert Givan, Thomas L. Dean, and Matthew Greig. 2003. Equivalence Notions and Model Minimization in Markov Decision Processes. *Artificial Intelligence* 147, 1-2 (2003), 163–223. https://doi.org/10.1016/S0004-3702(02)00376-4

[23] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5767–5777. https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html

[24] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5767–5777.

[25] Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. 2021. Mastering Atari with Discrete World Models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=0oabwyZbOu

[26] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2024. Mastering Diverse Domains through World Models. arXiv:2301.04104 [cs.AI]

[27] Bojun Huang. 2020. Steady State Analysis of Episodic Reinforcement Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/69bfa2aa2b7b139ff581a806abf0a886-Abstract.html

[28] Mete Kemertas and Tristan Aumentado-Armstrong. 2021. Towards Robust Bisimulation Metric Learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 4764–4777.

[29] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved Variational Inference with Inverse Autoregressive Flow. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4743–4751.

[30] Kim Guldstrand Larsen and Arne Skou. 1991. Bisimulation through Probabilistic Testing. 94, 1 (1991), 1–28. https://doi.org/10.1016/0890-5401(91)90030-6

[31] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529–533. https://doi.org/10.1038/nature14236

[32] Alfred Müller. 1997. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability* 29, 2 (1997), 429–443. http://www.jstor.org/stable/1428011

[33] D. Revuz. 1984. *Markov Chains* (second (revised) ed.). Elsevier Science Publishers B.V. https://books.google.be/books?id=zIDWlAEACAAJ

[34] Sahand Rezaei-Shoshtari, Rosie Zhao, Prakash Panangaden, David Meger, and Doina Precup. 2022. Continuous MDP Homomorphisms and Homomorphic Policy Gradient. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 20189–20204.

[35] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. 2020. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature* 588, 7839 (2020), 604–609. https://doi.org/10.1038/s41586-020-03051-4

[36] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). MIT press, Cambridge, MA.

[37] Jonathan Taylor, Doina Precup, and Prakash Panagaden. 2008. Bounding Performance Loss in Approximate MDP Homomorphisms. In *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), Vol. 21. Curran Associates, Inc.

[38] Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. 2018. Wasserstein Auto-Encoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=HkL7n1-0b

[39] Elise van der Pol, Thomas Kipf, Frans A. Oliehoek, and Max Welling. 2020. Plannable Approximations to MDP Homomorphisms: Equivariance under Actions. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1431–1439.

[40] Cédric Villani et al. 2009. *Optimal Transport: Old and New*. Vol. 338. Springer.

[41] Chenjun Xiao, Yifan Wu, Chen Ma, Dale Schuurmans, and Martin Müller. 2019. Learning to Combat Compounding-Error in Model-Based Reinforcement Learning. *CoRR* abs/1912.11206 (2019). arXiv:1912.11206

[42] Kenny Young and Tian Tian. 2019. MinAtar: An Atari-Inspired Testbed for Thorough and Reproducible Reinforcement Learning Experiments. arXiv:1903.03176 [cs.LG] https://arxiv.org/abs/1903.03176

[43] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. 2021. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *International Conference on Learning Representations*.

## A PROOF OF THEOREM 6.1.

We first restate Theorem 6.1 for clarity:

Theorem A.1. *For any policy $\overline{\pi}$ with local losses $L_{\mathcal{R}}$, $L_{\mathbf{P}}$ and $L_V$, we obtain for any rollout horizon $H$,*

$$\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\left[\left|\overline{V}^{\overline{\pi}}_{\theta,H}(\phi(s)) - V^{\overline{\pi}}(s)\right|\right] \leq \frac{1-\gamma^H}{1-\gamma}\left(L_{\mathcal{R}} + \gamma K L_{\mathbf{P}}\right) + \gamma^H L_V \tag{19}$$

*where $K$ is a constant related to the model-based value function.*

Proof. Our proof follows a similar pattern as the proof presented by Xiao et al. [41]. Let us define $U_h$ as follows, noting that $\mathbf{P}^0_{\overline{\pi}} = s$

$$U_h(s) = \sum_{t=0}^{h-1}\gamma^t \mathop{\mathbb{E}}_{s_t,a_t\sim\mathbf{P}^t_{\overline{\pi}}}[\mathcal{R}(s_t,a_t)] + \sum_{t=h}^{H-1}\gamma^t \mathop{\mathbb{E}}_{\tilde{s}_t,a_t\sim\bar{\mathbf{P}}^{t-h}_{\overline{\pi}}\circ\mathbf{P}^h_{\overline{\pi}}}\left[\overline{\mathcal{R}}(\tilde{s}_t,a_t)\right] + \gamma^H \mathop{\mathbb{E}}_{\tilde{s}_H,a_H\sim\bar{\mathbf{P}}^{H-h}_{\overline{\pi}}\circ\mathbf{P}^h_{\overline{\pi}}}\left[\overline{V}^{\overline{\pi}}_{\theta}(\tilde{s}_H)\right] \tag{20}$$

In what follows, we are going to make a telescopic sum argument. For this, we first demonstrate two ways of writing $U_h$:

$$U_h(s) = \sum_{t=0}^{h-1}\gamma^t \mathop{\mathbb{E}}_{s_t,a_t\sim\mathbf{P}^t_{\overline{\pi}}}[\mathcal{R}(s_t,a_t)] + \gamma^h \mathop{\mathbb{E}}_{s_h\sim\mathbf{P}^h_{\overline{\pi}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h}(\phi(s_h))\right] \tag{21}$$

and also as,

$$U_h(s) = \sum_{t=0}^{h-1}\gamma^t \mathop{\mathbb{E}}_{s_t,a_t\sim\mathbf{P}^t_{\overline{\pi}}}[\mathcal{R}(s_t,a_t)] + \gamma^h \mathop{\mathbb{E}}_{s_h,a_h\sim\mathbf{P}^h_{\overline{\pi}}}\left[\overline{\mathcal{R}}(\phi(s_h),a_h)\right] + \gamma^{h+1} \mathop{\mathbb{E}}_{\tilde{s}_{h+1}\sim\bar{\mathbf{P}}\circ\mathbf{P}^h_{\overline{\pi}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}_{h+1})\right] \tag{22}$$

Let us use the first one for $U_{h+1}$ and the second one for $U_h$:

$$\begin{aligned}
U_h(s) - U_{h+1}(s) &= \sum_{t=0}^{h-1}\gamma^t \mathop{\mathbb{E}}_{s_t,a_t\sim\mathbf{P}^t_{\overline{\pi}}}[\mathcal{R}(s_t,a_t)] + \gamma^h \mathop{\mathbb{E}}_{s_h,a_h\sim\mathbf{P}^h_{\overline{\pi}}}\left[\overline{\mathcal{R}}(\phi(s_h),a_h)\right] && (23)\\
&\quad + \gamma^{h+1} \mathop{\mathbb{E}}_{\tilde{s}_{h+1}\sim\bar{\mathbf{P}}\circ\mathbf{P}^h_{\overline{\pi}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}_{h+1})\right] && (24)\\
&\quad - \left(\sum_{t=0}^{h}\gamma^t \mathop{\mathbb{E}}_{s_t,a_t\sim\mathbf{P}^t_{\overline{\pi}}}[\mathcal{R}(s_t,a_t)] + \gamma^{h+1} \mathop{\mathbb{E}}_{s_{h+1}\sim\mathbf{P}^{h+1}_{\overline{\pi}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\phi(s_{h+1}))\right]\right)\\
&= \gamma^h \mathop{\mathbb{E}}_{s_h,a_h\sim\mathbf{P}^h_{\overline{\pi}}}\left[\overline{\mathcal{R}}(\phi(s_h),a_h) - \mathcal{R}(s_h,a_h)\right] && (25)\\
&\quad + \gamma^{h+1}\left(\mathop{\mathbb{E}}_{\tilde{s}_{h+1}\sim\bar{\mathbf{P}}\circ\mathbf{P}^h_{\overline{\pi}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}_{h+1})\right] - \mathop{\mathbb{E}}_{s_{h+1}\sim\mathbf{P}^{h+1}_{\overline{\pi}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\phi(s_{h+1}))\right]\right) && (26)\\
&= \gamma^h \mathop{\mathbb{E}}_{s_h,a_h\sim\mathbf{P}^h_{\overline{\pi}}}\left[\overline{\mathcal{R}}(\phi(s_h),a_h) - \mathcal{R}(s_h,a_h)\right] && (27)\\
&\quad + \gamma^{h+1} \mathop{\mathbb{E}}_{s_h\sim\mathbf{P}^h_{\overline{\pi}}}\left[\mathop{\mathbb{E}}_{\tilde{s}_{h+1}\sim\bar{\mathbf{P}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}_{h+1})\right] - \mathop{\mathbb{E}}_{s_{h+1}\sim\mathbf{P}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\phi(s_{h+1}))\right]\right] && (28)
\end{aligned}$$

$$\begin{aligned}
U_0(s) - U_H(s) &= \sum_{h=0}^{H-1}(U_h(s) - U_{h+1}(s)) && (29)\\
&= \sum_{h=0}^{H-1}\left(\gamma^h \mathop{\mathbb{E}}_{s_h,a_h\sim\mathbf{P}^h_{\overline{\pi}}}\left[\overline{\mathcal{R}}(\phi(s_h),a_h) - \mathcal{R}(s_h,a_h)\right]\right. && (30)\\
&\quad \left. + \gamma^{h+1} \mathop{\mathbb{E}}_{s_h\sim\mathbf{P}^h_{\overline{\pi}}}\left[\mathop{\mathbb{E}}_{\tilde{s}_{h+1}\sim\bar{\mathbf{P}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}_{h+1})\right] - \mathop{\mathbb{E}}_{s_{h+1}\sim\mathbf{P}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\phi(s_{h+1}))\right]\right]\right) && (31)
\end{aligned}$$

This conveniently leads to the following values for $U_0$ and $U_H$ for any state $s$,

$$U_0(s) = \overline{V}^{\overline{\pi}}_{\theta,H}(\phi(s)) \tag{32}$$

$$U_H(s) = V^{\bar{\pi}}(s) - \gamma^H \mathop{\mathbb{E}}_{s\sim \mathbf{P}^H_{\bar{\pi}}} \left[ V^{\bar{\pi}}(\phi(s_H)) - \bar{V}^{\bar{\pi}}_{\theta}(\phi(s_H)) \right] \tag{33}$$

From these definitions, we find that the expected difference between $U_0$ and $U_H$ is as follows,

$$U_0(s) - U_H(s) = \bar{V}^{\bar{\pi}}_{\theta,H}(\phi(s)) - \left( V^{\bar{\pi}}(s) - \gamma^H \mathop{\mathbb{E}}_{s\sim \mathbf{P}^H_{\bar{\pi}}} \left[ V^{\bar{\pi}}(s_H) - \bar{V}^{\bar{\pi}}_{\theta}(\phi(s_H)) \right] \right) \tag{34}$$

$$= \bar{V}^{\bar{\pi}}_{\theta,H}(\phi(s)) - V^{\bar{\pi}}(s) + \gamma^H \mathop{\mathbb{E}}_{s\sim \mathbf{P}^H_{\bar{\pi}}} \left[ V^{\bar{\pi}}(s_H) - \bar{V}^{\bar{\pi}}_{\theta}(\phi(s_H)) \right] \tag{35}$$

by plugging Eqs. (31) and (35)

$$\bar{V}^{\bar{\pi}}_{\theta,H}(\phi(s)) - V^{\bar{\pi}}(s) = \sum_{h=0}^{H-1} \left( \gamma^h \mathop{\mathbb{E}}_{s_h,a_h\sim \mathbf{P}^h_{\bar{\pi}}} \left[ \bar{\mathcal{R}}(\phi(s_h),a_h) - \mathcal{R}(s_h,a_h) \right] \right. \tag{36}$$

$$\left. + \gamma^{h+1} \mathop{\mathbb{E}}_{s_h\sim \mathbf{P}^h_{\bar{\pi}}} \left[ \mathop{\mathbb{E}}_{\tilde{s}_{h+1}\sim \bar{\mathbf{P}}} \left[ \bar{V}^{\bar{\pi}}_{\theta,H-h-1}(\tilde{s}_{h+1}) \right] - \mathop{\mathbb{E}}_{s_{h+1}\sim \mathbf{P}} \left[ \bar{V}^{\bar{\pi}}_{\theta,H-h-1}(\phi(s_{h+1})) \right] \right] \right) \tag{37}$$

$$- \gamma^H \mathop{\mathbb{E}}_{s\sim \mathbf{P}^H_{\bar{\pi}}} \left[ V^{\bar{\pi}}(s_H) - \bar{V}^{\bar{\pi}}_{\theta}(\phi(s_H)) \right] \tag{38}$$

Let us add absolute values and expectations over the stationary distribution.

$$\mathop{\mathbb{E}}_{s\sim \xi_{\bar{\pi}}} \left[ \left| \bar{V}^{\bar{\pi}}_{\theta,H}(\phi(s)) - V^{\bar{\pi}}(s) \right| \right] = \mathop{\mathbb{E}}_{s\sim \xi_{\bar{\pi}}} \left| \sum_{h=0}^{H-1} \left( \gamma^h \mathop{\mathbb{E}}_{s_h,a_h\sim \mathbf{P}^h_{\bar{\pi}}} \left[ \bar{\mathcal{R}}(\phi(s_h),a_h) - \mathcal{R}(s_h,a_h) \right] \right. \right. \tag{39}$$

$$\left. \left. + \gamma^{h+1} \mathop{\mathbb{E}}_{s_h\sim \mathbf{P}^h_{\bar{\pi}}} \left[ \mathop{\mathbb{E}}_{\tilde{s}_{h+1}\sim \bar{\mathbf{P}}} \left[ \bar{V}^{\bar{\pi}}_{\theta,H-h-1}(\tilde{s}_{h+1}) \right] - \mathop{\mathbb{E}}_{s_{h+1}\sim \mathbf{P}} \left[ \bar{V}^{\bar{\pi}}_{\theta,H-h-1}(\phi(s_{h+1})) \right] \right] \right) \right. \tag{40}$$

$$\left. - \gamma^H \mathop{\mathbb{E}}_{s_H\sim \mathbf{P}^H_{\bar{\pi}}} \left[ V^{\bar{\pi}}(s_H) - \bar{V}^{\bar{\pi}}_{\theta}(\phi(s_H)) \right] \right| \tag{41}$$

$$= \mathop{\mathbb{E}}_{s\sim \xi_{\bar{\pi}}} \left| \sum_{h=0}^{H-1} \gamma^h \mathop{\mathbb{E}}_{s_h,a_h\sim \mathbf{P}^h_{\bar{\pi}}} \left[ \bar{\mathcal{R}}(\phi(s_h),a_h) - \mathcal{R}(s_h,a_h) \right] \right. \tag{42}$$

$$\left. + \sum_{h=0}^{H-1} \gamma^{h+1} \mathop{\mathbb{E}}_{s_h\sim \mathbf{P}^h_{\bar{\pi}}} \left[ \mathop{\mathbb{E}}_{\tilde{s}_{h+1}\sim \bar{\mathbf{P}}} \left[ \bar{V}^{\bar{\pi}}_{\theta,H-h-1}(\tilde{s}_{h+1}) \right] - \mathop{\mathbb{E}}_{s_{h+1}\sim \mathbf{P}} \left[ \bar{V}^{\bar{\pi}}_{\theta,H-h-1}(\phi(s_{h+1})) \right] \right] \right. \tag{43}$$

$$\left. - \gamma^H \mathop{\mathbb{E}}_{s_H\sim \mathbf{P}^H_{\bar{\pi}}} \left[ V^{\bar{\pi}}(s_H) - \bar{V}^{\bar{\pi}}_{\theta}(\phi(s_H)) \right] \right| \tag{44}$$

$$\leq \mathop{\mathbb{E}}_{s\sim \xi_{\bar{\pi}}} \mathop{\mathbb{E}}_{s_h\sim \mathbf{P}^h_{\bar{\pi}}} \sum_{h=0}^{H-1} \gamma^h \left| \bar{\mathcal{R}}(\phi(s_h),a_h) - \mathcal{R}(s_h,a_h) \right| \tag{45}$$

$$+ \mathop{\mathbb{E}}_{s\sim \xi_{\bar{\pi}}} \mathop{\mathbb{E}}_{s_h\sim \mathbf{P}^h_{\bar{\pi}}} \sum_{h=0}^{H-1} \gamma^{h+1} \left| \mathop{\mathbb{E}}_{\tilde{s}_{h+1}\sim \bar{\mathbf{P}}} \left[ \bar{V}^{\bar{\pi}}_{\theta,H-h-1}(\tilde{s}_{h+1}) \right] - \mathop{\mathbb{E}}_{s_{h+1}\sim \mathbf{P}} \left[ \bar{V}^{\bar{\pi}}_{\theta,H-h-1}(\phi(s_{h+1})) \right] \right| \tag{46}$$

$$+ \gamma^H \mathop{\mathbb{E}}_{s\sim \xi_{\bar{\pi}}} \mathop{\mathbb{E}}_{s_H\sim \mathbf{P}^H_{\bar{\pi}}} \left| V^{\bar{\pi}}(s_H) - \bar{V}^{\bar{\pi}}_{\theta}(\phi(s_H)) \right| \tag{47}$$

By definition of the stationary distribution, $\mathbb{E}_{s\sim\xi_{\bar{\pi}}} \mathbb{E}_{s_h\sim \mathbf{P}^h_{\bar{\pi}}} X = \mathbb{E}_{s\sim\xi_{\bar{\pi}}} X$ and so we have,

$$\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\left[\left|\overline{V}^{\overline{\pi}}_{\theta,H}(\phi(s)) - V^{\overline{\pi}}(s)\right|\right] \leq \mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\sum_{h=0}^{H-1}\gamma^h\left|\overline{\mathcal{R}}(\phi(s),a) - \mathcal{R}(s,a)\right| \tag{48}$$

$$+\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\sum_{h=0}^{H-1}\gamma^{h+1}\left|\mathop{\mathbb{E}}_{\tilde{s}'\sim\overline{\mathbf{P}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}')\right] - \mathop{\mathbb{E}}_{s'\sim\mathbf{P}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\phi(s'))\right]\right| \tag{49}$$

$$+\gamma^H\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\left|V^{\overline{\pi}}(s) - \overline{V}^{\overline{\pi}}_{\theta}(\phi(s))\right| \tag{50}$$

$$=\frac{1-\gamma^H}{1-\gamma}\left(\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\left|\overline{\mathcal{R}}(\phi(s),a) - \mathcal{R}(s,a)\right|\right. \tag{51}$$

$$\left.+\gamma\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\left|\mathop{\mathbb{E}}_{\tilde{s}'\sim\overline{\mathbf{P}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}')\right] - \mathop{\mathbb{E}}_{s'\sim\mathbf{P}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\phi(s'))\right]\right|\right) \tag{52}$$

$$+\gamma^H\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\left|V^{\overline{\pi}}(s) - \overline{V}^{\overline{\pi}}_{\theta}(\phi(s))\right| \tag{53}$$

$$=\frac{1-\gamma^H}{1-\gamma}\left(L_{\mathcal{R}}\right. \tag{54}$$

$$\left.+\gamma\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\left|\mathop{\mathbb{E}}_{\tilde{s}'\sim\overline{\mathbf{P}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}')\right] - \mathop{\mathbb{E}}_{s'\sim\mathbf{P}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\phi(s'))\right]\right|\right) \tag{55}$$

$$+\gamma^H L_V \tag{56}$$

We now upperbound Eq. (55). First, we note that $\overline{\mathcal{R}}$ is Lipschitz from the fact that the latent MDP is discrete. Furthermore, this also implies that $\overline{V}^{\overline{\pi}}$ is Lipschitz continuous as well. From this, we have that $\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}')$ is Lipschitz continuous with some constant $K$,

$$\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\left|\mathop{\mathbb{E}}_{\tilde{s}'\sim\overline{\mathbf{P}}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\tilde{s}')\right] - \mathop{\mathbb{E}}_{s'\sim\mathbf{P}}\left[\overline{V}^{\overline{\pi}}_{\theta,H-h-1}(\phi(s'))\right]\right| \tag{57}$$

$$\leq \mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\sup_{f\in\mathcal{F}_K}\left|\mathop{\mathbb{E}}_{\tilde{s}'\sim\overline{\mathbf{P}}}\left[f(\tilde{s}')\right] - \mathop{\mathbb{E}}_{s'\sim\mathbf{P}}\left[f(\phi(s'))\right]\right| \tag{58}$$

$$= KL_{\mathbf{P}} \tag{59}$$

This leads to our final inequality,

$$\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}}\left[\left|\overline{V}^{\overline{\pi}}_{\theta,H}(\phi(s)) - V^{\overline{\pi}}(s)\right|\right] \leq \frac{1-\gamma^H}{1-\gamma}\left(L_{\mathcal{R}} + \gamma K L_{\mathbf{P}}\right) + \gamma^H L_V \tag{60}$$

□

## B PROOF OF THEOREM 6.3.

Restating Theorem 6.3 for clarity:

Theorem B.1. *Let $\mathcal{M}$ be an ergodic MDP and assume that the latent policy $\overline{\pi}$ has full support in every state, i.e. $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$, $\overline{\pi}(a|\phi(s)) > 0$. For a latent MDP $\overline{\mathcal{M}}$, the local losses defined in Eq.* (4) *are zero if and only if $\mathcal{M}\sim\overline{\mathcal{M}}$.*

Proof. We first show that for the markov chain $\mathcal{M}_\pi$ induced by the lifted policy $\pi(s,a) = \overline{\pi}(\phi(s),a)$, every state $s\in\mathcal{S}$ has strictly positive probability in the stationary distribution, i.e. $\xi_\pi(s) > 0$. Since every state in the MDP $\mathcal{M}$ is accessible, the Markov chain $\mathcal{M}_\pi$ induced by the policy $\pi$ is irreducible.

Therefore, there exists a path $\rho(s_I, s)$ from $s_I$ to any other state $s$. Due to the assumption of full support, this path has a positive probability of occurring. The ergodicity of the MDP guarantees that $s_I$ will be visited infinitely often. Hence, , will also be visited infinitely often and therefore $\xi_\pi(s) > 0$.

We now show that when the local losses are zero, we recover a bisimulation relation between the real MDP $\mathcal{M}$ and latent MDP $\overline{\mathcal{M}}$. We know by [12] that $\forall s_1, s_2$ such that $\phi(s_1) = \phi(s_2)$ it is guaranteed that,

$$\tilde{d}_{\overline{\pi}}(s_1,s_2) \leq \left[L_{\mathcal{R}} + \frac{\gamma L_{\mathbf{P}}}{1-\gamma}\right](\xi^{-1}_{\overline{\pi}}(s_1) + \xi^{-1}_{\overline{\pi}}(s_2)). \tag{61}$$

Obviously, when $L_{\mathcal{R}} = 0$ and $L_{\mathbf{P}} = 0$ this implies that $\tilde{d}_{\overline{\pi}}(s_1, s_2) = 0$.

( $\Longrightarrow$ ) Since for all states $\xi_{\overline{\pi}}^{-1}(s) > 0$ and the local losses are zero, we have that,

$$\mathop{\mathbb{E}}_{s,a\sim\xi}\left[\left|\mathcal{R}(s,a) - \overline{\mathcal{R}}(\phi(s),a)\right|\right] = 0. \tag{62}$$

This implies that $\forall s, a \in \xi : \left|\mathcal{R}(s,a) - \overline{\mathcal{R}}(\phi(s),a)\right| = 0$ and therefore that $\mathcal{R}(s,a) = \overline{\mathcal{R}}(\phi(s),a)$.

Additionally, since the local transition loss is also zero, we have,

$$\mathop{\mathbb{E}}_{s,a\sim\xi}\left[\mathcal{W}_{d_{\bar{S}}}\left(\phi(\mathbf{P}(\cdot|s,a)), \bar{\mathbf{P}}(\cdot|\phi(s),a)\right)\right] = 0. \tag{63}$$

which analogously implies $\phi(\mathbf{P}(\cdot|s,a)) = \bar{\mathbf{P}}(\cdot|\phi(s),a)$. Note that the metric used here is not a pseudometric but specifically $\tilde{d}_{\bar{S}} = \mathbf{1}_{\neq}$. This guarantees that we learned an MDP homomorphism and from [37] this further implies bisimulation.

( $\Longleftarrow$ ) In the other direction, we aim to show that when the model learns exact lax bisimulation, the local losses are zero. This follows by definition. □